Towards Systematic Vision: Limitations of Convolutional Neural Networks and Future Directions in Oscillatory Coding

By

Matthew Ricci

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

Brown University
Department of Cognitive, Linguistic and Psychological Sciences

October 2020

This dissertation by Matthew Ricci is accepted in its present form by the Department of Cognitive, Linguistic and Psychological Sciences as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date _____ _____

Thomas Serre, Advisor

Recommended to the Graduate Council

Date _____ _____

Elie Bienenstock, Reader

Date _____ _____

Michael Frank, Reader

Date _____ _____

Matthew Harrison, Reader

Approved by the Graduate Council

Date _____ _____

Andrew G. Campbell, Dean of the Graduate School

Curriculum Vitae

Matthew Ricci received a B.A. and M.A. in mathematics as well as a B.A. in the history of music from the University of Pennsylvania. After a year working as a research assistant for Randy Gallistel, he started graduate studies in the laboratory of Thomas Serre at Brown University where he was both an NIH training fellow and NSF graduate research fellow. He is currently a postdoctoral associate advised by Stuart Geman at Brown University's Data Science Initiative.

ACKNOWLEDGMENTS

This manuscript has benefited invaluably from the counsel of my advisor and thesis committee. Thomas Serre has been a relentless, energetic and energizing mentor who has encouraged me to develop what were initially vague intuitions into the fuller scientific stories which fill these pages. I am grateful for his ongoing advice. I have also drawn a great deal of inspiration from many long conversations with Elie Bienenstock, who taught me from early in my graduate education to think critically about scientific trends and their alternatives. He has been very generous with his time, and for this I am enduringly grateful. Much of the technical material in this document owes its origin to discussions with Matt Harrison, whose notes on the Kuramoto model I inherited from Charles Windolf. Matt very graciously listened to me ramble about my still-young ideas and helped me rigorize and develop them. I owe my understanding of the curse of dimensionality and importance sampling, among other topics, to him. As is often the case, some of the most insightful advice I have received has come from experts outside my precise discipline, notably from committee member Michael Frank. Early discussions we had during Brown's "Beyond Deep Learning" conference as well as comments he offered on drafts of this thesis have proved very valuable.

I must also extend my thanks to faculty mentors outside of Brown and indeed from before the start of my graduate studies. Important among these other mentors is Randy Gallistel, from whom I have learned a tremendous amount and with whom I have enjoyed a fruitful scientific collaboration. Randy once told me that he did not get to where he is today by worrying about appearing like a fool, and I try to practice this advice at least daily. Tony Kroch also deserves a great deal of thanks for introducing

me to Randy and for serving as my introduction to cognitive science as a whole. He has since served as a valuable counselor. I would also like to thank faculty members from my undergraduate studies who helped shaped my scientific and philosophical thinking, including Philip Gressman, David Harbater, Arman Schwartz and Emily Dolan. I think of my experience as their student often and fondly.

I was very lucky to develop the ideas in this thesis alongside truly thoughtful and insightful peers. My co-author and friend Junkyung Kim served as a tireless conversation partner throughout my graduate education, and this thesis is inhabited by his intellectual spirit. He has a well-deserved reputation as a consummate experimentalist, and I try to emulate his experimental design methods wherever I can. Drew Linsley remains an inspiring academic role model, and I am very pleased to continue my scientific collaboration with him in the coming years. It is rare to encounter such a breadth of knowledge, from neurophysiology to machine learning, and I count him both among my friends and my teachers. Yarden Katz has also been a steadfast friend and confidant during the latter half of my graduate studies, and I am thankful for his continuing advice on both academic and socio-political matters. Brown University has proved an endless supply of though-provoking friends and collaborators for whom I am very thankful, including Charles Windolf, Yuwei Zhang, Aneri Soni, Mathieu Chalvidal, and Minju Jung. This work is as much theirs as mine.

This thesis was written during a very brief period from mid-April to mid-June 2020, and it would not have been possible to write so quickly without a good deal of subconscious inspiration. I would therefore like to thank my "musical collaborators", Phish and Sandy Denny, for providing me with enough albums to spin long into the night during bleary-eyed writing sessions. The curious reader might try to decode their influence from these pages.

The deepest influence on this manuscript, albeit an indirect one, is due to my parents, Gabriel Ricci and Patricia Likos Ricci. I have inherited from them a burning

conviction that knowledge should be pursued for its own sake, and that a society's neglect of disinterested investigation is a dire sign. I was very lucky to have two scholars from the humanities as parents, and, consequently, this thesis was self-consciously written as a contribution to the liberal arts. I give them my love and thanks in equal measure.

Every thought expressed in this document has, by one method or another, been influenced by fiancée, Edwige Crucifix, whose boundless patience, understanding and affection have sustained me throughout my graduate studies. I can admit without hyperbole that I could not have successfully completed this PhD without her support, provided in the form of constant academic and social advice as well as persistent reminders that life is more important than impact factors. It is not possible to conceive of a better intellectual and romantic companion. I send her my love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

To Edwige Crucifix, for teaching me to think about the humanities like a scientist.

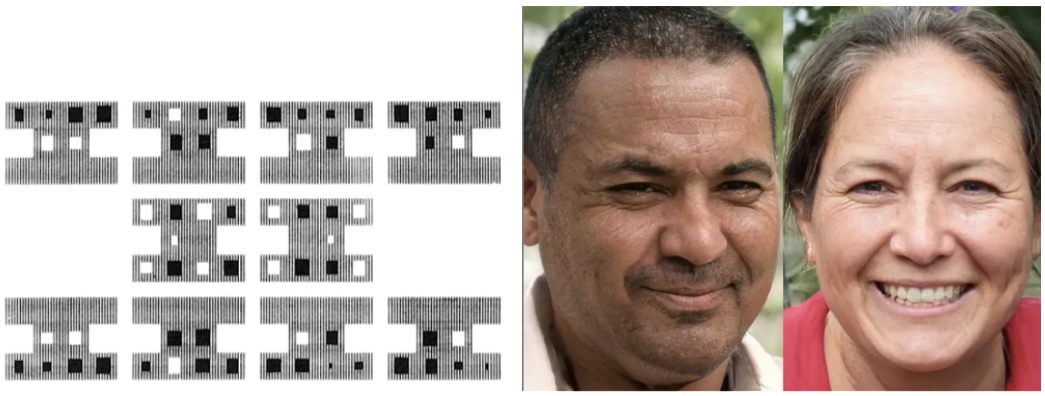
To my parents, for teaching me to think about science like a human.

Chapter One

Introduction

That there have been significant practical advances in machine learning and so-called "artificial intelligence" in the last decade is undeniable. For instance, Fig. 1a shows a schematic of a generative neural model from Ackley, Hinton, and Sejnowski, 1985 trained to encode 4-bit sequences, a significant achievement for the time. Fig. 1b, on the other hand, shows a pair of hyper-realistic human faces generated by a neural network from Karras, Laine, and Aila, 2019. These practical marvels are too numerous to count. That these advances represent a deepening of theoretical understanding over and above the claims of connectionists in the 1980s is, however, less assured. Here, the machine learning discipline is split: one camp argues that contemporary neural models are largely turbo-charged versions of their twentieth-century ancestors, incidentally accelerated by modern parallel computing hardware and the availability of huge, labeled data sets (Russakovsky et al., 2010); a second camp suggests that these incidental developments have transported us to a new theoretical world in which these incredible neural marvels must be studied by abstruse new paradigms (Mallat, 2016) and intense collective effort. Yet, what is so damning about the possibility that contemporary neural architectures are supped-up versions of their ancestors is not that the new theoretical tools brought to bear for their examination are incorrect or unimportant (quite the contrary), but rather that fundamental attacks levied against and unresolved questions posed to these systems from earlier periods might still hold

water. In short, machine learning practitioners and scholars must once again ask themselves, "What is intelligence? And is the current neural modeling paradigm the correct way of studying it?" (see Chollet, 2019).



(a) **(b)**

Figure 1.1 *Neural networks, past and present. (a)* A three-layer neural network encoding 4-bit strings (Ackley, Hinton, and Sejnowski, 1985). "I"-shapes depict synaptic weights connecting to given location with color of squares indicating sign (white positive, black negative) and size indicating magnitude. The bottom two layers can be used to model a density on these strings. *(b)* Two artificial faces generated by a generative adversarial network (Karras, Laine, and Aila, 2019).

The reader will be happy to know that I do not pretend to answer either of these questions, even partially, though this thesis is very much inspired by those fundamental attacks once levied against connectionist thinking, particularly proponents of the "computational theory of mind" (Rescorla, 2020): Jerry Fodor, Randy Gallistel, Noam Chomsky and their fellow travelers traced back to Descartes and Kant. These criticisms will be discussed in some detail later, but for now I will summarize them as "routine precedes representation"[1]. That is, mentation is best described as the flexible and hypothetically unbounded combination of atomic representations according to certain routines or processes rather than the construction of fixed but arbitrarily complex

[1]By representation, I mean the very narrow sense in which it is used in contemporary neural modeling as more or less a pattern of synaptic weights or features. Naturally, Fodor and Gallistel are staunch cognititivists and, as such, defend the general notion of "representation" tirelessly.

representations. Contemporary machine learning in all its varied forms does not commit to either side of this argument, though it is not controversial to say that the so-called "deep learning" sub-discipline, with its emphasis on complex representations of high-dimensional data learned by many-layered ("deep") networks, leans towards the latter, "representation-first", view.[2] This is quite a ringing endorsement, given the recent successes of deep learning.

However, in a spirit of Fodorian contrarianism and inspired by the possibility that contemporary deep learning is largely a recap of somewhat ancient science, this dissertation seeks first and foremost to consolidate and promote a body of psychological and machine learning literature in which the "routine precedes representation" way of thinking seems obviously correct. Given that machine learning has seen its most spectacular advances on visual tasks, I have chosen a body of literature from vision science, namely the "visual relations" literature. This collection of results concerns the ability of agents (biological or otherwise) to discern when objects in a visual scene obey a given relation; e.g., a human subject presented with an image of two shapes might be asked "Are these two shapes the same up to rotation?" or "Are these two shapes arranged more vertically or horizontally?" We will mostly focus on the relation expressed in the first question, that of "same-different", for its simplicity and fundamental role in visual reasoning more generally. Crucially, in the words of animal psychologist Juan Delius, the accurate assessment of this and other visual relations "... implies the operation of a unitary comparison operation issuing a variable signaling [the relation] regardless of the *particular qualities* of the stimuli" (Emphasis mine) (Delius, 1994, p. 27). That is, the mental routine assessing the relation should function properly regardless of the visual representations of the objects under the relation: "routine precedes representation". In many ways, Delius's claim is a restatement of

[2]This dissertation will assume a basic familiarity with artificial neural networks and their training.Goodfellow, Bengio, and Courville, 2016 is a good modern review.

Fodor and Pylyshyn, 1988's defense of the compositionality of thought[3]. As we will see, across several numerical simulations, contemporary machine learning models do not always meet this standard. The poor performance of these models on tasks which are trivially easy for the biological visual system adds to the growing list of perplexing deficiencies in contemporary machine learning models (see Szegedy, Zaremba, and Sutskever, 2013; Recht et al., 2019 for some interesting examples.)

Of course, it is one thing to open old wounds from the 1980s regarding the compositionality of thought, but it is quite another to propose a new model whose flexible routines are able to support relation detection at the level of biological vision. Such a model, inasmuch as biology is a proper guide, should proceed in several steps. The first step is to flexibly organize bundles of features into unified perceptual objects (Treisman, Kahneman, and Burkell, 1983; Kahneman and Treisman, 1984) so that the relations can be assessed between appropriate image parts (as opposed to, for example, coarse combinations of objects and their backgrounds). Second, spatial attention is shifted to another image region where subsequent objects forming a part of the relation might be found. Third, and throughout the grouping and shifting process, perceptual objects should be stored in a working memory buffer (Clevenger and Hummel, 2014) so that, after all relevant objects in the scene are exhausted, the relationships among stored objects can be ascertained. Very few contemporary neural networks are explicitly designed for visual relation detection per se, though the few that do exist of any relevance to this task typically only involve the attentive and mnemonic mechanisms of steps two and three. Notable recent examples come from the so-called "Visual Question Answering" (VQA) literature, in which neural models answer queries about images provided in the form of text strings.

[3]In the context of natural language, Fodor writes "Similarity of constituent structure accounts for the semantic relatedness between systematically related sentences only to the extent that the semantical properties of the shared constituents are independent."

Very few indeed are the modern models that address step one, the issue of perceptual organization. On this subject, critical engineering questions abound with little biological evidence as a guide. Should relevant features simply be enhanced while irrelevant ones are attenuated (Roelfsema and Houtkamp, 2011)? Or should a multiplexing scheme be employed in which different objects in a scene are "tagged" so they can each be individually selected for later processing (Fries, 2005; McLelland and VanRullen, 2016)? What principles should guide the grouping in the first place (Grossberg, Mingolla, and Ross, 1997)? And how, of course, should answers to any of these questions actually be implemented in neural hardware amenable to the modern statistical learning techniques to which machine learning practitioners have grown accustom?

This dissertation picks up on one persistent theme running through the perceptual organization literature, especially that portion of the literature dealing with the relations among objects in neural systems: temporal coherence among dynamical neurons has an important role to play. This idea is most strongly associated with the "correlation theory of brain function" due to von der Malsburg (Malsburg, 1994) and the "binding by synchrony" results made briefly famous by (Gray and Singer, 1989; Gray et al., 1989), which contend that perceptual organization is mediated by the temporal coordination of spiking neurons, either at the population or unit level. Other manifestations come from (Fries, 2015) who argued in his "communication by coherence" hypothesis that the phase of neural activity with respect to a macroscopic neural oscillation could serve as a tag for selective attention to grouped features. More recent evidence for the role of temporal coherence in perceptual organization comes from Brzezicka, Mamelak, and Rutishauser, 2020 who found that the phase of neural activity could index visual representations in working memory and from Alamia et al., 2019 who found that distinctive oscillatory regimes emerged in human EEG data when subjects performed a visual relation detection task.

Anyone wishing to transform these physiological intuitions into a bona fide machine learning model would do well to seek out that branch of scholarship where the issues of temporal coherence in dynamical systems and the correlations among units in a distributed system are treated with mathematical seriousness: the study of systems of coupled oscillators (Strogatz, 2000; Wang and Slotine, 2005). The study of these objects originated with Huygens' informal description in the seventeenth century but took on a new life with the advent of modern mathematical biology in the 1960s. It was only then that physicists, mathematicians and mathematical biologists, like Kuramoto, Malkin and Winfree developed the formal theory of these systems in their attempts to study the emergence of collective ("synchronous") behavior in a population of heterogeneous individuals (anything from voters with their own political opinions in a democratic society to interacting neurons each with their own intrinsic dynamics in a neural network). By far, the central model in the study of these systems is the "Kuramoto model" first proposed by the eponymous physicist in a beautiful and hugely influential two-page report from 1975 (Kuramoto, 1975). Since this early work, the study of synchrony in complex systems and the emergence of coordinated behavior in heterogenous populations has become a science in itself.

So far, I have suggested that 1) contemporary neural networks underperform on tasks involving the detection of relations among arbitrary objects, 2) that there exists a compelling neuroscientific literature implicating coherence of neural activity as a fundamental mechanism not only in the perceptual organization presumably required for relation detection but also flexible reasoning more generally and 3) that the study of systems of coupled oscillators is, more or less, a rigorous formulation of the intuitions from this neuroscientific literature. An obvious question emerges: can one build a model using the mathematics of (3), which captures the dynamics explored in (2) and solves the cognitive problems of (1)? This is not an easy question. First of all, it presupposes that oscillatory systems can "learn" how to solve problems or at least

exhibit target behavior. Very little is understood about these systems in general (see Arenas and Albert, 2008 for some open problems), and less still is known about how to control them. Further, a direct application of oscillatory systems to the problems of perceptual organization and relation detection would require integrating these systems into larger networks of modules carrying out those other processes suspected to underlie flexible cognition.

Instead, this dissertation takes the simpler route of laying the groundwork for a fuller exploration of these questions at a later date. Its general purpose is to convince scientists that "representation-first" machine learning is not always advisable and that somewhat heterodox literature of von der Malsburg, Fries and others should provide us with inspiration for the creation of future "routine-first" systems. Its main technical contribution is the development of several learning algorithms, based on contemporary methods from machine learning, specifically tailored to oscillatory systems. I will deploy these algorithms on systems attempting to learn grouping abilities which I believe to be pre-requisite behaviors for flexible reasoning and relation detection. Initial results are encouraging, and future work will scale these findings to more complex problems and realistic scenarios.

The subsequent chapters of this dissertation will be

- **Chapter 2**: A review the literature on visual relations in psychology, neuroscience and machine learning followed by numerical experiments. I will argue that this literature exemplifies the benefits of a "routine precedes representation" approach to visual cognition. I will show across several computer simulations that feedforward models of the visual cortex which are otherwise powerful models of visual processing fail to model relational understanding in efficient and biologically relevant ways.

- **Chapter 3**: A review of oscillatory coding for visual cognition followed by a

technical review of oscillatory systems. A particular focus will be the nascent literature on the optimization of oscillatory systems. I will also lay out how, depending on the setting of various hyperparameters, these systems take on either equilibrium or non-equilibrium characteristics and therefore enable or prohibit the application of different training methods. I will spell out these methods in detail in the following chapters.

- **Chapter 4**: The development of a maximum likelihood technique for the training of an equilibrium oscillatory system. I will show how conditional inference in these systems can be interpreted as perceptual grouping. I will also demonstrate through a mean field reduction that variants on these systems have unexpected fast synapse dynamics speculated upon in Malsburg, 1994. I will also show how selective attention in these systems takes the form of Bayesian inference.

- **Chapter 5** The development of a supervised learning technique for the generally non-equilibrium Kuramoto model and application of this method to a quasi-same-different detection task. I will show that the quenched random parameters of the Kuramoto model can be construed as samples of an adaptable distribution whose parameters can be adjusted by error minimization. Various applications of this framework will be explored.

1.1 Contributions

This thesis adds to knowledge in several areas. A principal cognitive contribution is the dissertation's formalization of the notion of systematicity in machine vision and a subsequent analysis of this behavior in contemporary neural networks. This analysis attempts to synthesize several hitherto disparate areas of study: the design of feedforward neural networks, the computational theory of mind as advocated by Jerry Fodor and others, and the psychology of visual reasoning. It is the hope of the author

that other scholars will begin to formalize and quantify systematicity in their own ways and this measurement will mark another shared area of research for cognitive scientists and machine learning practitioners.

The conceit of systematicity is used in this thesis to motivate a renewed focus on oscillatory phenomena in visual cortex. In particular, I will present new experimental evidence collected with colleagues from Centre de Recherche Cerveau & Cognition, Université Toulouse III, implicating oscillatory mechanisms in visual reasoning. This finding sheds new light on the critically important but deeply mysterious phenomenon of cortical rhythms.

Inspired by this result, I develop the early stages of a new machine learning paradigm based on oscillatory systems. The initial material I have provided represents only the smallest possible steps of what could very likely be a large-scale research program, but it is my desire that what few threads I offer on this subject will be taken up by machine learning practitioners, cognitive neuroscientists and computational physicists interested in these matters. If the last decade's incredibly fruitful dialogue between computer vision and the study of the visual cortex is to be trusted, the development of machine learning techniques directly relevant to cortical rhythms could prove extremely valuable for cognitive neuroscientists in particular.

Wherever relevant, I have noted where dissertation material has or will soon appear in publication.

Chapter Two

Systematicity and visual relations

According[1] to an already outdated account of Dean, Patterson, and Young, 2018, about 100 machine learning papers are published on the preprint website ArXiV.org each day, and a great deal of these probably concern computer vision, in many ways the crown jewel of the last decade's "deep learning" revolution. This number has probably increased exponentially since it was estimated in 2017. It would be safe to guess that the overwhelming majority of this material consists of incremental developments of pre-existing ideas within the albeit very fruitful paradigm of deep learning while comparatively little is devoted to a deeper understanding of the paradigm itself, its fundamental concepts and boundaries, where it succeeds and where fails[2]. That these analyses are somewhat rare is understandable, since they typically entail both a difficult meta-scientific discussion of what objects in the world lend themselves to a fruitful, explanatory theory and the even trickier task of constructing a technical vocabulary with which to build that theory. Rarer still are those areas in the literature where fundamental analyses of the paradigm are brought into contact with concrete material and experiments from either deep learning or those branches of brain science

[1]The material from this chapter has been published in Ricci, Kim, and Serre, 2018 and Kim et al., 2018. A forthcoming publication, Ricci, Cadene, and Serre, Submitted, addresses a similar topic with a greater focus on "visual question answering" (VQA) and recurrent models.

[2]There are of course very notable recent exceptions, both defending and critiquing deep learning, for instance, LeCun, Bengio, and Hinton, 2015 and Lake et al., 2016, respectively.

which are supposedly its inspiration.

This chapter intends to do just that, by situating some recent developments in the emulation of biological visual reasoning by deep networks in the context of earlier theoretical debates most closely associated with Fodor and Pylyshyn, 1988. Many of the issues raised by Fodor and Pylyshyn have either never been resolved or are too mired in controversy to be of practical use to the machine learning scholar, though this chapter will contend that the insistence by Fodor and Pylyshyn, 1988 on the productivity, compositionality and especially systematicity of thought has serious and concrete implications for the creation of thinking visual machines. By "systematicity", we roughly mean the fact that minds seem to automatically generalize their understanding of a complex expression to all syntactically identical but semantically different expressions: a classic example holds that there is no one who understands "John loves Mary" but not "Mary loves John". Throughout a review of the psychological and neuroscientific literature on visual relations, we will argue that visual relation detection is a hitherto neglected but strongly corroborating example of systematic cognition in the Fodorian sense. We will focus largely on one type of visual relation, that of identity up to transformation or the "same-different" relation. Further, we will outline how the cognitive mechanisms supporting this behavior, properly construed, align quite closely with those which "classical" cognitive scientists have traditionally argued underlie representations with a combinatorial syntax. Put more plainly, we will show how biological agents use a combination of attention and working memory to sequentially construct representations of the relations among visual objects as opposed to representing the relation as an atomic feature.

Finally, we will demonstrate through several numerical simulations how the standard model of visual processing, the convolutional neural network (CNN), which lacks these "classical" cognitive mechanisms, exhibits decidedly unsystematic behavior when trained on a visual relations task. The key assumption underlying these experiments

is that a truly systematic machine should be able to learn two instances of otherwise syntactically identical visual rules with equal ease. We find that this is not the case for CNNs. This simulation also implies the existence of a visual relation "Chomsky Hierarchy" which taxonomizes visual rules according to which machines express them efficiently. We also demonstrate that, though the addition of a so-called "relation network" (RN) module helps make CNNs more systematic by improving their ability to generalize visual rules to objects in arbitrary locations, unsystematic overfitting to particular object features remains. What's more, we show how the RN sidesteps the problem of feature binding by assuming visual objects are of stereotyped shape and scale. Collectively, our results computationally recapitulate the consensus of the psychological and neuroscientific literature that systematic relation detection at the level of biological vision requires a perceptual grouping/feature binding mechanism for the formation of perceptual objects followed by manipulation of these objects by attention and working memory.

2.1 Fodor & Pylyshyn: Then and Now

Despite its labyrinthine prose, Fodor and Pylyshyn, 1988 remains a thrilling and inspiring injunction to investigate the conceptual foundations of neural networks whose pugnacity has been rarely matched[3]. Its authors mount a subtle and multifaceted critique of the Second Great Awakening of neural network modeling of the mind in the 1980s, the first being that of McCulloch and Pitts, Hebb, Minsky, Rosenblatt and others. It is clear that the authors intend some of their critiques to be fundamental and some of them to be corollary, and indeed the most fundamental of them has neither been fully understood let alone resolved. We will outline both types of critiques below and insist on the importance of the latter. At the outset, we must note that

[3]See Marcus, 2018

the source of much confusion about their polemic is Fodor's and Pylyshyn's claim that the neural networks of the 1980s were deficient at the *cognitive level*, by which they meant the level of analysis whose fundamental items were semantically evaluable representations and the processes acting on them. Fodor and Pylyshyn were adamant that these networks were not deficient at the *implementational level*, since networks of neurons with spreading activation can easily implement the "classical" machines to which the authors are obviously wedded, namely the Turing and von Neumann machines[4]. It is rather important keep the distinction between critiques at the cognitive and implementational levels in mind while evaluating the soundness of Fodor's and Pylyshyn's argument.

What is the nature of this deficiency? According to Fodor and Pylyshyn, 1988, representations in biological minds exist in two sorts: an atomic sort, which is not reducible to semantically evaluable sub-parts, and a complex sort which is recursively assembled from atomic representations. Representation in biological cognition, in their view, therefore comprises a combinatorial syntax and the production rules of this syntax are what we have called "routines" above, in an homage to visual perception (Ullman, 1984). In subsequent work Fodor and Pylyshyn, 2014, Fodor and Pylyshyn make the stronger claim that it is in fact *only* these routines and the manner in which they mechanically respond to the syntactical structure of complex mental expressions that endow representations with meaning. This syntactical reductionist claim à la Chomsky (Chomsky, 2009) is a form of the "routine precedes representation" maxim used above. Fodor and Pylyshyn emphasize that complex representations, those assembled from atomic representations, must literally contain those atoms as constituents and that this part-whole relationship must have a structural analogue in the physical implementation of the cognitive architecture. This distinguishes complex

[4]See Hyötyniemi, 1996 for a constructive proof of the equivalence of recurrent neural networks and Turing machines

representations from neural network features which are excited by the activity of afferent units. For instance, Fodor and Pylyshyn would consider the Boolean formula $A \wedge B$ to be a complex representation with real constituents $\{A, B\}$ (assuming A and B are meaning-bearing symbols) but would consider the activation of a neuron representing a face as a result of afferent activity in mouth- and eye-representing units to be merely an atomic feature without real constituents. In the former case, the logical deduction $A \wedge B \implies A$ can be drawn by the mechanical operation of a classical computer which can physically access the constituent "A", whereas the afferent eye and mouth features causally connected to the face unit remain largely inaccessible. Here, Fodor and Pylyshyn join Gallistel and King, 2009 in insisting that cognition hinges on the "computational accessibility" of representations.

Perhaps most contentious is Fodor's and Pylyshyn's central claim that the mind cannot be "both a combinatorial representational system and a Connectionist architecture at the cognitive level" (Fodor and Pylyshyn, 1988, p. 17) The authors are not particularly forthcoming with their definition of "connectionist architecture at the cognitive level," though one surmises they mean something like a system in which meaning-bearing representations are manipulated and evaluated by spreading activation in a neural network. As we have already mentioned, it is perfectly possible to create a combinatorial representational system in connectionist hardware, but Fodor and Pylyshyn contend that a cognitive theory of such a system cannot be grounded in a connectionist vocabulary. The authors are, therefore, primarily concerned with the meta-scientific project of choosing a level of analysis and a technical vocabulary for the interpretation of cognitive systems as opposed to the construction of those systems in the first place.

Far more relevant to the contemporary machine learning practitioner is the interconnected set of cognitive behaviors which, according to Fodor and Pylyshyn, thinking machines (be they classical or connectionist) must strive to emulate: pro-

ductivity, systematicity and compositionality. We say that a cognitive architecture is "productive" if, despite having a finite vocabulary of atomic representations, "[t]he representational capacities of such a system are, by assumption, unbounded under appropriate idealization" (Fodor and Pylyshyn, 1988, p. 21). These infinite abilities with finite means are typically explained by a recursive syntax. Consequently, he quintessential example of representational productivity comes from natural language: intuitively, there is no longest sentence, and the ability of humans to produce or interpret a sentence is limited only by practical factors like working memory and attentional capacity. These limitations can, in turn, be incrementally alleviated with special tools and methods (e.g., a pencil for writing). Second, a cognitive architecture is "systematic" if the ability to represent given expressions automatically entails the representation of perhaps an infinite set of other expressions. Continuing the linguistic examples, Fodor and Pylyshyn note that there is no one who understands "John loves Mary" without understanding "Mary loves John" or "Billy loves Alice", for that matter. Finally, a cognitive architecture is "compositional" if the meaning of complex expressions is determined by the meaning of the constituents, excepting idioms (e.g. Compare the meanings of "He kicked the wall" to "He kicked the fence" to "He kicked the bucket".).

Of these three properties, systematicity is perhaps easiest to quantify and therefore lends itself without much difficulty to experimental investigation. Indeed, it is superficially similar to the familiar concept of invariance from perception and neural network design (Riesenhuber and Poggio, 1999), though there are key differences. For their part, Fodor and Pylyshyn explain the systematicity of thought in classical terminology: representational systems automatically understand systematically related expressions because 1) the expressions share a syntactical form and 2) the constituents of the expressions roughly maintain their meaning independent of context. A systematic representational system is easy enough to imagine in classical computing

hardware where syntax is explicit model[5], and Fodor and Pylyshyn, 1988 spend a good deal of time explaining how this behavior is supported by the classical computer's sequential "storing, retrieving, or otherwise operating on structured symbolic expressions". The appearance of systematic representation in more neurally plausible systems, however, has been the subject of some debate. For example, Chalmers, 1993 has argued that Fodor and Pylyshyn, 1988's line of reasoning forbids all neural networks from being truly systematic, even if they are made to implement classical machines, and subsequent critics (Aizawa, 1997; Matthews, 1997) pointed out that classical machines themselves are not necessarily systematic. The rebuttal mounted by Fodor and colleagues has typically been that neural networks, so far, only exhibit systematic behavior as a matter of chance and not of logical necessity (Fodor and McLaughlin, 1990), compared to biological minds and classical computers, where, they argue, systematicity seems to come "for free". For instance, it is known that children as young as seven months can recognize phrases with familiar structures, even if the phrase constituents are novel (Marcus, Fernandes, and Johnson, 2007). In machine learning terms, systematicity represents an extreme form of "out-of-distribution" (Krueger et al., 2020) generalization, and, indeed, there have been numerous attempts to model systematicity as generalization "for free" in neural networks (see Jansen and Watter, 2012 and (Calvo and Symons, 2014) for somewhat recent reviews). Needless to say, the ability of neural networks to generalize to truly novel inputs, including those that share a "syntactic structure" with training data, is not entirely understood and numerous interesting failures continue to be reported (e.g. Szegedy, Zaremba, and Sutskever, 2013), even in networks designed explicitly for systematic cognition (Loula, Baroni, and Lake, 2019).

[5]For instance, in the computer language Python, the expression '1+1' is systematically related to '2+2' since 1) they share a syntactic structure defined by the '+' operator and the semantics of Python ensure that every integer type is interpreted equivalently by '+'.

It could be argued that adequate modeling of systematicity is only a serious issue for uncontroversially rule-based domains, like language, and that traditional neural network architectures are perfectly sufficient for other domains, like vision. After all, vision models which are openly sympathetic to the classical cognitive position of Fodor and Pylyshyn, 1988, like visual grammars (Zhu and Mumford, 2006), parts-based models (Biederman, 1987; Yuille, Hallinan, and Cohen, 1992) and compositional machines (Jin and Geman, 2006), are hopelessly outperformed on recognition tasks by today's deep networks, notably the CNN, which are in many ways straightforward extensions traditional networks of the 1980s (notably Fukushima, 1980). However, it could very well be the case that machine vision has simply not progressed, despite all appearances and announcements to the contrary, to a stage in which questions of complex syntactical expressions and deduction are even tractable. As we have already suggested, most machine vision research in the past decade has been dedicated to the problem of image classification, and indeed there is strong evidence that the CNN is a good model of pre-attentive visual processing (Serre, Oliva, and Poggio, 2007)[6]. Despite some important but imperfect attempts to model more complex visual reasoning (see Ricci, Cadene, and Serre, Submitted for a review), image classification still dominates the attention of machine learning theorists and experimentalists.

Nevertheless, very little of our intuitive visual experience involves the classification of static images, though this is impossible to quantify precisely. A rather larger portion of our conscious visual experience consists of dynamic interaction with unfamiliar objects in a structured environment. We may walk down the street, see a large piece of machinery perched on a ledge above us and know we must take another route to avoid danger. From inspecting the shape of a keyhole, we might realize that we have

[6]Though the CNN is still considered the standard model of biological visual recognition, there is somewhat recent work showing that task-optimized recurrent models are better predictors of primate neural activity (Yamins et al., 2014)

been using the wrong key to enter our home or the wrong key upside-down. In short, the visual world has structure, just as language has structure, and the combinatorial explosion resulting from the numerous objects of the visual world and the many relations among them necessitates a serious discussion of systematic cognition. Today, neural networks excel at labeling individual objects in an image, but this is nothing more than the construction of "atomic" representations in the sense of Fodor and Pylyshyn, 1988: far more interesting and relevant to visual experience is the way in which we dynamically combine these atoms into structured representations. And, unless we want to embark on the quixotic journey of learning a "grandmother neuron" (Quiroga et al., 2005) for every possible relationship among any number of objects[7], we should take seriously Fodor's and Pylyshyn's intuition about mechanisms for "storing, retrieving, or otherwise operating on structured symbolic expressions", even for static images.

Just as CNN-based image classification mirrored a long and important literature on rapid visual classification in the mammalian cortex inspiration, it will be very useful to curate a corresponding literature for the creation of visual machines where systematicity is more obviously required. To that end, we note that the typical examples of systematicity in language involve binary relations among lexical items which are largely understood by force of the relation's syntax and not by recourse to the meaning of the constituents: we understand "X loves Y", where X and Y are people's names if we understand "John loves Mary". Are there similar examples in visual cognition? The answer is emphatically "yes", though the literature on so-called "visual relations" has rarely been brought into contact with the theoretical material discussed here. In the next section, we will review this literature in detail. Following that, we will propose several quantitative measures of systematicity in neural networks

[7]Via personal correspondence, Randy Gallistel has described this as the "Learning one representation for a red Volkswagen and one for a red Volkswagen with a dent in the hood" approach.

and explore how these networks fare by this measure on a series of visual relations tasks.

2.2 Visual Relations: A Review

In this section, we will briefly outline the basic results from the psychophysical, neuroscientific and machine learning literatures on visual relations. A few persistent themes will be observed: serial vs. parallel relation processing, the explicitness or implicitness of relational representation, various roles for attention and working memory, etc. The notion of systematicity is not often explicitly invoked by scholars studying visual relations, but we will note wherever the concept is secretly at play. Our goal is to curate a set of benchmarks for systematic relational understanding in biological intelligence so that the deficits we will subsequently find in machines seem all the starker.

2.2.1 Psychophysics

Below, we review a set of psychophysics experiments organized by the types of visual relations they investigate. We want to emphasize the formal properties of relations themselves since, as we will argue in a later section, we suspect these properties will play an important role in the development of a theory of visual relations. Here, the three types of relations we explore are 1) same-different relations, 2) same-different relations under transformation and 2) spatial relations.

Same-Different Relations The capacity to judge whether items in a scene are the same or different is fundamental to relational processing. William James noted the centrality of same-different relations to human reasoning when he claimed "the sense of sameness is the very keel and backbone of our thinking" (James, 1890). This

sentiment was reiterated by (Nickerson, 1978), who stated "without the ability to make such [same-different] comparisons there could be no perception as we know it". Further, as the "sameness" of two objects in an image is a rather abstract property, determined "regardless of the *particular qualities* of the stimuli" (Delius, 1994), it is unlikely that the mechanism underlying same-different judgments relies on the glorified template matching found in contemporary feedforward vision modeling. Instead, an intuitive mechanism for computing same-different judgments is some algorithm which dynamically selects objects in a scene, stores their representations in memory, and then compares them. Psychophysical investigation of human same-different processing has probed for the existence of such a comparative mechanism and has tested its speed and performance in numerous stimulus conditions.

In a typical same-different task, subjects view an image containing two or more items and are asked to determine whether or not all of the items are "the same". The definition of sameness varies between experiments: items are either completely identical (Donderi and Zelnicker, 1969; Donderi, 1983) or the same on some subset of cued dimensions, like color or shape (Eriksen, O'Hara, and Eriksen, 1982; Eviatar, Zaidel, and Wickens, 1994). Stimuli from a classic experiment of (Donderi and Zelnicker, 1969) are depicted in Fig. 2.1. Items can either be presented simultaneously in one stimulus or sequentially, across several frames. Here, we only consider the case of simultaneously presented items.

Early psychophysical work on same-different processing in humans was explicitly cast as a form of visual search (Sternberg, 1966; Treisman and Gelade, 1980) in which subjects had to determine whether or not all items in a stimulus were identical. As in visual search, the central phenomenon of interest was the reaction time in determining the homogeneity or heterogeneity of items in the visual field. However, unlike the typical search paradigm, subjects are only tasked with detecting the uniformity of all stimulus items and not the presence of a particular feature. This raises some interesting

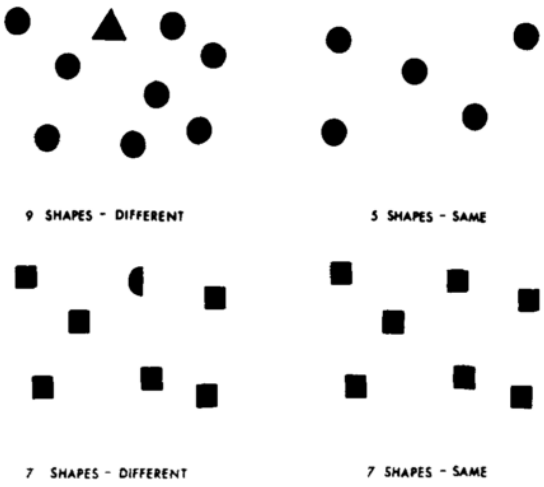


Figure 2.1 (Donderi and Zelnicker, 1969) tested subjects on a rapid same-different task. Pictured are four stimuli, two different (left panels) and two same (right panels). Notice that the different item in the lower-left stimulus actually shares a feature (flat right side) with the other items. This was true of several stimuli, but the authors report no effect from these shared features. Figure taken from (Donderi and Zelnicker, 1969)

questions about the processing of abstract visual properties. For example, can an abstract relation like sameness "pop out" in the manner of a shape or color? Further, are the visual processes underlying search in the traditional setting identical to those in relation detection or are there new roles for mechanisms like working memory? For example, a traditional visual search task requires that subjects simply remember throughout a trial whether they have seen a target stimulus, but a same-different task intuitively requires that features from past items in a scene be kept in memory for comparison to new items.

The key dependent variables in the same-different paradigm are subjects' accuracy and reaction time (RT). Subjects are typically told to respond by declaring "same" or "different" as fast as possible. For example, (Donderi and Zelnicker, 1969) tested subjects on a same-different task with visual stimuli having 2 to 13 items (e.g. Fig.2.1, bottom panels, shows 7 items) at a stimulus onset asynchrony (SOA) of 50 ms. They found that accuracy was high for all numbers of items and, importantly, reaction time for correct decisions was constant at about 1.3 s. (Donderi and Zelnicker, 1969)

concluded from their subjects' constant reaction time that same-different judgments occur in parallel across the visual field, without the need for visual search, at least in the case of what the authors called "codable" items. A codable image item, according to the authors, is one that is easily nameable or recognizable (e.g. like a letter or digit). In their words, "[a] highly codable stimulus has a consistent association with a single verbal response" (Donderi and Zelnicker, 1969, p. 197). The authors do not elaborate further, but they likely want to distinguish between their stimulus items (simple, easily describable shapes) and random curves, polygons, etc. In the parlance of contemporary psychophysics, we might instead refer to codable items as having hardwired features in the visual system. This distinction between codable and non-codable stimuli was given a feature-based interpretation by (Treisman and Gelade, 1980) in the context of visual search. Note that (Donderi and Zelnicker, 1969) did not systematically vary the types of items used in stimuli and so could not probe the robustness of subjects to variance in item shape, color, etc. In short, systematicity was not yet an object of experimental interest to (Donderi and Zelnicker, 1969) or indeed many subsequent researchers.

(Donderi and Zelnicker, 1969) noted that subjects were faster on correct "different" declarations than on correct "same" declarations, except when stimuli had 11 items. The relative quickness of "different" judgments seems intuitive. Whatever parallel process checks for "same" vs "different" can terminate as soon as a single mismatching item is found. The relative fastness of correct "different" responses was later found, surprisingly, to be somewhat anomalous by (Nickerson, 1972) and others. Often, "same" judgments are faster than "different" judgments, though subjects are more likely to make false "different" responses to "same" pairs than false "same" responses to "different" pairs. Work by (Nickerson, 1968; Krueger, 1973b; Thomas, 1974) suggested that this "fast-same" effect might arise from priming. For example, in a same-different task with stimuli consisting of either 10 As or 9 As and one B, a

subject will simply be exposed to the A item more, and may be primed to respond more quickly when viewing a "same" stimulus, as it contains only the primed item. However, even when stimuli are balanced so that items appear with equal frequency throughout the task, the fast-same effect persists (Nickerson, 1968; Krueger, 1973b).

These paradoxical findings were explained in part by a model of (Krueger, 1978) grounded in signal detection theory. Krueger hypothesized that in a single glance, subjects can contrast two neighboring stimulus items to check for mismatching features. The number of mismatching features is compared to two thresholds (see Fig. 2.2), corresponding to three possible actions. If the number of mismatches is lower than both thresholds, the observer will detect that the two items are the same and either proceed to the next pair of items or declare the stimulus to be "same" if all previous pairs have been checked. If the number of mismatches is higher than both thresholds, the observer will detect that the two items are different and automatically declare the image to be "different". However, if the number of detected mismatches is between both thresholds, the pair of items is ambiguous and must be rechecked for mismatches. Subjects in same-different tasks are known to recheck their decisions based on spurious mismatches (Burrows, 1972; Krueger, 1973a; Silverman, 1973). Subjects recheck on both "same" and "different" stimuli, as evidenced by the fact that subjects very often know when they have made an error (Laming, 1968; Bamber, 1972; Bamber and Paine, 1973; Snodgrass, 1972).

(Krueger, 1978) used this signal-detection framework to explain the fast-same effect by noting the way noise differentially affects "same" and "different" images. For example, if both "same" and "different" images have the same distribution of perceived differences but with different modes (Fig. 2.2a), then rechecking will be equally likely for both types of images. This is because the proportion of samples from both "same" and "different" stimuli in the ambiguous inter-threshold region is equal. In the case of high perceptual noise, the distribution of perceived differences for "same"

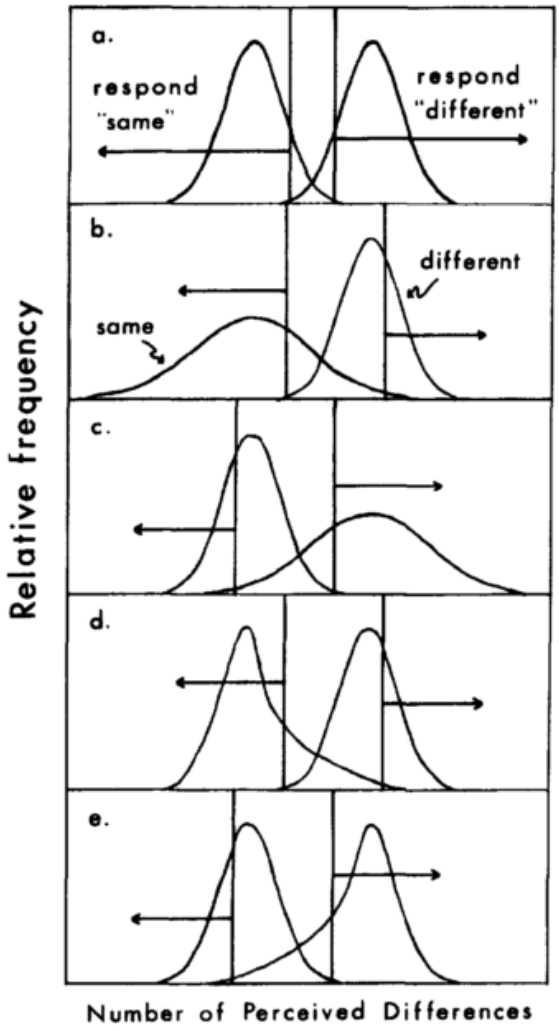


Figure 2.2 Impact of noise on a same-different observer in a signal detection paradigm. (Krueger, 1978) explains how comparatively rapid same vs different judgments can occur in light of both high variance in "same" images, from perceptual noise, and high variance in "different" images, from experimental manipulation. The case in panel d is argued to underlie the fast-same effect: the bulge in the "same" distribution makes observers shift their thresholds rightward. The "same" distribution will likely have a positive skew since these images have no true differences, only perceived ones. Since the skew causes an asymmetry in the "same" distribution, more of the "different" distribution than the "same" distribution will move into the ambiguous region. This will cause more rechecking on "different" images. Figure taken from (Krueger, 1978).

images has high variance (Fig. 2.2b), forcing the observer to slide his thresholds rightward, thereby ambiguating more "different" samples and disambiguating more "same" samples. The disambiguation of more "same" samples should result in less rechecking of "same" stimuli and therefore faster "same" responses. Further, the intrusion of the "different" distribution into the inter-threshold region will give rise to more false "different" responses on "same" images. Fig. 2.2c depicts the opposite case, in which experimentally-controlled variance in "different" stimuli is high, resulting in faster "different" responses. Fig. 2.2d shows the realistic case of a positively-skewed "same" distribution, with a bulge near the low end of perceived differences. Again, this case will give rise to a faster "same" response and increased false "different" responses because of the ambiguity of the "different" distributions. This case is realistic since "same" images, by nature, will produce a small number of perceived differences, having literally identical items. The opposite case is shown in Fig. 2.2e.

(Krueger, 1978) argues that the fast-same effect obtains in cases like Fig. 2.2d. Subjects are prone to recheck items in both "same" and "different" stimuli, depending on the number of perceived differences, but perceptual noise uniquely produces low numbers of mismatched features in the "same" case, since identical items are physically equivalent in the first place. This gives "same" a strong RT advantage over "different", though it produces more false "different" responses on "same" stimuli.

Krueger used this theory to create a model which recapitulated known RT and accuracy effects from the same-different literature on a data-set of binary images. Like Krueger's imagined observer, the model noisily samples items from a stimulus, counts their mismatching pixels, and returns a judgment according to its thresholds. Krueger's model samples pairs of items in sequence, and thus runs somewhat counter to (Donderi and Zelnicker, 1969)'s claims regarding the parallel nature of same-different judgments. Further, Krueger sought to explain the fast-same effect, whereas (Donderi and Zelnicker, 1969)'s produced the opposite result. On the issue of rapid parallel

processing in same-different tasks, Krueger's theory is silent. However, the fast-different effect observed in (Donderi and Zelnicker, 1969) comports with Krueger's model. In Fig. 2.2d, we saw how high perceptual noise would favor a fast-same judgment. Conceivably, this is just the type of noise that would arise in the case of stimulus items for which there are no good hardwired features in the visual system, i.e., for non-"codable" items in the sense of (Donderi and Zelnicker, 1969). This suggests the appearance of the fast-same effect for synthetic or generally unfamiliar items and the fast-different effect for natural or familiar items. More recent evidence (Wiley, Wilson, and Rapp, 2016) indicates that practice can improve subjects' performance on a same-different task with unfamiliar stimuli, effectively "coding" the stimulus items over the course of the experiment.

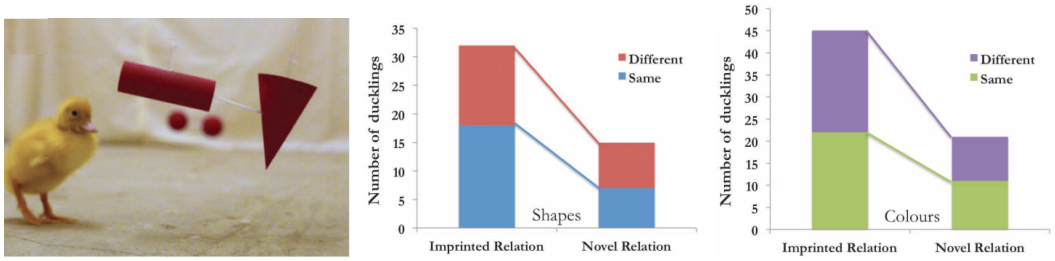
The signal-detection approach of (Krueger, 1978) has received substantial empirical support (Donderi, 1983), but it glosses over the mechanism used for comparing individual items. Krueger's model simply takes the Hamming distance between two binary arrays and compares this distance to a threshold. However, it could be that the two items agree on some "dimensions" (e.g. shape, color, etc.) but disagree on others, and this partial matching might have an effect on the same-different judgment. It is known, for example, that when subjects are asked to make a same-different judgment according to a cued dimension, the task-irrelevant dimension can affect performance on the task-relevant dimension (Hawkins and Shigley, 1972; Egeth, 1966; Hawkins, McDonald, and Cox, 1973; Williams, 1974; Overmyer and Simon, 1985). These studies indicate that RT for "same" judgments for a cued shape dimension increases when the two items differ on the task-irrelevant color dimension. Likewise, reaction times for "different" judgments for a cued shape dimension decreases when the two items are different colors compared to when they are the same color.

Two models have been offered to explain the effect of task-irrelevant dimensions on same-different tasks. (Eriksen and Schultz, 1979) proposed a "response competition"

model, in which similarity between items across all dimensions, task-relevant or -irrelevant, are computed gradually and in parallel. The similarity between two items on a given dimension is conceived of as some continuous quantity which may exceed a response threshold. A same-different judgment is produced as soon the relevant similarity measure crosses threshold. If the comparisons on task-relevant and -irrelevant dimensions produce incongruent results (i.e., "same" on one dimension vs. "different" on the other), then a competing response is simultaneously primed. This competing response delays the correct response.

(Eviatar, Zaidel, and Wickens, 1994) proposed an alternative "confluence" account, which posited same-different judgments occur according to discrete stages instead of a continuous flow of computations. Like the response competition model, items are compared in parallel across all dimensions. However, only after comparisons across all dimensions are complete is a second decision stage reached, in which response times are facilitated by congruent comparisons and inhibited by incongruent comparisons.

The models agree that comparisons across all dimensions are computed simultaneously and in parallel, but they disagree about when the final same-different response is activated, as measured, for example, by EEG (Zhang et al., 2013). There seems to be more evidence in support of the confluence rather than the response competition model (Donderi, 1983; Zhang et al., 2013), though for our purposes it suffices to note the connection between dimensional congruency and RT. Same-different perception is not unique to humans, and it is in frankly in non-human studies where the emergence of systematicity is clearest. Same-different detection has been observed in non-human animals, namely birds and primates(Katz and Wirght, 2006; Daniel, Wright, and Katz, 2015; Martinho III and Kacelnik, 2016), honeybees (Giurfa et al., 2001), and rats (Wasserman, Castro, and Freeman, 2012). For a more detailed review of same-different reasoning in animals, see Wasserman, Castro, and Freeman, 2012. Most of these studies involved some type of reinforcement learning, though the newly hatched ducklings



(a) **(b)** **(c)**

Figure 2.3 *Same-different reasoning in ducks (a)* Martinho III and Kacelnik, 2016 tested the ability of day-old ducklings to imprint on the abstract relation of visual sameness. During a training phase, newly hatched ducks were placed in a chamber with a single pair of simple 3D shapes hanging from strings. The objects either had the same/different color or same/different shape. Later, the ducks were placed in the chamber with a novel pair of objects either obeying or disobeying the observed relation from the training phase. *(b,c)* Number of ducklings to follow the observed versus a novel relation in the testing phase. In both the color and shape conditions, ducklings were significantly more likely to follow the imprinted relation, even though the objects in the testing phase were different than those of the training phase and despite having imprinted on one example.

used in (Martinho III and Kacelnik, 2016) acquired a notion of same vs different from a single training example. After imprinting on a single pair of same/different objects, the ducklings preferred to follow novel objects obeying the same relation Fig. 2.3. This indicates the ducklings could generalize a same-different rule to novel items with essentially no training; that is, the ducks had a systematic understanding of an abstract visual relation "for free". A particularly striking example of systematic same-different understanding comes from the bee study of (Giurfa et al., 2001). A population of honeybees was trained on a simple same-different task involving the matching of colors or odors. Bees lighted at the base of a Y maze, where a visual or olfactory stimulus was placed (Fig. 2.4). At the fork in the Y, a second stimulus was placed which was either the same or different color/odor. The bees were reinforced to take one leg of the fork depending on whether or not the observed relation was "same" or "different". Gradually, the bees learned the pattern and were, remarkably, able to transfer their knowledge to novel colors and odors. Much more remarkable, however, was the fact

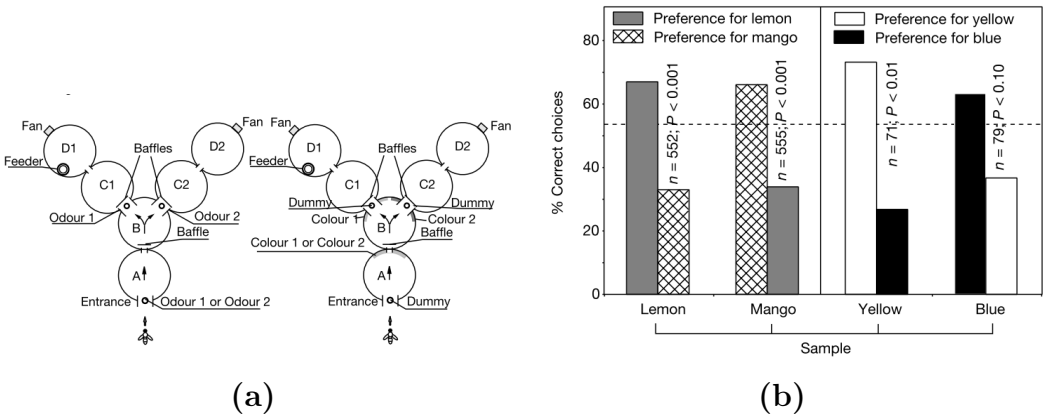


Figure 2.4 *Systematic relation transfer in bees. (a)* Bees entered a Y maze and were confronted with an olfactory (left) or visual (right) stimulus. At the fork in the Y, the same or a different stimulus was present. Depending on the condition, bees either had to pursue the branch marked by the same or the different stimulus to receive a reward. *(b)* Bees first learned the olfactory condition and demonstrated a significant propensity to follow the cued relation. Then, during a testing phase, bees were placed in the visual condition and rapidly developed the correct preference for the cued relation, indicating they had transferred relational knowledge from the olfactory to visual domain.

that bees could rapidly transfer their knowledge from olfactory same-different *to visual same-different*. This suggests honeybees are capable of representing an extremely abstract notion of sameness, one we might even call systematic. Keep in mind that bees have neither a hippocampus nor an entorhinal cortex.

Same-Different Relations Under Transformation In a pure same-different task, items in a visual scene are declared the "same" if they are physically identical along a cued dimension. The ability to recognize identical items in a scene is evidently very useful, though it is limited to very artificial settings. Most natural same-different judgments involve items that are only identical up to changes in rotation, scale, illumination or other transformations. For example, two road signs might seem identical up to scale due to distance from a driving observer. While a portion of the observer's sameness judgment likely arises from the translation and scale invariance of the ventral visual stream (Riesenhuber and Poggio, 1999), it cannot be completely

explained by such invariances. For, invariant object recognition requires invariant features for *particular* objects, whereas, as we will see below, humans seem to make same-different judgments under transformation for arbitrary, novel objects. Instead, the ability to detect same-different relations under transformation is akin to the systematic understanding of a more-or-less universally applicable rule. Moreover, when humans detect that objects lie in a given same-different relationship up to transformation, they can still recognize that the items are not visually identical. Hence, unlike in the case of pure invariance, no information is lost. In Fodorian language, the constituents of the complex expression (that is, the relation) remain computational accessible.

Classic work by (Shepard and Metzler, 1971) studied the capacity of humans to detect identity under rotation for line drawings of 3D objects. Eight subjects were each shown 1600 pairs of images (Fig. 2.5) depicting strings of 10 cubes in a 3D space and asked to determine if the strings were congruent up to rotation. Matching image pairs were constructed by rotating cube-strings either in the 2D viewing plane ("plane" condition) or in depth ("depth" condition). Non-matching image pairs were constructed by both a rotation and a reflection, so that the objects could not be rotated into congruency. Subjects were asked to respond as quickly as possible.

Overall, accuracy was quite high, at 96.8%, and although the authors do not report accuracy for the "plane" vs "depth" condition, these marginal accuracies must have both been over 90%. The key finding from this study was that subject RT per stimulus pair was a robustly linear function of the angular difference between each image in both the "plane" and "depth" conditions. Somewhat surprisingly, the function relating RT to angular difference in the "depth" condition had lesser slope than the corresponding function in the "plane" condition, indicating that subjects could detect identity under depth rotation at large angular deviations faster than under plane rotation.

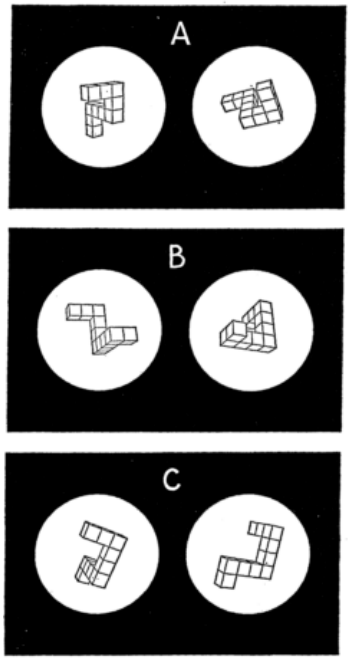


Figure 2.5 *Mental rotation.* Three stimuli from a rotation matching experiment. In the top panel, the two block-strings can be rotated in the picture plane into congruence. In the middle panel, the two images can be rotated in depth into congruence. In the lower panel, the images are congruent, but only up to a rotation *and* a reflection. Figure taken from (Shepard and Metzler, 1971).

Shepard and Metzler conclude that subjects mentally rotate images in real time to test for congruency. Indeed, their subjects reported in post-experiment interviews that "they attempted to rotate one end of one object into congruence with the corresponding end of the other object [and] discovered that the two objects were different when, after this 'rotation,' the two free ends still remained noncongruent". (Shepard and Metzler, 1971, p.703). The linearity of RT, indicating active mental rotation, is a rather robust finding, having been replicated by (Cooper, 1975; Cooper, 1976; Cooper and Shepard, 1973; Shepard and Cooper, 1982). These studies further demonstrated that the linearity of RT was independent of subjects' being explicitly told to use mental rotation strategies.

Despite these numerous replications, mental rotation and mental imagery in general remain somewhat controversial. For example, these phenomena cannot be completely "visual" *per se* as similar effects have been observed in congenitally blind human

subjects (Marmor and Zaback, 1976; Carpenter and Eisenberg, 1978). Furthermore, attempts to replicate mental imagery phenomena in non-human animals, like pigeons, parrots and monkeys, have been largely unsuccessful (Hollard and Delius, 1982; Burmann, Dehnhardt, and Mauck, 2005; Nekovarova et al., 2013), except in the case of sea lions (Mauck and Dehnhardt, 1997; Stich, Dehnhardt, and Mauck, 2003).

A more recent exploration of the ability to detect same-different relations under transformation was conducted by (Fleuret et al., 2011). Fleuret and colleagues compared humans and machines on a synthetic visual reasoning test (SVRT) involving classes of stimuli each structured by one of twenty-three visual relations (see Fig. 2.6 for examples). Only some of these classes of stimuli involved "same-different under transformation" rules, and we focus on only these stimuli here. For example, one of the test's "problems" involved distinguishing stimuli featuring two items that were the same up to translation, scaling and rotation from stimuli disobeying this rule (Fig. 2.6). In general, all stimuli involved black closed curves with wiggly edges on a white background, and each of the 23 problems was designed to require relational- rather than feature-based judgments. That is, Fleuret and colleagues attempted to make the stimuli varied enough so that problems could not be solved by simple template matching strategies. In other words, the authors were implicitly probing what we have called "systematic" relational understanding.

Twenty subjects took the SVRT by using the computer interface shown in Fig. 2.6c. On each trial, a new stimulus was displayed in the top panel and subjects were told to place each new stimulus into one of the two category boxes in the lower panels. Trials were unspeeded. Feedback was provided after each trial according to a subject's categorization consistency. Correctly classified stimuli remained in the category boxes as a mnemonic device. Trials continued either until the subject made seven correct responses in a row, resulting in "success", or until the subject saw 35 trials without

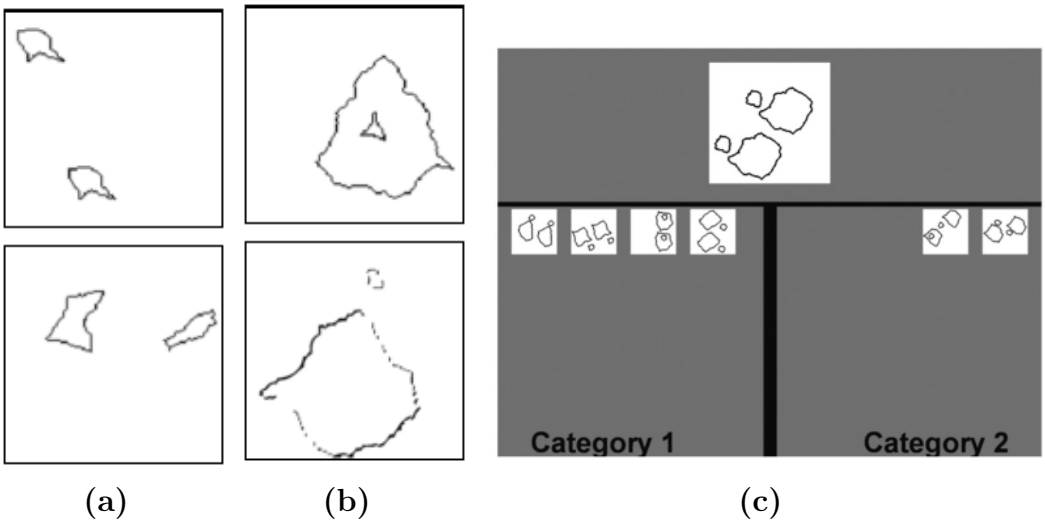


Figure 2.6 *The Synthetic Visual Reasoning Test.* Stimuli and user interface from (Fleuret et al., 2011). *a*) Stimuli from the "same-different" problem (SVRT 1). In the top panel, the two items are congruent. Items in the bottom panel are incongruent. *b*) Stimuli from the "inside-outside" problem (SVRT 4), with the top panel as a positive example and the bottom, negative. Note that this problem is not a "same-different up to transformation" problem. *c*) Stimuli appeared in the top section of this user interface. Subjects had to drag the stimulus into one of two boxes labeled Category 1 and Category 2. They were provided with feedback as to the consistency of their decisions. The last several correct classifications were displayed in the lower boxes as a mnemonic aid. All panels taken from (Fleuret et al., 2011).

"success", resulting in "failure". Each of the twenty subjects did all 23 problems.

Four problems proved particularly difficult for subjects, though performance on the remaining 19 was strong, taking an average of 6.27 ± 0.85 trials before learning. Out of all 23 problems, 7 of them were "same-different up to transformation" problems: problems 1, 5, 16, 19, 20, 21 and 22. Of these problems, only two, problems 16 (12 subjects failed) and 21 (9 subjects failed), were hard for subjects. Problem 16 used stimuli with bilateral symmetry about a vertical axis and problem 21 used stimuli in which two items were identical up to scale, translation and rotation.

(Fleuret et al., 2011) also tested two standard machine learning algorithms, Adaboost (Freund and Schapire, 1995) and a support vector machine (SVM) with a Gaussian kernel, on the SVRT. Only Adaboost (see Fig. 2.7) performed somewhat acceptably. The authors measured the testing error rate of Adaboost as a function of the number of training samples (Fig. 2.7a) and also as a function of three increasingly sophisticated pre-processing techniques (Fig. 2.7b): 1) numbers of pixels in rectangular windows of varying size, 2) edge statistics and 3) wavelet coefficients. Though performance improved with more samples and richer features, only 7 problems had less than 6% error at 10,000 examples with the most sophisticated features. Further, 4 of the "same-different up to translation" problems noted above remained above 25% error at this number of training examples and feature type. Fleuret and colleagues noted the special difficulty of these problems and emphasized the comparative success of human subjects.

Yet, it is somewhat difficult to compare human and machine performance on this task, since there is no balanced measure of learning rate. Arguably, human subjects have had a lifetime of experience solving such problems, so they may have come to the experiment having already seen millions of samples. Moreover, the human visual system likely incorporates far more sophisticated features than those used by Adaboost

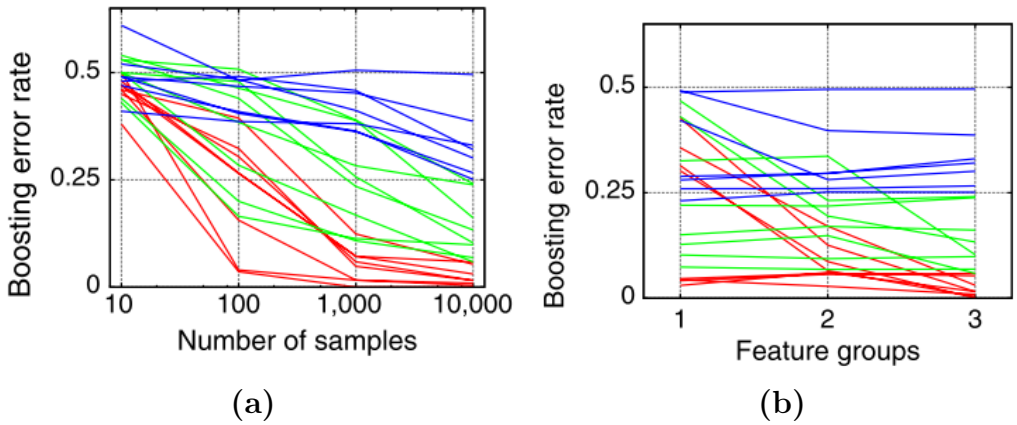


Figure 2.7 *Blackbox algorithms on SVRT.* Boosting performance on the SVRT from (Fleuret et al., 2011). *a*) The left panel depicts the error rates for Boosting using feature group 1 (see main text) for various training set sizes on a log scale. The authors grouped problems into three clusters (blue, green, red), corresponding to their difficulty. *b*) The right panel shows boosting performance with 10,000 training samples for three increasingly rich feature representations (see main text). Again, problems are grouped into three sets by difficulty. Figures taken from (Fleuret et al., 2011)

in (Fleuret et al., 2011).

As we will see in Sec. 2.3, the result of (Fleuret et al., 2011) generalizes to more sophisticated learning machines inspired by the feedforward sweep of the visual cortex. The large difference between humans and machines on this task, suggests that humans may detect relations using mechanisms wholly different from those employed by traditional statistical learning methods, i.e., feature-based classifiers. Two candidate mechanisms, attention and working memory, are explored in more detail in following section.

Spatial Relations Two items in a scene can be the same or different independent of their location. For example, (Donderi and Zelnicker, 1969) used stimuli with items scattered at random locations (Fig. 2.1), whereas (Shepard and Metzler, 1971) placed items at two fixed locations (Fig. 2.5). Simply put, a same-different relation incorporates no spatial information.

However, there are many visual relations for which spatial information is key. When one notices a pen on a table, or a cat beneath a chair, one has recognized

that items in a visual scene exist in a particular spatial configuration. We refer to this subset of visual relations as *spatial relations*. Spatial relations disregard the sameness/difference of the items in the relation and simply rely on the given items' satisfying a spatial rule. Moreover, as any two items can conceivably exist in an infinite number of spatial relations, whatever mechanism humans employ to detect such relations must be extremely flexible. It must be able to apprehend arbitrary objects in more-or-less arbitrary relationships.

There is substantial evidence that the apprehension of spatial relations requires attention (Hummel and Stankiewicz, 1998; Logan, 1994; Logan and Sadler, 1996). For example, (Logan, 1994) asked subjects to detect one of four spatial relations between items in a cluttered scene. Subjects were asked questions like "Is there a dash above a plus?" before viewing an image containing dashes and pluses bundled into pairs. Stimuli having the cued relation contained at least one dash above a plus. 24 subjects were randomly asked to detect *above*, *below*, *left*, and *right* relations over the course of 512 trials. The display size and number of items were varied together. (Logan, 1994) chose dash and plus items for stimuli since they have been argued to "pop out" preattentively in a field of distractors (Treisman and Patterson, 1984; Treisman and Souther, 1985; Treisman and Gormican, 1988). However, (Logan, 1994) found that when subjects were asked to detect spatial relations between these items, instead of just the items themselves, RT increased linearly with display size, indicating subjects employed attentional search. In a control experiment for which subjects simply had to detect the presence of a dash among distractors, RT remained constant with display size. Additionally, RT tended to decrease when subjects were cued to the location of the pair of stimulus items obeying the target relation. Logan concluded that attention was required to detect spatial relations, a claim later supported by (Moore, Elsinger, and Lleras, 1994; Rosielle, Crabb, and Cooper, 2002; Evans and Treisman, 2005) and (Holcombe, Linares, and Vaziri-Pashkam, 2011).

Although (Logan, 1994) and subsequent studies demonstrated attention is necessary to detect spatial relations, they remained largely agnostic on exactly how such an attentional mechanism should function. Having shown that relation detection requires serial search through a cluttered scene, these authors implicitly invoked a moving window or "spotlight" of attention, but they only describe its movement through a scene and not its local relation detection mechanism. For, even if an attentional window is correctly placed on two items potentially obeying a spatial rule, there must be some subsequent routine designed to recognize the given relation.

A taxonomy of such routines was provided by (Franconeri et al., 2012) (see Fig. 2.8). Franconeri and colleagues grouped routines for spatial relation detection into two types depending on whether the attentional window selected one or multiple items at a time. For instance, suppose the cued relation is "Cross Left of X" (Fig. 2.8a). An attentional mechanism selecting multiple items simultaneously could place a spotlight over both the cross and X and then compare this attended region to a stored template (Fig. 2.8b). This type of attention would rely on templates composed of hardwired features that happen to resemble the given instance of the spatial relation. However, detecting relations purely with templates is rather inflexible, requiring one template for every pair of objects in every spatial relation. This combinatorial explosion of templates is a classic flaw of connectionist systems trying to encode feature relations (Fodor and Pylyshyn, 1988). So, while there is some evidence for the existence of hardwired visual features for certain frequently occurring spatial relations (VanRullen, 2009), the template approach seems critically limited by its inflexibility and wastefulness.

This wastefulness can be slightly reduced by attaching a coordinate system to the attentional spotlight. In this case, multiple items could be selected by the spotlight, but only one needs to be matched to a template. For example, if the window of attention is fixed on both the cross and the X, then the window's coordinate system

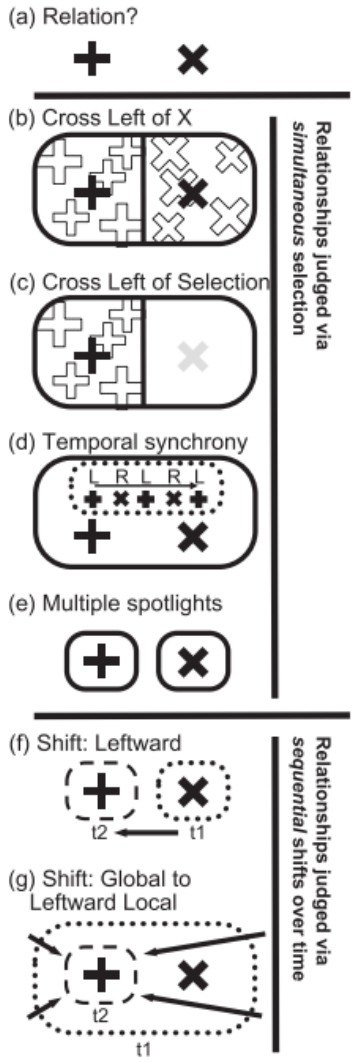


Figure 2.8 *Shifting attention for flexible visual relation processing.* Six attentional window strategies according to (Franconeri et al., 2012). *a*) A window will be centered on one of two items, a cross and an x, to test for a left-right relation. *b*) One solution is to make the attentional window encompass both items and match the items to a template resembling their arrangement. *c*) If the window has a coordinate reference frame, then the origin can be placed over one item and template matching can be performed on the other. *d*) To avoid using templates for the relation, the window could instead allow features from the two items to bind-by-synchrony with units representing the relation. *e*) If multiple spotlights are available, each one could focus on an item and index its location. *f*) However, it the window can only encompass a single item at a time, it might instead shift from one item to another and store the direction of the shift in working memory. *g*) Or, it might encompass both items to get coarse arrangement information (horizontal vs vertical) and then zoom to one item, recording the direction of the zoom in working memory. Figure taken from (Franconeri et al., 2012)

can encode which side of the window is the left (Fig. 2.8c). Whatever item occurs on the left of the window can then be identified with a template as in Fig. 2.8a. This hybrid of templates with spatial reference frames could presumably be implemented by lateral occipital complex (LOC) neurons which are sensitive to both object identity and location (Edelman and Intrator, 2000; DiCarlo and Maunsell, 2003; DiCarlo and Cox, 2007). Like the pure template matching approach in Fig. 2.8a, the hybrid approach still requires that the attentional window select both objects at once since the relation between the items is only defined with respect to the window's reference frame. Note that this method remains quite inflexible: it still requires templates of particular objects in a given relation.

A flexible representation of spatial relations could be achieved by a dynamic network with separate detectors for object identity and location. This network would bind the location of an object to its identity by synchronizing the activity of the corresponding object/location detectors (Fig. 2.8d). After objects are bound to their location by synchrony, a subsequent process computes the spatial relation. This "binding-by-synchrony" approach only needs as many templates as the total number of objects and relations to be represented, whereas the pure template model needs as many templates as the *product* of the number of objects and relations (Hummel and Biederman, 1992; Malsburg, 1999). Observe that the binding interpretation of spatial relations still requires attention to be directed both objects simultaneously. Only then can the process of binding-by-synchrony select the precise objects to be compared among distractors in a cluttered scene. A similarly flexible method of spatial relation detection was proposed by (Pylyshyn, 1989a) in the form of the FINST model and its multiple spotlights (Fig. 2.8e). Instead of using a separate collection of location detectors, the multiple spotlights model can tag the locations of multiple objects by tracking the location of several attentional windows.

(Franconeri et al., 2012) distinguishes the above interpretations of attention from

those which only select one object at a time. Some studies have supported the single-object hypothesis by showing that the ability to localize an object in the first place requires focused attention (Treisman and Gelade, 1980; Evans and Treisman, 2005) and that focused attention tends to suppress information from outside the spotlight (Luck et al., 1997; Treisman, 1996; Reynolds, Chelazzi, and Robert Desimone, 1999). An attentional mechanism only focusing on one object at a time must shift its spotlight in order to compute a spatial relation. (Franconeri et al., 2012) argues these "shift" models can work in two ways. First, the spotlight can shift from one object to another while the direction of the shift is stored in working memory (Fig. 2.8f). In the example of the cross to the left of an X, the ordered sequence of events "X feature activation", "leftward attentional shift" and "cross feature activation" could be stored in working memory, after which a secondary executive process could deduce from this sequence the cued spatial relation. The direction of the attentional shift could be rightward, but only if the executive decision process can flip the sequence of events to appropriately match the target relation. Second, the attentional window could instead begin with a large aperture containing both objects and then zoom in to one object. The initial, global interpretation of the scene could extract coarse information about the object identities and the overall shape of their arrangement. Next, the spotlight would shrink its aperture and move leftward, storing the change in spotlight size and direction in working memory. If, after this move to local attention, a plus feature detector is activated, an executive process could extract the spatial relation by querying the stored information in working memory (Fig. 2.8g). Both the simple left-right shift and the global-local shift detect spatial relations with a combination of a moving attentional spotlight and working memory. Like binding-by-synchrony, these shift accounts create flexible representations of spatial relations between arbitrary objects.

(Franconeri et al., 2012) investigated these different attentional mechanisms in four EEG experiments, in which subjects were asked to detect left-right relationships.

There is substantial evidence that shifts of attention to one side of the visual field result in greater negativity in electrode sites on the contralateral side (Luck, 2012). (Franconeri et al., 2012) showed that this electrophysical correlate of attentional shifts accompanied their subjects' relational judgments, even after they demonstrated that the individual stimulus items (colored circles) could be detected preattentively. This distinctive EEG signal occurred even when subjects were dissuaded from using attentional shifts by a difficult dual task. Contralateral negativity during relation detection was even stronger on a harder task involving left-right relations between more complicated shapes.

The authors conclude that left-right relation detection relies on attentional shifts. Further support for a shifting spotlight of attention was provided in a recent follow-up experiment by (Yuan, Uttal, and Franconeri, 2016). These results run counter to the conclusions of (Hayworth, Lescroart, and Biederman, 2011), who used an fMRI adaptation paradigm to claim that spatial relations were represented without attentional shifts in the manner of binding-by-synchrony, although they did not specifically look for electrophysiological correlates of attention. (Franconeri et al., 2012) defend the global-to-local interpretation of attention (Fig. 2.8g) over the simple left-right shift. They argue that such an interpretation is consistent with global processing biases (Navon, 1977) and with our intuition that spatial relations are detected by selecting relevant objects simultaneously, since this is true for the first stage of the global-to-local method. Yet, it is unclear how a global-to-local attentional routine could extract accurate object identities in the first, global step, in the presence crowding (Whitney and Levi, 2011) effects and illusory conjunctions (Treisman and Gelade, 1980). Asserting that the initial joint representation of both objects can disentangle their respective features leads to a reduction to the template matching case of Fig. 2.8a.

One observer lacking any attentional mechanism at all was the Adaboost algorithm

taking the SVRT in (Fleuret et al., 2011). Perhaps in consequence, its average performance on those 16 problems incorporating spatial relations (e.g. Fig. 2.6b) lagged behind that of human subjects. Interestingly, Adaboost's average performance on same-different-up-to-transformation problems (see Sec. 2.2.1) and spatial relation problems were roughly equivalent, indicating that these types of problems were essentially of equal difficulty for the algorithm. Humans, on the other hand, performed slightly significantly better on spatial relation problems (average of 5.68 trials until learning) than on same-different problems (average of 6.67 trials until learning).

Often, in a spatial relations psychophysics experiment, subjects are asked to compare a visual scene with a sentence. This experimental design is natural since linguistic syntax and the semantics of words like *above* and *below* can be intuitively mapped to the "visual syntax" of a structured scene. Inspired by the psycholinguistic literature on spatial words (Clark, 1973; Miller and Johnson-Laird, 1976; Talmy, 1983; Levelt, 1984; Herskovits, 1986; Garnham, 1989; Vandaloise, 1991), numerous studies have compared the linguistic description of spatial relations with their manifestation in visual stimuli.

Early work by (Clark, 1972) compared RT of subjects performing an above-below task to the predictions of models designed to match pictures with sentences. Subjects were given sentences like "A is above B" or "B isn't above A" and asked to determine the truth or falsity of the sentence from a visual scene with objects A and B. Clark concluded from 11 experiments that sentences and pictures were both encoded like logical propositions and that subject RT resulted from a serial process of comparing corresponding parts between sentence and picture propositions. Importantly, no evidence was found that subjects' reaction time resulted from transforming a "base" sentence into a new version. For example, the authors claimed that subject RT on the "A isn't above B" condition did not result from, say, checking for "B is above A" and then transforming it to the negative form for comparison. This distinguishes the

results of (Clark, 1972) from work in transformational grammar (Miller, 1962) which showed grammaticality judgments take more time as sentences deviate from a "base" representation by grammatical transformations. This work, too, was later called into question when it was found that RT was largely predicted by sentence length, and not by number of intermediate transformations (Bever, 1988).

Later, various studies compared the way different spatial relations (e.g. absolute vs relative location) are apprehended in both vision and psycholinguistics (Logan, 1995; Logan and Sadler, 1996), how the acceptability of linguistic descriptions of spatial relations is influenced by scene structure (Regier and Carlson, 2001), and how the correspondence between linguistic and visual representations changes in the presence of distractors (Carlson and Logan, 2001). A prominent focus in contemporary research on this topic is the way in which language and vision both encode spatial relationships *asymmetrically*. This asymmetry is easy to see in the linguistic case. For instance, in the sentence "The square is to the left of the circle," one item, the square, is singled out as the subject as the sentence, whereas the circle is only defined in relation to the subject. Asymmetry is also manifest in the preference of certain descriptions of a spatial relation over others: it feels more natural to say "The bike is to the right of the building" than "The building is to the left of the bike" (Talmy, 1983).

(Roth and Franconeri, 2012) studied how language could bias the movement of an attentional window in a spatial relation detection task. Forty-five subjects participated in two experiments in which a given relation was cued with a sentence (e.g. "The red circle is left of the green circle"), one of the relation items was briefly displayed on the screen, followed shortly by the second item, after which the subject was told to check for the truth or falsity of the sentence as fast as possible (see Fig. 2.9). Left-right and above-below relations were tested. The authors found that RT was significantly faster when the first item to appear during a task was the sentence's subject. For example, the response to "The red circle is left of the green circle" is facilitated if the first item to

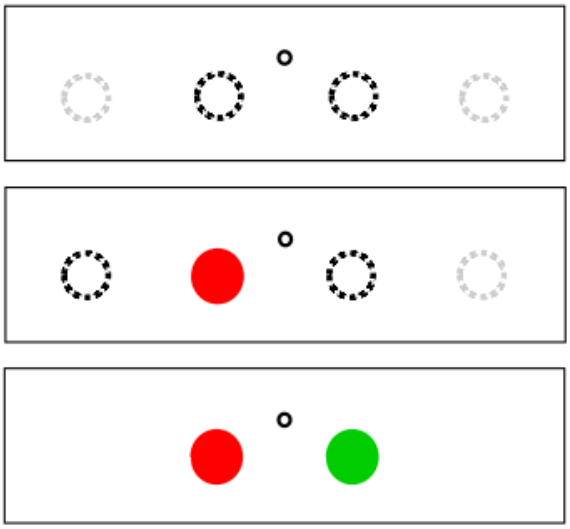


Figure 2.9 *Language and visual relations.* Stimuli from a spatial relations task testing for linguistic biases. Subjects were given a sentence describing a spatial relation. First, a fixation dot appeared for subjects (top panel) Then, a colored circle would appear at one of the two central spots. The two possible locations for this initial dot removed confounds of absolute location. Finally, a second colored dot appeared, and subjects were asked to judge whether the cued relation obtained. Subjects showed an RT advantage when the cued dot was the subject from the cued sentence. Figure taken from (Roth and Franconeri, 2012)

appear is the red circle. In other words, a shift of attention from red to green resulted in faster relation detection than vice versa. (Roth and Franconeri, 2012) concludes that the visual representations of spatial relations, like their linguistic counterparts, are encoded asymmetrically. This result was replicated in subsequent studies by (Michal and Franconeri, 2014) and (Michal et al., 2016). (Roth and Franconeri, 2012) hints at a stronger claim as well, stating that the detection of visual relations might be impossible without a linguistic faculty to back it up (see (Loewenstein and Gentner, 2005) and (Dessalegn and Landau, 2008) for details).

Our sudden focus on language may seem a bit out of place. However, a visual relation is a syntactically structured expression: it has real constituents (the objects in the relation) and the meaning of the full expression is compositionally determined by the identities of the objects and the rule connecting them. Formally speaking, an image depicting (CIRCLE left of SQUARE) is not so different from the utterance

"circle left of square". The express purpose of this chapter, in fact, is to subsume the study of visual relations into the framework of (Fodor and Pylyshyn, 1988), who considered language the most obviously systematic aspect of human behavior, but whose right to this phenomenon was not exclusive. Is it possible that the faculty of language, specialized as it is to deal with syntax, is also involved in the processing of other "syntactical" expressions from other sensory modalities? In fact, there is a burgeoning branch of machine learning devoted to the learning of joint linguistic-visual representations, the discipline of visual question answering (VQA). Though it is still in its infancy, the co-study of language and visual relations could be very fruitful (see Ricci, Cadene, and Serre, Submitted for a review)

Although we have only begun to emphasize attention in the current section, it is equally relevant to the previous discussions of various same-different relations. Shifting attention in the same-different paradigm would store two items in working memory, and, instead of extracting a spatial arrangement from attentional movement, could compare the attended items before and after the shift.

Equally important to all of the visual relations discussed here is the role of working memory. Indeed, it is difficult to imagine how a shifting spotlight of attention could detect relations without working memory. Models of visual working memory typically agree that mnemonic capacity can be quantified both by the number of items stored in memory and the precision with which each item is represented. However, they differ in how memory resources are allocated to these two demands. Slot-based models posit two distinct types of mnemonic resources, a fixed number of registers in which items can be stored and a precision of item representation within each slot. These resources are separate, so that a fixed maximum number of items, typically about 4 (Luck and Vogel, 1997), can be stored in memory, regardless of each object's complexity (Miller, 1956; Cowan, 2001; Awh, Barton, and Vogel, 2007; Zhang and Luck, 2008). Alternative models describe the number of storable items and the precision of item

representation in terms of one continuously divisible resource (Wilken and Ma, 2004; Alvarez and Cavanagh, 2004; Fougnie, Asplund, and Marois, 2010).

The allocation of memory resources is particularly important in the study of the mnemonic representation of spatial relations, since it is unclear if these relations, perhaps encoded by attentional shifts, should be placed in a memory slot alone or somehow attached to the objects in the relation. In the case of Fig. 2.8f, are "plus activation", "leftward attentional shift", and "cross activation" given three separate slots in memory? Or is the attentional shift simply attached to the activation of one or both object detections?

(Clevenger and Hummel, 2014) addressed this question for the restricted case of slot-based models in a psychophysics experiment testing for visual relational memory capacity. The authors tested three models of relational memory, each predicting a different number of allocated slots, s, given the same number of objects and relations. According to the first model, every item in a scene is multiplexed with its whole set of relations to every other scene item. If a scene has n items, then by this account, memorizing all relations in the scene should take $s = n$ slots. Each slot would hold information like "Item 1; larger, left, ...". This model assumes that working memory load is independent of the number of item relations to be memorized, a somewhat dubious claim. According to the second model, slots in working memory hold ordered pairs of items with a relational tag: "left(Item 1, Item 2)" would encode "Item 1 is to the left of Item 2". If there are n items in a scene, r relations to be recognized, then this account demands as many slots as there are combinations of item pairs and relations, $s = r(n^2 - n)$. A final model, the so-called "stacked relation" model, assumes that every pair of items is multiplexed with every relevant relation. When relations are "stacked" with every pair of items, then every slot can store, for example, "left+larger(Item 1, Item 2)". This single slot now encodes the fact that Item 1 is both left of and larger than Item 2. In this manner, the number of slots needed for

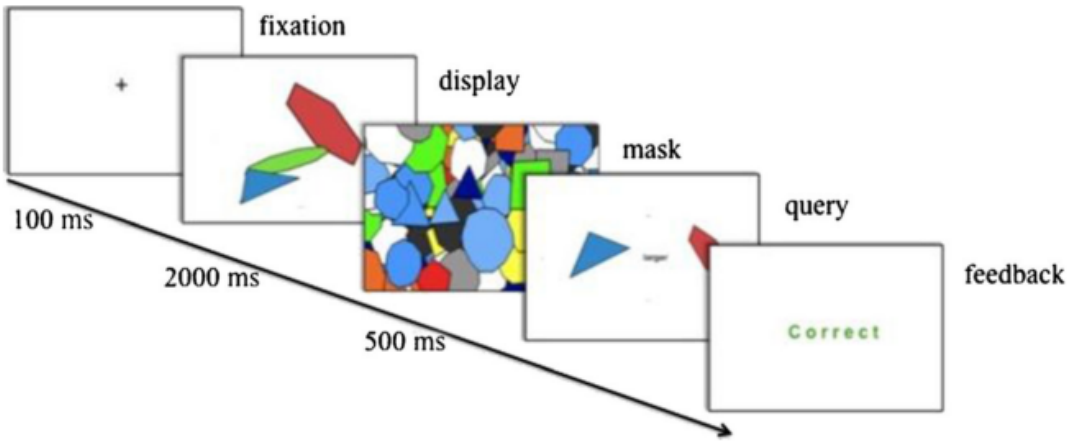


Figure 2.10 *Working memory for visual relations.* Stimulus presentation from (Clevenger and Hummel, 2014). After a 100 ms fixation period, a scene with 2-4 colored polygons appeared for 2000 ms. Following a 500 ms polygonal mask, subjects were questioned about a spatial relationship (pictured " blue polygon right of red polygon"), allowed to respond and given feedback. Note that the 2000 ms SOA was more than long enough for the deployment of attention and even multiple serial eye fixations.

relational processing reduces to the number of item pairs, $s = n^2 - n$.

Next, (Clevenger and Hummel, 2014) defined the probability that an observer with working memory could correctly recall a given visual relation as a function of s and compared human performance on a relational memory task to each of the three above accounts. 63 participants viewed a scene depicting 2, 3 or 4 colored polygons in a spatial arrangement for 2000 ms, followed by a 500 ms polygonal mask (Fig. 2.10). After the mask, subjects were queried about a particular relationship between two of the scene's polygons. Accuracy was measured as a function of the total number polygons in the scene. The authors then fit the three candidate models to subject performance and found a substantially better fit for the stacked relation model. (Clevenger and Hummel, 2014), verified that their subjects actually stored relations in memory, instead of computing them after the fact from absolute locations, in a second control experiment.

(Clevenger and Hummel, 2014) concluded that their subjects stored pairs of polygons in working memory slots together with their relevant relations. However, the study was limited in several ways. First, the authors never explicitly tested for

the recall of more than $n = 2$ objects or more than $r = 1$ relations at a time. It could be argued that subjects must have remembered $n > 2,\ r > 1$ objects and relations since they were asked to recall random relations on each trial and still performed better than chance, indicating they remembered more than single pairs. Yet, the study as it was conducted only fully probed the "writing" aspect of working memory, by demonstrating arbitrary binary relations could had been stored in memory on each trial. It did not, however, fully probe the "reading" aspect of visual memory. For example, perhaps many stacked relation items can be stored in memory but recalling more than a single relationship at once is exceptionally difficult. Perhaps this difficulty could be alleviated by certain task designs, like with different attentional cues before stimulus presentation. Moreover, when the authors construct a function for probability of recall, they assume that the mnemonic load imposed by a given scene only depends on the number of objects and relations and ignore ensemble statistics of the scene. What if, for instance, remembering a scene of 4 green polygons is harder than remembering scene of 4 multi-colored polygons? By focusing exclusively on the number of objects and possible relations and ignoring actual image features, (Clevenger and Hummel, 2014) only partially addressed the question of memory resource allocation. The trade-off in resource allocation between slots and precision likely depends on more than relations.

Additional aspects of this trade-off were explored by (Franconeri, Alvarez, and Enns, 2007; Brady, 2011; Brady and Tenenbaum, 2013), and (Brady and Alvarez, 2015). (Franconeri, Alvarez, and Enns, 2007) found subjects could attend to a variable number of locations in a scene, depending on the spatial precision demanded by the task. When a search task required fine spatial resolution and a cue indicated target location with high precision, then only a few objects could be attended to simultaneously. When the task required coarser resolution and the cue indicated target location with lower precision, more locations could be attended to. Importantly, the total number of objects in the scene remained constant. Despite this constancy,

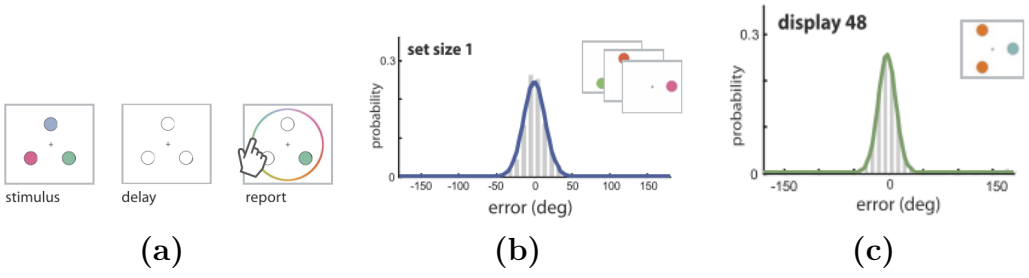


Figure 2.11 *Context effects in working memory for composite images.* Stimulus and error patterns from (Brady and Alvarez, 2015). *a*) Subjects were shown a set 1-6 of colored disks equidistant from a central point. The disks disappeared briefly, and subjects were then asked to recall the color of each disk in a random order by using a slider. *b*) Pattern of errors for a particular 1-disk stimulus. Errors are measured in degree of deviation from true values in HSV color space. *c*) The same plot, but for a particular 3-disk stimulus. Notice that the angular variance for subject responses is much narrower. Brady and colleagues seek to explain how context among multiple items affects memory allocation. Figure taken from (Brady and Alvarez, 2015).

the division of attention could be flexibly modified by different cues. While neither strictly about working memory nor relations, the study of (Franconeri, Alvarez, and Enns, 2007) does demonstrate the importance of cue type for the visual analysis of a cluttered scene. As attention is speculated to interface with working memory in order to detect spatial relations, cue type could have an indirect effect on working memory capacity as well. The experiments of (Clevenger and Hummel, 2014), modified to incorporate the cueing paradigm of (Franconeri, Alvarez, and Enns, 2007), might demonstrate a cue-dependent flexibility of mnemonic resources.

Furthermore, (Brady and Alvarez, 2015) provide evidence that ensemble statistics of a visual scene affect how it is stored in working memory. Subjects were briefly shown images of 1-6 colored circles and then asked to report the exact color of each of the circles in a random order. Their subjects made errors highly inconsistent with a model of working memory based only on the number of scene items. For example, subjects consistently found certain 3-item stimuli to be much easier to recall than others. Moreover, certain 3-item stimuli were easier to recall than some 1-item stimuli. The pattern of errors was neither explained by a model based on numbers of individual items, nor on models based on a chunking strategy that unitized multiple items into

one memory representation (Halford, Wilson, and Phillips, 1998). Instead, errors were best explained by a hierarchical model incorporating both information about individual items *and* how they could be grouped into various chunks. Further evidence that visual memory stores items not as indivisible units but as hierarchically structured objects was provided by (Brady, 2011) and (Brady and Tenenbaum, 2013). While these results are not strictly about spatial relations, it is easy to see how they could be applied to a spatial rule paradigm. How, for instance, would the results of (Clevenger and Hummel, 2014) have changed if polygon colors had been systematically varied in light of (Brady and Alvarez, 2015)'s findings? And what would a model of working memory look like that allowed for structured, hierarchical representations of objects?

The psychophysics literature reviewed so far suggests important roles for attention and memory in the detection of visual relations. Though some relations, seemingly, can be detected pre-attentively (recall (Donderi and Zelnicker, 1969)), most cases require serial attention and the storage of intermediate computations in working memory. Both attention and working memory have extensive neuroscientific literatures, though this material does not typically explore relational processing. We review some of the neuroscientific studies dealing with visual relations in the next section.

2.2.2 Neuroscience

There is a substantial literature on the neural basis of attention (see (Petersen and Posner, 2012) for review and (Harris and Thiele, 2011) for related literature on cortical rhythms as a basis for attention) and working memory (see (Constantinidis and Klingberg, 2016)), though a comparatively little amount of this literature deals with visual relations specifically. Moreover, most studies of the neural foundation of relational processing rely on somewhat coarse imaging data rather than circuit-level or single-cell analyses, so neuroscientific theories of visual reasoning are typically only based on brain regions activated during relational psychophysics. However, there

is still much for the modeler to glean from this literature. The three main topics discussed in this section are *1*) whether or not different brain regions are activated during different types of visual relations, *2*) to what extent these different regions are hemispherically organized, and *3*) how much of this hemispheric organization arises from different receptive field sizes in the ventral and dorsal streams of the visual cortex.

In the last section, we saw how subjects in (Brady and Alvarez, 2015) performed an experiment in which locations had to be matched to colors (Fig. 2.11). A natural question is how information from two modalities, spatial location and color, can be integrated into one judgment. There is evidence that posterior cortical regions are specialized to hold only particular modalities in working memory (Zeki and Shipp, 1988; Jonides et al., 1993; Courtney et al., 1996). These specialized regions then converge (Fuster, 1997) on prefrontal cortex (PFC) where it is speculated that diverse modalities are integrated in working memory. For example, prefrontal cortex has been shown to hold both spatial and object information in working memory (Rao, Rainer, and Miller, 1997; Rainer, Asaad, and Miller, 1998b; Rainer, Asaad, and Miller, 1998a).

Prabhakaran et al. (2000) tested the role of PFC as a spatial and verbal integrator in an fMRI study with 6 subjects which involved some simple relational processing. In a key experiment from this study, subjects viewed a stimulus with four pairs of parentheses and four capital letters (Fig. 2.12). The stimulus disappeared for either 5000 ms (the "delay" condition) or 250 ms (the "no-delay") condition, during which time subjects were told to keep the previous display in memory. After this delay, one pair of parentheses enclosing a letter reappeared on the display. Subjects were asked both whether the reappeared parentheses were at some parentheses location from the previous display and if the new letter was in the previous display, regardless if it reappeared in the right location. In the "bound" condition, the initial display depicted parentheses enclosing letters across the visual field. In the "separate" con-

dition, the initial display simply had letters arranged horizontally at the center of the visual field. In the "bound" condition, letters and parentheses were bound by a spatial relation, namely, co-location. (Prabhakaran et al., 2000) were curious which brain regions were activated when the task required recognition of this relation and the subsequent "binding" of verbal to location information. They found that the maintenance of working memory in the "bound" condition resulted in significantly more right prefrontal activation than in posterior regions. Control experiments testing for preferential activation of prefrontal cortex to either spatial or verbal information showed no effect. The opposite was found for posterior regions, subregions of which would selectively activate to different stimulus modalities. The emphasis placed by (Prabhakaran et al., 2000) on integrating verbal and spatial information is interesting in light of the aforementioned results (Section 2.2.1) connecting linguistic and visual representations of spatial relations.

It could be argued that the results of (Prabhakaran et al., 2000) do not directly bear on visual relations since the task only required memory and not any reasoning. However, a similar result was found by (Kroger et al., 2002) with a task for which relational complexity could be parametrically varied. Subjects were shown Raven's progressive matrices which varied in relational complexity and difficulty (Fig. 2.13). Relational complexity was varied by making the solution to the matrix problem depend on more items (e.g. Fig. 2.13a depends on one item, while Fig. 2.13b depends on two). Relational complexity was balanced with difficulty, which was controlled by adding more distractors (see Fig. 2.13c). The authors found that both increased relational complexity and difficulty led to increased activation in parietal and dorsolateral prefrontal cortex. However, only high levels of relational complexity selectively activated anterior left prefrontal cortex. They concluded that this specific brain region was involved in the integration of complex relations among stimuli, though their result runs somewhat counter to that of (Prabhakaran et al., 2000) who implicated right

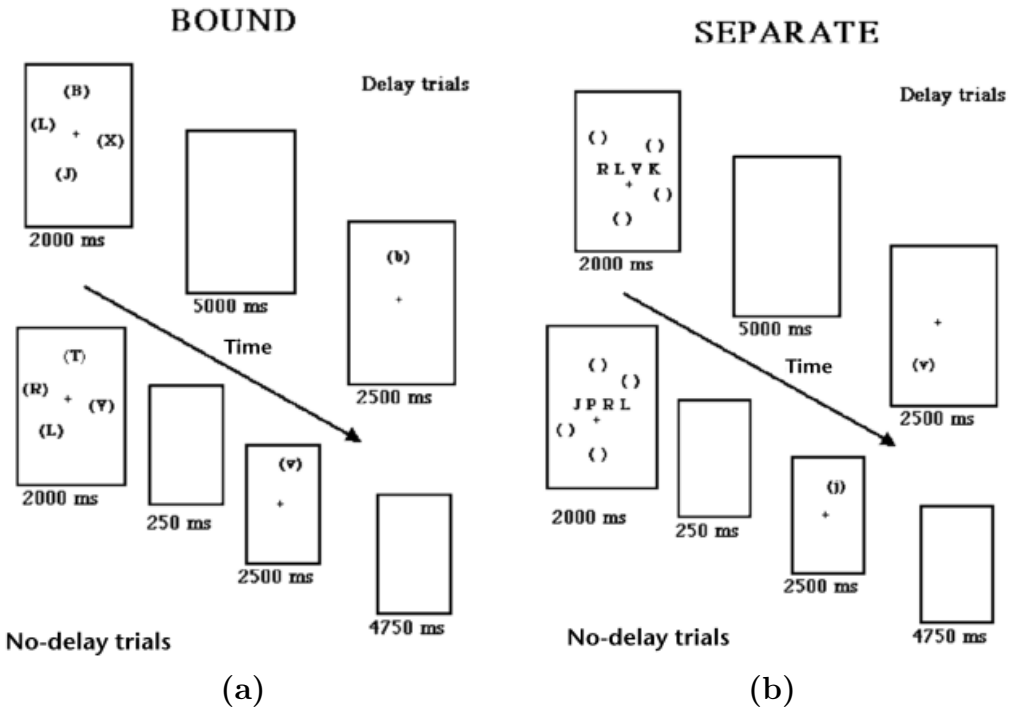


Figure 2.12 *Pre-frontal cortex as spatial and verbal integrator.* Example stimuli from a two-factor spatial relations design of (Prabhakaran et al., 2000). Subjects had to remember 4 spatial locations marked by parentheses and 4 letter identities from an initial stimulus displayed for 2000 ms. In the "bound" condition, parentheses and letters appeared at the same location. In the "separate" condition, parentheses could appear anywhere on the stimulus, but letters were always centrally displayed along the horizontal axis. Either a 250 ms or 5000 interstimulus period occurred, according to the delay vs no-delay condition, followed by the appearance of a letter and a single pair of parentheses. The appearing letter was always enclosed by the appearing parentheses, even in the "separate" condition. Subjects were asked to determine if both the new parentheses and new letter had appeared in the first stimulus. Notice that, when letters reappear, they are lower case, so letter identity just be remembered, and not just a particular orthographic symbol. Figure taken from (Prabhakaran et al., 2000).

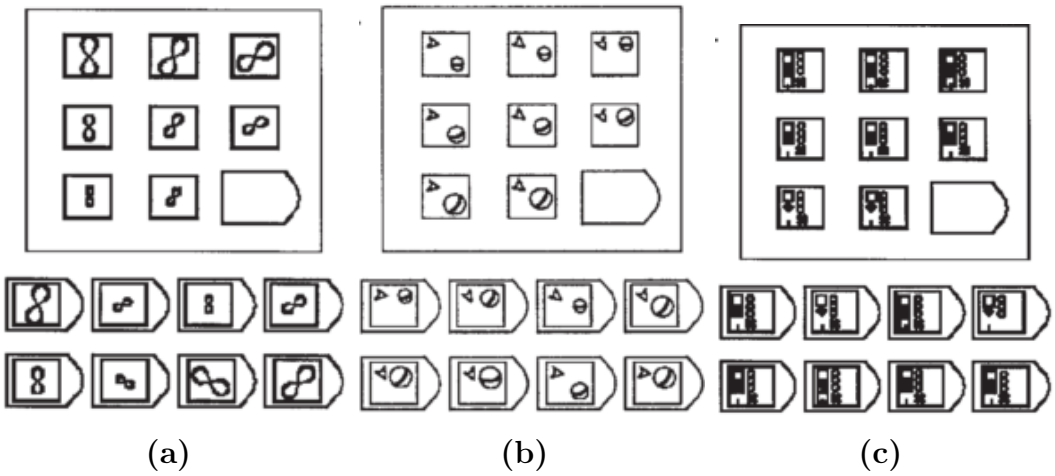


Figure 2.13 *Visual relations in progressive matrices.* Raven's progressive matrices used in the spatial relations task of (Kroger et al., 2002). *a*) Each cell in the matrix contains a single item and the angular relations among these single items across the matrix comprise the relationship to be completed by the subject. *b*) Same as *a*, but with two items per cell. Now there are two relations to be recognized: the angle of the triangular item and the vertical position of the ball. *c*) Panels *a* and *b* varied by relational complexity, but panel *c* shows how difficulty can be increased while maintaining relational complexity. There are numerous items per cell, but only one of them is meaningful for the inter-cell relations. However, subjects still need to search for relations among all items, and spurious relations could still be detected. Figure taken from (Kroger et al., 2002).

prefrontal cortex instead of left. It may be that (Prabhakaran et al., 2000)'s result depends on the integration of specifically verbal and spatial information, or it may be that the experiment in (Kroger et al., 2002) was not properly balanced for difficulty, since subjects would have to check distractors for possible relational relevance as well.

Later work by (Golde, Cramon, and Schubotz, 2010) further explored the role of PFC as a relational integrator, this time compared to the premotor cortex (PMC). Like (Prabhakaran et al., 2000), this study sought to distinguish two brain regions according to their computational role and sensory preference. While the PMC is traditionally conceived of as dealing largely with action-related information, it has been implicated in some abstract reasoning tasks (Knauff et al., 2003; Goel and Dolan, 2004). (Golde, Cramon, and Schubotz, 2010) imaged the PFC and PMC with fMRI while subjects completed two types of Raven's progressive matrices. For some of these matrices, the pattern depended on the serial ordering of matrix elements. For others,

solution depended on information distributed non-sequentially throughout the matrix. The authors also varied whether or not the matrix depicted concrete (photographic) or abstract (synthetic) images. Though both PFC and PMC were equally activated by the solution of matrices depicting abstract and concrete relations, only the PFC was strongly activated by solutions requiring the integration of relational information. The PMC, on the other hand, was activated by solutions that required serial reasoning. The authors conclude that the PMC is specialized for the concatenation of elements in a visual relation, whereas the PFC is required for integrating relations across a visual scene. Thus, claim the authors, PFC and PMC are differentiated by computational function and not by domain selectivity.

The above studies compared brain regions along an integrative/non-integrative task axis. Another axis along which visual relations tasks are divided is the so-called *categorical-coordinate* distinction. A categorical relation is a spatial relation that requires the observer to recognize general arrangements like *above, below*, etc. Coordinate relations, on the other hand, require precise metrical judgments. Noticing that a fork is to the left of a knife is the detection of a categorical relation; noticing that the fork and knife are roughly separated by 20 cm is the detection of a coordinate relation. Humans have shown performance advantages for categorical judgments over precise, metrical judgments (Kosslyn et al., 1989; Hummel and Holyoak, 2003; Gentner, 2010).

There is evidence that the detection of categorical and coordinate relations relies on two different hemispherically segregated networks. (Kosslyn et al., 1989) reported that their subjects displayed a left hemispheric advantage for categorical relations and a right hemispheric advantage for coordinate relations in an experiment with stimuli presented in either visual hemifield. Though (Kosslyn et al., 1989) backed up their findings with computer simulations, their results were subsequently called into question as artifacts of task difficulty (Slotnick et al., 2001) or spatial resolution (Sergent, 1991).

Yet, strong evidence for the hemispheric dichotomy has come from studies showing impairment to coordinate processing after damage to right parietal cortex (Hannay, Varney, and Benton, 1976; Laeng, 1994) and impairments in left-right judgments following left posterior parietal lesions (Mayer et al., 1999). A joint fMRI/lesion study by (Amorapanth, Widick, and Chatterjee, 2010) supported the hemispheric dichotomy, though the authors emphasized that the distinction was one of degree rather than kind.

This dichotomy is thought to arise from each hemisphere's spatial frequency tuning and the way this tuning recruits neurons of different receptive field (RF) sizes. (Hinton, McClelland, and Rumelhart, 1986) and (Eurich and Schwegler, 1997) have argued that the spatial resolution of a neural network is increased to the degree that the borders of its neurons' RFs overlap the interiors of other RFs. Conversely, networks with little RF border overlap will benefit categorical judgments over precise metrical judgments (Prinzmetal, 2005). Consider for example, RFs depicted in Fig. 2.14. Each gray disk represents the RF of a single neuron and the ensemble of 3 neurons divides the visual field into 7 regions labeled $a - g$. Consequently, a small stimulus traversing the path γ would elicit 7 different population responses indicating stimulus location. This division of space arises, for example, from neuron 3's boundary overlapping the interiors of neurons 1 and 2. The ensemble in Fig. 2.14 is relatively suited for coordinate relational detection because of its high spatial resolution, much like the neurons with large, overlapping RFs in monkey parietal cortex (Motter and Mountcastle, 1981; Motter et al., 1987). The extreme case of little boundary overlap would occur when all neurons have perfectly overlapping RFs. In this case, space would only be divided into 3 regions corresponding the interior of the overlapping receptive fields and the exteriors marked by a and g in Fig. 2.14. This new ensemble would be relatively suited for categorical judgments since it functions like a digital position detector: the ensemble is activated if a stimulus on γ lies in the overlapping RFs and is inactive otherwise. The ensemble

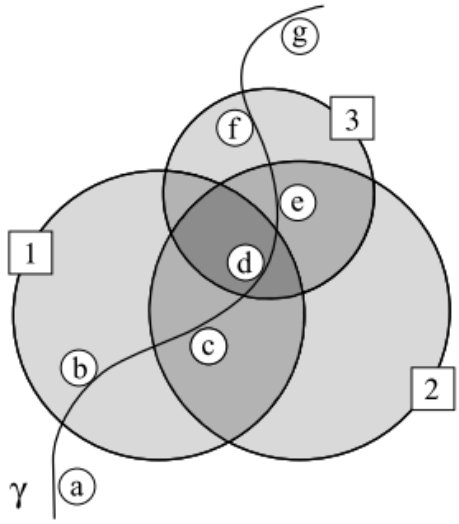


Figure 2.14 *Spatial acuity from highly overlapping receptive fields.* Three receptive fields demonstrating high spatial resolution for overlapping RF boundaries. RFs overlap to create 7 different population responses, including a and g for responses above and below the ensemble. A particle moving along the path γ will elicit 7 responses according to its spatial location. Figure from (Eurich and Schwegler, 1997).

could therefore make categorical, rather than graded, judgments. Additionally, it has been argued that ensembles with large, overlapping RFs[8] tend to be sensitive to low spatial frequencies, whereas those with little overlap select for high frequencies (Laeng et al., 2011). So, we should expect, somewhat paradoxically, fine spatial resolution in ensembles sensitive to low frequencies and coarse spatial resolution (and thus better categorical detection) in networks selecting for high frequencies.

(Ivry and Roberston, 1998) has argued that the left and right hemispheres select for high and low spatial frequencies, respectively. Thus, the lateral dichotomy for categorical vs coordinate relations could arise from the spatial frequency preferences and RF sizes of each hemisphere. Moreover, (Laeng et al., 2011) has proposed that attentional cues can encourage left or right hemispheric processing leading to facilitation of categorical or coordinate relation detection. (Laeng et al., 2011) tested this

[8]An ensemble does not need to have large RFs in order have high spatial resolution (Eurich and Schwegler, 1997), though increased RF size does promote boundary overlap under reasonable assumptions.

hypothesis an a categorical/coordinate relation experiment based on evidence that attention to larger/smaller image portion selectively recruits the right/left hemisphere (Delis, Robertson, and Efron, 1986; Fink et al., 1996; Fink et al., 1997; Hübner, 1998; Hübner and Studer, 2009; Robertson and Kim, 1999; Yamaguchi, Yamagata, and Kobayashi, 2000). Subjects were asked to make either categorical (rotated vs non-rotated) or coordinate (distance between two image items) judgments while RT was measured. First, an image appeared on a display depicting two animals. Then, the animals disappeared and were replaced by a square which either accurately or inaccurately predicted the location of the animals when they were to reappear. The squared cue was varied in size, with the small square intended to promote a small attentional aperture, and the large one intended for a larger aperture. When the animals reappeared on the screen, subjects on categorical trials were told to report whether the animals remained in a given categorical relation seen in the initial presentation. On coordinate trials, they had to report whether the distance between the animals changed. The authors found that small cues decreased RT of correct judgments on categorical trials, and increased RT on coordinate trials. Large cues had the opposite effect. (Laeng et al., 2011) argued that this effect arose from the way attention recruited different hemispheres during the task. Big cues, they argued, engaged the low-frequency selective right hemisphere preferentially. This spatially distributed attention recruited more overlapping RFs resulting in higher spatial resolution and faster RT. On the other hand, small cues promoted more focused attention, led to more left hemisphere activation and the recruitment of fewer, less overlapping RFs, and therefore better categorical detection.

These results were bolstered by further work by (Ham et al., 2012) who found a connection between categorical vs coordinate processing and the size of an attentional aperture as measured by BOLD signal in early visual cortex in humans. Unlike in the experiment of (Laeng et al., 2011), subjects were given no exogenous attentional

cue. Yet, (Ham et al., 2012) still found that cortical activity was diffuse during a coordinate task and focused during a categorical task. This effect was the strongest in V3, but insignificant in earlier areas.

2.2.3 Computational Models

These physiological studies largely support the psychophysical literature we discussed earlier: visual relations can be detected with an attentional window with neural correlates as early as V3 together with working memory in PFC and PMC. Further, the work of (Eurich and Schwegler, 1997; Laeng et al., 2011) regarding receptive field sizes and attentional apertures, sheds light on how one of the basic requirements for relational detection, localization, might function. However, physiological evidence is only as good as the computational model that explains it. For example, the study of biological object recognition has benefited immensely from fruitful comparisons to models like HMAX (Riesenhuber and Poggio, 1999; Serre, Oliva, and Poggio, 2007) and feedforward neural networks in general (Hong et al., 2016).

Yet, there have been surprisingly few attempts at the modeling of visual relation detection, perhaps because a definition of good performance has not yet emerged. There are notable exceptions, which we review here. The next section of this chapter will be a detailed study of CNNs and so-called "relation networks", so we will save a description of those models for later. For now, we will simply describe some models of mostly historical interest.

Explicit relational units Encoding visual relations with neurons selecting for particular objects in particular relations is inflexible and computationally expensive. In such a model, relations are only encoded "implicitly" in every output neuron. A more flexible solution is to use two sets of neurons, one for objects and one explicitly

for relations, whose combined activity represents a given scene structure. For instance, to detect a cat above a dog, the "explicit relation" neural network would activate its cat, dog and "above" units simultaneously. Whereas the implicit approach would require $rn(n-1)$ output neurons to detect r (asymmetric) binary relations among n objects, the implicit model reduces the number to $n(n-1)+r$, similarly to the working memory model of (Clevenger and Hummel, 2014).

A notable (not necessarily visual) formalism incorporating explicit relation units is the semantic cognition model of (Rogers and McClelland, 2004) (Fig. 2.15). The model is a 4-layer feedforward neural network with two separate sets of input units representing semantic items and binary relations between them. The goal of the model is to complete relational statements like "Robins can _" with a correct attribute like "fly". Middle layers of the network encode both a representation of each scene item and their relational semantics. The whole model is trained by supervision with backpropagation of error. Training examples of relations are provided to the model by simultaneously activating an input item and an input relation together with the correct completion attribute. After training, the model can represent similarities between categories, so that selecting either "Robin is _" or "Salmon is _" will both result in "living".

However, without an explicit formal semantics and corresponding notions like set inclusion, the model cannot answer questions like "If there are both boys and girls in city X, are there more children or more boys?" (Halford, Wilson, and Phillips, 2010). Further, the model struggles to represent composite propositions like "Robin ISA _ AND Salmon ISA _". The simultaneous output activation of "bird" and "fish" leaves ambiguous the question of which, between robins and salmon, are birds and which are fish. Composite propositions are notoriously difficult to represent in connectionist systems lacking a combinatorial syntax and semantics (Fodor and Pylyshyn, 1988). There are some connectionist systems, like the structured tensor analogical reasoning

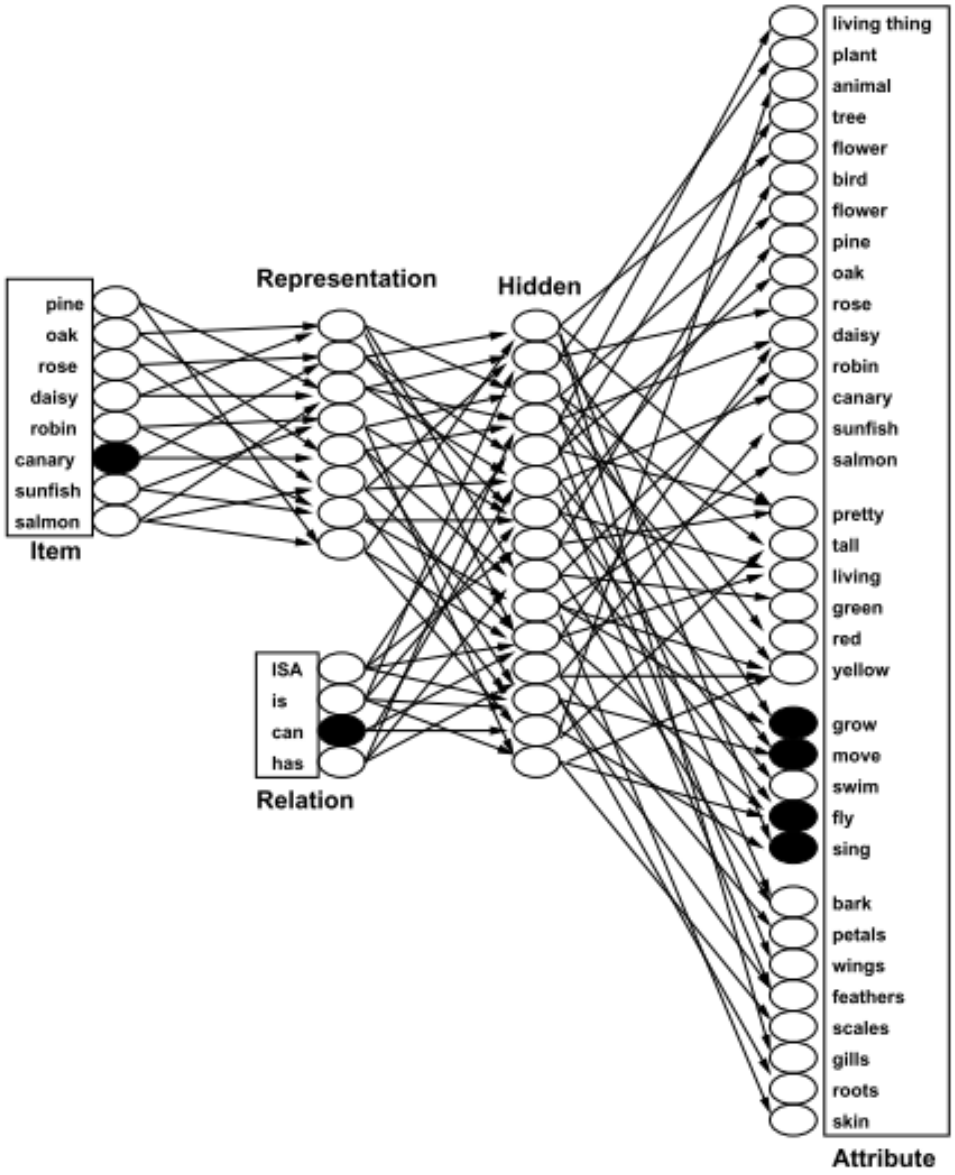


Figure 2.15 *Hardcoded relational processing in a neural network.* A four-layer feedforward network for learning relations from (Rogers and McClelland, 2004). The network is designed to complete relational expressions like "Canary can _" with an attribute like "fly". The model is then said to understand that "canary" and "fly" obey the relationship "can". The model depends on two types of input units, one for item labels and one for explicit relational units. Training consists of activating a unit from each of the input types and providing the correct output. Intermediate layers learn to correctly associate items, relations and attributes. However, the model has trouble in representing composite relationships simultaneously (see main text). Figure from (Rogers and McClelland, 2004).

model (STAR) of (Halford, Wilson, and Phillips, 1998), that can represent composite propositions, but only by resorting to the wasteful, inflexible templates that plague feedforward networks. Additionally, because relations and scene items are always provided to the semantic cognition model as simultaneous inputs during training, the network can only represent relations with respect to particular objects. It cannot detect a familiar relation between new objects. This suggests such a model would be critically limited in the detection of spatial relations, which often involve completely novel objects, but in familiar arrangements.

Relations from binding The computational expense of a connectionist model is typically measured in the number of neurons and training samples it needs to achieve a desired accuracy. As we have emphasized, connectionist systems for relation detection have high costs by both these measures. One method of alleviating this cost is to encode relations not by neurons but by new types of neuronal activity. If neurons are allowed dynamic activity, for example in the form of spikes, then relations could be represented by synchronizing the activity of units representing different items in the relation. This is the basis for so-called *binding-by-synchrony* models, which we first described with respect to the shifting attention experiment of (Franconeri et al., 2012). Biological neurons are known to synchronize their firing rates in various circumstances, and this correlated activity has been speculated to play a role in attention (Robertson, 2003), perceptual grouping, (Farid, 2002) and consciousness (Leeuwen, 2007). Greater detail will be provided in Ch. 3.

There are numerous computational models which attempt to represent relational structure by binding-like mechanisms, for example (Doumas, Hummel, and Sandhofer, 2008), (Shastri, 1999), (Sougné, 1999) (Hummel and Biederman, 1992), (Lades et al., 1993), and (Hummel and Holyoak, 2003). Binding models represent a somewhat underdeveloped area of research. As of yet, no binding-by-synchrony model has been

shown to work for natural images. Moreover, the synchronization of neural firing is an emergent, parallel process across a network, and does not involve a serially shifting window, as we have seen from psychophysical and neuroscientific evidence. And, though neuronal synchrony has been implicated in many important psychological phenomena, no rigorous computational demonstration linking synchrony to behavior has yet been offered. We take some steps at alleviating this in Chs. 4 and 5.

Relations from attentional routines Computational studies more in line with contemporary psychophysics assert that relations should be extracted by a serial attentional routine. Theoretical work by (Ullman, 1984) emphasizes attentional routines as a key middle ground between perception and higher order cognition, in both humans and potential computational models.

Ullman defends a set of basic attentional operations which he claims must underlie relational reasoning. First, attention must be able to shift its focus from one image region to another. The studies of (Franconeri et al., 2012) and (Yuan, Uttal, and Franconeri, 2016) support this claim, and evidence from patients with simultagnosia shows that the inability to shift attention causes severe deficits in visual reasoning (Hécaen and Ajuriaguerra, 1954). Ullman also emphasizes the role of shifting attention in controlling RT in visual search and recognition (Eriksen and Schultz, 1977; Posner, Nissen, and Ogden, 1978; Tasal, 1983), visual change sensitivity (Shulman, Remington, and Mclean, 1979), and short-term visual working memory (Sperling, 1960; Shiffrin, McKay, and Shaffer, 1976).

Second, locations across a scene must be indexed in a manner that can guide a shifting focus. By indexing, Ullman refers to some process by which spatial locations are highlighted for distinguishing properties and assigned a coordinate which can be fed to an attentional window. For example, features that pop-out in a cluttered scene (Treisman and Gelade, 1980) are rapidly assigned indices readable by attention,

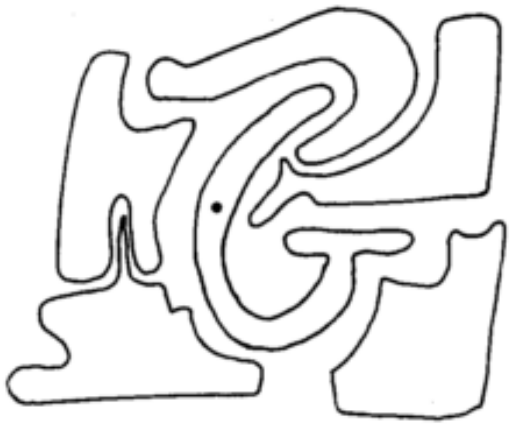


Figure 2.16 *Perceptual grouping as a visual routine.* Six interlocking objects. (Ullman, 1984) uses this example to explain why perceptual grouping must be an elemental operation for any attentional routine. Without the ability to disentangle these interlocking curves into six discrete objects, numerosity judgments will be impaired. Figure from (Ullman, 1984).

regardless of their location in the visual field. Other features are only assigned an index when the focus of attention is close to them.

Next, Ullman claims that image regions must be perceptually grouped or bound. For example, to count the number of closed curves in Fig. 2.16, it helps to unitize each enclosed region into a single object. Non-grouping strategies, like simple shifts of attention, are unsuited for images like Fig. 2.16, where the borders of objects intertwine in complicated ways. Uniting these curves requires grouping spatially disparate regions of each curve into a cohesive figure. It is difficult to see, for example, how visual relations depending on numerosity could be detected without a perceptual grouping mechanism. Numerosity requires the detection of unitized "wholes", instead of unbound bags of features. A related requirement for an attentional routine, says Ullman, is the ability to trace contours. This ability would aid in the detection of closed curves and in texture segregation (Jolicoeur, Ullman, and Mackay, 1984).

Finally, says Ullman, the computations from these attentional operations must be stored ("marked", in his words), in memory. When the complete attentional routine

finishes scanning a scene, the intermediate results in memory can be integrated into a final judgment, similar to the way the PFC is thought to integrate visual information in working memory in relational detection (Golde, Cramon, and Schubotz, 2010; Kroger et al., 2002).

We have seen several of these theoretical requirements appear before in practical relation processing experiments: (Franconeri et al., 2012) found evidence for the serial shifting of attention, the visual search experiments of (Donderi and Zelnicker, 1969) must require some sort of location indexing, and integration of information in working memory played key part in the work of (Golde, Cramon, and Schubotz, 2010). Others of Ullman's desiderata seem less necessary. While perceptual grouping might aid in numerosity judgments, it seems like unbound bags of features might suffice for spatial relation detection, especially when objects are spatially separated. The same could be said of contour tracing. From the literature we have reviewed so far, the hard core of Ullman's list seems to be 1) shifting attention, 2) location indexing and 3) working memory.

While (Ullman, 1984) outlined some key theoretical components of an attentional system for visual reasoning, he did not actually construct a functioning model. Such a model was later designed by (Kopp, 1994) and used to assign linguistic descriptions to scenes using a shifting saccadic mechanism. The model (Fig. 2.17) can learn spatial relations like "right" and "above" by supervision on simple binary scenes paired with linguistic descriptions. The scene items, though just simple binary patterns, are given fanciful names, like CAT. Like the so-called "what and where" pathways of the visual cortex (Schneider, 1969; Ungerleider and Haxby, 1994), semantic and spatial information flow through model in two segregated channels. The "what" channel takes in a cropped version of the scene, according to the current focus of attention, and subsamples it. This coarse object representation is then encoded by a two-layer neural network whose output is fed to a correlation matrix which matches visual/spatial

activity to simultaneous linguistic input.

The "where" channel first locates the two objects by finding regions of the retina with maximal activity. Once the centroids of these regions have been detected by lateral inhibition, an attentional window likened to saccadic eye movement randomly jumps back and forth between them. The direction of saccadic movement is constrained to only move up, down, left, right or not at all. Saccades fix a small attentional window on the retina which crops the scene so that it can be fed to the semantic pathway. Upon each saccade, the coordinates of the attentional window are fed to a second neural network where the direction of saccadic movement is classified. The output of this classification then converges on the correlation matrix. The model is trained by providing simple binary scenes with strings like "DOG RIGHT_OF CAT", which are parsed by a simple encoding network and then fed to the correlation matrix. The correlation matrix encodes the co-occurrence of linguistic and visual/spatial information.

During a test phase, the model is run in time and allowed to "describe" a scene by finding the linguistic template that has the highest correlation with the visual/saccadic information in the matrix (Fig. 2.17). Note that, like many of the authors we have encountered so far, (Kopp, 1994) has made an explicit connection between visual relations and linguistic description. His model only works because of a learned correlation between linguistic representations of direction and saccadic movement. The model also has a very simple form of memory, in the form of the correlation matrix, which, like memory models discussed earlier, multiplexes relational (in this case locational) information with object information. Further, Kopp's model focuses an attentional window on one object at a time, just like the human visual system, according to (Franconeri et al., 2012). However, because saccades are randomly generated and training data is balanced so that visually-identical but linguistically-different scene descriptions appear equally often (e.g. there are as many "DOG RIGHT_OF CAT

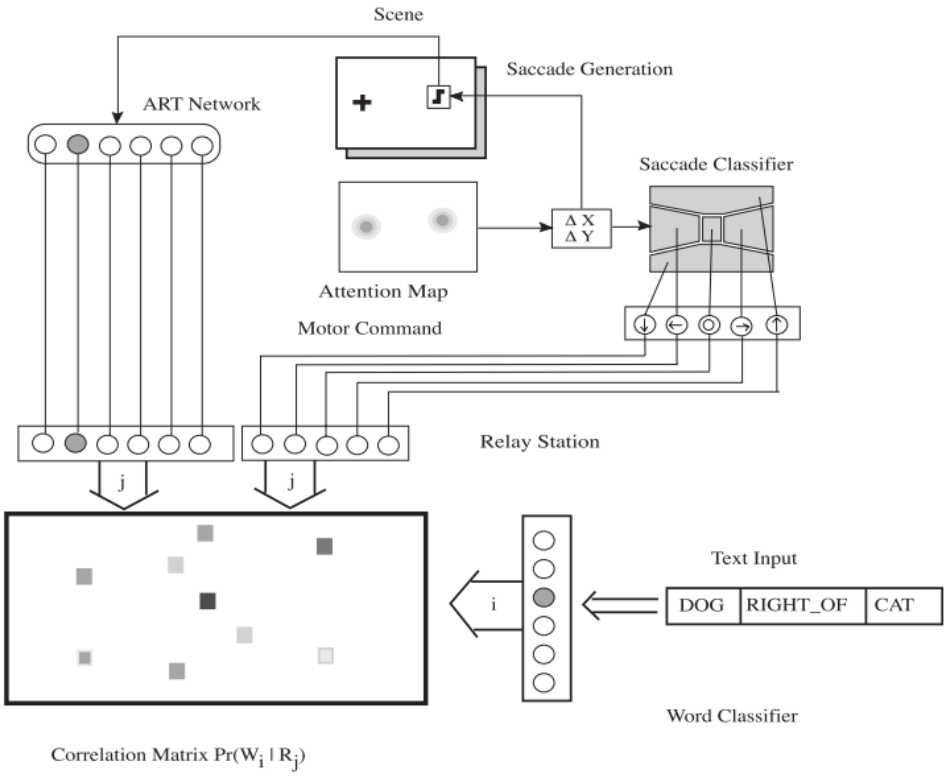


Figure 2.17 *An early attentional model of visual reasoning.* The spatial relation detection model of (Kopp, 1994), designed to match strings of English with simple binary scenes. The model is roughly inspired by the ventral and dorsal streams of the visual cortex. The model's "what" path (beginning top leftward arrow) encodes the pixel data from an attentional window on the scene with a two-layer neural network terminating on the columns of an array designed to correlate visual activity with linguistic representations. The dorsal stream locates scene items by lateral inhibition, randomly saccades between the items in one of 5 pre-wired directions, then uses saccade coordinates to classify the directions of attentional shifts. These directions are encoded by a 1-layer neural network and then passed to the correlation matrix. Strings of English from the output layer are trivially parsed and encoded with 1-layer network, now terminating on the rows of the correlation matrix. The model is run in time and the matrix row with maximal activity after a given time is chosen as the most likely scene interpretation. Figure from (Kopp, 1994).

as CAT LEFT_OF DOG), it can only represent symmetric relations. The saccadic system of the model does not incorporate the linguistic asymmetry of subject vs object, and therefore we would not expect it to recapitulate the RT results of (Roth and Franconeri, 2012). Moreover, the system can produce redundant descriptions like "CAT RIGHT_OF FROG UNDER FLY OVER FROG", since there is no executive control of saccades. This severely limits the system's use in cluttered scenes where, not only are relevant objects difficult to isolate, but executive attentional control is required to avoid task-irrelevant comparisons.

Nevertheless, the model of (Kopp, 1994) incorporates many of (Ullman, 1984)'s desiderata and much of the psychophysical/neuroscientific literature we reviewed earlier. Another model, radically different in construction, but similar in principle, was offered by (Chikkerur et al., 2010). Their Bayesian inference model of attention was not explicitly designed to detect spatial relations, but simply to simultaneously identify and localize objects in a cluttered scene. The authors showed how, merely by performing inference on the joint distribution of object identity and location, they could explain both spatial and feature-based attention. The former emerged from their Bayesian theory as a means of reducing uncertainty in shape information; the latter, as a means of reducing spatial uncertainty. They used this model to explain some key findings in the attentional literature: pop-out effects, attentional modulation of neuronal tuning curves, and patterns of eye fixations. (Chikkerur et al., 2010), like (Kopp, 1994) emphasized the role of attention in serially locating single objects at a time. Moreover, their model was designed to work in clutter, which we already noted as a key limitation in (Kopp, 1994)'s early model. Combining the Bayesian model of (Chikkerur et al., 2010) with a working memory could result in a better-functioning, more mathematically principled version of Kopp's original model.

2.3 Systematicity deficits in feedforward neural networks

Consider the images on Figure 2.18(a). These images were correctly classified as two different breeds of dog by a state-of-the-art convolutional neural network (CNN; He et al., 2015). This is quite a remarkable feat because the network must learn to extract subtle diagnostic cues from images subject to a wide variety of factors such as scale, pose and lighting. The network was trained on millions of photographs, and images such as these were accurately categorized into one thousand natural object labels, surpassing, for the first time, the accuracy of a human observer for the recognition of one thousand image categories on the ImageNet classification challenge (Deng et al., 2009).

Now, consider the SVRT image on the left side of Figure 2.18(b). On its face, it is quite simple compared to the images on Figure 2.18(a). It is just a binary image containing two three-dimensional shapes. Further, it has a rather distinguishing property: both shapes are the same up to rotation. The relation between the two items in this simple scene is rather intuitive and obvious to human and non-human observers. Recall the striking example from Martinho III and Kacelnik (2016), in which newborn ducklings were shown to imprint on an abstract concept of "sameness" from a single training example at birth (Figure 2.18(b), right panel). Yet, as we will show in this study, CNNs struggle to *systematically* learn this simple visual relation.

Why is it that a CNN can accurately categorize natural images while struggling to recognize a simple abstract relation? That such task is very difficult for contemporary computer vision algorithms is known. Earlier, we discussed the results of Fleuret et al. (2011) who showed that black-box classifiers fail on most tasks from the synthetic visual reasoning test (SVRT), a battery of twenty-three visual-relation problems, despite massive amounts of training data. More recent work has shown how CNNs, including

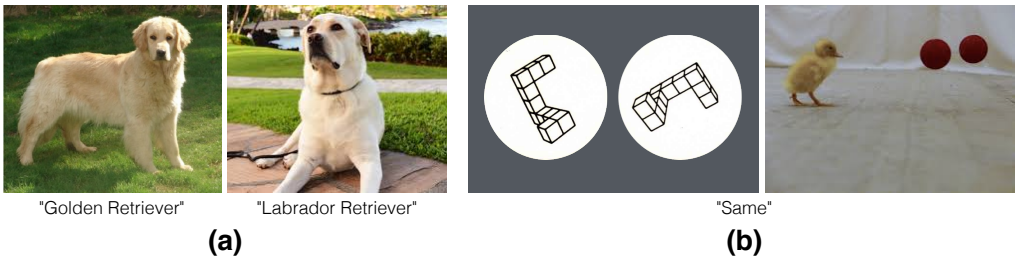


Figure 2.18 *Object recognition and visual reasoning in machines and minds.(a)* State-of-the-art convolutional neural networks can learn to categorize images (including dog breeds) with high accuracy even when the task requires detecting subtle visual cues. The same networks struggle to learn the visual recognition problems shown in panel (b). *(b)* In addition to categorizing visual objects, humans can also perform comparison between objects and determine if they are identical up to a rotation (left). The ability to recognize "sameness" is also observed in other species in the animal kingdom such as birds (right). The geometric figures are adapted from (Shepard and Metzler, 1971), and the image with a duckling is taken with permission from Martinho III and Kacelnik (2016).

variants of the popular LeNet (LeCun et al., 1998) and AlexNet (Krizhevsky, Sutskever, and Hinton, 2012) architectures, could only solve a handful of the twenty-three SVRT problems (Ellis, Solar-lezama, and Tenenbaum, 2015; Stabinger, Rodríguez-Sánchez, and Piater, 2016). Similarly, Gülçehre and Bengio (2013), after showing how CNNs fail to learn a same-different task with simple binary "sprite" items, only managed to train a multi-layer perceptron on this task by providing carefully engineered training schedules.

However, these results are not entirely conclusive. First, each of these studies only tested a small number of feedforward architectures, leaving open the possibility that low accuracy on some of the problems might simply be a result of a poor choice of model hyper-parameters. Second, while the twenty-three SVRT problems represent a diverse collection of relational concepts, the images used in each problem are also visually distinct (e.g., some relations requiring stimuli to have three items, while other require two). This makes a direct comparison of difficulty between different problems

challenging because the performance of a computational model on a given problem may be driven by specific features in that problem rather than the underlying abstract rule. To our knowledge, there has been no exploration of the limits of contemporary machine learning algorithms on relational reasoning problems. Additionally, the issue has been overshadowed by the recent success of novel architectures called "relational networks" (RNs) on seemingly challenging "visual question answering" benchmarks (Santoro et al., 2017).

In this section[9], we probe the limits of feedforward neural networks, including CNNs and RNs, on visual-relation tasks. Our goal is to demonstrate that these neural architectures fail to learn and understand these relations *systematically*. That is, learning and generalization of these relations depends strongly on nuisance parameters having no bearing on the relation. In Experiment 1, we perform a performance analysis of CNN architectures on each of the twenty-three SVRT problems, which reveals a dichotomy of visual-relation problems: hard same-different problems and easy spatial-relation problems. This result suggests that systematicity deficits depend on the nature of the relation. In Experiment 2, we introduce a novel, controlled, visual-relation challenge called parametric SVRT (PSVRT), which we use to demonstrate that the unsystematic sensitivity of CNNs to nuisance parameters in visual relations. In particular, we suggest that CNNs solve these problems via rote memorization of all possible spatial arrangements of individual items. In Experiment 3, we examine two models, the RN and a novel Siamese network, which simulate the effects of perceptual grouping and attentional routing to solve visual relations problems. We find that the former struggles to learn the notion of sameness and tends to overfit to particular relation features, but that the latter can render seemingly difficult visual reasoning problems rather trivial.

[9] A shorter version (Ricci, Kim, and Serre, 2018) of this section is to appear in the Proceedings of the *40th Annual Conference of the Cognitive Science Society*.

Overall, our study suggests that a critical reappraisal of the capability of current machine vision systems is warranted. Naturally, for each visual relation problem, there exists *some* feedforward network which solves the problem, owing to the universal approximation abilities of these models (Cybenko, 1989). Therefore, our critique is ultimately cast in terms of efficiency. We argue that quantitative measurement of systematicity in neural systems should rest on a comparison of architectural or learning efficiency between humans and machines. We further argue that mechanisms for individuating objects and manipulating their representations, presumably through feedback processes that are absent in current feedforward architectures, are closing this efficiency gap and more closely approximating the systematic visual reasoning abilities of biological intelligence.

2.3.1 Experiment 1: A dichotomy of visual-relation problems

The SVRT challenge

The Synthetic Visual Reasoning Test (SVRT) is a collection of twenty-three binary classification problems in which opposing classes differ based on whether or not images obey an abstract rule (Fleuret et al., 2011) (for a discussion of human ability on this task, see Sec. 2.2.1). For example, in problem number 1, positive examples feature two items which are the same up to translation (Figure 2.19), whereas negative examples do not. In problem 9, positive examples have three items, the largest of which is in between the two smaller ones. All stimuli depict simple, closed, black curves on a white background.

For each of the twenty-three problems, we generated 2 million examples split evenly into training and test sets using code made publicly available by the authors of the original study at http://www.idiap.ch/~fleuret/svrt.

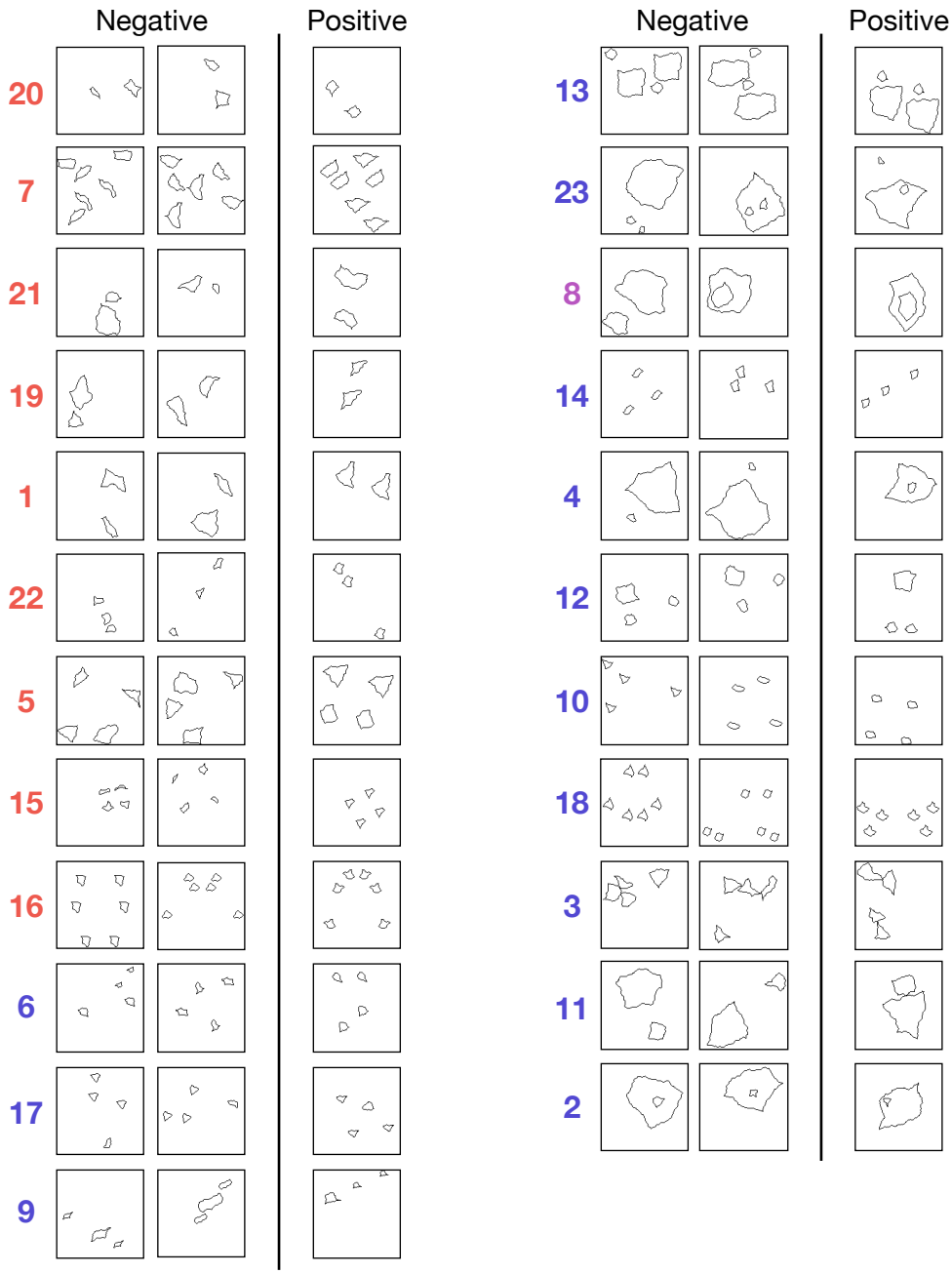


Figure 2.19 *Sample images from the twenty-three SVRT problems.* For each problem, three example images, two negative and one positive, are displayed in a row. Problems are ordered and color-coded identically to Figure 2.20. Images in each problem all respect a certain visual structure (e.g., in problem 9, three objects, identical up to a scale, are arranged in a row.). Positive and negative categories are then characterized by whether or not objects in an image obey a rule (e.g., in problem 3, an image is considered positive if it contains two touching objects and negative if it contains three touching objects.). Descriptions of all problems can be found in (Fleuret et al., 2011).

Hyper-parameter search

We tested nine different CNNs of three different depths (2, 4 and 6 convolutional layers) and with three different convolutional filter sizes (2×2, 4×4 and 6×6) in the first layer. This initial receptive field size effectively determines the size of receptive fields throughout the network. The number of filters in the first layer was 6, 12 or 18, respectively, for each choice of initial receptive field size. In the other convolutional layers, filter size was fixed at 2×2 with the number of filters doubling every layer. All convolutional layers had strides of 1 and used ReLU activations. Pooling layers were placed after every convolutional layer, with pooling kernels of size 3×3 and strides of 2. On top of the retinotopic layers, all nine CNNs had three fully connected layers with 1,024 hidden units in each layer, followed by a 2-dimensional classification layer. All CNNs were trained on all problems. Network parameters were initialized using Xavier initialization (Glorot and Bengio, 2010) and were trained using the Adaptive Moment Estimation (Adam) optimizer (Kingma and Ba, 2015) with base learning rate of $\eta = 10^{-4}$. All experiments were run using TensorFlow (Abadi et al., 2016).

Figure 2.20 shows a ranked bar plot of the best-performing network accuracy for each of the twenty-three SVRT problems. Bars are colored red or blue according to the SVRT problem descriptions given in (Fleuret et al., 2011). Problems whose descriptions have words like "same" or "identical" are colored red. These *Same-Different* (SD) problems have items that are congruent up to some transformation. *Spatial-Relation* (SR) problems, whose descriptions have phrases like "left of", "next to" or "touching," are colored blue. Figure 2.19 shows positive and negative samples for each of the corresponding twenty-three problems (also sorted by network accuracy from low to high).

The resulting dichotomy across the SVRT problems is striking. CNNs fare uniformly worse on SD problems than they do on SR problems. Many SR problems were learned satisfactorily, whereas some SD problems (e.g., problems 20, 7) resulted in accuracy

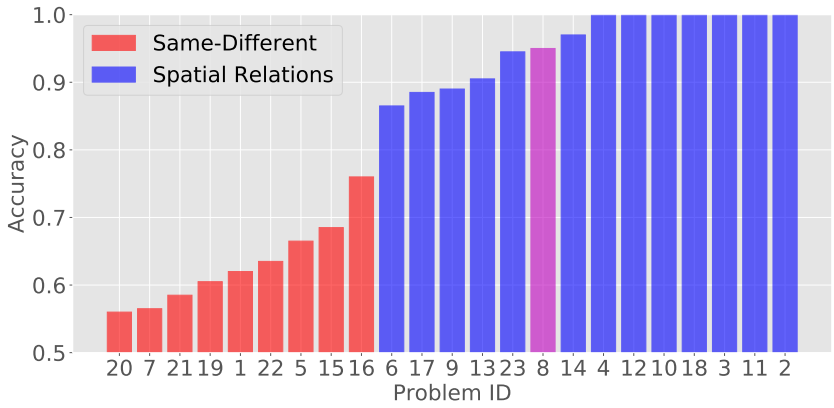


Figure 2.20 *SVRT results.* Multiple CNNs with different combinations of hyper-parameters were trained on each of the twenty-three SVRT problems. Shown are the ranked accuracies of the best-performing network optimized for each problem individually. The x-axis shows the problem ID. CNNs from this analysis were found to produce uniformly lower accuracies on same-different problems (red bars) than on spatial-relation problems (blue bars). The purple bar represents a problem which required detecting both a same-different relation and a spatial relation.

not substantially above chance. From this analysis, it appears as if SD tasks pose a particularly difficult challenge to CNNs. This is consistent with results from an earlier study by Stabinger, Rodríguez-Sánchez, and Piater (2016).

Additionally, our search revealed that SR problems are equally well-learned across all network configurations, with less than 10% difference in final accuracy between the worst and the best network. On the other hand, deeper networks yielded significantly higher accuracy on SD problems compared to smaller ones, suggesting that SD problems require a higher capacity than SR problems. Experiment 1 corroborates the results of previous studies which found feedforward neural networks performed badly on many visual-relation problems (Fleuret et al., 2011; Gülçehre and Bengio, 2013; Ellis, Solar-lezama, and Tenenbaum, 2015; Stabinger, Rodríguez-Sánchez, and Piater, 2016; Santoro et al., 2017) and suggests that low accuracy cannot be simply attributed to a poor choice of hyper-parameters (like learning rate, etc.).

We will emphasize again that results like this do not suggest that SD problems are impossible to learn in CNNs, as long as the architectures available to the model are sufficiently varied. Our results so far concern less a property of neural architectures than a property of visual relations themselves. Naturally, there is a connection between the nature of visual relations and the structure of machines capable of understanding them, a connection strongly reminiscent of the correspondence of formal languages with the automata capable of recognizing them (see Hopcroft, Motwani, and Ullman, 2008 for a detailed introduction). Future work should formalize this symmetry.

Though useful for surveying many types of relations, the SVRT challenge has two important limitations. First, different problems have different visual structure. For instance, Problem 2 (*"inside-outside"*) requires that an image contain one large item and one small item. Problem 1 (*"same-different up to translation"*), on the other hand, requires that an image contain two items, identically sized and positioned without one being contained in the other. In other cases, different problems simply require different number of items in a single image (two items in Problem 1 vs. three in Problem 9). This confound leaves open the possibility that image features, not abstract relational rules, make some problems harder than others. This confound makes it difficult to tease apart deficits in systematic understanding from perceptual deficits[10]. Instead, a better way to compare visual-relation problems would be to define various problems on the *same* set of images. Second, the *ad hoc* procedure used to generate simple, closed curves as items in SVRT prevents quantification of image variability and its effect on task difficulty. As a result, even within a single problem in SVRT, it is unclear whether its difficulty is inherent to the classification rule itself or simply results from the particular choice of image generation parameters unrelated to the rule.

[10]For example, perhaps some objects in the SVRT data set are hard to perceive for a given architecture because of poor visual acuity caused by an infelicitous choice of subsampling rate, kernel size, etc.

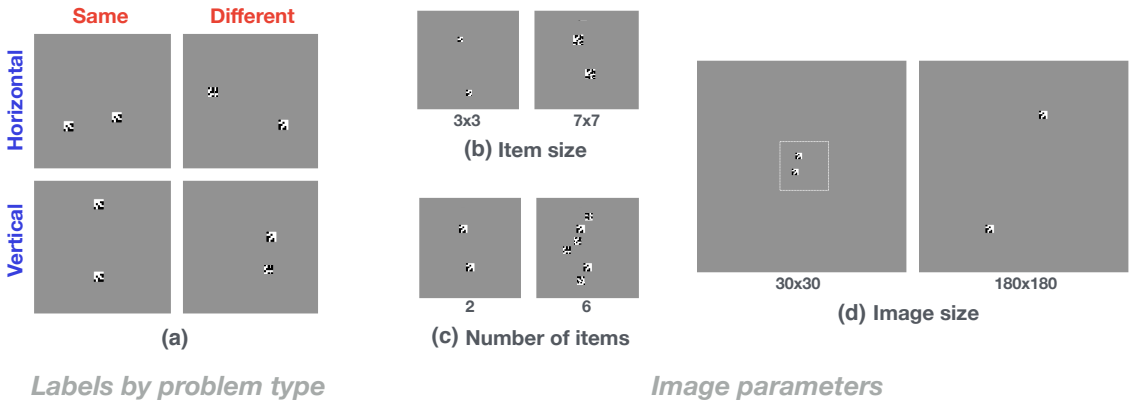


Figure 2.21 *The PSVRT challenge. (a)* Four images show the joint categories of SD (grouped by columns) and SR (grouped by rows) tasks. Our image generator is designed so that each image can be used to pose both problems by simply labeling it according to different rules. An image is *Same* or *Different* depending on whether it contains identical (left column) or different (right column) square bit patterns. An image is *Horizontal* (top row) or *Vertical* (bottom row) depending on whether the orientation of the displacement between the items is greater than or equal to 45°. These images were generated with the baseline image parameters: $m = 4$, $n = 60$, $k = 2$. *(b), (c), (d)* Six example images show different choices of image parameters used in our experiment: item size *(b)*, number of items *(c)* and image size (*(d)*, the size of an invisible central square in which items are randomly placed). All images shown here belong to *Same* and *Vertical* categories. When more than 2 items are used, SD category label is determined by whether there are at least two identical items in the image. SR category label is determined according to whether the average orientation of the displacements between all pairs of items is greater than or equal to 45°.

2.3.2 Experiment 2: Quantitative measurement of the systematic understanding of spatial-relation and same-different problems

The PSVRT challenge

To address the limitations of SVRT, we constructed a new visual-relation benchmark consisting of two idealized problems (Figure 2.21) from the dichotomy that emerged from Experiment 1: *Spatial Relations* (SR) and *Same-Different* (SD). Critically, both problems used exactly the same images, but with different labels. Further, we parameterized the dataset so that we could systematically control various image parameters, namely, the size of scene items, the number of scene items, and the size of the whole image. Items were binary bit patterns placed on a blank background.

For each configuration of image parameters, we trained a new instance of a single CNN architecture and measured the ease with which it fit the data. Our goal was to examine how hard it is for a CNN architecture to learn relations for visually different but conceptually equivalent problems. For example, imagine two instances of the same CNN architecture, one trained on a same-different problem with small items in a large image, and the other trained on large items in a small image. If the CNNs can truly learn the "rule" underlying these problems, then one would expect the models to learn both problems with more-or-less equal ease. However, if the CNNs only memorize the distinguishing features of the two image classes, then learning should be affected by the variability of the example images in each category. For example, when image size and items size are large, there are simply more possible samples, which might put a strain on the representational capacity of a CNN trying to learn by rote memorization. In other words, we equate a sensitivity in learning ability to non-rule-related image properties with a deficit in visual systematicity.

In rule-based problems such as visual relations, these two strategies can be dis-

tinguished by training and testing the same architecture on a problem instantiated over a multitude of image distributions. Here, our main question is not whether a model trained on one set of images can accurately predict the labels of another, unseen set of images sampled from the same distribution. Rather, we want to understand whether an architecture that can easily learn a visual relation instantiated from one image distribution (defined by one set of image parameters) can also learn the same relation instantiated from another distribution (defined by another set of parameters) with equal ease by taking advantage of the abstractness of the visual rule. Evidence that CNNs use rote memorization of examples was found in a study by Stabinger and Rodriguez-Sanchez (2017), who tested state-of-the-art CNNs on variants of same-different visual relation using a dataset of realistically rendered images of checkerboards. Stabinger and Rodriguez-Sanchez (2017) found that CNN accuracy was lower on data sets whose images are rendered with higher degrees of freedom in viewpoint. In our study, we take a similar approach while using much simpler synthetic images where we can explicitly compute intra-class variability as a function of image parameters. This way, we do not introduce any additional perceptual nuisances such as specularity or 3D rotation whose contribution to image variability and CNN performance is difficult to quantify. Because PSVRT images are randomly synthesized, we generate training images on-line without explicitly reusing data, and there is no hold-out set in this experiment. Thus, we use training accuracy to measure the ease with which a model learns a visual-relation problem.

Methods

Our image generator produces a gray-scale image by randomly placing square binary bit patterns (consisting of values 1 and -1) on a blank background (with value 0). The generator uses three parameters to control image variability: the size (m) of each bit pattern or item, the size (n) of the input image and the number (k) of items in

an image. Our parametric construction allows a dissociation between two possible factors that may affect problem difficulty: classification rules vs. image variability. To highlight the parametric nature of the images, we call this new challenge the *parametric SVRT* or *PSVRT*.

Additionally, our image generator is designed such that each image can be used to pose both problems by simply labeling it according to different rules (Figure 2.21). In SR, an image is classified according to whether the items in an image are arranged horizontally or vertically as measured by the orientation of the line joining their centers (with a 45° threshold). In SD, an image is classified according to whether or not it contains at least two identical items. When $k \geq 3$, the SD category label is determined by whether or not there are *at least 2* identical items in the image, and the SR category label is determined according to whether the *average* orientation of the displacements between all pairs of items is greater than or equal to 45°. Each image is generated by first drawing a joint class label for SD and SR from a uniform distribution over $\{\textit{Different}, \textit{Same}\} \times \{\textit{Horizontal}, \textit{Vertical}\}$. The first item is sampled from a uniform distribution in $\{-1, 1\}^{m \times m}$. Then, if the sampled SD label is *Same*, between 1 and $k - 1$ identical copies of the first item are created. If the sampled SD label is *Different*, no identical copies are made. The rest of k unique items are then consecutively sampled. These k items are then randomly placed in an $n \times n$ image while ensuring at least 1 background pixel spacing between items. Generating images by always drawing class labels for both problems ensures that the image distribution is identical between the two problem types.

We trained the same CNN repeatedly from scratch over multiple subsets of the data in order to see if learnability depends on the dataset's image parameters. CNNs were trained on 20 million images and training accuracy was sampled every 200 thousand images. These samples were averaged across the length of a training run as well as over multiple trials for each condition, yielding a scalar measure of learnability

called "mean area under the learning curve" (mean ALC). ALC is high when accuracy increases earlier and more rapidly throughout the course of training and/or when it converges to a higher final accuracy by the end of training.

First, we found a baseline architecture which could easily learn both same-different and spatial-relation PSVRT problems for one parameter configuration (item size $m = 4$, image size $n = 60$ and item number $k = 2$). Then, for a range of combinations of item size, image size and number of items, we trained an instance of this architecture from scratch. If a network learns the underlying rule of each visual relation, the resulting representations will be efficient at handling variations unrelated to the relation (e.g., a feature set to detect *any* pair of items arranged horizontally). As a result, the network should be equally good at learning the same problem in other image datasets with greater intra-category variability. In other words, ALC will be consistently high over a range of image parameters. Alternatively, if the network's architecture doesn't allow for such representations and thus is only able to learn prototypes of examples within each category, the architecture will be progressively worse at learning the same visual relation instantiated with higher image variability. In this case, ALC will gradually decrease as image variability increases.

The baseline CNN we used in this experiment had four convolutional layers. The first layer had 8 filters with a 4×4 receptive field size. In the rest of convolutional layers, filter size was fixed at 2×2 with the number of filters in each layer doubling from the immediately preceding layer. All convolutional layers had ReLU activations with strides of 1. Pooling layers were placed after every convolutional layer, with pooling kernels of size 3×3 and strides of 2. On top of retinotopic layers, all nine CNNs had three fully connected layers with 256 hidden units in each layer, followed by a 2-dimensional classification layer. All network parameters were initialized using Xavier initialization (Glorot and Bengio, 2010) and were trained using the Adaptive Moment Estimation (Adam) optimizer (Kingma and Ba, 2015) with base learning

rate of $\eta = 10^{-4}$. All experiments were run using TensorFlow (Abadi et al., 2016). To understand the effect of network size on learnability, we also used two control networks in this experiment: (1) a 'wide' control that had the same depth as the baseline but twice as many filters in the convolutional layers and four times as many hidden units in the fully connected layers and (2) and a 'deep' control which had twice as many convolutional layers as the baseline, by adding a convolutional layer of filter size 2×2 after each existing convolutional layer. Each extra convolutional layer had the same number of filters as the immediately preceding convolutional layer.

We varied each of three image parameters separately to examine its effect on learnability. This resulted in three sub-experiments (n was varied between 30 and 180 while m and k were fixed at 4 and 2, respectively; m was varied between 3 and 7, while n and k were fixed at 60 and 2, respectively; k was varied between 2 and 6 while n and m were fixed at 60 and 4, respectively). To use the same CNN architecture over a range of image sizes n, we fixed the actual input image size at 180 by 180 pixels by placing a smaller PSVRT image (if $n < 180$) at the center of a blank background of size 180 by 180 pixels. The baseline CNN was trained from scratch in each condition with 20 million training images and a batch size of 50.

Results

In all conditions, we found a strong dichotomy in the observed learning curves. In cases where learning occurred, training accuracy abruptly jumped from chance-level and gradually plateaued. We call this sudden, dramatic rise in accuracy the "learning event". When there was no learning event, accuracy remained at chance throughout a training session and the ALC was 0.5. Strong bimodality was observed even within a single experimental condition in which the learning event took place in only a subset of 10 randomly initialized trials. This led us to use two different quantities for describing a model's performance: (1) mean ALC obtained from *learned* trials (in which accuracy

crossed 55%) and (2) the number of trials in which the learning event never took place (*non-learned*). Note that these two quantities are independent, computed from two complementary subsets of 10 trials.

In SR, across all image parameters and in all trials, the learning event immediately occurred at the start of training and quickly approached 100% accuracy, producing consistently high and flat mean ALC curves (Figure 2.22, blue dotted lines). In SD, however, we found that the overall ALC was significantly lower than SR (Figure 2.22, red dotted lines).

In addition, we have also identified two main ways in which image variability affects learnability. First, among the trials in which the learning event did occur, the final accuracy achieved by the CNN at the end of training gradually decreased as image size (n) of the number of items (k) increased. This caused ALC to decrease from around 0.95 to 0.8. Second, increasing image size (n) also made the learning event decreasingly likely, with more than half of the trials failing to escape chance level when image size was greater than 60 (Figure 2.22, gray bars). We call this systematic degradation of performance accompanied by the increase in image variability the *"straining effect"*. In contrast, increasing item size produced no visible straining effect on the CNN. Similar to SR, learnability, both in terms of the frequency of the learning event as well as final accuracy, did not change significantly over the range of item sizes we considered.

The fact that straining is only observed in SD, and not in SR and that it is only observed along some of the image parameters, n and k, suggests that straining is not simply a direct outcome of an increase in image variability. Using a CNN with more than twice the number of free parameters (Figure 2.22, purple dotted lines) or with twice as many convolutional layers (Figure 2.22, brown dotted lines) as a control did not qualitatively change the trend observed in the baseline model. Although increasing network size did result in improved learned accuracy in general, it also made learning

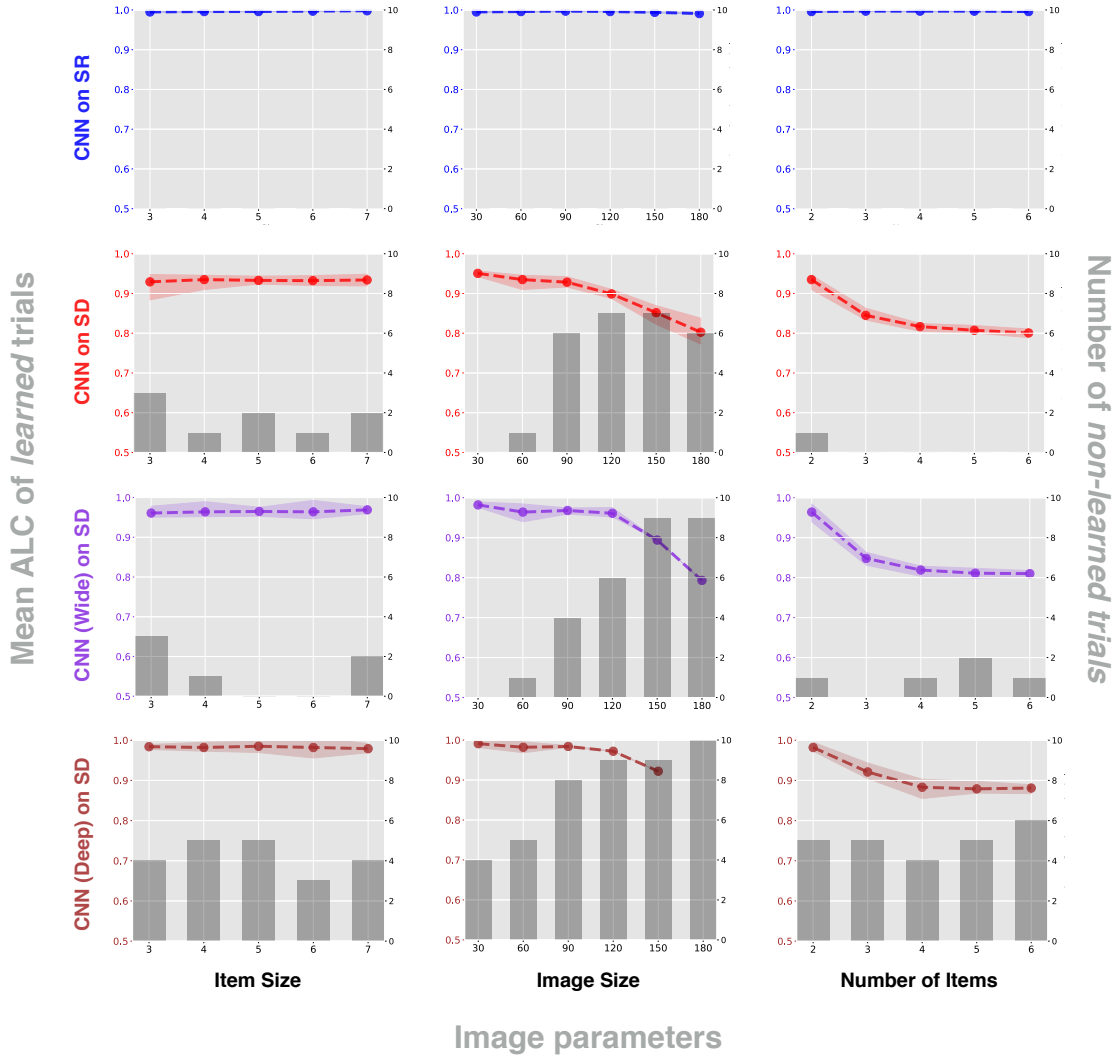


Figure 2.22 *Mean area under the learning curve (ALC) over PSVRT image parameters.* ALC is the normalized area under a training accuracy curve over the course of training on 20 million images. Colored dots are the mean ALCs of learned trials (trials in which validation accuracy exceeded 55%) out of 10 randomly initialized trials. Shaded regions around the colored dots indicate the intervals between the maximum and the minimum ALC among learned trials. Gray bars denote the number of non-learned trials, out of 10 trials, in which validation accuracy never exceeded 55%. Three model-task combinations (CNN on SR (blue), CNN on SD (red), wide CNN control on SD (violet) and deep CNN control on SD (brown)) are plotted, and each combination is explored over three image variability parameters: item size, image size and number of items.

less likely, yielding more non-learned trials than the baseline CNN.

We also rule out the possibility of the loss of spatial acuity from pooling or subsampling operations as a possible cause of straining for two reasons. First, our CNNs achieved the best overall accuracy when image size was smallest. If the loss of

spatial acuity was the source of straining, increasing image size should have improved the network's performance instead of hurting it because items would have tended to be placed farther apart from each other. Second, as we will show in Experiment 3.2, an identical convolutional network where objects are forcibly separated into different channels does not exhibit any straining, suggesting that it is not the loss of spatial acuity per se that makes the SD problem difficult, but rather the fact that CNNs lack the ability to spatially separate representations of individual items in an image.

We hypothesize that these straining effects reflect the way positioning of each item contributes to image variability. A little arithmetic shows that image variability is an exponential function of image size as the base and number of items as the exponent. Thus, increasing image size while fixing the number of items at 2 results in a quadratic-rate increase in image variability, while increasing the number of items leads to an exponential-rate increase in image variability. Image variability is also an exponential function of item size as the exponent and 2 (for using binary pixels) as the base.

The comparatively weak effects of item size and item number shed light on the computational strategy used by CNNs to solve SD. Our working hypothesis is that CNNs learn "subtraction templates", filters with one positive region and one negative region (like a Haar or Gabor wavelet), in order to detect the similarity between two image regions. A different subtraction template is required for each relative arrangement of items, since each item must lie in one of the template's two regions. When identical items lie in these opposing regions, they are effectively subtracted by the synaptic weights. This difference is then used to choose the appropriate same/different label. Note that this strategy does not require memorizing specific items. Hence, increasing item size (and therefore total number of possible items) should not make the task appreciably harder. Further, a single subtraction template can be used even in scenes with more than two items, since images are classified as "same" when they

have *at least* two identical items. So, any straining effect from item number should be negligible as well. Instead, the principal straining effect with this strategy should arise from image size, which increases the possible number arrangements of items.

Taken together, these results suggest that, when CNNs learn a PSVRT condition, they are simply building a feature set tailored to the relative positional arrangements of items in a particular data set, instead of learning the abstract "rule" per se. If a network is able to learn features that capture the visual relation at hand (e.g., a feature set to detect *any* pair of items arranged horizontally), then these features should, by definition, be minimally sensitive to the image variations that are irrelevant to the relation. This seems to be the case only in SR. In SD, increasing image variability lowered ALC for the CNNs. This suggests that the features learned by CNN are not invariant rule-detectors, but rather merely a collection of templates covering a particular distribution in the image space.

2.3.3 Experiment 3: Is object individuation needed to solve visual relations?

Our main hypothesis is that CNNs struggle to learn systematically visual relations in part because they are feedforward architectures which lack a mechanism for grouping features into individuated objects. Recently, however, Santoro et al. (2017) proposed the relational network (RN), a feedforward architecture aimed at learning visual relations without such an individuation mechanism. RNs are fully-connected feedforward networks which operate on pairs of so-called "objects" (Figure 2.23; for concision, we will refer to a neural network consisting of a CNN feeding into an RN as just an "RN"). These objects are simply feature columns from all retinotopic locations in a deep layer of a CNN, similar to the feature columns found in higher areas of the visual cortex (Tanaka, 2003). These feature vectors will sometimes represent parts of the background, incomplete items or even multiple items because the network does

not explicitly represent individual objects. This makes the "objects" used by an RN rather different from those discussed in the psychophysical literature, where perceptual objects are speculated to obey gestalt rules like boundedness and continuity (Spelke et al., 1994). Santoro et al. (2017) emphasize that their model performed well even though it employs this highly unstructured notion of object: "A central contribution of this work is to demonstrate the flexibility with which relatively unstructured inputs, such as CNN or LSTM [long short-term memory] embeddings, can be considered as a set of objects for an RN."

In particular, the RN was able to outperform a baseline CNN on the "sort-of-CLEVR" challenge, a visual question answering task using images with simple geometric items (see Figure 2.24(a) for examples of sort-of-CLEVR items). In sort-of-CLEVR, scenes contain up to six items, each of which has one of two shapes and six colors. The RN was trained to answer both relational questions (e.g., *"What is the shape of the object that is farthest from the gray object?"*) and non-relational questions (e.g., *"Is the red object on the top or bottom of the scene?"*).

However, the "sort-of-CLEVR" tasks suffers from three important shortcomings. First, the number of possible items is exceedingly small (6 colors $\times$ 2 shapes $=$ 12 items). Combined with the fact that the authors used rather small (75 $\times$ 75) images, this means the total number of sort-of-CLEVR stimuli was rather low, at least compared to PSVRT stimuli. The small number of samples in sort-of-CLEVR might have encouraged the RN to use rote memorization instead of actually learning relational concepts. Second, while the authors trained the RN to compare the attributes of scene items (e.g., *"How many objects have the same shape as the green object."*), they did not examine if the model could learn the concept of sameness, per se (e.g., *"Are any two items the same in this scene?"*). Detecting sameness is a particularly hard task because it requires matching all attributes between all pairs of items. Third, sort-of-CLEVR stimuli are not parameterized as they are in PSVRT; one cannot

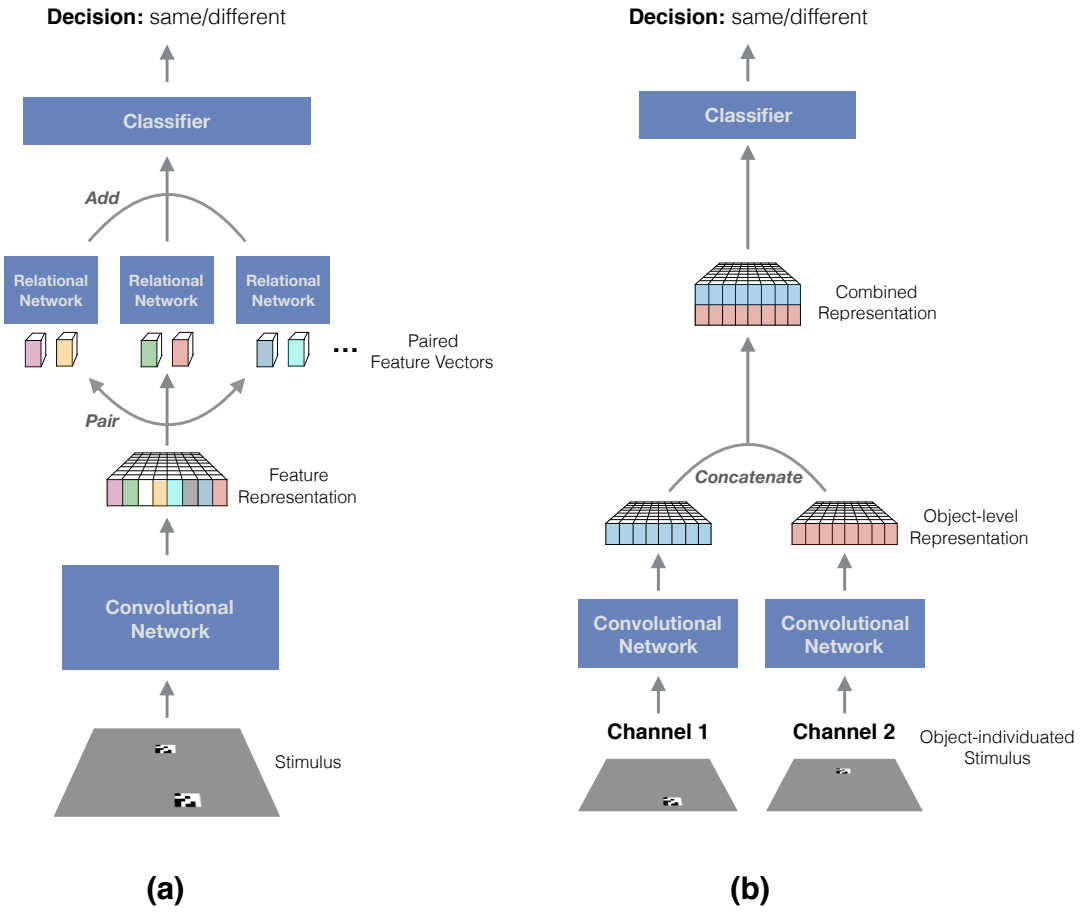


Figure 2.23 *A comparison between a relational network and the proposed Siamese architecture.* (*a*) A relational network (panel (*a*), top half) is a fully-connected, feedforward neural network which accepts pairs of CNN feature vectors as input. First, the image is passed through a CNN to extract features. Every pair of feature activations ("objects") at every retinotopic location in the final CNN layer is passed through the RN. The outputs of the RN on every pair of activations is then summed and passed through a final feedforward network, producing the decision. Depending on the spatial resolution of the final CNN layer and the receptive field of each unit, the object representations of an RN may correspond to a single scene item, multiple items, partial items or even the background. (*b*) In contrast, objects in our Siamese network are forced to contain a single item. First, we split stimuli into several images, each containing a single item. Then, each of the images is passed through a separate CNN (here, Channel 1 and Channel 2), producing a representation of a single object. These objects are then combined by concatenation into a single representation and passed through a classifier. The network simulates the effects of the attentional and perceptual grouping processes suspected to underlie biological visual reasoning (see Discussion).

systematically vary image features while keeping the abstract rule fixed. Thus, it is difficult to say whether the success of RNs arises from their ability to flexibly learn relations among arbitrary objects (as is hypothesized for humans (Franconeri et al., 2012)) or rather their ability to fit particular image features.

Crucially, without a parameterized dataset, it is difficult to evaluate the authors' claim regarding the efficacy of "relatively unstructured" objects in visual reasoning problems. Since the objects used by RNs are simply feature columns, they have a fixed receptive field. Thus, the success of RNs on sort-of-CLEVR might be due to felicitously sized and arranged items instead of actual relational learning. For, if image features are allowed to parametrically vary, such spatially rigid representations might fail to correctly encode individual objects whenever, for instance, multiple, small and tightly-arranged items fall within the same receptive field or when a large, irregularly-shaped item spans multiple receptive fields.

Our goal in Experiment 3 was to re-evaluate relational networks on sort-of-CLEVR when these handicaps are removed. To that end, we performed four sub-experiments. First, we trained RNs on a bona fide same-different task using versions of sort-of-CLEVR missing certain color-shape combinations in order to see if the model would overfit to training item attributes (see (Johnson et al., 2017) for a similar demonstration in a different visual reasoning problem). Such over-fitting would indicate that the RN merely memorizes particular item combinations instead of learning systematic rules. Second, we trained RNs on a sort-of-CLEVR same-different task in which certain locations for objects were left out in order to see if the model would overfit to training item locations. Third, we tested an RN on PSVRT in order to evaluate the ease with which the model can fit data when scene items systematically vary in appearance and arrangement. As in Experiment 2, we measured mean ALC in order to see if the RN's object representations alleviated the straining found in CNNs.

Finally, we compared the performance of the RN on PSVRT to that of an idealized

model using ground-truth object individuation. Our new model is a "Siamese" network (Bromley et al., 1994) which processes each scene item in a separate (CNN) channel and then passes the processed items to a single classifier network. This model simulates the effects of attentional selection and perceptual grouping by segregating the representations of each item. Unlike an RN, whose object representations may in fact contain no item, multiple items or incomplete items, object representations in the Siamese network contain exactly one item.

Methods

Sub-experiment 3.1: Relational transfer to novel attribute combinations Here, we sought to measure the ability of an RN to transfer the concept of sameness from a training set to a novel set of objects, a classic and very well-studied paradigm in animal psychology (see (Wright and Kelly, 2017) for a review) and thus an important benchmark for models of visual reasoning. We used software for relational networks publicly available at https://github.com/gitlimlab/Relation-Network-Tensorflow. Like the original architecture used by Santoro et al. (2017), our RN had four convolutional layers with ReLU non-linearities and batch normalization. We used 24 features for each convolutional layer, fewer than those used by (Santoro et al., 2017), but sufficient for good training accuracy. These convolutional layers were followed by two four-layer MLPs, both with ReLU non-linearities. These MLPs had 256 features each, again fewer than those in (Santoro et al., 2017), but sufficient for fitting the data. The final classification layer had a softmax nonlinearity and the whole network was optimized with a cross-entropy loss using an Adam optimizer with learning rate $\eta = 10^{-4}$ and mini-batches of size 64. The original authors did not report receptive field sizes or strides. Our RN used receptive field sizes of 5×5 throughout the convolutional layers and had strides of 3 in the first two convolutional layers and strides of 2 in the next two. There was no pooling. We confirmed that this model was able to reproduce the

results from (Santoro et al., 2017) on the sort-of-CLEVR task.

We constructed twelve different versions of the sort-of-CLEVR dataset, each one missing one of the twelve possible color $\times$ shape attribute combinations (see Figure 2.24(a)). Images in each dataset only depicted two items, randomly placed on a 128×128 background. Half of the time, these items were the same (same color and same shape). For each dataset, we trained the RN architecture to detect the possible sameness of the two scene items while measuring validation accuracy on the left-out images. We then averaged training accuracy and validation accuracy across all of the left-out conditions.

Sub-Experiment 3.2: Relational transfer to novel locations Next, we measured the ability of a relational network to transfer the concept of sameness to novel spatial arrangements of objects. We used exactly the same network as in Sub-experiment 3.1 except that strides were set to 1 to reduce information loss throughout processing. However, we now generated two new versions of sort-of-CLEVR in which each item was entirely contained in one high-level receptive field of size 17×17. The "far" data set forced each of the two items in the same-different relation to lie in different receptive fields. The "close" data set forced the items to be in same high-level receptive field. We hypothesized that the RN could not generalize to either of these data sets if trained on the other. Because the RN's comparator mechanism is wholly dependent on the arbitrary retinotopic grid created by the convolutional layers, there is no inherent reason the RN should systematically transfer its knowledge about objects with one relation to the grid to another set of objects with a differing relation to this grid.

Further, to confirm that the grid of high-level receptive fields was truly the determining factor in RN performance rather than the combined CNN-RN system, we trained the model to detect the sameness of objects (not constrained by a retinotopic

grid) while keeping the convolutional layers *random*.

Sub-experiment 3.3: Relational Networks on PSVRT For this experiment, we trained an RN on our Experiment 2 with PSVRT stimuli and observed whether the straining effect found in CNNs was alleviated in RNs. For this sub-experiment, we used the exact architecture from sub-experiment 3.1, but increased the number of units to the original values from (Santoro et al., 2017) in order to give the RN the best possible chance of learning the very difficult PSVRT task. The convolutional layers had 32, 64, 128 and 256 features, the first MLP had 2,000 units in each layer, and the final MLP had 2,000, 1,000, 5,000 and 100 units in its four layers. We focused only on same-different learning and only varied image size from 30 to 180 pixels since this produced the strongest straining effect in CNNs. Item size was fixed at 4 and the number of items was fixed at 2. We trained on 20 million images, using ten randomly initialized trials. As in Experiment 2, we measured mean ALC as well as number of non-learned trials. Before training on the whole spectrum of image sizes, we ensured that the RN was capable of fitting the data when item size was 4 and image size was 60.

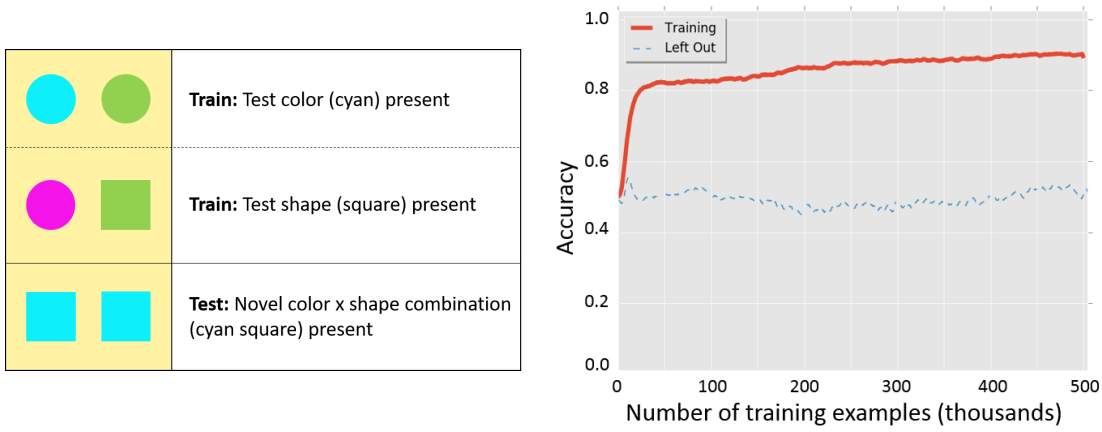
Sub-experiment 3.4: The need for perceptual grouping and object individuation Here, we introduce a Siamese network which processes scene items individually in separate CNN "channels" (Figure 2.23(a)). First, we manually split each PSVRT stimulus into several images, each of which contained a single item. These images were then individually processed by two copies of the same network (mimicking, in a sense, the process of sequentially attending to individuated objects). For example, if one stimulus contained two objects in the original PSVRT, our new stimulus would be presented to the Siamese network as two separate images. The scene items retained their original location in each image so that item position varied just as widely as in the original PSVRT. These images were then individually processed by each CNN channel,

using the same architecture as in Experiment 2. This resulted in two object-separated feature maps in the topmost retinotopic layer (Figure 2.23(b)). These feature maps were then concatenated before being passed to the fully-connected classifier layers.

This Siamese configuration is essentially an idealized version of the kinds of object representations resulting from psychological processes such as perceptual grouping and attentional selection. Because convolutional layers in this configuration are now constrained to process only one object at a time, regardless of the total number of objects presented in an image, the network can completely disregard the positional information of individual objects and only preserve information about their identities under comparison.

Results

Sub-experiment 3.1: Relational transfer to novel attribute combinations From the sort-of-CLEVR transfer task, we found that the RN does not generalize on average to left-out color-shape attribute combinations (Figure 2.24). Since there are only 11 color-shape combinations in any given setup, the model did not need to learn to generalize across many items. As a result, the RN learned orders of magnitude faster than the CNNs in Experiment 2; e.g., average training accuracy (solid red) exceeded 80% within 50,000 examples. However, while the average training accuracy curve rose rapidly to around 90%, the average validation accuracy remained at chance. In other words, there was no transfer of same-different ability to the left-out condition, even though the attributes from that condition (e.g., cyan square) were represented in the training set, just not in that combination (e.g., cyan circle and green square; Figure 2.24a).



(a) (b)

Figure 2.24 *Attribute hold-out reveals lack of systematic relational reasoning in RNs. (a) Sample items used during training and testing in Experiment 3.* We trained relational networks (RNs) on twelve two-item same-different data sets each missing one color-shape combination from sort-of-CLEVR (2 shapes $\times$ 6 colors). Then, we tested the model on the left-out combination. The top and middle rows of panel *(a)* show two possible pairs of items when the left-out combination is "cyan square". Row 1 shows a cyan circle and row 2 shows a green square. However, only in the test set is the model queried about images involving a cyan square (e.g., the "same" image in row 3). Note that, during training, the model observes each left-out attribute, just not in the left-out combination. *(b)* Averaged accuracy curves of an RN while being trained on the sort-of-CLEVR data sets missing one color-shape combination. The red curve shows the training accuracy. The blue dashed line shows the accuracy on validation data with the left-out items.

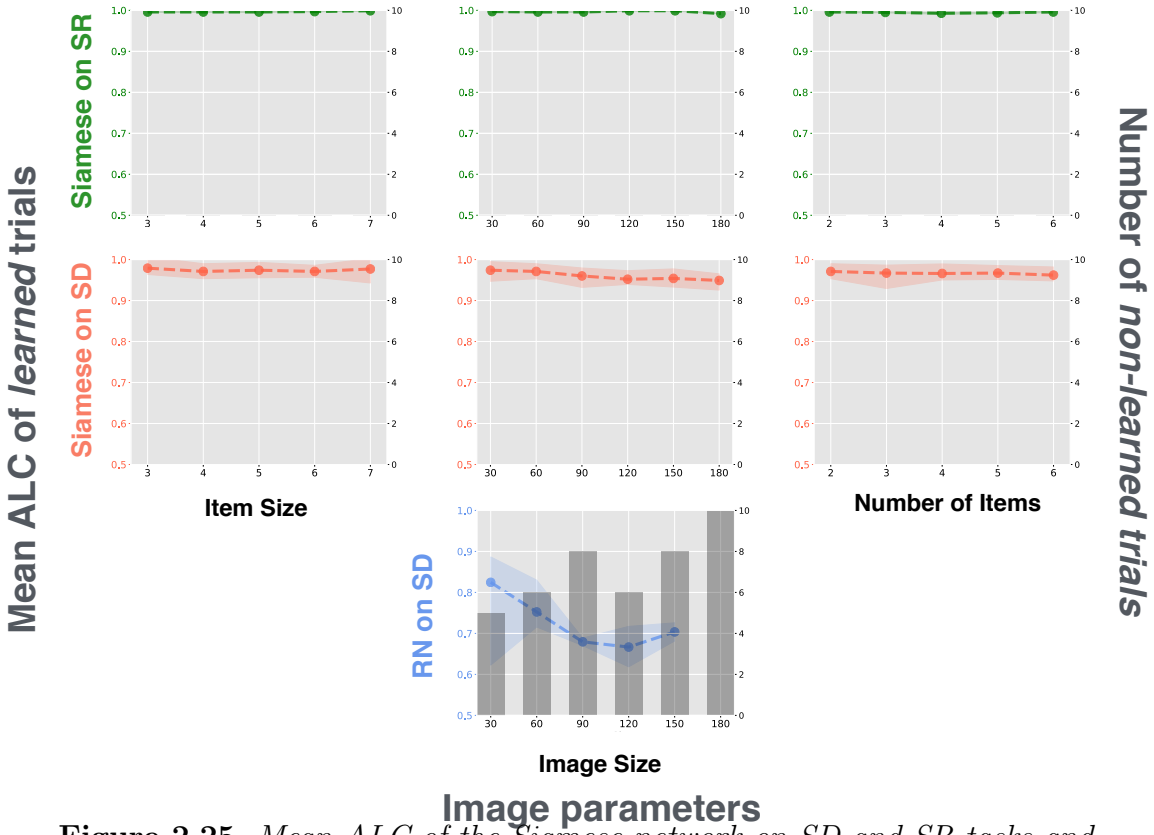
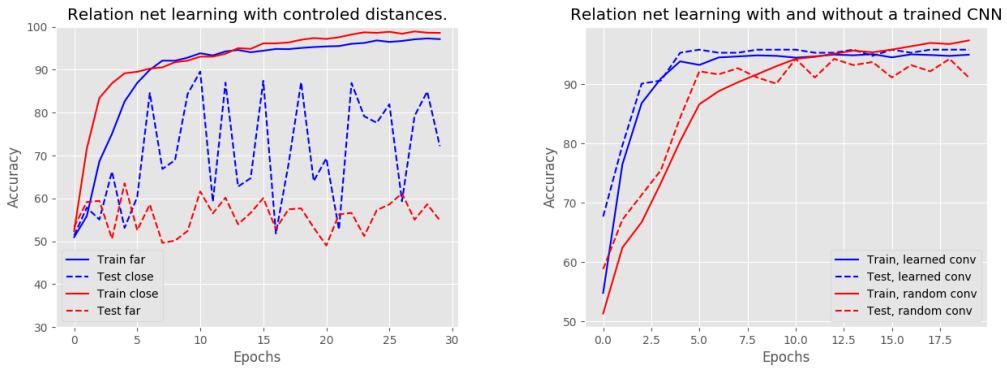


Figure 2.25 *Mean ALC of the Siamese network on SD and SR tasks and the RN on SD over image sizes.* Unlike for CNNs, mean ALC curves of the Siamese network exhibit no significant straining. The network learns equally well on all datasets with different image variability parameters. The significant difference between SD and SR conditions observed with CNNs is no longer present in the Siamese network. In contrast, the RN exhibits a strong straining effect that is qualitatively similar to CNNs, with the average ALC as well as the probability of learning decreasing as image size increases.

Sub-experiment 3.2: Relational transfer to novel object locations We found that the RN struggles to transfer its knowledge about sameness in particular locations to new locations. Fig. 2.26a depicts the training and testing accuracy of the RN in two conditions, one in which the model was trained on the "far" data and tested on the "close" data (blue curves) and the opposite setting (red curves). Naturally, the RN can fit the training data in both cases. However, when the model tried to generalize to data in which objects were contained within a single receptive field, accuracy oscillated erratically between 50 and 90 %. This is to be expected, since the RN comparator operates *between* high-level feature columns and not *within* them. Nevertheless, the fact that accuracy was often significantly above chance in this condition means that some information about the sameness relation must be encoded in the convolutional features, at least when the system is trained in this manner. Transfer from the "close" to the "far" condition is much worse, hovering between 55 and 60 % for the duration of training. One interpretation of this result is that the RN learns to extract the sameness signal with the convolutional layers and simply pass it unperturbed to the final decision layer, essentially transforming the rest of the system into a glorified identity function. The comparator operating between high-level feature columns is thereby disabled, preventing systematic transfer to the held-out locations. Next, to emphasize that the RN is simultaneously too elaborate (in the sense that the "perceptual" convolutional layers are irrelevant) and too weak (in the sense that the network displays the deficits in systematicity we have already noted) for use in relational reasoning, we trained the network while keeping its convolutional layers random. Fig. 2.26b shows the model learning the same-different task of Sub-Experiment 3.1 (without held-out attribute combinations) when the convolutional layers were learned (blue curves) and left random (red curves). There is evidently no need to train the convolutional layers, since performance is functionally identical in both conditions. Again, this is to be expected, since any injective mapping from



(a) **(b)**

Figure 2.26 *Location systematicity deficits in RNs. (a)* A relational net was trained on one of two same-different sort-of-CLEVR data sets: a "far" set in which the two objects were forced to lie entirely in two separate high-level receptive fields and a "close" set in which the objects were constrained to lie within the same receptive field. The same model was trained on one (solid lines; blue for "far" and red for "close") and tested on the other (dashed lines; blue for "far" and red for "close"). In neither case could the network generalize to held out data, indicating that the system's comparator mechanism is arbitrarily constrained by the arrangement of its high-level receptive fields. *(b)* To emphasize that an RN is really only as good as its comparator, we also trained the system on same-different sort-of-CLEVR without any held-out data but while keeping the CNN part fixed. The system learns as easily as before.

the input to the comparator mechanism will, by definition, preserve the veridical same-different relation between the items. Even a random mapping will do.

Sub-experiment 3.3: Relational Networks on PSVRT We found that the RN exhibits a qualitatively similar straining effect to increasing image size (Figure 2.25, pale blue dotted lines). Similar to CNNs in Experiment 2, the mean ALC of learned trials gradually decreased as image size increased, together with the observed likelihood of learning out of 10 restarts. Since the top retinotopic feature vectors that are treated as "object representations" in RN have rather large, fixed and highly overlapping receptive fields, the RN is strained just as easily as regular CNNs. In order to accommodate this fixed architecture, the RN must learn a dictionary of features

that captures all arrangements of items for a given image size condition. This is an increasingly difficult feat as the image size grows, straining the model heavily until it simply cannot learn at all in the final condition (image size 180×180).

Sub-experiment 3.4: The need for perceptual grouping and object individuation The mean ALC curves for the Siamese network on PSVRT were strikingly different from those of the CNN in Experiment 2 (Figure 2.25, first and second rows). Barely any straining effect was observed on the SD task, and the model learned within 5 million examples across all image size parameters in either the SD or SR tasks. In SD, since objects are individuated by fiat, the network need not learn all possible spatial arrangements of items. The network must simply learn to compare whichever two items reach the classifier layers through the two CNN channels. This greatly simplifies the SD problem, alleviating straining. In both SD and SR, the Siamese network can learn to flexibly represent the task-relevant properties of each object such that learnability is not at all influenced by image variability. In other words, a feedforward network, once endowed with object individuation, can easily construct invariant feature representations with which arbitrary objects can be related.

This result implies that object individuation makes visual relation a rather trivial problem for feedforward networks. In informal experiments (data not shown) we found that even very shallow Siamese networks (e.g. with one convolutional layer) could still learn SD much faster than baseline CNNs. Naturally, we do not intend our Siamese network as a bona fide solution to visual reasoning, but rather as a proof of the efficacy of object individuation in visual reasoning problems. A genuine visual reasoning model would be able to dynamically select and group features in the scene (see Discussion section).

Discussion

Recent progress in computational vision has been significant. Modern deep learning architectures can discriminate between one thousand object categories (He et al., 2015) and identify faces among millions of distractors (Kemelmacher-Shlizerman et al., 2016) at a level approaching – and possibly surpassing that of human observers. While these neural networks do not aim to mimic the organization of the visual cortex in detail, they are at least partly inspired by biology. Modern deep learning architectures are indeed closely related to earlier hierarchical models of the visual cortex albeit with much better categorization accuracy (see Serre, 2015; Kriegeskorte, 2015, for reviews). Further, CNNs have been shown to account well for monkey inferotemporal data (Yamins et al., 2014) and human lateral occipital data (Khaligh-Razavi and Kriegeskorte, 2014; Guclu and Gerven, 2015). In addition, deep networks have been shown to be consistent with a number of human behaviors including rapid visual categorization (Kheradpisheh et al., 2016; Eberhardt, Cader, and Serre, 2016), image memorability (Dubey et al., 2015), typicality (Lake, Salakhutdinov, and Tenenbaum, 2015a) as well as similarity (Peterson, Abbott, and Griffiths, 2016) and shape sensitivity (Kubilius, Bracci, and Op de Beeck, 2016) judgments.

Concurrently, a growing body of literature has been highlighting key dissimilarities between current deep network models and various aspects of visual cognition. One prominent example is adversarial perturbation (Goodfellow, Shlens, and Szegedy, 2015), a type of structured image distortion that asymmetrically affects CNNs and humans. Although barely perceptible to a human observer, adversarial perturbation renders an image unrecognizable to a CNN, even though the same CNN can correctly recognize the unperturbed image with high confidence. Another example is the poor generalization of CNNs in conditions that pose no difficulty to human observers, such as learning novel object categories with minimal supervision or when the parts of a familiar object are shown in unfamiliar but realistic configurations (Lake, Salakhutdinov, and

Tenenbaum, 2015b; Saleh, Elgammal, and Feldman, 2016; Erdogan and Jacobs, 2017). Direct evidence for qualitatively different feature representations used by humans and CNNs was shown in (Ullman et al., 2016; Linsley et al., 2017).

The present study adds to this body of literature by demonstrating feedforward neural networks' fundamental inability to efficiently and robustly learn visual relations. Our results indicate that visual-relation problems can quickly exceed the representational capacity of feedforward networks. While learning feature templates for single objects appears tractable for modern deep networks, learning feature templates for *arrangements* of objects becomes rapidly intractable because of the combinatorial explosion in the requisite number of templates. That notions of "sameness" and stimuli with a combinatorial structure are difficult to represent with feedforward networks has been long acknowledged by cognitive scientists (Fodor and Pylyshyn, 1988; Marcus, 2001).

Compared to the feedforward networks in this study, biological visual systems excel at detecting relations. Fleuret et al. (2011) found that human observers are capable of learning rather complicated visual rules and generalizing them to new instances from just a few training examples. Participants could learn the rule underlying the hardest SVRT problem for CNNs in our Experiment 1, problem 20, from an average of about 6 examples. Problem 20 is rather complicated as it involves two shapes such that *"one shape can be obtained from the other by reflection around the perpendicular bisector of the line joining their centers."* In contrast, the best performing CNN model for this problem could not get significantly above chance from one million training examples.

This failure of modern computer vision algorithms is all the more striking given the widespread ability to recognize visual relations across the animal kingdom. Previous studies showed that non-human primates (Donderi and Zelnicker, 1969; Katz and Wirght, 2006), birds (Daniel, Wright, and Katz, 2015; Martinho III and Kacelnik, 2016), rodents (Wasserman, Castro, and Freeman, 2012) and even insects (Giurfa et al.,

2001) can be trained to recognize abstract relations between training objects and then transfer this knowledge to novel objects. Contrast the behavior of the ducklings in (Martinho III and Kacelnik, 2016) with the RN of Experiment 3, which demonstrated no ability to transfer the concept of same-different to novel objects (Figure 2.24) even after hundreds of thousands of training examples.

There is substantial evidence that visual-relation detection in primates depends on re-entrant/feedback signals beyond feedforward, pre-attentive processes. It is relatively well accepted that, despite the widespread presence of feedback connections in our visual cortex, certain visual recognition tasks, including the detection of natural object categories, are possible in the near absence of cortical feedback – based primarily on a single feedforward sweep of activity through our visual cortex (Serre, 2016). However, psychophysical evidence suggests that this feedforward sweep is too spatially coarse to localize objects even when they can be recognized (Evans and Treisman, 2005). The implication is that object localization in clutter requires attention (Zhang et al., 2011). It is difficult to imagine how one could recognize a relation between two objects without spatial information. Indeed, converging evidence (Logan, 1994; Moore, Elsinger, and Lleras, 1994; Rosielle, Crabb, and Cooper, 2002; Holcombe, Linares, and Vaziri-Pashkam, 2011; Franconeri et al., 2012; Ham et al., 2012) suggests that the processing of spatial relations between pairs of objects in a cluttered scene requires attention, even when individual objects can be detected pre-attentively.

Another brain mechanism implicated in our ability to process visual relations is working memory (Kroger et al., 2002; Golde, Cramon, and Schubotz, 2010; Clevenger and Hummel, 2014; Brady and Alvarez, 2015). In particular, imaging studies (Kroger et al., 2002; Golde, Cramon, and Schubotz, 2010) have highlighted the role of working memory in prefrontal and pre-motor cortices when participants solve Raven's progressive matrices which require both spatial and same-different reasoning.

What is the computational role of attention working memory in the detection of

visual relations? One assumption (Franconeri et al., 2012) is that these two mechanisms allow flexible representations of relations to be constructed *dynamically* at run-time via a sequence of attention shifts rather than *statically* by storing visual-relation templates in synaptic weights (as done in feedforward neural networks). Such representations built "on-the-fly" circumvent the combinatorial explosion associated with the storage of templates for all possible relations, helping to prevent the capacity overload that plagues feedforward neural networks.

Humans can easily recognize when two objects are the same up to some transformation (Shepard and Metzler, 1971) or when objects exist in a given spatial relation (Fleuret et al., 2011; Franconeri et al., 2012). More generally, humans can effortlessly construct an unbounded set of structured descriptions about their visual world (Geman et al., 2015). Mechanisms in the visual system such as perceptual grouping, attention and working memory exemplify how the brain learns and handles combinatorial structures in the visual environment with small amount of experience (Tenenbaum et al., 2011). However, exactly how attentional and mnemonic mechanisms interact with hierarchical feature representations in the visual cortex is not well understood. Given the vast superiority of humans over modern computers in their ability to detect visual relations, we see the exploration of these cortical mechanisms as a crucial step in our computational understanding of systematic visual reasoning.

Chapter Three

Neural mechanisms of systematicity: Oscillatory systems and visual cognition

We[1] have already learned from Fodor and Pylyshyn, 1988 that the modeling of systematicity in neural systems is principally a problem of implementation. *That* systematic behavior can arise in neural networks is assured[2], but *how* that behavior comes about is a mystery. We have already argued that attention and working memory must be key components, and, indeed, there are already interesting new models exploiting these mechanisms for rule-like generalization on static images (Linsley et al., 2018). Before these mechanisms can operate properly, however, we must ensure that visual representations of items in a complex scene are sufficiently organized to avoid the catastrophic failures found in "attentional" models like the RN (See Sec. 2.3.3). Perceptual grouping is in many ways a prerequisite of visual reasoning. Reasoning, after all, acts on objects.

To that end, we will pick up on a particular historical trend in the neural modeling of

[1]Sec. 3.1 of this chapter has been published as Alamia et al., 2019.

[2]That is, unless one believes that the hardware of the mind is not made of neurons. See Gallistel and King, 2009 for interesting counterproposals.

perceptual grouping and of cognition more generally. This is the trend, beginning with Edelman, von der Malsburg, Crick and others, which contends that a temporal code created by dynamic neurons is the ideal means for the flexible organization of perceptual objects. Their rejection of purely rate-based codes proceeds like this. Modulo extra-classical receptive field effects (Angelucci and Shushruth, 2013), an increased spiking rate of a neuron indicates the presence of a given stimulus (see Tsodyks and Markram, 1997 for exceptions), and this rate is largely determined by the pattern of anatomical connections defining the network in which the neuron is embedded. These connections, however, only change at a timescale slower than that of active behavior. How then are flexible, context-dependent notions like the demonstrative "this" (i.e. "No, not that object. *This* object!") encoded in neural activity? More generally, how does neural activity flexibly recombine the raw features output by anatomical connections at a timescale fast enough to support active cognition?

According to proponents of temporal coding, it is the higher order temporal statistics of dynamic neurons that encode these on-the-fly properties, not just the rate. A common claim by this camp is that the "glue" of perceptual organization, the thing that binds raw features into perceptual objects, is the temporal correlations either between spiking neurons or between a spiking neuron and a macroscopic signal like LFP. These neural activities and macroscopic signals in turn display rhythmic or oscillatory activity thought to structure not only perceptual grouping, but also attention and working memory. Being transient and context-dependent, as opposed to anatomical, these oscillatory mechanisms could potentially implement the flexible, rule-based cognitive routines theoretically required for systematic thought.

The debate between rate-coding theorists and temporal-coding theorists is long, technical and ongoing, and our goal is not to recapitulate it in detail here. Nevertheless, the idea that cognitive processes are grounded in unexplored neural territory is enticing, particular for the computational modeler, since it opens to the door to models

categorically different from the rate-based formalisms which continue to dominate both computational neuroscience and machine learning. So far, the temporal coding proponents have competed with their rate coding counterparts at a disadvantage, not having a robust and fully-developed machine learning apparatus at their side in the way that the latter scholars inherited rate coding/McCulloch Pitts models like the CNN.

The rest of this dissertation will be devoted to justifying and building machine learning formalisms based on the principles of oscillatory coding. We will begin by connecting the previous chapter to the current one by presenting an experiment implicating oscillatory coding in visual reasoning. This experiment shows that the dichotomy previously explored between spatial relation and same-different problems manifests as different oscillatory signatures in human EEG recordings. Next, we will briefly discuss the neuroscientific literature on oscillatory coding, particularly its role in "binding-by-synchrony", "communication-through-coherence" and visual working memory. The rest of the chapter will propose a general theoretical framework for the modeling of these phenomena in the form of the Kuramoto model (Kuramoto, 1975). A technical exposition of this model will be provided followed by a description of relevant learning algorithms. This technical diversion will provide the groundwork for subsequent chapters, in which the Kuramoto model is used to model the oscillatory phenomena described here including those related to visual reasoning.

3.1 Oscillatory dynamics in visual relation detection.

In Chapter 1, we hypothesized that feedforward networks generally performed better on spatial relation (SR) vs same-different (SD) detection since the latter intuitively places a greater strain on attention and working memory. In this section, we will test this hypothesis directly for human subjects. We proceed in two steps. First, we

recapitulate the results of Sec. 2.3.2 on a new stimulus set suitable to human subjects and used in previous studies (Hollard and Delius, 1982; Delius and Hollard, 1987). Second, we recorded EEG responses in human subjects performing SR and SD tasks on this data while ensuring tasks stimuli in both conditions were balanced for difficulty. We found that the SD condition resulted in both different evoked potentials and higher activity in oscillatory components in the occipital-parietal areas associated with attentional and mnemonic processes. These findings provide a biological reflection of our computational intuition from Sec. 2.3.3. Further, though there have been previous electrophysiology studies concerning visual reasoning (Franconeri et al., 2012; Luck, 2012; Hayworth, Lescroart, and Biederman, 2011; Yuan, Uttal, and Franconeri, 2016; Golde, Cramon, and Schubotz, 2010; Amorapanth, Widick, and Chatterjee, 2010; Zhang et al., 2013), the current experiment is the first to our knowledge to explicitly compare spatial vs same-different judgments.

3.1.1 Experiment 1: Spatial relations vs Same-different on perceptually relevant stimuli

We first extended the results of Kim et al., 2018 for our novel stimulus set: we trained two separate CNN architectures to solve an SD and an SR task using the same stimulus set. The input to these networks was a 50×50 image in which two hexominoes (width and height of 2 to 5 pixels) were displayed at opposite sides of the screen (see Fig. 3.1)a. We chose these shapes since they exhibit enough variability to be difficult to memorize but are not completely random like PSVRT shapes. They have also been used in pigeon studies on same-different reasoning by Hollard and Delius, 1982 and Delius and Hollard, 1987.

The network consisted of 6 convolutional layers. Each layer contained 4 channels of size 2x2, with stride of 1. All convolutional layers used a ReLu activation function with stride of 1 and were followed by pooling layers with 2x2 kernels and a stride of

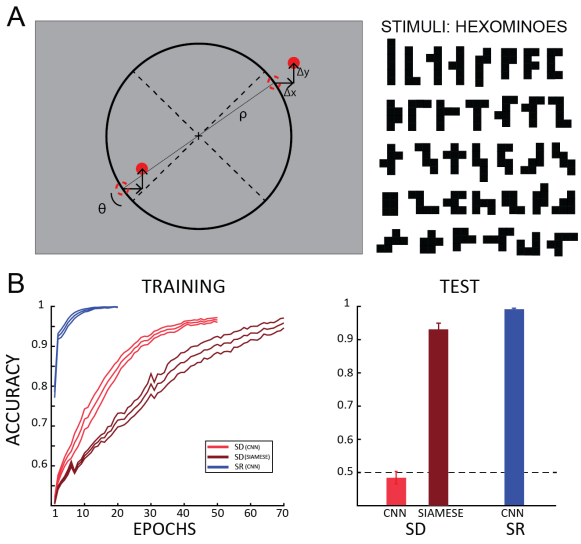


Figure 3.1 *Network performance on perceptually meaningful visual relations problems.*Simulations stimuli and results. A) The stimuli were the same in the simulations and in the human experiments. The items were displayed at opposite sides of the screen (either 45° and 225° or -45° and -225°). Both item positions were jittered by a random amount in both the x and y axis (Δx and Δy in the picture) to make the task non-trivial for human participants (i.e. preventing participants from performing the SR task considering only the relative position of one item and its closest diagonal, thus ignoring the spatial relationship between the two items). The items used are hexominoes (right panel). Minimum and maximum item height and width are 1.2° – 3.6° and 1.2° – 2.7° of visual angle respectively, and 2 to 5 pixels used for the simulations (image size was 50 × 50 pixels). B) Accuracy of the CNN network on the same-different (SD; light red) and spatial relationship (SR; blue) tasks, and of a Siamese network trained on the SD task (dark red). The Siamese network mimics segmentation in a feedforward networks, by separating the items in two distinct channels of the network. The left panel shows the training curves for each network (accuracy over epochs during training); we stopped the training when the validation accuracy reached 90%. In the right panel the test accuracy, evaluated using novel items never used for training, reveals that the CCN seem to only learn the abstract rule for the SR but not for the SD task, as shown in a previous study. Conversely, the Siamese network can solve the SD task, demonstrating that segmentation is the missing process in the CNN to successfully accomplish the task. In both panels we show average values ± SE over 10 repetitions using different random initializations.

1. Eventually, two fully connected layers with 128 units preceded a two-dimensional classification layer with a sigmoid activation function. As a regularizer we set a dropout rate of 0.3 in each layer of the network. We used binary cross-entropy as a loss function, the Adaptive Moment Estimation (Adam) optimizer 36 and a learning rate of 10e-4. Each simulation was run over 20 epochs with batch size of 50. All simulations were run in TensorFlow.

The networks were trained to classify whether the two hexominoes were the same or not (SD task) or whether they were aligned more vertically or more horizontally with respect to the midline (SR task).

Data was split into 1000 training and 1000 testing stimuli and networks were trained on 10 random initializations. We report the mean accuracy and standard deviation over these 10 repetitions in Fig. 3.1. Our results are consistent with those from Kim et al., 2018: a CNN appears to be able to learn the abstract rule (as measured by the network's ability to generalize beyond the shapes used for training) for SR tasks much more easily than SD tasks. The effortless ability of humans and other animals to learn SD tasks suggest the possible involvement of additional computations that are lacking in CNN, like object individuation, mediated by attention and working memory.

This hypothesis is supported by the results obtained by our simulations with a Siamese architecture, in which each item is processed separately and eventually combined before being passed to a classifier, similar to that of Sec. 2.3.3. The Siamese network had the same exact convolutional architecture as described above; additionally, the difference between features-vectors of each separate input was feed to the classifier to perform the SD task. This model mimics the effect of selective attention and item segregation by feeding to the network each item separately. Performance is depicted in Fig. 3.1b. Evidently, these mechanisms, automated by the Siamese network, can be very effective.

Next, we test the prediction that SD and SR tasks relies on different computational

mechanisms in humans, by recording EEG signals from a pool of 28 participants (14 of which tested on a pilot experiment –fig. S5) performing the same SD and SR tasks.

3.1.2 Experiment 2: EEG markers of visual reasoning

Participants completed 16 blocks using the same stimuli as those used to train CNNs 3.1: in half of the blocks they were asked to report whether the two hexominoes were the same or not (SD conditions), in the other half whether the hexominoes were more vertically or horizontally aligned (SR conditions). The two conditions were interleaved in a block design. Participants were required to answer after one second from stimulus onset in order to disentangle motor from visual components in the EEG recordings (Fig. 3.2a). The QUEST algorithm was used to assure that participants' accuracy was matched between the two tasks and remained constant throughout the whole experiment. This was done by adjusting two experimental parameters trial by trial (i.e., the hexominoes eccentricity in SD blocks, ρ, and the angle from the diagonal in SR blocks, θ; see Fig. 1A and S1). Maintaining a comparable accuracy between the two tasks reduces the potential for confounds in the electrophysiological analysis due to differences in performance. We confirmed the absence of any substantial behavioral difference between the SD and SR tasks (Fig. 3.2b) with a Bayesian ANOVA on both accuracy (BF10 = 0.361, error $<$ 0.001%) and RT (BF10 = 0.317, error $<$ 0.89%). In addition, we also investigated each condition separately (Fig. 2B), comparing the difference between 'same' and 'different' trials (in SD blocks) and 'vertical' and 'horizontal' trials (in SR blocks) in both RT and accuracy. All comparisons revealed overall no differences between tasks, except for the accuracy of vertical and horizontal trials in the SR condition, in which the BF proved inconclusive (accuracy: SD - BF10 = 0.39, error $<$ 0.012%; SR - BF10 = 1.80, error $<$ 0.001%; RT: SD - BF10 = 0.333, error $<$ 0.01%; SR - BF10 = 0.34, error $<$ 0.01%).

After having confirmed that performance was equal in the two tasks, we charac-

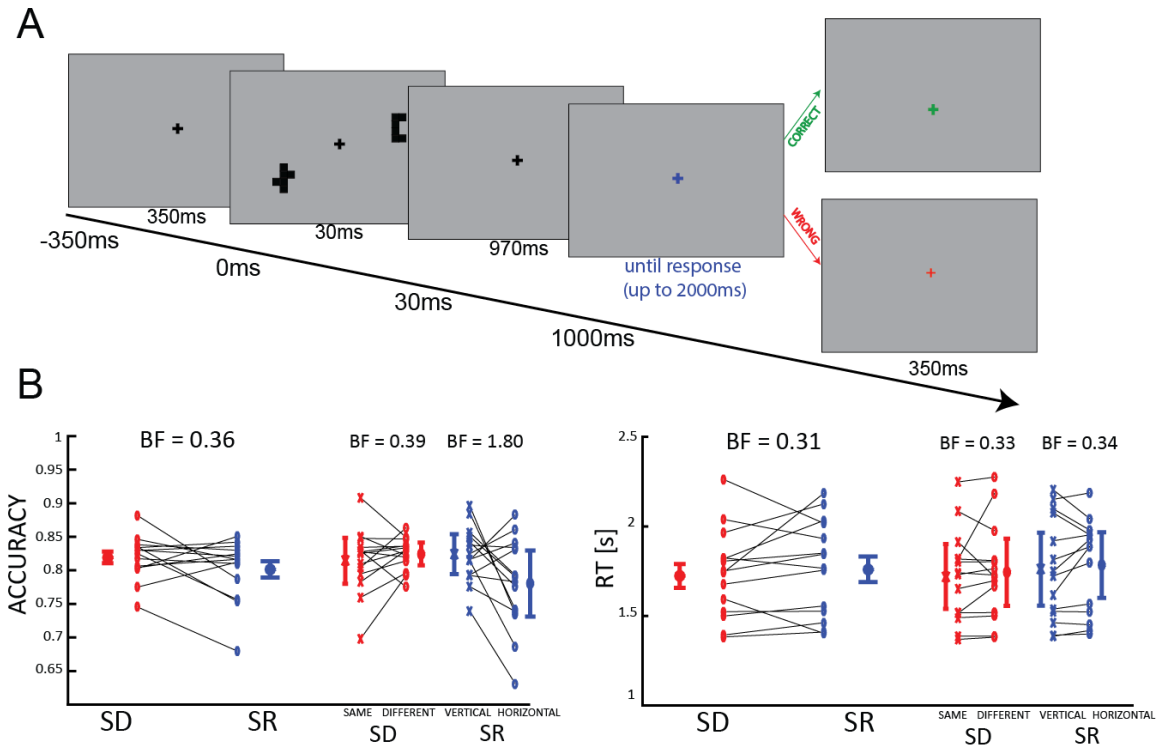


Figure 3.2 *Comparative psychophysics on two visual relations problems.* Participants experimental design and behavioral results. A) At the beginning of each trial a black fixation cross was displayed for 350ms. After 2 stimuli were shown for 30ms, participants waited an additional 970ms before providing the answer. The response was cued by the fixation cross turning blue. After the response, the color of the fixation cross provided feedback: green if the response was correct, red otherwise. B) Humans perform the SD and SR tasks with comparable levels of performance. In the left and right panels are shown the averages ± SE for accuracy and reaction times, respectively. Each pair of connected markers represent an individual subject. The results for the same-different (in red) and spatial relationship (in blue) conditions are further break down for each condition separately (same-different and vertical-horizontal). BF indicates the Bayes factor against the null hypothesis (difference between the two conditions).

terized the evoked potentials (EP) in each task. First, we estimated the difference between SR and SD conditions considering 7 midline electrodes (Fig. 3.3). The results of a point-by-point t-test corrected for multiple comparisons revealed a significant difference in central and posterior electrodes (mostly Pz and CPz) between 250ms after the onset of the stimuli and the response cue, and the opposite effect in frontal electrodes (FCz and Fz) from 750ms to 1000ms, as confirmed by the topography (Fig. 3.3). Overall, these results indicate larger potentials in visual areas during the SD task than in the SR. Previous studies have shown a relation between EP amplitude (particularly P300 and late components) with attention (Krusemark, Kiehl, and Newman, 2016; Van Voorhis and Hillyard, 1977) and visual working memory (Kok, 2001; McEvoy, Smith, and Gevins, 1998; Fabiani, Karis, and Donchin, 1986). Our results are thus consistent with a larger involvement of executive functions in the SD vs. SR task. In the following we investigated whether this hypothesis is corroborated by corresponding oscillatory effects in the time-frequency domain.

We performed a time-frequency analysis to try to identify differences between conditions observed in specific frequency bands commonly related to executive functions (e.g., visual working memory). For this purpose, we computed a baseline-corrected log-scale ratio between the two conditions (as shown in Fig. 3.4a), averaging over all electrodes. Remarkably, a point-by-point 2-tailed t-test corrected with cluster-based permutation test 16 revealed a significantly larger activity in the low beta-band (16-24Hz) in the SD condition between 250 and 950ms after stimuli onset (Fig. 3.4b). We further quantify the magnitude of the effect by computing the effect size of a one sample t-test against zero averaging per each participant the values within the significant region ($t(13)=2.571$, $p=0.023$, Cohen's $d=0.687$). The topography of the effect spread mostly over parietal and occipital regions Fig. 3.4c), mimicking the topography of the EPs analysis. As previously, these results confirm the prediction that the SD task may involve additional computational mechanisms beyond feedforward

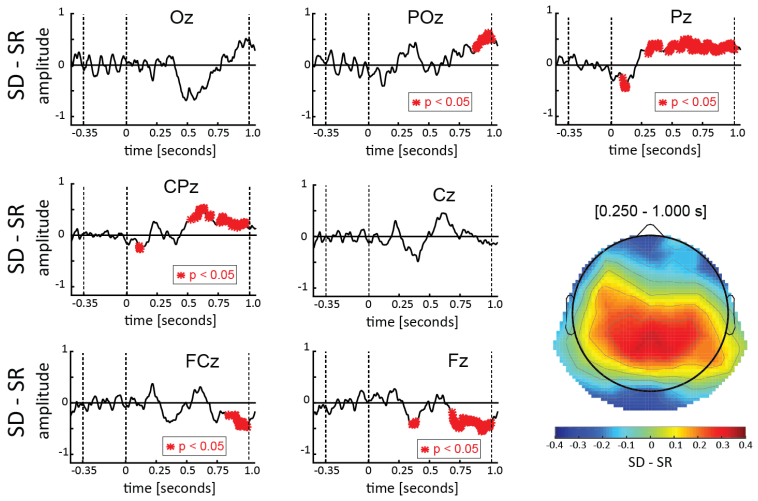


Figure 3.3 *ERP results.* Each panel represents the difference between ERPs elicited in the SD and SR conditions for the 7 midline electrodes. Shown in red are the points for which a difference was found that differs significantly from zero. The results reveal a significant difference from 250ms after stimuli onset until the response cue (at 1000ms) in central parietal regions, and an opposite effect after 750ms in frontal regions. In the bottom-right panel the topography, computed over the 250ms – 1000ms interval, confirmed a larger activity in the SD than in the SR condition (positive difference, warmer colors) in the central-parietal regions, and an opposite effect (negative difference, colder colors) in the frontal regions (and –although not significant– also included occipital regions).

computations, possibly indexed by the oscillatory processes identified here.

3.2 Oscillatory coding in visual cognition generally

3.2.1 Binding by synchrony

The visual world is made of objects and those objects are made of features. Features are often shared between objects, yet we have little trouble saying which features are bound to which things, given enough time (Treisman and Gelade, 1980). Yet, how the appropriate assignment of features to objects occurs in neural machinery is something of a mystery, and this mystery has come to be known as "the binding problem" (see Treisman, 1996 for an old but valuable review). The solution offered by contemporary computer vision models is to simply learn many feature conjunctions so that, for example, a scene depicting "red circle and green square" can always be distinguished from one depicting "red square and green circle" with the use of shape-color conjunction detectors. However, as we have seen in Ch. 1, the strategy of learning numerous conjunctions that define a relation is the very definition of non-systematic cognition, and this strategy has a concrete effect on learning performance.

An alternative solution is that features are bound together when the neurons representing those features coordinate their temporal activity. In the simplest conception, "red square" would be represented by synchronous spiking of a "red" neuron and a "square" neuron. This is the so-called "binding-by-synchrony" (BBS) hypothesis. Following theoretical speculation by Edelman, 1978 and Pearson, Finkel, and Edelman, 1987, this hypothesis found empirical support from Gray and Singer, 1989, who simultaneously measured single-unit activity and LFP and cat striate cortex during the viewing of an oriented bar. The authors found that neural populations excited by a preferred stimulus fired in rhythmic bursts (Fig. 3.5a), each internally around 35-45 Hz, and that these bursts were in tight anti-phase with an oscillating LFP signal. The

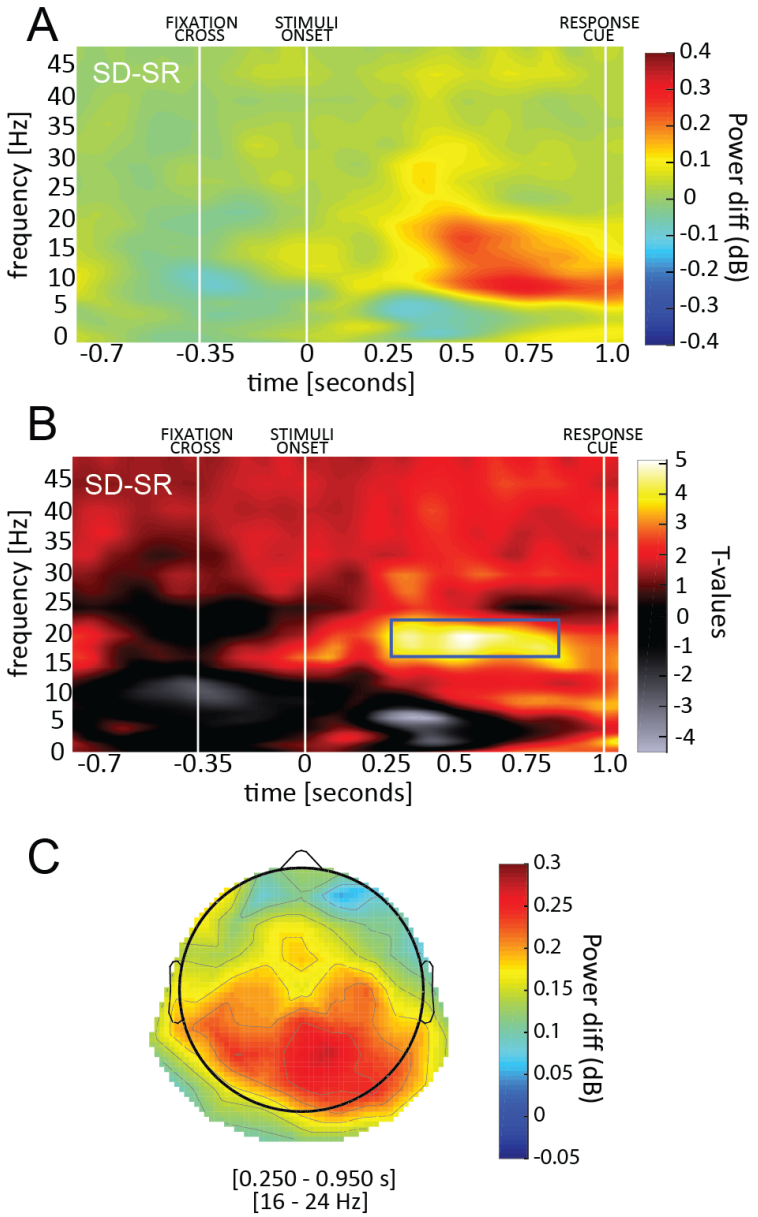


Figure 3.4 *Time-frequency results.* A) The difference between SD and SR power spectra is shown in the first panel. White lines indicate the onset of the fixation cross, the stimuli and the response cue. B) The second panel shows the corresponding t values (when testing the difference against zero). We observed a significant region in the low beta band (16-24Hz), between 250ms and 950ms after stimulus onset. C) The topography of the significant time-frequency window reveals the involvement of occipital-parietal regions.

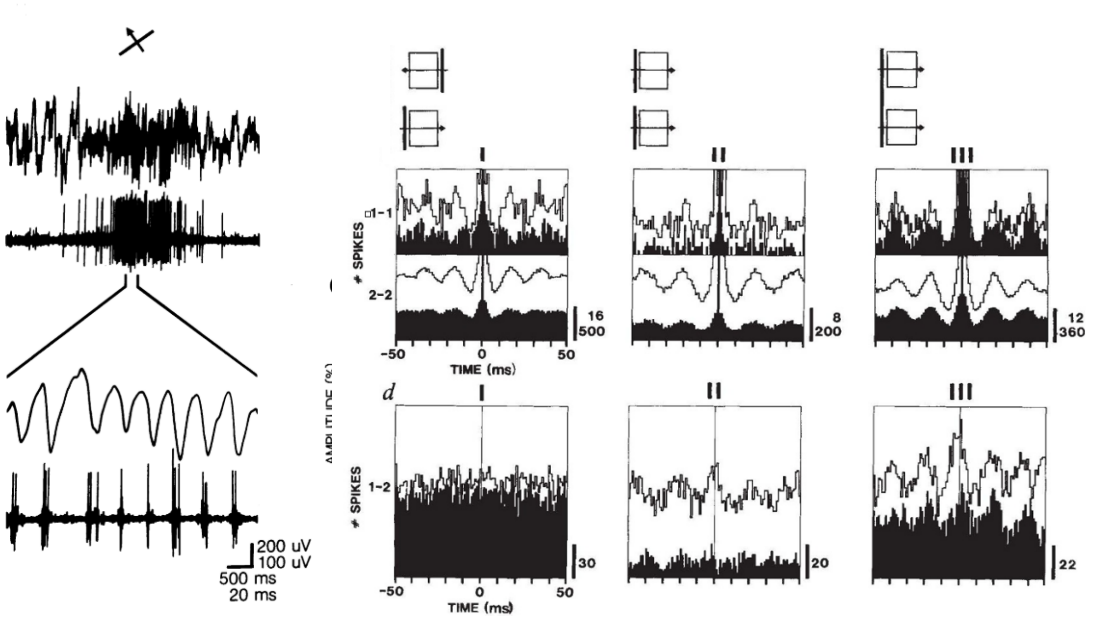


Figure 3.5 *Synchrony for long-distance neuronal communication in cat visual cortex. (a)* Multi-unit array (MUA) and local field potential (LFP) responses to an optimal bar stimulus. Lower traces are peak activity at an expanded timescale. Observe oscillation in LFP with anti-correlated bursting in MUA recording. *(b)* In a later study, Gray and colleagues examined the emergence of this oscillatory effect as a function of the gestalt quality of the observed stimulus. Columns correspond to stimulus depicted in small squares at the top: broken bars moving in opposite directions; broken bars moving in the same direction; a single elongated bar. Rows depict autocorrelation and cross-correlation for recorded units in each condition. Observe the pronounced periodic correlations in the second and third conditions but not the first.

authors speculated that this oscillatory activity could arise from mutual excitation and recurrent inhibition (see Buzsaki and Wang, 2012 for implementations) and further that, "the phase of the oscillatory response may be used as a further dimension of coding in addition to the amplitude and duration of the response." The authors later explicitly cast their result as a form of binding by synchrony when they showed that the degree of synchrony in response to a stimulus with two bars increased to the extent that those bars seemed to form a single object (Fig. 3.5b).

These and other results played into a later theoretical framework developed in

Malsburg, 1994, the so-called "correlation theory of brain function". Here, von der Malsburg formalized the role of spike-time correlations as a binding mechanism and claimed the mechanism only functioned properly in the presence of dynamical, "fast" synapses. These dynamical synaptic conductances would change at a much faster scale than learning in the anatomical or "slow" synapse, thereby enabling the type of flexible representation of information on the fly we argued above could support systematic cognition. Von der Malsburg would go on to demonstrate the efficacy of these dynamical synaptic models over the course of several models (Bienenstock et al., 1987; Lades et al., 1993), arguing that the evolving neural topologies caused by dynamical synapses were themselves interesting data structures capable of resolving the binding problem. Fast synaptic plasticity continues to be an active area of theoretical research (Masse et al., 2019).

The binding-by-synchrony hypothesis is not without its critics (see Shadlen, Movshon, and Hughes, 1999 for a skeptical review). Indeed, the requisite synchrony for bound objects is not always electrophysiologically forthcoming, as demonstrated by Lamme and Spekreijse, 1998; Roelfsema, Lamme, and Spekreijse, 2004; Palanca and DeAngelis, 2005; Chen et al., 2014. Nevertheless, more recent results from Martin and Heydt, 2015 take steps to resolve these findings with the older theories of Gray and others, providing evidence from macaque visual cortex that synchronous binding occurs instead for "proto-objects", simple representations indicating only the presence of a given object in a rough location and distinguished form other proto-objects by a tag or index manifest as oscillatory phase (for a series of interesting psychophysical studies on proto-objects, see Pylyshyn and Storm, 1988; Pylyshyn, 1989a; Pylyshyn, 1989b; Pylyshyn, 1994; Burkell and Pylyshyn, 1997; Feldman and Tremoulet, 2006; Pylyshyn, 2004).

3.2.2 Communication through coherence

In Ch. 1, we used the relation net to demonstrate the pitfalls of attention without grouping. The features of objects within a single high-level receptive field tend to be merged and relations between the objects are subsequently lost. Since anatomical hardware is limited, one solution to this problem is to use a single neural population to represent multiple objects at different *times*. Conceivably, a downstream population could selectively attend to a single object as long as it was attentive at the correct moment. This selective routing could in turn support the type of flexible cognition which we have associated with systematic reasoning.

This is the essence of Pascal Fries' "communication by coherence" (CTC) theory, which claims that brain "communication structure is mechanistically implemented by the pattern of coherence among neuronal groups, that is, the pattern of phase-locking among oscillations in the communicating neural groups." The essence of the idea is the following. Neural populations activated during focused attention have been reported to oscillate at the so-called "gamma" rhythm (30-70 Hz) (Fell et al., 2003). This oscillation in turn modulates the input gain of the population, controlling when it will be sensitive to incoming information (Cardin et al., 2009; Siegle, Pritchett, and Moore, 2014). Any other population wishing to transmit information to this rhythmically active ensemble must therefore time its spikes to the moments of maximal sensitivity. Hence, optimal communication should occur when both the sending and receiving populations are internally synchronized (e.g. in the case of V1 and V2 in visual cortex, (Jia, Tanabe, and Kohn, 2013; Zandvakili and Kohn, 2015)) and externally synchronized with one another (Fig. 3.6).

Though a compelling hypothesis, CTC has not proven easy to computationally model, since it has traditionally conflicted with other theories about cortical oscillations and their role in attention (Lisman and Idart, 1995). A promising recent model of McLelland and VanRullen, 2016 has partially resolved these competing theories with a

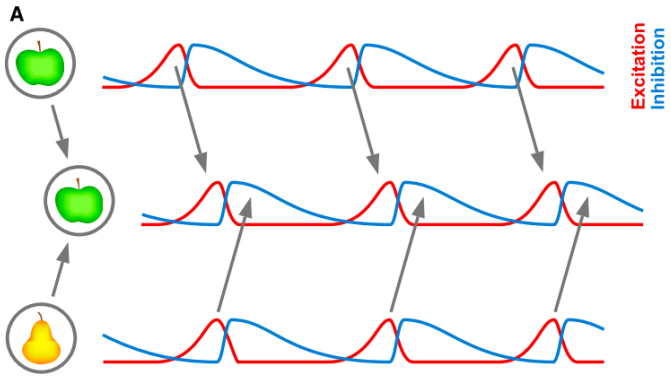
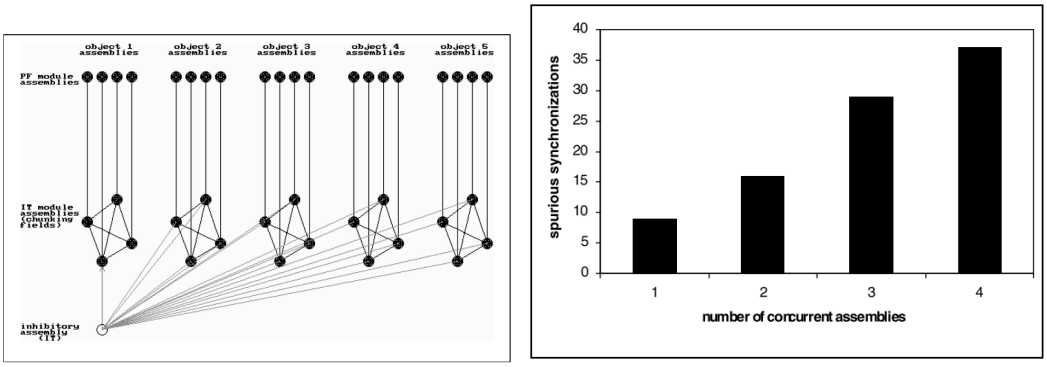


Figure 3.6 *Communication through coherence.* Each row depicts overlapping excitatory (red) and inhibitory (blue) traces for three neuronal populations. The top and bottom traces are for populations representing an apple and a pear, respectively. The center traces are for a third population seeking to rout information from the other two populations. The peak of excitation in the apple-representing population is phase-locked with that of the selecting population so that information can be routed before the onset of inhibition decreases the output of the former and the receptivity of the latter. Excitation in the pear population, however, is too late in the cycle, so that the selector is insensitive to its output.

demonstration of visual attention based on oscillatory coding in a network of integrate-and-fire neurons. It is difficult, however, to see how such a model could be adapted to more complex stimuli or be trained for other tasks.

3.2.3 Working memory

Both BBS and CTC employ the phase of rhythmic activity as a sort of tag or index indicating which neurons correspond to a given thing in the world. We end this section by briefly noting the utility of this indexing mechanism for working memory, where it would function as a memory address or hash. A notable model of this sort was proposed by Raffone and Wolters, 2001. Their system (Fig. 3.7) consisted of several excitatory pyramidal assemblies encoding features of a given object which were connected through a hidden layer to collection of IT-like assemblies with global inhibitory connections. The authors found that emergent rhythmic activity in the network tended to segment



(a) **(b)**

Figure 3.7 *Phase coding for working memory. (a)* A hierarchical network is constructed consisting of several internally excitatory assemblies acting as feature detectors (diamond ensembles with afferents) subject to global inhibition from high-level units (white units below; only one shown). An impinging stimulus sets the network into persistent rhythmic activity, with populations representing individual objects coming to occupy different phase "slots" due to inhibitory competition in the manner of Malsburg and Schneider, 1986 (See. Sec. 5.1 for details). *(b)* As the number of simultaneously activated assemblies is increased, it becomes more difficult for the network to apportion activity into temporal slots and eventually representations begin to interfere, similar to the case of slot-based working memory.

object representations into different temporal slots. The fidelity of retrieved object representations from recurrent activity decreased as the number of stored objects increased, indicating a capacity of about 3-4 objects consistent with slot-based theories of visual working memory (Luck et al., 1997). Electrophysiological evidence for a synchrony-based mechanism for visual working memory has not been forthcoming, though there have been promising recent results implicating combined rate-phase coding (Brzezicka, Mamelak, and Rutishauser, 2020).

3.3 Kuramoto: Paradigm of synchrony

Is there a general computational framework that could capture these neuroscientific phenomena? Innumerable dynamical systems give rise to synchronous behavior, but not all of them are suitable to the type of neuroscientific and machine learning settings

which interest us. As far as neuroscience is concerned, we must find a model which could presumably support the types of phase-coding theorized by Fries, 2005 and which would ideally be able to describe a wide range of neuro-dynamical regimes. From a machine learning perspective, it is paramount that we avoid complicated biophysical modeling, not knowing in advance what aspects, for example, of neuron morphology are actually computationally useful. In short, we must determine the smallest extension to the standard artificial neural network toolbox which still allows us to model interesting neural-dynamical phenomena and hopefully make interesting machines that alleviate the limitations discussed in Ch. 1.

Luckily, oscillatory phenomena in complex systems is a vast area of research, so we are not at a loss for inspiration. Even luckier, a collection of important results due to Winfree, Malkin, Kuramoto (see Izhikevich, 2007, ch. 10 for a review) and others ensures that a huge swath of these systems can be reduced to a *single* archetype. Through the so-called "phase reduction" method, it can be shown that complicated oscillatory systems can be reduced to a canonical, easily parametrizable format in which, remarkably, every oscillator can be described by a one-dimensional phase variable. From its conception, phase reduction has become an indispensable method for the study of dynamical systems, including in computational neuroscience (Ermentrout and Kopell, 1984; Ermentrout and Kopell, 1991; Guckenheimer and Holmes, 1991; Tass, 20007; Brown, Moehlis, and Holmes, 2004). One particularly simple version of a "phase-reduced" model is the celebrated Kuramoto model (Kuramoto, 1975), variations of which will occupy us for the next two chapters.

3.3.1 Phase reduction

Consider a dynamical system of the form

$$\dot{x} = f(x, t), \tag{3.1}$$

where $x \in \mathbb{R}^m$ and f is a generally non-linear equation of motion. If the system has a single limit cycle $\gamma = x_0(t)$, then we refer to the system as an *oscillator* and we refer to $\mathbb{R}^m$ as the oscillator's state space. Denote the period of the oscillator by T and its frequency by $\omega = \frac{1}{T}$. Oscillators are ubiquitous in the natural world, notably in neuroscience, where the famous integrate-and-fire, Hodgkin-Huxley and FitzhHugh-Nagumo neuron formalisms (see Izhikevich, 2007) are tried and true models of oscillatory neural dynamics. In neuroscientific applications, m can be quite large. It is not uncommon for single units to be modeled with hundreds or even thousands of compartments. Many of these dimensions will be of little use to us, since we are principally concerned with simple scalar measures of overall coherence in a population of neurons.

Luckily, the periodicity of the oscillator simplifies the situation considerably as long as we remain near the limit cycle. To that end, we choose an arbitrary point $x' \in \gamma$ and define the *phase function* $\theta = \Theta(x)$ for all $x \in \gamma$ by setting $\Theta(x') = 0$ and declaring that θ advances at a constant rate as x winds around γ:

$$\theta(t) = \omega t. \tag{3.2}$$

Because θ advances at a constant rate around γ, we can use θ to denote a given state of the oscillator on the limit cycle by writing $x(\theta)$ (as opposed to $x(t)$). See Fig. 3.8 for a diagram of conversion to a phase variable on an oscillator's limit cycle.

A point on γ can be pushed off the limit cycle if it is acted upon by some external perturbation, for instance, interaction with other oscillators, so we must therefore extend $\Theta(x)$ to points away from γ. To that end, we will assign the same θ to all points x in the basin of attraction of γ which eventually converge to the same $x_0(\theta(t))$. We call a set of points converging to the same phase on the limit cycle an *isochron*.

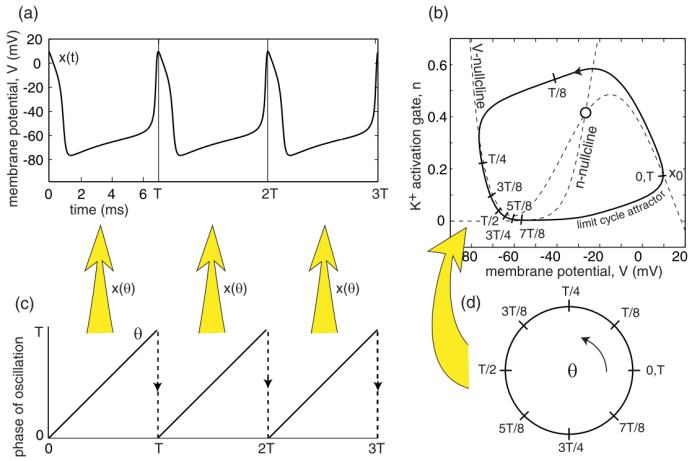


Figure 3.8 *Phase reduction for a single oscillator. (a)* The membrane potential of a FitzHugh-Nagumo oscillator set into periodic activity by sufficient input current. *(b)* The state of the neuron has two dimensions, membrane potential (x axis) and the status of a potassium activation gate (y axis). The activity of the neuron through time can be represented by a closed curve (limit cycle) in this state space since the unit is periodic. *(c)* If the period of the neuron is T, then we can imagine the phase θ of the unit advancing at a constant rate between an arbitrary point on the periodic limit cycle, for instance the spike. *(d)* This correspondence between high(er)-dimensional neuron state and phase indicating only relative position on the limit cycle has the effect of "de-warping" the limit cycle into a circular phase cycle moving at constant speed. Figure taken from Izhikevich, 2007

Following this logic, we differentiate $\theta(t)$ to get

$$\begin{aligned}\dot{\theta} &= \frac{d}{dt}\Theta(x(t)) \\ &= \nabla_{x=x(t)}\Theta(x) \cdot \frac{d}{dt}x(t) \\ &= \nabla_{x=x(t)}\Theta(x) \cdot f(x(t)) \end{aligned} \tag{3.3}$$

Therefore, we must choose $\Theta(x)$ so that $\nabla_{x=x(t)}\Theta(x) \cdot f(x(t)) = \omega$ everywhere in the basin of attraction. Defined in this way, the phase of all points in the basin advances at the constant frequency $\dot{\theta} = \omega$ and the previously m-dimensional dynamics of the oscillator have been reduced to a single dimension. A closed form for $\Theta(x)$ is often difficult to acquire except in very simple cases, so numerical approximations are typically used (Kralemann, Pikovsky, and Rosenblum, 2011; Kralemann, Pivovsky, and Rosenblum, 2014).

Oscillators are interesting objects on their own, but our concern here is less single oscillators and more the dynamics of interconnected systems of n oscillators. In these systems, the behavior of a single oscillator is determined not just by its own intrinsic equation of motion f but also the influence of its neighbors. Does the reduction to a single phase variable for a given oscillator still hold under the influence of the rest of the system? $\Theta(x)$ is not generally injective, so that we cannot reconstruct a given state x_j from its phase alone. However, if x remains near the limit cycle, we can approximate it by $x_0(\theta)$. Hence, if the interaction with the rest of the system is weak enough that no x is displaced too far from γ, we can always reduce the system from $n \times m$ (number of oscillators $\times$ dimension of oscillator state) dimensions to $n \times 1$ (number of oscillators $\times$ one phase dimension) dimensions. This is the essence of phase reduction theory.

Considering a single oscillator with state x, we may formalize this intuition by introducing the *phase response function*

$$g(\theta; \delta) = \Theta(x_0(\theta) + \delta) - \theta \tag{3.4}$$

which measures the deviation in phase resulting from the perturbation of a point on the limit cycle $x_0(\theta)$ by an impulse $\delta \in \mathbb{R}^m$. The new phase of the perturbed oscillator is well approximated by the Taylor series

$$\begin{aligned}\Theta(x_0(\theta)+\delta) &= \Theta(x_0(\theta)) + \nabla_x \Theta(x)|_{x=x_0(\theta)} \cdot \delta + O(|\delta|^2) \\ &= \theta + \nabla_x \Theta(x)|_{x=x_0(\theta)} \cdot \delta + O(|\delta|^2) \end{aligned} \tag{3.5}$$

when the magnitude of the impulse $|\delta|$ is small. Substituting into Eq.3.4 gives

$$\begin{aligned} g(\theta;\delta) &\approx \nabla_x \Theta(x)|_{x=x_0(\theta)} \cdot \delta \\ &= Z(\theta) \cdot \delta, \end{aligned} \tag{3.6}$$

where $Z(\theta)$ is the *phase sensitivity function.* In words, the change in phase caused by a small impulse to the oscillator at its limit cycle is approximated by a projection of the impulse onto the gradient of Θ. Again, the exact calculation of $Z(\theta)$ is difficult and but good numerical methods are available (Ermentrout, 1996).

Now, suppose we have a dynamical system n interacting oscillators described by the combined state vector $X \in \mathbb{R}^{n \times m}$ and obeying the equation of motion

$$\dot{X} = F(X,t) + \epsilon P(X,t) \tag{3.7}$$

where the components of F represent the intrinsic dynamics of each of the n oscillators, P represents the external influence of the system on each oscillator, and ϵ is a scalar small enough so that the influence of P on any oscillator is not so strong as to perturb it far from its limit cycle. The dynamics of the system's phases are then given by

$$\begin{aligned} \dot{\theta(t)} &= \frac{d}{dt}\Theta(X(t)) \\ &= \nabla_X \Theta(X)|_{X=X(t)} \cdot (F(X(t)) + \epsilon P(X(t),)) \\ &= \omega + \epsilon \nabla_X \Theta(X)|_{X=X(t)} \cdot P(X(t),t), \end{aligned} \tag{3.8}$$

where ω is the vector of intrinsic frequencies according to Eq. 3.3. Note that final expression still depends on the original state variables X. If we assume that ϵ is small

enough, then we can approximate $X(t)$ by a nearby point $X_0(\theta(t))$ on the limit cycle and having the same phase as $X(t)$; i.e. $\Theta(X(t)) \approx \Theta(X_0(t))$. Then

$$\begin{aligned}\nabla_X \Theta(X)|_{X=X(t)} \cdot P(X(t),t) &\approx \nabla_X \Theta(X)|_{X=X_0(\theta(t))} \cdot P(X_0(\theta(t),t)) \\ &== Z(\theta(t)) \cdot P(X_0(\theta(t)),t), \end{aligned} \tag{3.9}$$

where Z is the phase sensitivity function defined above. Putting this all together, we have

$$\begin{aligned}\dot{\theta} &= \omega + \epsilon Z(\theta) \cdot P(\theta,t) \\ &= \omega + \epsilon \chi(\theta,t). \end{aligned} \tag{3.10}$$

We have thereby reduced the $n \times m$-dimensional system in Eq. 3.7 to an $n \times 1$ dimensional system in phase alone. The dynamics are governed by ω, a vector of intrinsic frequencies indicating the rate at which each oscillator would advance around the unit circle in the absence of the rest of the network, and χ, a coupling term which controls how the phase an oscillator is influenced by those of its neighbors.

Despite its lower dimensionality, the phase-reduced system can still be probed for interesting behaviors in the original system. See Nakao, 2016 for an excellent review of phase reduction.

The Kuramoto Model

The coupling term χ in Eq. 3.10 is very general, so it is useful to study some special cases. A typical case studied by Hoppensteadt and Izhikevich, 1997 and others is when χ is a sum of odd functions depending on pairwise phase differences $\theta_i - \theta_j$:

$$\chi(\theta) = \sum_{i,j} \chi_{ij}(\theta_i - \theta_j) \tag{3.11}$$

In that case, we can express χ by a sine series

$$\chi(\theta) = \sum_{i,j} K_{c,ij} \sum_{c=1}^{\infty} \sin(c(\theta_i - \theta_j)) \tag{3.12}$$

where $K_{c,ij}$ is the c^{th} Fourier coefficient of χ_{ij}. If we only take the first Fourier coefficient and substituting into Eq. 3.10, we arrive at the very simple oscillatory system

$$\dot{\theta_j} = \omega_j + \epsilon \sum_{i,j} K_{ij} \sin(\theta_i - \theta_j). \tag{3.13}$$

This is the celebrated *Kuramoto model*, the subject of enormous interest in computational physics, computational neuroscience and mathematical biology since its proposal in Kuramoto, 1975. The system has remained so captivating for the manner in which it describes the emergence of coordinated (i.e. synchronous) activity in a generally heterogeneous population. Specifically, under certain conditions on the intrinsic frequencies, ω, and the coupling matrix, K, the system gradually evolves from a random initial state towards one in which all oscillators share a common frequency and similar phase. Indeed, this sort of "order from chaos" is very much the guiding spirit of the whole study of complex systems, where the Kuramoto model has played a central role. We briefly review some of the interesting behaviors of the Kuramoto model here, though the model is still an active area of research.

Kuramoto himself studied a simple mean-field version of the model in which $K_{ij} = \frac{K}{n}$ was constant and ω was a random sample from n i.i.d. random variables with unimodal density $g(\omega)$ symmetric about 0. He then defined the so-called *global order parameter*

$$r = \left| re^{\psi} \right| = \left| \frac{1}{n} \sum_{j} e^{i\theta_j} \right| \tag{3.14}$$

which is 1 exactly when all oscillators share a common phase and approaches 0 for large n when the θ_j are uniformly distributed. Clearly, when $K = 0$, $r \approx 0$ for all t since there is no pressure to synchronize. Conversely, when K is very large, the influence of the rest of the network will dominate the intrinsic behavior of single oscillators, resulting in complete phase locking, $r = 1$, as $t \to \infty$. Where does the transition from $r = 0$ to $r = 1$ occur?

Taking n large and using a continuity equation (Griffiths, 1999) description of the

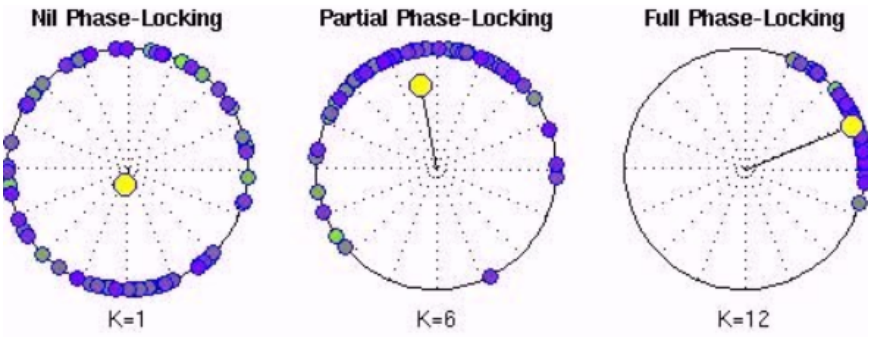


Figure 3.9 *Phase locking in the mean-field Kuramoto model.* The instantaneous state of a Kuramoto model with mean-field coupling is depicted for three coupling strengths K. The yellow-tipped vector is the mean field, whose length, r, is the global order parameter measuring the coherence of the system. For low K, the system is disordered forever. For middling K, the system begins to frequency-lock, with a macroscopically significant number of oscillators advancing around the unit circle at the same rate and similar phase. For large K, all units couple into this cluster.

system, Kuramoto solved for the stationary density on $\rho(\theta)$ and from this determined

$$\begin{cases} r \to 0 & \text{if } K \leq \frac{2}{\pi g(0)} \\ r \to 1 & \text{if } K > \frac{2}{\pi g(0)} \end{cases} \tag{3.15}$$

Therefore, the mean-field Kuramoto model undergoes a phase transition at $K = K_c = \frac{2}{\pi g(0)}$ from disorder when coupling is weak to global synchrony when coupling is strong. K_c is large when the mode $g(0)$ is small, which is precisely the case that intrinsic frequencies have large variance. Intuitively, these heterogeneous oscillators should be hard to synchronize, so the coupling strength must be made correspondingly strong (see Acebrón et al., 2005 for a detailed derivation of Kuramoto's original result). Surprisingly little is known about the emergence of synchrony in the general model of Eq. 3.13. Numerous studies have considered the onset of synchrony on complex graphs with non-constant coupling (Arenas and Albert, 2008; Lopes et al., 2016), uniform negative couplings (Ciobotaru et al., 2018), mixed positive and negative couplings (Hong and Strogatz, 2011), purely local couplings (Ha, Li, and Xue, 2013), adaptive Hebbian interactions (Timms and English, 2014; Ha, Noh, and Park, 2016;

Ha, Lee, and Li, 2018; Bronski et al., 2016), stochastic Langevin dynamics (Sakaguchi, 1988; Sakaguchi, Shinomoto, and Kuramoto, 1988), linearized interactions (Roberts, 2008), and quenched random couplings (Daido, 1987; Daido, 1992; Daido, 2000). By "quenched" we mean a type of randomness which is not subject to model dynamics. For instance, Daido was the first to study the case of the Kuramoto model with Gaussian interactions, a system which he analogized to spin glasses in the study of magnetism (Nishimori, 2010). The validity of this analogy is the subject of ongoing controversy (Bonilla, Pérez Vicente, and Rubí, 1993; Iatsenko, McClintock, and Stefanovska, 2014; Ottino-l and Strogatz, 2018). Nevertheless, quenched random couplings are still a good model of oscillatory interactions in numerous real-world systems, especially neural networks, where disordered couplings are the product of learning and the encoding of information.

Global synchrony is by no means the only macroscopic behavior of interest in the Kuramoto model. In particular, numerous studies have examined the emergence of multi-polar synchrony, in which the Kuramoto model separates into clusters which are internally synchronous but mutually desynchronized (Schaub et al., 2016; Smith and Gottwald, 2019). Multi-polar synchronization is particularly interesting for our case since the phases of synchronous clusters could act as tags or indices for the type of "communication-by-coherence" mechanism theorized by Fries, 2015.

The study of macroscopic behaviors in the Kuramoto model would be considerably easier if the density of phases after a burn-in period assumed some simple form. A magnetic spin system, like an Ising model, assumes a Gibbs density when the system reaches equilibrium, and phase transitions in this system occur where derivatives of the normalizing constant of this density (the partition function) are discontinuous (Ruelles, 2004). Certain similarities between the Kuramoto model and these magnetic formalisms prompted (Hemmen and Wreszinski, 1993) to pose an equilibrium analysis of the Kuramoto model in which the system was said to minimize a given energy

function during its evolution. Perez-Vicente and Ritort, 1997, for their part, noted that the equilibrium formulation of Hemmen and Wreszinski, 1993 was only valid as long as probability currents at the boundaries of $[-\pi, \pi]$ were zero[3]. This issue seems to have been mostly resolved by Gupta, Campa, and Ruffo, 2014a; Gupta, Campa, and Ruffo, 2014b who situated the Kuramoto model in a more general system with inertia and noise:

$$m\ddot{\theta_j} = -\mu\dot{\theta_j} + \sum_i K_{ij} \sin(\psi - \theta_j) + \mu\omega_j + \sqrt{\mu}\eta_j(t), \tag{3.16}$$

where m is a scalar mass, μ is a friction constant, r is the global order parameter, ψ is the phase of the mean field (see Eq. 3.14), and η_j is an uncorrelated noise process with variance T:

$$\langle \eta_i(t), \eta_j(t') \rangle = 2\frac{T}{\sqrt{m}}\delta_{ij}\delta(t - t'). \tag{3.17}$$

As before, we let $\omega \sim g$ for a symmetric, unimodal density g with variance σ. Then, the triple (m, T, σ) defines a point in the generalized model parameter space depicted in Fig. 3.10. When $m = 0$, we are in somewhat familiar Kuramoto territory: we either have the original formulation for $T = 0$ or the stochastic Langevin model cited earlier for $T > 0$ (Sakaguchi, 1988; Sakaguchi, Shinomoto, and Kuramoto, 1988). The authors demonstrate the model only obeys detailed balance and reaches equilibrium in the case that $\sigma = 0$. In this case, the model is equivalent to a classical XY system of magnetic spins (Tong, 2012) and has the fabled Gibbs distribution at equilibrium. Detailed balance is violated when $\sigma > 0$ in which case the system is out of equilibrium. One may imagine ω as an external driver of the system, forcing it away from equilibrium.

The complicated Eq. 3.16 is likely of interest to physicists on its own merit, but we mainly include here for the implications it has about *learning* in these systems. Namely, when $\sigma = 0$, the system has a Gibbs distribution at equilibrium, and we are

[3]In other words, one can cast the Kuramoto model as an equilibrium model with Gibbs-distributed phases as long as the system always evolves from a state in which all phases lie in one half-circle.

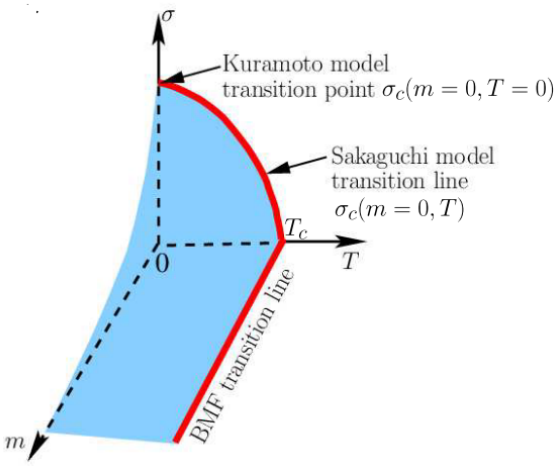


Figure 3.10 *Parameter regimes of a generalized Kuramoto model.* Regimes of the Kuramoto-type model given by Eq. 3.16 are plotted in a three-dimensional parameter space. The model is synchronized inside the region bounded by the blue surface and is desynchronized elsewhere. Phase transition boundaries for different model regimes (details in Gupta, Campa, and Ruffo, 2014a, Fig. 1, caption) are denoted by red lines. When σ is zero, the model reduces to the Brownian Mean Field (Chavanis, 2014) system, in this case an equilibrium XY spin model interacting with a heat bath. For non-zero σ, the system transitions to the non-equilibrium Kuramoto ($T = 0$) or noisy Sakaguchi ($T > 0$) model. Figure taken from Gupta, Campa, and Ruffo, 2014a.

free to use the unsupervised density fitting methods of the probabilistic graphical modeling literature. We are out of luck in the case that $\sigma > 0$ and must resort to supervised learning and direct backpropagation of error through the flow of θ. This distinction forms the theoretical dichotomy of Chs. 4 and 5.

3.3.2 Neuroscientific applications of the Kuramoto model

More often than not, studies of the Kuramoto model begin with a reflexive statement on the system's ubiquity in the natural world. These statements would grow tiresome if the omnipresence of the model were not so striking and so surprising. Our focus in this section is a brief review of the purely neuroscientific applications of the Kuramoto model, but a listing of other sorts of applications is warranted. See Arenas and Albert, 2008 for summaries of Kuramoto applications to genetic networks (Garcia-Ojalvo, Elowitz, and Strogatz, 2004), circadian rhythms (Fukuda et al., 2007), population dynamics (Blasius, Huppert, and Stone, 1999), parallel computing (Korniss et al., 2000), data mining (Miyano and Tsutsui, 2007), consensus formation (Pluchino, Latora, and Rapisarda, 2004), wireless communication networks (Diaz-Guilera et al., 2009), decentralized logistics (Lämmer et al., 2006), power-grids (Filatrella, Nielsen, and Pedersen, 2008), laser arrays (Strogatz and Mirollo, 1993), finance (Onnela et al., 2003), and world trade (Garlaschelli et al., 2007).

Neuroscientific applications of the Kuramoto model often rely on the emergence of synchrony to test for the functional connectivity of brain regions. For example, Honey and Sporns, 2008 used a Kuramoto dynamics with weights taken from macaque connectivity data (Fig. 3.11) to investigate the influence of specific lesions. Synchrony in the system was used to "[understand] how rapidly, and in which region order, the cortical network [would] synchronize from a randomly perturbed initial state." Their study indicated that brain regions with a large number of out-going connections stabilized earlier than loosely connected regions. The authors found that correlations

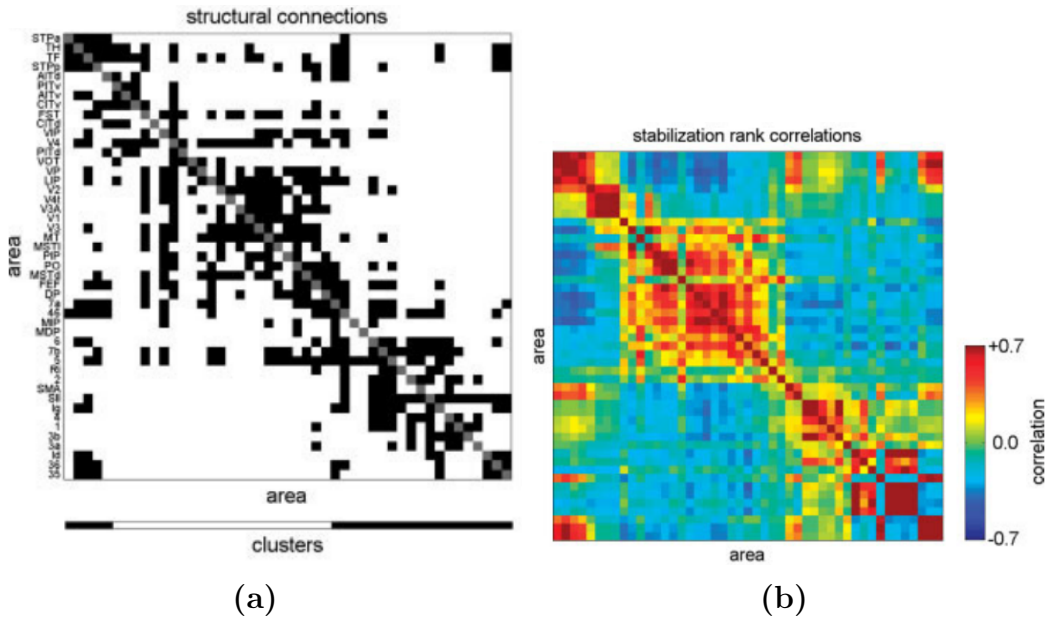
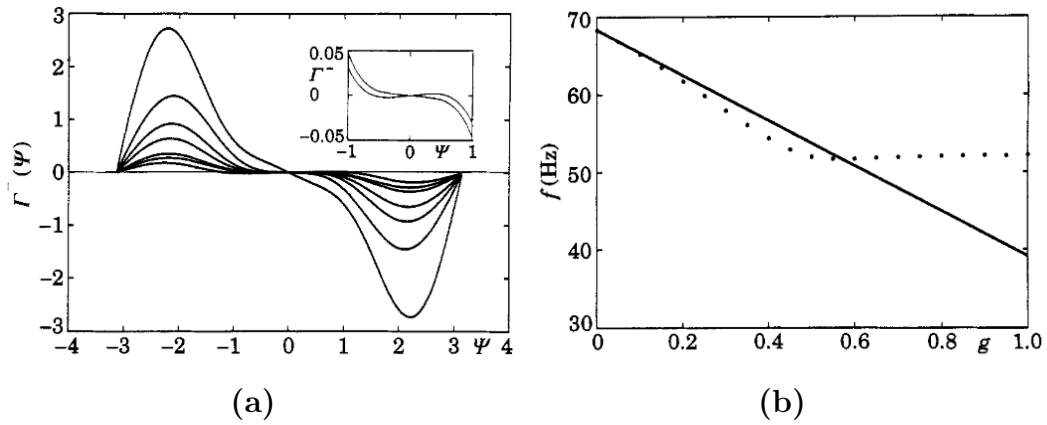


Figure 3.11 *The Kuramoto model for the understanding functional connectivity in cortex. (a)* Connectivity matrix for macaque cortex. *(b)* Correlations in the time-course of synchronization between cortical regions. This correlation matrix was found to agree closely with similar correlations in BOLD signal from monkey cortex.

between the order in which regions synchronized agreed closely with similar correlation measures from BOLD measurements in monkey cortex. Later studies examined similar dynamics in cat (Gómez-Gardeñes et al., 2010) and human (Villegas, Moretti, and Muñoz, 2014; Schmidt et al., 2015) cortex. Several studies have used the Kuramoto model to study patterns in functional connectivity in human cortex during resting state. Cabral et al., 2012, for example, simulated neural activity and the resultant hemodynamic response with a Kuramoto-like model and found that weak coupling created patterns of functional connectivity similar to those reported in schizophrenic patients. A subsequent study by the same first author (Cabral et al., 2014) found that the Kuramoto model could be used to simulate transient frequency synchronization at between 8-30 Hz seen in human MEG data during a resting state. Schmidt et al., 2015, for their part, used the Kuramoto model to demonstrate how overlapping neighborhood structure in complex networks could mimic patterns in functional connectivity as



(a) **(b)**

Figure 3.12 *Phase reduction on a Hodgkin-Huxley system. (a)* Hansel and colleagues used four Fourier modes for approximating the phase coupling function in a system of two Hodgkin-Huxley units. Depicted here is the effective coupling function for several values of input current. For low currents, the neurons will synchronize. Beyond a critical input current, however, an extra fixed point emerges (inset) and the neurons will not synchronize. Phase reduction has been used here to predict an unintuitive phase transition. *(b)* Firing rate of the system as a function of coupling strength. Dots are from numerical simulations and the straight line is the analytical prediction from phase reduction. The two diverge beyond a critical coupling since the weak coupling condition on which phase reduction depends is broken.

measured in human fMRI and DTI data.

Other studies use Kuramoto oscillators to model individual neurons. For instance, Hansel, Mato, and Meunier, 1993 used phase reduction to model networks of Hodgkin-Huxley neurons as Kuramoto-like oscillators. The authors used four Fourier modes from Eq. 3.10 in their approximation and showed how the reduced system could predict the onset of bistability and out-of-phase locking (Fig.3.12) Later, Kitzbichler et al., 2009, showed that the patterns of transient phase-locking in the Kuramoto model near the critical points strongly resembled the correlations in human MEG data taken during resting state. Breakspear, Heitmann, and Daffertshofer, 2010 went on to argue that the Kuramoto model could be used to model behavior characteristic of cortical activity in the beta regime (Freyer et al., 2009), that regime which characterized same-different processing in Sec. 3.1.

3.3.3 Current approaches to optimizing the Kuramoto model

Analytical expressions for the degree of synchrony in the general Kuramoto model–with potentially random, frustrated (Zanette, 2005) and arbitrary frequency distributions–are not forthcoming[4]. As a result, there have been numerous attempts to instead develop empirical methods for the determination of parameter regimes giving rise to global synchrony. Often, these methods treat the study of global synchrony as an optimization problem in which intrinsic frequencies and couplings must be gradually adapted to maximize a synchrony-dependent fitness (or minimize a desynchrony-dependent loss).

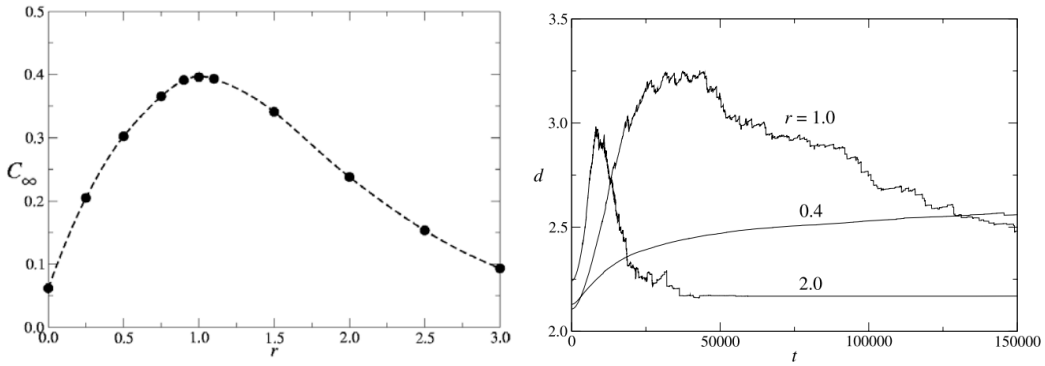
An early study by Gleiser and Zanette, 2006 sought to maximize phase-locking in the Kuramoto model by rewiring an Erdös-Rényi graph (Erdös and Rényi, 1959)s. Every T time-steps, the empirical frequency of each oscillator was calculated

$$\Omega_j = \frac{1}{T - t'} \int_t^{T+t} \dot{\theta}(t') \, dt' \tag{3.18}$$

Then, for each oscillator indexed by i, the link (i, j) to the unit j having the farthest empirical frequency from i as measured by $|\Omega_i - \Omega_j|$ was replaced with (i, k) where k had the closest intrinsic frequency. The coupling strength on these links was constant and experimentally varied. The authors found that the system gradually evolved towards a small-world structure as it adapted to produce greater global synchrony (Fig. 3.13). Interestingly, the model was found to synchronize at much lower coupling strengths than in the all-to-all coupling case (e.g. 3.15).

Later work by Tanaka and Aoyagi, 2008 sought to optimize both couplings and intrinsic frequencies using gradient descent, assuming positive, symmetric, couplings. When intrinsic frequencies and total coupling strength were held fixed, the authors found that a randomly initialized network would gradually develop links between

[4]However, see Ott and Antonsen, 2008 and Gottwald, 2015 for some exciting recent developments on that front.



(a) **(b)**

Figure 3.13 *Early results concerning optimal graphs for synchrony. (a)* Every T steps of a Kuramoto dynamics, Gleiser and Zanette adjusted the adjacency matrix in the underlying graph according to the nearness of oscillator intrinsic frequencies. Plotted here is the clustering coefficient of the graph for large t as a function of coupling strength r. Note that maximal clustering occurs for middling coupling strengths. Clustering is lesser for weak and strong coupling strengths because, in the former case, rewirings which would increase clustering have little effect on synchrony anyway, and, in the latter case, rewiring is unnecessary since strong couplings enforce synchrony easily. *(b)* The graph distance, d, between oscillators is plotted as a function of time for different coupling strengths. In all cases, d decreases asymptotically, indicating the graph is becoming more small-world like.

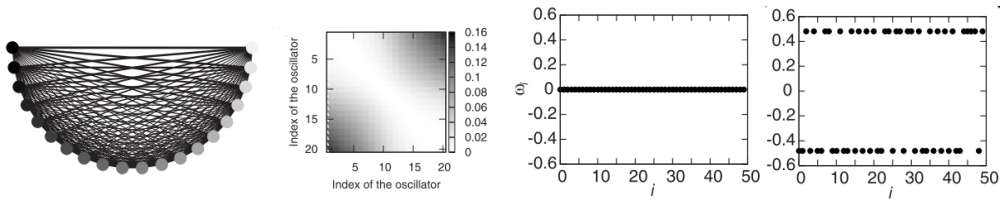


Figure 3.14 *Couplings between disparate oscillators encourages synchrony. (a)* A small network of oscillators was constructed with all-to-all coupling. The network is depicted here with oscillators ordered from left to right by intrinsic frequency, as indicated by grayscale value. The weight matrix optimized for synchrony is plotted in the second panel. Note that oscillators with greatly different intrinsic frequencies tend to be strongly connected. *(b)* In a second demonstration, the authors also optimized intrinsic frequencies. When coupling strength was low, oscillators converged to a single value (left panel), and when coupling strength was high, oscillators converged to two clusters. Figure from Tanaka and Aoyagi, 2008

oscillators having disparate intrinsic frequencies, prefiguring numerous future results in the literature (Fig. 3.14). If instead both couplings and intrinsic frequencies were optimized, a large total coupling strength encouraged intrinsic frequencies to split into two groups, while a small coupling strength forced frequencies to converge to a single mode.

The first result of Tanaka and Aoyagi, 2008 was replicated in several experiments by Markus Brede (Brede, 2008a; Brede, 2008b; Brede, Stella, and Kalloniatis, 2018). Contrary to previous work, Brede optimized complex graphs encouraging synchrony by random search. Starting from a random initial connectivity, an average global order parameter was calculated by

$$\bar{r} = \frac{1}{T - t'} \int_{t'}^{T} r(t)\, dt. \tag{3.19}$$

A random link was then switched and $\bar{r}$ was recalculated. If the resulting average global order was higher than before the switch, it was kept, and it was otherwise ignored. Intrinsic frequency distributions were fixed. Consistently, oscillators of disparate intrinsic frequencies came to link to one another over the course of the

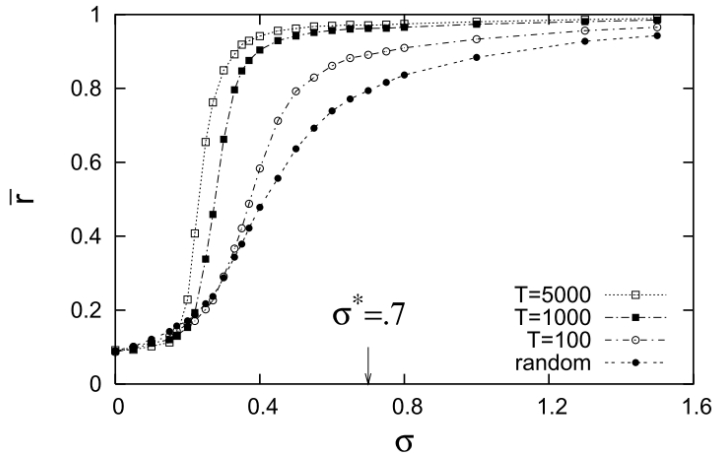


Figure 3.15 *Structured graphs encourage synchrony across a range of coupling strengths.* Brede optimized a complex graph with 100 nodes for which the probability of connection between two nodes was .035. Coupling strength was fixed at $\sigma^* = .07$. The graph was optimized for a single vector of intrinsic frequencies. Then, he plotted the average coherence of the system (Eq. 3.19) across a range of coupling strengths compared to a random baseline. Increased training iterations, T, created uniformly more coherent dynamics across the range of couplings considered.

optimization. Brede found that, although the network was optimized for a single coupling strength, it still gave rise to greater coherence across a range of coupling strengths (Fig. 3.15).

Optimization of Kuramoto parameters marks an important step in the empirical understanding of oscillatory systems and their potential role in neural function. The introduction of optimization techniques into the study of these systems is especially tantalizing since it mimics the development of learning methods for feedforward networks which helped establish today's fruitful and exciting dialogue between deep networks and visual cortex (Eberhardt, Cader, and Serre, 2016; Yamins et al., 2014; Nayebi et al., 2018). Nevertheless, the methods we have so far described either rely on certain model restrictions like symmetric, positive couplings which eliminate the more exotic regimes of Kuramoto dynamics (e.g. potential glassy states) or employ indirect techniques like random search.

Our intention in the subsequent chapters will be the development of direct learning

methods relying on the gradient of certain macroscopic system quantities with respect to model parameters and which do not place any restrictions on those parameters. In short, we will exploit the prevalence and strength of contemporary machine learning methods to do deep learning on the Kuramoto model. The technical description of the Kuramoto model we have just provided has laid the groundwork for our ultimate goal of making a learning system which could conceivably model those mysterious oscillatory functions described in this chapter's beginning: binding by synchrony, communication by coherence, and phase coding more generally. We turn to this sort of modeling next.

Chapter Four

Kosterlitz machines

The[1] formal models of phase-coding we have reviewed, descriptive and informative as they are, cannot learn. This is a serious limitation for any model with pretenses towards biological plausibility, since the best predictors of neural activity, especially in visual cortex, are invariably neural networks optimized for task performance and not hardwired to mimic cortical activity directly (Yamins et al., 2014; Nayebi et al., 2018). Our challenge is then to devise a neural architecture whose dynamics mimics the CTC mechanism of gating by coherence and can be adapted to a given task by standard machine learning methods.

This chapter proposes one solution to this challenge, what we will call the *Kosterlitz machine*[2] (KM), based on models from the equilibrium regime of Eq. 3.16 (see Ch.3. Fig. 3.10). Relying on the fact that such a system obeys detailed balance and is therefore Gibbs-distributed at equilibrium, we can attempt to model this distribution directly so that it comes to resemble a target distribution of phases. In the CTC framework, this type of learning would be of use to downstream neural populations looking to model the distribution of phases of afferent populations to which the

[1]Material from this chapter was presented as Ricci, Windolf, and Serre, 2019 and is the subject of a manuscript in preparation with co-authors Charles Windolf and Thomas Serre.

[2]The model we propose is based on the classical XY model from statistical physics, one of whose famous investigators is Brown University's Michael Kosterlitz.

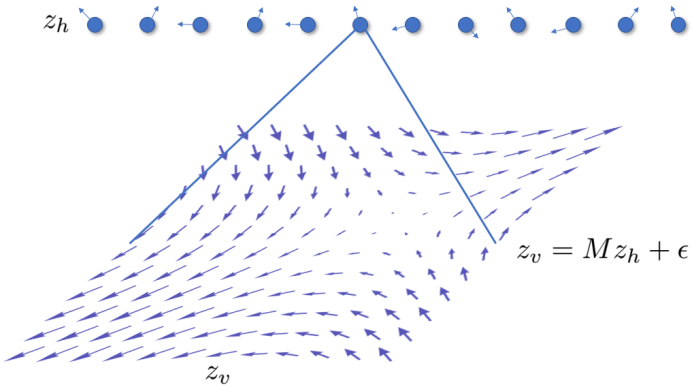


Figure 4.1 *Learning a generative model of vector arrays.* Our goal in this chapter is to construct a hierarchical graphical model in stochastic neurons have both a rate and a phase. The result will be a system for which hidden units function as latent phase-based codes for generating angular data.

downstream ensemble could be entrained. More generally, the method outlined in this chapter can be applied to the unsupervised learning of arbitrary 2D vector fields (Fig. 4.1). A Kosterlitz Machine is at first glance a 2D generalization of the famous Boltzmann Machine (Ackley, Hinton, and Sejnowski, 1985), itself an adaption of the Ising and n-vector models of statistical physics. An important difference between the KM and these other formalisms is that neurons in the KM have both a magnitude and direction. In statistical terms, this magnitude encodes the confidence that a given feature is present in the neuron's receptive field; in physical terms, as we will see, this confidence is a dynamical coupling which can force the system to separate into non-interacting ensembles in the manner of Malsburg, 1994. This is the first of several peculiarities that differentiates the KM from its analogues in statistical physics.

A principal benefit of the KM is the manner in which it intuitively embeds *mechanistic* descriptions of phase coding, like gating by coherence, within the *statistical* framework of probabilistic graphical models. Casting oscillatory neural dynamics in probabilistic terms allows us to abstract away from the complicated microscopic behavior of individual neurons and focus instead on high-level macroscopic processes, like inference. For example, the level of coherence of units within a given receptive field

is closely related to the statistical confidence the target unit has in the presence of a given feature appearing in the data. Further, we will show that the decoupling behavior induced by the neural amplitudes can be cast as a form of variational inference.

The Kosterlitz Machine is in many ways a synthesis of existing models, so we will do well to review those systems first. We will then lay out two versions of the system depending on which types of neuron magnitudes are allowed. We will demonstrate how these systems can be efficiently trained and stacked to form deep hierarchical neural networks. Across several experiments, we will show how perceptual grouping in the KM can be cast as conditional inference and communication through coherence takes the form of a prior on latent phases. Finally, some future directions will be discussed.

4.1 Equilibrium oscillatory systems in image processing

Though it lacks any notion of phase, perhaps the most famous machine learning system inspired by equilibrium statistical mechanics is the celebrated Boltzmann Machine (BM) (Ackley, Hinton, and Sejnowski, 1985), a classical method of non-parametric density estimation which remains the subject of a good deal of interest for its connection to the study of spin glasses (Nishimori, 2010) and Hebbian learning in neural systems (Reichert, 2014). Most of theoretical framework for the phase-based systems we will describe are inherited from the BM, so we will briefly review this framework here. The interested reader can consult Fischer and Igel, 2014 for more details on BMs and Wainwright and Jordan, 2008 for a voluminous review of probabilistic graphical models more generally.

A Boltzmann Machine (Fig. 4.2) is a Markov random field (MRF) taking random values on $\{0, 1\}^n$ (or $\{\pm 1\}$, depending on the author) on an undirected graph $G =$

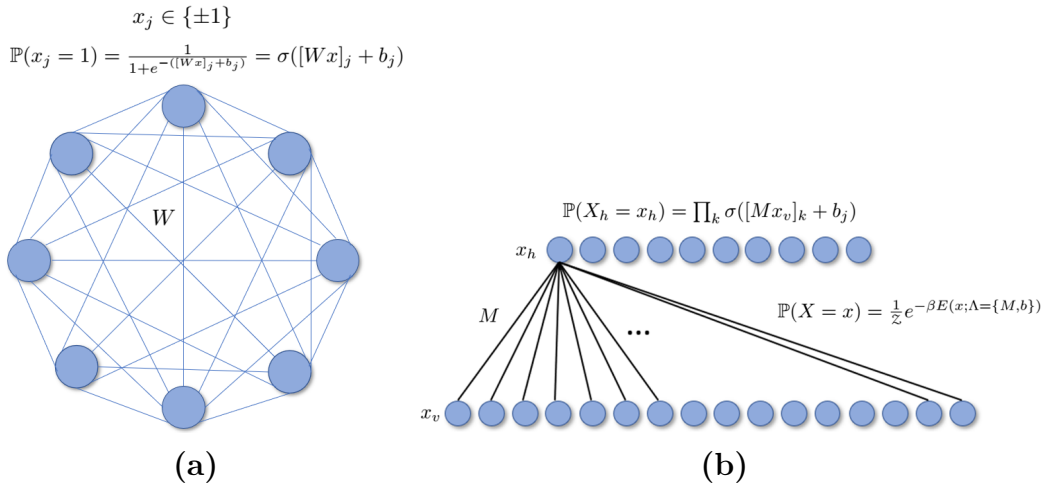


Figure 4.2 *Boltzmann machines. (a)* A Boltzmann machine is a binary-valued Markov random field parameterized by a weight matrix W and a vector of biases b. The Gibbs density on model configurations implies that the conditional distribution on a single unit's state given the rest of the network is Bernoulli distributed with parameter controlled by a sigmoidal function of the unit's input (second equation). *(b)* A restricted Boltzmann machine (RBM) splits the graph into two layers so that units are conditionally independent within a layer and the "hidden" layer functions like latent codes for generating "visible" data according to a learned dictionary of features, M.

$\{V, E\}$. The log-likelihood of a random configuration X is given by

$$\begin{aligned} \log p(X = x|\Lambda = \{W, b\}) &= -\beta(x^T W x + b^T x) - F(W, b) \\ &= -\beta E(x; W, b) - F(W, b), \end{aligned} \tag{4.1}$$

where $W \in \mathbb{R}^{n \times n}$ is a symmetric matrix of synaptic weights or couplings, $b \in \mathbb{R}^n$ is a vector of biases or firing thresholds, F is a constant called the system's *free energy*, and β is a free parameter controlling the entropy of X. The parameters of the BM are collected into Λ. We call the component of the log-likelihood depending on x the *energy* of x and denote it by $E(x)$.

Suppressing the dependence on Λ, we note that the joint distribution on X is given

by

$$p(X = x) = e^{-\beta E(x) - F}$$
$$= \frac{1}{\mathcal{Z}} e^{-\beta E(x)}, \tag{4.2}$$

where $F = \log \mathcal{Z}$. Eq. is therefore a Gibbs or Boltzmann distribution with inverse temperature parameter $\beta = \frac{1}{T}$ and partition function $\mathcal{Z}$, which contains all of the essential information about X's distribution and normalizes its density:

$$\mathcal{Z} = \int_x e^{-\beta E(x)} \, dx. \tag{4.3}$$

The physicist will now recognize a BM machine as simply an Ising model with an external field parametrized by b and potentially disordered interactions, W.

A Boltzmann machine is used to model a target distribution D by adapting Λ by maximum likelihood estimation:

$$\Lambda^* = \arg\max_{\Lambda} \mathbb{E}_{x \sim D} \left[\log p(x; \Lambda) \right] \tag{4.4}$$

Unfortunately, computing Λ^* by gradient ascent involves calculating the partition function $\mathcal{Z}$ and the complexity of this calculation is generally exponential in the number of units in the MRF. Instead, the gradient is approximated by either sampling or variational methods which we will discuss in more detail below.

Typically, these approximations are facilitated by constraining Λ to lie in some feasible set which expedites sampling or variational inference. The most common restriction is to assume that X is partitioned into two or more ensembles which are internally conditionally independent. For instance, we may partition X by $X = [X_v : X_h]$, where : denotes concatenation, and call the corresponding sets of vertices, $V_v, V_h \subset V$, the *visible* and *hidden* partitions, respectively. We may then restrict W to be an off diagonal block-matrix of the form

$$W = \left[\begin{array}{c|c} 0 & \frac{1}{2}M \\ \hline \frac{1}{2}M^T & 0 \end{array} \right] \tag{4.5}$$

where $M \in \mathbb{R}^{n \times m}$. Biases are partitioned similarly. In this case, weights between units in a partitioned are removed, G is made bipartite, and one can check that $p(X_v|X_h) = \prod_j p(X_j|X_h)$ and vice versa for X_h. We call this system a *restricted Boltzmann machine* (RBM). This can be generalized to any number of partitions.

RBMs are in general much easier to train than general Boltzmann machines and are additionally interesting for their interpretation as a multi-layer generative neural network (Fig. 4.2b). For example, if a data distribution $X \sim D$ takes values on $\{0,1\}^n$, then we can construct a hierarchical generative model of the data with a visible partition X_1 taking values in $\{0,1\}^n$ and hidden partitions $\{X_\ell\}_{\ell=2}^L$ by performing maximum likelihood estimation on the marginal distribution

$$p(X_1 = x_1|\Lambda) = \int_{x_2,\ldots,x_L} p(x_1,\ldots,x_L;\Lambda)\, dx_2\ \ldots dx_L.$$

Now, the random variables X_2 function as latent factors generating the data X_1, X_3 functions as latent factors generating X_2, and so on. When $L > 2$ we refer to this system as a *deep* restricted Boltzmann machine (DRBM)[3].

Though they are currently outperformed by newer density estimation methods, like Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), Boltzmann machines and their derivatives have remained an important object of study for both machine learning practitioners and computational neuroscientists for their intuitive connection to neural systems. For example, just from Eq. 4.1 we can demonstrate that each unit in a BM functions as a linear-nonlinear neuron (Ostojic and Brunel, 2011) with a sigmoidal activation function. Further, the derivative of the network's log-likelihood with respect to a given synapse is a function of only the activities on the pre- and post-synaptic sides of that weight, making the learning algorithm for Boltzmann machines entirely *local* in addition to being unsupervised. This circumvents

[3]This is not the same as a deep belief network (DBN), which is a hierarchical generative model on a directed graph. See Salakhutdinov and Larochelle, 2010 for a comparison between DBNs and DBMs.

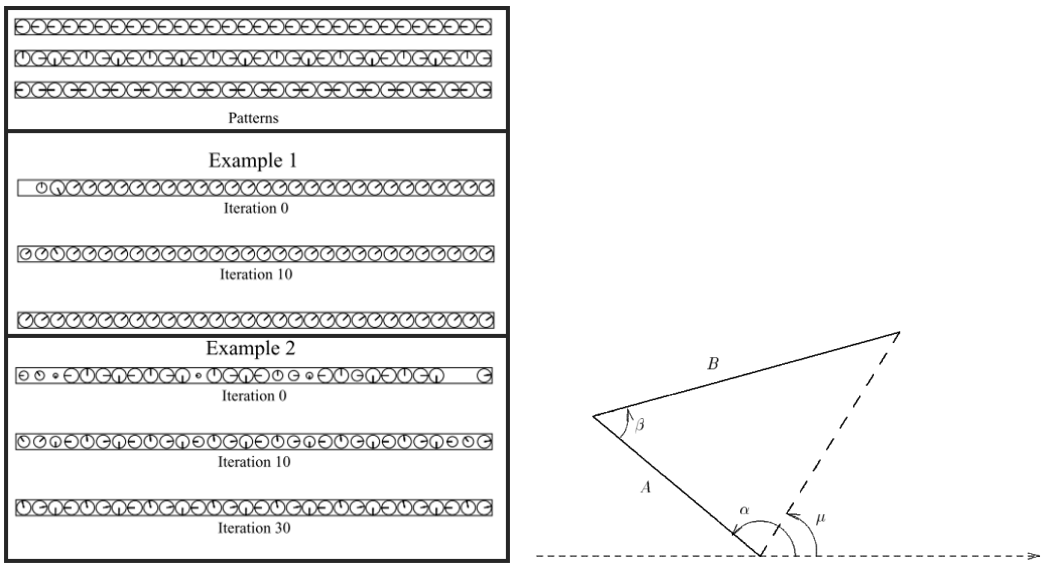


Figure 4.3 *Directional unit Boltzmann machines. (a)* A directional unit Boltzmann machine (DUBM) is a probabilistic graphical model in which units take values on the circle. Zemel and colleagues trained their system to memorize three simple patterns (top panel) and then retrieved the stored inputs by annealing the system from a random starting configuration. The middle panel shows convergence from a random state to the first pattern. The bottom panel shows convergence to the second pattern. Notice that uniform rotations of a configuration are equally likely. *(b)* In a second demonstration, the authors trained the model to learn the angular position, μ, of a robot arm with forearm A and upper arm B. Angles α and β were provided as input and μ was the target output.

the use of a global supervisor (as in the case of CNNs) or an additional discriminator network (as in the case of GANs).

The earliest adaptation of the Boltzmann machine framework to a phase-based system came in the form of directional-unit Boltzmann machines (DUBMs) (Zemel, Williams, and Mozer, 1995), which replaced the binary neural states of the BM with circular states represented by a phase variable $\theta \in [0, 2\pi]$ with periodic boundaries. Just as the Boltzmann machine is the data scientific analogue of the Ising model, the DUBM is analogous to classical XY model described in Sec. 3.3.1. Zemel and colleagues trained a small, two-layer DUBM's on a pattern completion task and a task involving the rotation of a robot arm (Fig. 4.3), emphasizing that the magnitude of a

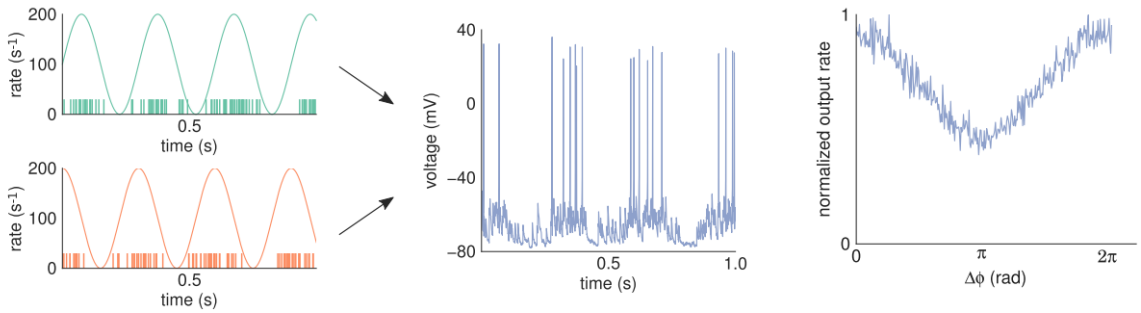


Figure 4.4 *Phase as an inhibitory envelope.* Reichert and Serre conceived of neural phase in their system as an inhibitory envelope suppressing spikes, as in the left panel. When two of these bursting units integrate at a synaptic terminal (middle panel), their resulting postsynaptic rate is a function of their phase difference (right panel). Minimal rate is achieved when afferents are π out of phase.

unit's expected value increased to the degree that afferent units "agreed" in phase. The magnitude of the unit's expected state can therefore be interpreted as a measure of confidence in predicted the phase of units in its receptive field which is close to 0 when the units are completely desynchronized and 1 when the units are synchronized. However, since the unit states themselves always have magnitude 1, this $[0, 1]$-valued confidence measure does not actually play a role in the evolution of the stochastic system.

Nearly twenty years later, Reichert and Serre, 2013 proposed a similar system in which unit states had both a phase and bona fide binary magnitude. Such a system might be expected to display categorically different behavior than that of Zemel, Williams, and Mozer, 1995 since neural magnitudes and the confidences they encode more directly influence the system's evolution. The authors likened the magnitudes of their model neurons to the rate of a traditional McCulloch-Pitts unit and analogized the phase variable to the phase of a periodic inhibitory envelope on the unit's spikes. The system was trained only indirectly: real-valued weights were simply taken from a traditional DRBM trained on binary scenes of a few shapes or digits. The authors then affixed the trained system with additional phase variables on every unit. Reichert

and Serre, 2013 proposed a stochastic update rule by which each unit would update its extended state given activity from neighboring units. This update rule was a complicated function of both the magnitude of the unit's weighted inputs and the sum of pointwise moduli of the afferent activities.

The system was evaluated by "clamping" a given binary stimulus to the visible layer of the network, initializing it with random phases and then updating the stochastic state of the whole network in parallel using the aforementioned update rule. Surprisingly, although Reichert and Serre, 2013 had not trained the system for segmentation, the authors found that stochastic updating in this manner caused the clamped image to spontaneously group into perceptually meaningful segments (Fig. 4.5), a phenomenon they likened to the hypothesized "binding by synchrony" mechanism described in Sec. 3.2.1. Further, the authors noted that their interpretation of neural phase the timing of an inhibitory envelope provided a natural gating mechanism during synaptic integration: units would destructively interfere to the degree that their envelopes were out of phase. Reichert and Serre provide little explanation of their

Figure 4.5 *Spontaneous "binding by synchrony" in a phase-based deep network.* Reichert and Serre found that when a deep Boltzmann machine was trained on binary images like those in the left panels, its weights could be used to encourage a spontaneous "binding by synchrony" in a similar system (right panels).

remarkable finding, except to intuit that "units gradually decouple" (Reichert and Serre, 2013, p. 4) over the course of stochastic updating. Moreover, their two-step

procedure of training a binary DRBM and then affixing it with a novel stochastic updating procedure unfortunately divorced the system's microscopic dynamics from its macroscopic statistical meaning. For example, it is common to refer to a traditional Boltzmann machine as a stochastic recurrent neural network since samples of the system are often acquired by Gibbs sampling which requires the sequential activation of units which in turn send their states to other units, and so on. However, the sequential activation of units over the course of Gibbs sampling defines a Markov chain whose invariant distribution is known. The dynamics of Reichert and Serre, 2013, if they carry out Gibbs sampling at all, do so on a distribution which is unknown. In particular, no joint distribution or energy functional is provided by the authors.

Consequently, the behavior of the system, surprising as it is, is rather difficult to explain. Personal correspondence with the authors also indicated that the conditions under which satisfactory "binding by synchrony" would occur in the model were not always clear. This is to be expected, as the system was not trained end-to-end to perform perceptual grouping with phase. Our goal in the subsequent sections will be to rigorize the model of Reichert and Serre, 2013 in an attempt to both stabilize and explain their results. The result will be an end-to-end trainable model of phase-based perceptual grouping in which units have both a random phase and magnitude.

4.2 Kosterlitz Machines

We define a *Kosterlitz machine* (Fig. 4.6a) as a Markov random field with a random configuration Z taking values in $\mathbb{C}^{n\times m}$ according to the Gibbs density

$$p(Z = z; \Lambda) = \frac{1}{\mathcal{Z}} e^{-E(z;\Lambda)}, \tag{4.6}$$

where Λ is a set of parameters as before. Again, we partition Z into two sets $Z = [Z_v : Z_h]$, where $Z_v \in \mathbb{C}^n$ and $Z_h \in \mathbb{C}^m$ are called the "visible" and "hidden" states, respectively. Below, visible units will be indexed by j and hidden units will be

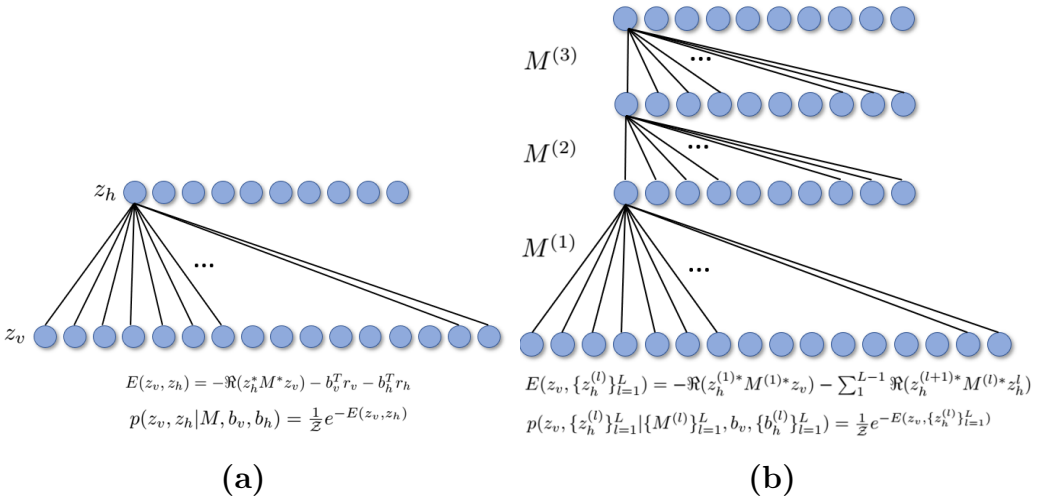


Figure 4.6 *Kosterlitz machines. (a)* A (restricted) Kosterlitz machine is a generalization of a DUBM in which units have amplitude in addition to a phase. *(b)* Deeper hierarchies can be trained for richer representations of data.

indexed by k. We may think of Z_h as a latent code serving to generate the data Z_v according to the model $p(Z_v|Z_h;\Lambda)$. As before, this partitioning can be generalized, resulting in a deep KM (Fig: 4.6b). Henceforth, we will drop the distinction between a random variable and its sample and simply denote all random states with lower case letters.

This generative model will be fit to samples X from some data distribution D taking values in $\mathbb{C}^n$ by computing the maximum likelihood parameters

$$\Lambda^* = \arg\max_{\Lambda} \mathbb{E}_{x\sim D} \left[\log p(z_v;\Lambda)\right]. \tag{4.7}$$

When we maximize marginal log-likelihood, we find that partial derivatives with

respect to parameters λ_j have the form

$$\begin{aligned}
\frac{\partial \log p(z_v;\Lambda)}{\partial \lambda_j} &= \frac{\partial}{\partial \lambda_j} \log\left(\frac{1}{\mathcal{Z}} \int_{z_h} e^{-E(z_v,z_h;\Lambda)}\ dz_h\right) \\
&= \frac{\partial}{\partial \lambda_j} \log\left(\int_{z_h} e^{-E(z_v,z_h;\Lambda)}\ dz_h\right) - \frac{\partial \log \mathcal{Z}}{\partial \lambda_j} \\
&= -\frac{1}{\mathcal{Z} p(z_v;\Lambda)} \int_{z_h} \frac{\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j} e^{-E(z_v,z_h;\Lambda)}\ dz_h + \frac{1}{\mathcal{Z}} \int_{z_v}\int_{z_h} \frac{\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j} e^{-E(z_v,z_h;\Lambda)}\ dz_h \\
&= -\int_{z_h} \frac{\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j} \frac{p(z_v,z_h;\Lambda)}{p(z_v;\Lambda)}\ dz_h + \int_{z_v}\int_{z_h} \frac{\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j} p(z_v,z_h;\Lambda)\ dz_v\ dz_h \\
&= \mathbb{E}_{p(z_h|z_v)}\left[\frac{-\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j}\right] - \mathbb{E}_{p(z_v,z_h)}\left[\frac{-\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j}\right] \qquad (4.8) \\
&= \left\langle -\frac{\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j} \right\rangle_{\text{data}} - \left\langle -\frac{\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j} \right\rangle_{\text{model}} \qquad (4.9)
\end{aligned}$$

The "data "expectation in Eq. (4.8) is taken with respect to the conditional distribution of the hidden units given the visible units and the "model" expectation is with respect to the full joint. Calculating the second expectation in (4.8) requires computing the partition function, which is typically intractable. Consequently, this expectation is usually replaced by a statistic acquired either by running a k-length Gibbs sampling chain or by variational approximation. The former procedure for computing Λ^* by approximating the gradient of log-likelihood with k steps of Gibbs sampling is called k-step *contrastive divergence* (CD-k).

So far, the basic framework we have described is identical to that of a Boltzmann machine, or indeed any MLE application for some marginal distribution. Next, we describe some peculiar features of KMs that arise from the multi-dimensional state space of their units and their energy in two cases, the first of which is a direct rigorization of Reichert and Serre, 2013 and the second of which is a generalization to non-binary unit magnitudes.

4.2.1 Bernoulli Kosterlitz Machines

Being a sort of Ising model, the traditional Boltzmann machine has Bernoulli-distributed units typically taking values in $\{0, 1\}$. These "up-down" states inherited from the physics literature have a satisfying statistical meaning in the context of the generative model described above: a hidden unit's achieving a 1/0 state represents the machine's confirmation/rejection of a particular latent hypothesis about the data. In this spirit, each unit in a Bernoulli Kosterlitz Machine (BKM) takes random values on $\mathbb{S}^1 \cup \{0\}$, and we represent its state by the complex number $z = re^{i\theta}$, where $r \in \{0, 1\}$. We call the modulus r_j the "rate" of z_j since if all phases are constant, then the neuron reduces to the traditional rate-coding McCulloch-Pitts unit with a sigmoidal activation. We may now think of r as encoding the confirmation or rejection of a latent hypothesis about the data and θ as encoding a particular instance of this hypothesis. In other words, r says "if" a hypothesis is true and θ says "how" it is true.

Next, we let $\Lambda = \{W, b\}$, where $W \in \mathbb{C}^{(n+m)\times(n+m)}$ is again a complex Hermitian block matrix of the form

$$W = \left[\begin{array}{c|c} 0 & M \\ \hline M^* & 0 \end{array}\right] \tag{4.10}$$

and $b \in \mathbb{R}^{n+m}$ is partitioned into two sets $[b_v : b_h]$. If $r = [r_v : r_h]$ is the vector of moduli of z (e.g. $|z_j| = r_j$), then we choose the BKM energy

$$\begin{aligned} E(z; \Lambda) &= -\frac{1}{2} z^* W z - b^T r \\ &= -\operatorname{Re}\left(z_h^* M^* z_v\right) - b_v^T r_v - b_h^T r_h. \end{aligned} \tag{4.11}$$

The block form of W ensures that all units within a visible/hidden partition are conditionally independent and gives a KM the structure of a two-layer neural network. When a KM has this bipartite structure, we call it a *restricted* KM (RKM). Below, it is understood that all KMs are RKMs.

Observe that, if a weight $m_{jk} = c_{jk}e^{i\gamma_{jk}}$ connects visible and hidden units $z_j = r_j e^{i\theta_j}$

and $z_k = r_k e^{i\theta_k}$, then this two-unit configuration decrements the BKM's energy by

$$r_j c_{jk} r_k \cos(\theta_j - \theta_k - \gamma_{jk}) + b_j r_j + b_k r_k = \begin{cases} 0 & r_j = r_k = 0 \\ c_{jk}\cos(\theta_j - \theta_k - c_{jk}) + b_j + b_k & r_j = r_k = 1 \\ b_i & r_i = 1, r_l = 0, i \neq l \end{cases}$$

Hence, the (j,k) pair is at its least energetic when both units are on and $\theta_j - \theta_k = \gamma_{jk}$. When M is real, this occurs when the (j,k) pair is synchronized.

Our goal is to learn maximum likelihood parameters for the marginal density $p(z_v; M, b)$. Letting $x_k = a_k e^{i\alpha_k} = \sum_j z_j \overline{m_{jk}}$ be the input to hidden unit k, we find this marginal density has the form

$$\begin{aligned}
p(z_v; M, b) &= \frac{1}{\mathcal{Z}} \int_{z_h} e^{-E(z_v, z_h; M, b)}\, dz_h \\
&= \frac{1}{\mathcal{Z}} \int_{z_h} e^{\mathrm{Re}\left(\sum_{j,k} \overline{z_k}\ \overline{m_{jk}} z_j\right) + \sum_j b_j r_j + \sum_k b_k r_k}\, dz_h \\
&= \frac{e^{\sum_j b_j r_j}}{\mathcal{Z}} \int_{z_h} e^{\sum_k r_k a_k \cos(\theta_k - \alpha_k) + \sum_k b_k r_k} \\
&= \frac{e^{\sum_j b_j r_j}}{\mathcal{Z}} \int_{z_1} \cdots \int_{z_m} \prod_k e^{r_k a_k \cos(\theta_k - \alpha_k) + b_k r_k}\, dz_1 \cdots\, dz_m \\
&= \frac{e^{\sum_j b_j r_j}}{\mathcal{Z}} \prod_k \int_{z_k} e^{r_k a_k \cos(\theta_k - \alpha_k) + b_k r_k}\, dz_k \\
&= \frac{e^{\sum_j b_j r_j}}{\mathcal{Z}} \prod_k \int_{\theta_k} \sum_{r_k} e^{r_k a_k \cos(\theta_k - \alpha_k) + b_k r_k}\, d\theta_k \\
&= \frac{(2\pi)^m e^{\sum_j b_j r_j}}{\mathcal{Z}} \prod_k (1 + e^{b_k} I_0(a_k)) \\
&= \frac{e^{\sum_j b_j r_j}}{\mathcal{Z}} \prod_k (1 + e^{b_k} I_0(a_k)) \\
&= \frac{1}{\mathcal{Z}} e^{\sum_j b_j r_j + \sum_k \log\left(1 + e^{b_k} I_0(a_k)\right)} \\
&= \frac{1}{\mathcal{Z}} e^{G(z_v; M, b)}
\end{aligned} \tag{4.12}$$

where we have absorbed constants into the partition function. G is typically called the model's *marginal free energy*.

M plays the role of a synaptic weight matrix and b a vector of biases or spiking thresholds in this network. This can be observed by denoting the input to a visible unit z_j by $x_j = a_j e^{i\alpha_j} = \sum_k z_k m_{jk}$ and the vector of visible units besides z_j by z_{-j} and calculating the conditional density $p(z_j|z_{-j}, z_h; \Lambda)$:

$$\begin{aligned} p(z_j|z_{-j}, z_h; \Lambda) &= \frac{p(z;\Lambda)}{\int_{z_j} p(z;\Lambda)\ dz_j} \\ &= \frac{\frac{1}{\mathcal{Z}} e^{-E(z_j, z_{-j}, z_h; \Lambda)}}{\frac{1}{\mathcal{Z}} \int_{z_j} e^{-E(z_j, z_{-j}, z_h; \Lambda)}\ dz_j} \\ &= \frac{e^{\mathrm{Re}\left(z_j \sum_k \overline{z_k m_{jk}}\right) + b_j r_j}}{\int_{z_j} e^{\mathrm{Re}\left(z_j \sum_k \overline{z_k m_{jk}}\right) + b_j r_j}\ dz_j} \\ &= \frac{e^{\mathrm{Re}(z_j \overline{x_j}) + b_j r_j}}{\int_{z_j} e^{\mathrm{Re}(z_j \overline{x_j}) + b_j r_j}\ dz_j} \\ &= \frac{1}{\mathcal{Z}_j} e^{r_j a_j \cos(\theta_j - \alpha_j) + b_j r_j}, \end{aligned} \tag{4.13}$$

where

$$\begin{aligned} \mathcal{Z}_j &= \int_{\theta_j=0}^{2\pi} \sum_{r_j=0,1} e^{r_j a_j \cos(\theta_j - \alpha_j) + b_j r_j}\ d\theta_j \\ &= 2\pi(1 + e^{b_j} I_0(a_j)), \end{aligned}$$

and I_0 is a modified Bessel function of the first kind of degree 0. Note that $z_j|z_h$ is nearly distributed as a von Mise random variable[4], save the mass at $r_j = 0$. Hence, we say that a random variable z_j with density Eq. 4.13 is *Bernoulli-von Mises* distributed and write $z_j \sim \mathrm{BVM}(\alpha_j, a_j)$. Neuroscientifically speaking, we may imagine the calculation of the conditional distribution as a form of synaptic integration (Fig. 4.7) in which the phases of pre-synaptic units constructively or destructively interfere at the post-synaptic terminus. Destructive interference is caused by a desynchronous

[4]We say that a random variable θ is von Mises distributed if it takes values on the circle according to the density $\frac{1}{2\pi I_0(a)} e^{a\cos(\alpha-\theta)}$ where I_0 is a modified Bessel function of the first kind with degree 0. The von Mises density is symmetric about the central angle α and its dispersion is controlled by the concentration parameter a.

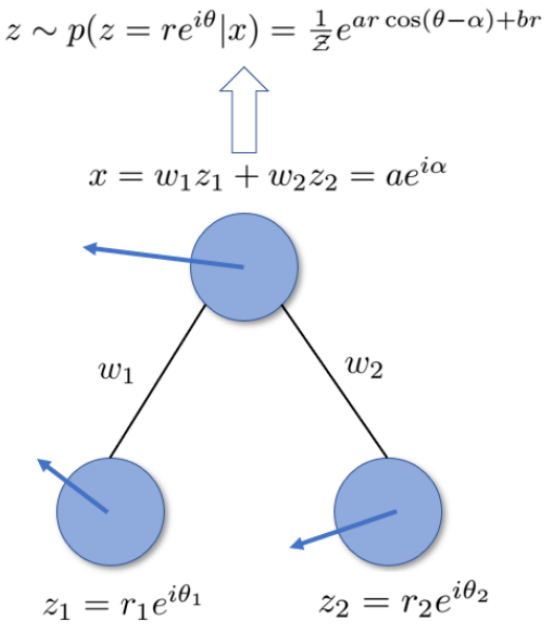


Figure 4.7 *Phase-based gating at the KM synapse.* Two units, z_1 and z_2 synapse onto a third. The random state of the target unit is parameterized by the integrated activity $x = ae^{i\alpha}$. These parameters define a Bernoulli-von Mises distribution from which the actual state is sampled.

receptive field and results in low post-synaptic amplitude, which Zemel, Williams, and Mozer, 1995 likened to a measure of statistical confidence.

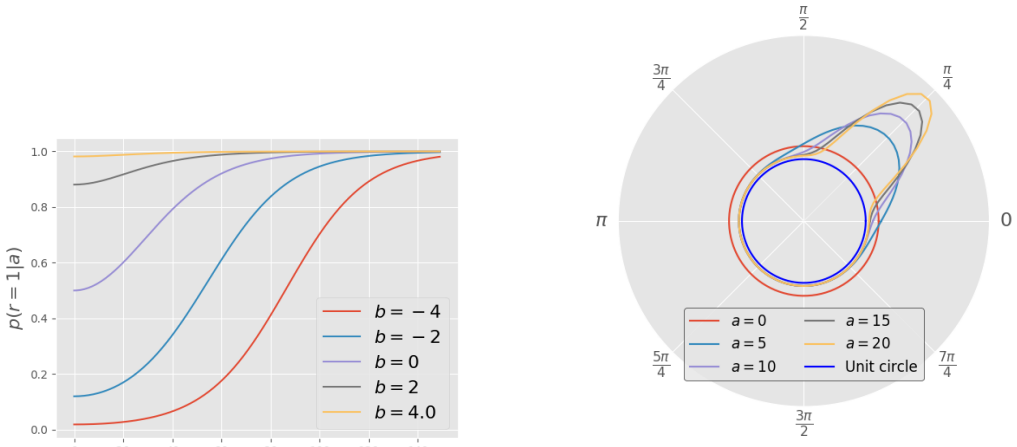
We can sample states of unit j in two steps. First, we sample from the rate marginal $p(r_j|x_j)$ with Bernoulli parameter

$$\begin{aligned} p(r_j|x_j) &= \int_{\theta_j=0}^{2\pi} p(r_j, \theta_j|x_j)\ d\theta_j \\ &= \frac{e^{b_j r_j} I_0(r_j a_j)}{(1 + e^{b_j} I_0(a_j))}. \end{aligned} \tag{4.14}$$

Then, conditioned on the r_j sample, we sample from θ_j:

$$\begin{aligned} p(\theta_j|r_j, x_j) &= \frac{p(r_j, \theta_j|x_j)}{\int_{\theta_j=0}^{2\pi} p(r_j, \theta_j|x_j)\ d\theta_j} \\ &= \frac{1}{2\pi I_0(r_j a_j)} e^{r_j a_j \cos(\theta_j - \alpha_j)} \end{aligned} \tag{4.15}$$

Note that $\theta_j|r_j$ is indeed von Mises distributed with central angle α_j and concentration parameter $r_j a_j$. In particular, θ_j is uniform on the circle when $r_j = 0$. The conditional



(a) **(b)**

Figure 4.8 *Marginal and conditional distributions of BKM units. (a)* Here we plot the likelihood that a given unit has state $r = 1$ given the magnitude of postsynaptic input, a, (the "concentration" parameter of the Bernoulli-von Mises density) for several values of the unit's bias, b. The likelihood of firing increases with a and the rate of increase is heightened with larger biases. *(b)* Next, we show the density on postsynaptic angle θ conditioned on the magnitude r. When $r = 0$, θ is uniform on the circle (although 0-magnitude units have no phase anyway). The mean angle, α, is the angle of integrated postsynaptic activity and its variance is controlled by a: larger a means less noise.

densities on hidden units are computed identically. Fig. (4.8a) shows how the 1 or "on" state is achieved only with strong input when biases become large. Fig. (4.8b) depicts the density on θ_j given $r_j = 1$ for various input magnitudes a_j. Observe that the mean angle for a unit is the angle of its post-synaptic activity. The variance on this angle decreases as the post-synaptic amplitude increases.

We now calculate the learning signal in Eq. (4.8)and observe that

$$\nabla_M E = -z_h^* z_v \tag{4.16}$$

$$\nabla_{b_v} E = -r_v \tag{4.17}$$

$$\nabla_{b_h} E = -r_h. \tag{4.18}$$

Substituting into Eq. (4.8) and denoting the difference in phase between unit j

and unit k by ϕ_{jk}, we find the magnitude and phase of the derivative of the BKM log-likelihood with respect to a specific weight $m_{jk} = c_{jk}e^{i\gamma_{jk}}$ to be

$$\left|\nabla_{m_{jk}} \mathbb{E}_{x\sim D}\left[\log p(x;\Lambda)\right]\right|^2 = \left(\langle r_j r_k \cos(\phi_{jk}-\gamma_{jk})\rangle_{\text{data}} - \langle r_j r_k \cos(\phi_{jk}-\gamma_{jk})\rangle_{\text{model}}\right)^2 + \left(\langle r_j r_k \sin(\phi_{jk}-\gamma_{jk})\rangle_{\text{data}} - \langle r_j r_k \sin(\phi_{jk}-\gamma_{jk})\rangle_{\text{model}}\right)^2 \tag{4.19}$$

$$\arg \nabla_{m_{jk}} \mathbb{E}_{x\sim D}\left[\log p(x;\Lambda)\right])] = \arctan\left(\frac{\langle r_j r_k \sin(\phi_{jk}-\gamma_{jk})\rangle_{\text{data}} - \langle r_j r_k \sin(\phi_{jk}-\gamma_{jk})\rangle_{\text{model}}}{\langle r_j r_k \cos(\phi_{jk}-\gamma_{jk})\rangle_{\text{data}} - \langle r_j r_k \cos(\phi_{jk}-\gamma_{jk})\rangle_{\text{model}}}\right) \tag{4.20}$$

Below, we will describe how to approximate this derivative either by using samples from the "data" and "model" distributions or with a variational mean field approach.

Even before laying out the details of these approximations, we can gain some insight into the learning rule by considering the first case of approximation by sampling. Suppose we have data-driven samples $z_v^{(d)} \sim D$ and $z_h^{(d)} \sim p(z_h|z_v)$ as well as $z_v^{(m)}, z_h^{(m)} \sim p(z_v, z_h)$ from the model's full just distribution. Then, writing the samples as complex exponentials, the derivative[5] at the weight m_{jk} is approximately

$$\bar{z_k^{(d)}} z_j^{(d)} - \bar{z_k^{(m)}} z_j^{(m)} = r_k^{(d)} r_j^{(d)} e^{i(\theta_k^{(d)} - \theta_j^{(d)})} - r_k^{(m)} r_j^{(m)} e^{i(\theta_k^{(m)} - \theta_j^{(m)})}$$

If units j and k are synchronized in both the data and model samples then the imaginary part of the derivative disappears and the update reduces to the regular Boltzmann machine learning rule whereby the weight is increased if $r_j^{(d)} r_k^{(d)} > r_j^{(m)} r_k^{(m)}$ and decreased otherwise. Intuitively, the learning rule is trying to increase the strength between data element j and latent unit k if empirical evidence from the data set indicates j and k are more correlated than expected by the model. The strength of the weight is decreased if they are less correlated than expected. If instead the magnitudes of the samples are all 1 but the samples are out of phase, the magnitude

[5]Note that we are using complex notation only for convenience and nowhere are we taking complex derivatives. For example, Eq. 4.11 is not holomorphic.

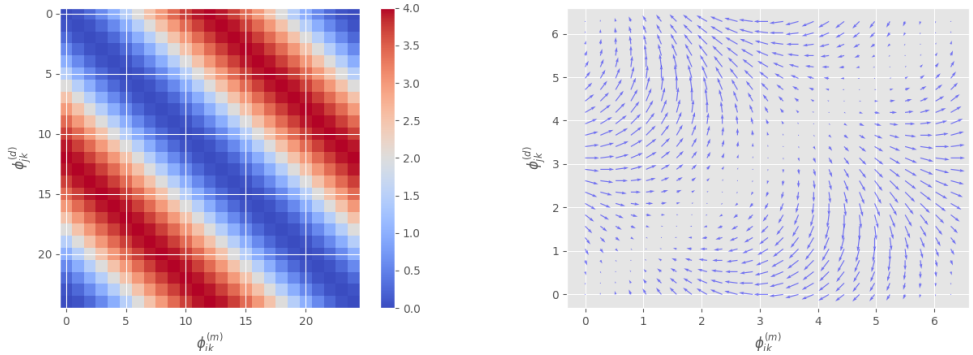


Figure 4.9 *Sampling approximations to the BKM gradient. (a)* Magnitude of weight derivative as approximated by samples from the data and model distributions.0 emph(b) Angle of weight derivative as approximated by samples from the data and model distributions.

of the derivative of the weight is approximately

$$2(1 - \cos(\phi_{jk}^{(d)} - \phi_{jk}^{(m)}))$$

which is minimal when the $\phi_{jk}^{(d)} = \phi_{jk}^{(m)}$ and maximal when $\phi_{jk}^{(d)} = \phi_{jk}^{(m)} + \pi$ (Fig. 4.9a). In other words, if the phase relationship between a data element and a latent factor differs depending on whether or not samples are driven by the data or the model, then the strength of the weight is increased. This underscores the fact that only *relative* phase relationships between the data and latent factors matter to the model. The likelihood of any sample is invariant up to a constant phase shift.

Finally, the argument of the weight derivative, assuming samples have magnitude 1, can be visualized as a two-dimensional vector field

$$\bar{z_k^{(d)}} z_j^{(d)} - \bar{z_k^{(m)}} z_j^{(m)} = e^{i(\theta_k^{(d)} - \theta_j^{(d)})} - e^{i(\theta_k^{(m)} - \theta_j^{(m)})} \quad = \left(\cos(\phi_{jk}^{(d)}) - \cos(\phi_{jk}^{(m)}), \sin(\phi_{jk}^{(d)}) - \sin(\phi_{jk}^{(m)})\right) \tag{4.21}$$

whereby the second equals sign we mean "is represented in $\mathbb{R}^2$ as". We plot this vector field in 4.9b. Again, when $\phi_{jk}^{(d)} = \phi_{jk}^{(m)}$, the argument of the derivative of 0.

4.2.2 Gaussian Kosterlitz Machines

Like hidden units in a Boltzmann machine, those in a BKM encode with their states the belief in the presence of a particular hypothesis about the data. The BKM goes one step further than a Boltzmann machine in also declaring that a hypothesis about the data can be true (i.e. $r_k = 1$) in any (continuous) number of qualitatively different ways ($z_k = e^{i\theta_k},\ \theta_k \in [0, 2\pi]$). We can extend this thinking yet again by replacing the binary truth values of model units with continuous variables encoding the graded confidence in the truth value of a particular hypothesis. To that end, we let $z = [z_v : z_h] \in \mathbb{C}^{n+m}$ such that $|z_j| = r_j \in \mathbb{R}$ for all j and define the energy

$$\begin{aligned} E(z_v, z_h; \Lambda) = \{W, b\}) &= \frac{1}{2}(z-b)^*(I-W)(z-b) \\ &= \mathrm{Re}\left((z_h - b_h)^* M^* (z_v - b_v)\right) + \qquad (4.22) \\ &\quad (z_h - b_h)^*(z_h - b_h) + (z_v - b_v)^*(z_v - b_v) \qquad (4.23) \end{aligned}$$

[Make note about variance] where W is block off-diagonal with blocks M, M^* as in Eq. (4.10) and $b = [b_v : b_h]$ is now a complex vector in $\mathbb{C}^{m+n}$. Note that the state space of each unit is now $\mathbb{C}$ instead of the semi-discrete space of BKM units. A Markov random field with this energy is distributed according to a multivariate complex normal distribution (Andersen et al., 1995) with mean b and precision $(I - W)$, and we therefore refer to this model as a *gaussian* Kosterlitz machine (GKM) and write $z \sim \mathbb{CN}_{m+n}(b, (I-W)^{-1})$.

In particular, each visible (correspondingly, hidden) unit conditioned on its input is a bivariate normal random variable in Cartesian coordinates with density,

$$p(z_j|x_j) = \frac{1}{2\pi} e^{-\frac{1}{2}\overline{(z_j - x_j)}(z_j - x_j)}, \qquad (4.24)$$

where $x_j = \sum_k m_{jk}(z_k - b_k) + b_j = a_j e^{i\alpha_j}$. In polar coordinates, this yields

$$p(r_j, \theta_j|x_j) = \frac{1}{2\pi} r_j e^{-(r_j^2 + a^2)/2 + r_j a_j \cos(\theta_j - \alpha_j)} \qquad (4.25)$$

As before, we are primarily interested in probability densities on r_j and $\theta_j|r_j$. Integrating Eq. (4.25) with respect to θ_j gives

$$p(r_j|x_j) = r_j e^{-(r_j^2+a_j^2)/2} I_0(r_j a_j), \tag{4.26}$$

which we recognize as the density of a Rice random variable. Probability densities of r_j for several values of the modulus of the input are depicted in Fig. 4.10. Rician distributions are used to model the phenomenon of multi-path interference in signal propagation, whereby a signal traveling to a receiver along several paths interferes with itself. We say that the signal is subject to "fading" when this interference is destructive and the magnitude of the signal at the receiver is attenuated. Under this interpretation, a receptive field in a GKM functions like a collection of paths by which a single phase message is to be sent to the post-synaptic unit. Disagreement among these channels functions like multi-path interference, subjecting the receiving unit to Rician fading (Rice, 1948).

The conditional density on phase is given by

$$\begin{aligned} p(\theta_j|r_j, x_j) &= \frac{p(r_j, \theta_j)}{p(r_j)} \\ &= \frac{\frac{1}{2\pi} r_j e^{-(r_j^2+a_j^2)/2 + r_j a_j \cos(\theta_j - \alpha_j)}}{r_j e^{-(r_j^2+a_j^2)/2} I_0(r_j a_j)} \\ &= \frac{e^{r_j a_j \cos(\theta_j - \alpha_j)}}{2\pi I_0(r_j a_j)}, \end{aligned}$$

so that $\theta_j|r_j$ is again von Mises distributed. We may calculate the learning signal for Eq. (4.8) by differentiating Eq. (4.22) with respect to M,

$$\nabla_M E = (z_h - b_h)^*(z_v - b_v) \tag{4.27}$$

and b_v, b_h,

$$\begin{aligned} \nabla_{b_v} E &= -M(z_h - b_h) - (z_v - b_v) \\ \nabla_{b_h} E &= -M^*(z_v - b_v) - (z_h - b_h). \end{aligned} \tag{4.28}$$

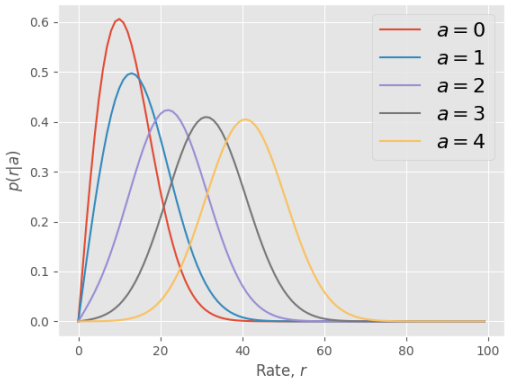


Figure 4.10 *Gaussian Kosterlitz Machines have Rician rates.* Here, we plot the density on a single GKM unit's magnitude, r, for several values of input magnitude, a. For the connection to Rician fading, see the main text.

As with the BKM, we approximate these gradients with the methods described in Sec. 4.2.4.

4.2.3 Mean field approximation and fast synapse dynamics

Sampling from an MRF in order to get approximate gradients like Eq. 4.2.4 or calculate the posterior in a multi-layer hierarchical system can be very slow. Instead of drawing a stochastic sample from the model, we can deterministically optimize a network configuration which is "close enough" to a true sample. If we constrain the optimization to be over distributions which are simpler than the true distribution, then this variational approach can not only accelerate inference and learning but also give us insight into model dynamics (Zemel, Williams, and Mozer, 1995; Salakhutdinov and Larochelle, 2010).

The variational approximation proceeds as follows. Suppose our model has a true density p which is hard to sample from. This typically arises because variables in the network interact in some nontrivial way that prevents the density from being factored. In the case of the Kosterlitz machine, this interaction comes from the term $\mathrm{Re}(\frac{1}{2}z^*Wz)$ which cannot be written as a sum of energies contributed by single phase variables.

Instead, we will choose some feasible set $\mathcal{C}$ of simpler distributions and optimize

$$\begin{aligned} q^*(z) &= \arg\min_{q\in\mathcal{C}} KL(q \parallel p) \\ &= \arg\min_{q\in\mathcal{C}} H(q,p) - H(q) \\ &= F_{MF} \end{aligned} \tag{4.29}$$

where $KL(\cdot \parallel \cdot)$ is the Kullback-Leibler divergence and $H(\cdot)$ and $H(\cdot,\cdot)$ are the entropy and cross entropy functionals respectively. In the context of this minimization problem, we refer to Eq. 4.29 as the *variational free energy*. If we specify that the feasible set for the optimization is the set of fully-factoring densities $\mathcal{C} = \{q \ : q(z) \geq 0, \int_z q(z) = 1, q(z) = \prod_j q_j(z_j)\}$ then we call q the *mean field approximation* to p. This name is inherited from the physics literature where an approximation involving non-interacting magnetic spins implicitly assumes spins are only influenced by the average (*mean*) behavior of their neighbors (Tong, 2012, Sec. 5.2.1).

For our purposes, however, this type of approximation is primarily interesting because of the insight it gives into the network's dynamics. In both the stochastic and deterministic settings, the model evolves to minimize its mean-field free energy, which we will calculate explicitly here for the case of the BKM. Our variational approximation will come from the set of fully factoring Bernoulli-von Mises distributions:

$$\mathcal{C} = \left\{ q(z) = \prod_j q_j(z_j) \ : \ q_j(z_j = r_j e^{i\theta_j}) = \frac{1}{1+I_0(a_j)} e^{a_j r_j \cos(\theta_j - \alpha_j)} \right\} \tag{4.30}$$

After some algebra and relying on the fact expectations over q factor nicely, we find that the variational mean-field free energy of a BKM is

$$F_{MF} = -\sum_{j,k} \frac{I_1(a_j)I_1(a_k)}{(1+I_0(a_j))(1+I_1(a_k))} \cos(\alpha_j - \alpha_k - \gamma_{jk}) + \frac{b_j I_0(a_j)}{1+I_0(a_j)} + \frac{b_k I_0(a_k)}{1+I_0(a_k)} + H(q_j(a_j,a_k)) \tag{4.31}$$

$$= -\sum_{j,k} f_1(a_j,a_k)\cos(\alpha_j - \alpha_k - \mu_{jk}) + f_2(a_j) + f_2(a_k),$$

where I_0 and I_1 are modified Bessel functions of degree 0 and 1 and where we have collected functions of mean field magnitudes a_j and a_k into multiplicative and additive components f_1 and f_2. Note that $0 \leq f_1 \leq 1$. Minimizing F_{MF} gives rise to the gradient system

$$\frac{\partial F_{MF}}{\partial \alpha_i} = \sum_k f_1(a_i, a_k) \sin(\alpha_i - \alpha_k - \gamma_{ik}) \tag{4.32}$$

$$\frac{\partial F_{MF}}{\partial a_i} = \sum_k \frac{\partial f_1(a_i, a_k)}{\partial a_i} \cos(\alpha_i - \alpha_k - \gamma_{ik}) + \frac{\partial f_2(a_i; b_i)}{\partial a_i}. \tag{4.33}$$

We recognize Eq. 4.32 as a form of the Kuramoto model with 0 intrinsic frequencies and random phase shifts arising from the angles of weights γ_{jk} (similar to Kundu et al., 2017). Most importantly, the Kuramoto flow has dynamical couplings in the form of $f_1(a_i, a_k)$ and resembling those of Picallo and Riecke, 2011; Timms and English, 2014; Ha, Noh, and Park, 2016; Ha, Lee, and Li, 2018. The dynamics of the couplings are given by Eq. 4.33.

We can better understand the relationship between synchrony in this system and the dynamical couplings by plotting the f_1-dependent part and f_2-dependent part of Eq. 4.33 separately. In Fig. 4.11a, the f_1-dependent part of the coupling dynamics is plotted as a function of the current strength of a mean field magnitude a_i and the current difference in phase between unit i and unit k. Two phenomena are apparent. Most notably, the function is minimized when units i and k are π out of phase. In this case, $\frac{\partial F_{MF}}{\partial a_i}$ is negative so that the coupling between i and k is weakened. In other words, units gradually decouple to the degree that those units tend to be out of phase. Conversely, the coupling is strengthened to the degree that units are in phase. These coupling dynamics are dampened as a_i grows large. Recall that a_i is inversely related to the circular variance of a Bernoulli-von Mises distribution. Hence, if q_i has low noise, then the dynamical coupling a_i remains largely constant; if q_i is noisy, then it is either attracted to or repulsed by unit k according to their phase difference. The f_2-dependent part of Eq. 4.33 has a similar effect (Fig. 4.11b). If i has a negative

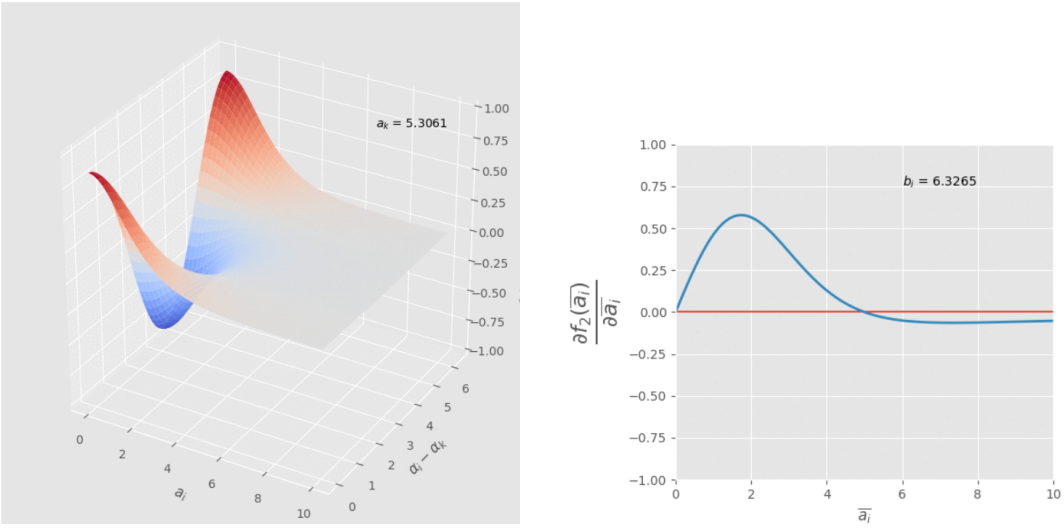


Figure 4.11 *Fast synapse dynamics in BKM evolution. (a)* Here, we plot the f_1-dependent part of the mean-field dynamics in Eq. 4.33. The value of $y_{ik} = \frac{\partial f_1(a_i,a_k)}{\partial a_i} \cos(\alpha_i - \alpha_k)$ is plotted as a function of a_i and $\alpha_i - \alpha_k$. As α_i and α_k approach π out of phase, y_{ik} becomes increasingly negative, decreasing a_i and thereby decoupling the phases of units i and k according to Eq. 4.32. As the phases synchronize, however, the opposite obtains, and the functional connection strengthens. The magnitude of either effect decays as a function of both a_i and a_k. *(b)* Next, we show the contribution to fast synapse dynamics from the f_2-dependent portion of Eq. 4.33 for $b_i \approx 6.3$. We see that the value of $\frac{\partial f_2(a_i;b_i)}{\partial a_i}$ decreases as a function of a_i, eventually becoming negative and serving to decouple the units as in the previous panel. As b_i decreases the initial hump flattens and eventually becomes concave up.

bias, the unit is decoupled from its neighbors; if it has a positive bias, it is attracted to its neighbors; and, in either case, the coupling is kept constant to the degree that i has low variance. In sum, the observation of Reichert and Serre, 2013 that units tend to gradually decouple into independent sub-groups was well-founded, and the mean-field dynamics helps explain why.

This mean-field approach casts new light on the model neuron used in Kosterlitz Machines and their predecessors in Reichert and Serre, 2013. Recall that the latter authors likened unit magnitudes to rates in the traditional McCulloch-Pitts neuron and unit phases to the timings of inhibitory envelopes for use in gating. Now, we see that unit magnitudes have yet another interpretation as dynamic or "fast" synapses.

Fast synapses have been the subject of numerous biological (Fiebig, Herman, and Lansner, 2020) and theoretical (Picallo and Riecke, 2011) investigations, including in the context of energy-based models (Bienenstock et al., 1987) like our current case. Dynamical synapses also played a famous role in Malsburg, 1994's correlation theory of brain function, in which they served to gradually decouple desynchronous neural assemblies, like our present case.

4.2.4 Approximating the gradient of log-likelihood

As we mentioned earlier, calculation of the exact log-likelihood gradient of a KM is typically intractable since the "model" expectation in Eq. is taken over all possible configurations of the network. It is typical in the energy-based model literature to approximate these gradients with sample statistics or variational approximations. The exact mix of stochastic and deterministic approximations to the gradient is more an art than a science, so we have simply adapted the one popular method for training deep Boltzmann machines (Salakhutdinov and Larochelle, 2010) to the particular case of KMs.

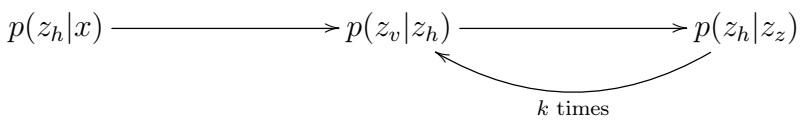
Recall that the derivative of the log likelihood of a KM with respect to an arbitrary parameter λ_j is

$$\frac{\partial \log p(z_v;\Lambda)}{\partial \lambda_j} = \left\langle -\frac{\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j} \right\rangle_{\text{data}} - \left\langle -\frac{\partial E(z_v,z_h;\Lambda)}{\partial \lambda_j} \right\rangle_{\text{model}}$$

The derivation is at Eq. 4.2.4. When training a deep model, we will replace the "data" expectation with statistics taken from an approximate mean-field posterior and we will replace the "model" expectation with statistics taken from a $k-$length Gibbs sampling chain. Variational inference for the first part of the gradient is typically more efficient, however, since we expect the posterior to be unimodal when the model is driven by the data so that the mean-field free energy to be minimized quickly. When approximating the full joint distribution, unimodality is less assured, and we resort to

sampling instead.

For a two-layer model, it suffices to use Gibbs samples to approximate both parts of the gradient; i.e. it suffices to use CD-K. Let $(z_h^* z_v)^{(k)}$ be the k-th sample of $z_h^* z_v$ from a Gibbs sampling chain (starting from a data point $x \sim D$) over the conditionals of $p(z_v, z_h)$. Since all z_j are conditionally independent given z_h and all z_k are conditionally independent given z_v, this k-th approximation can be acquired by block sampling the BKM layers:



Then, the CD-k approximation to the weight gradient of the log-likelihood of a BKM is

$$\nabla_M \log p(z_v; \Lambda) = (z_h^* x)^{(1)} - (z_h^* z_v)^{(k)}. \tag{4.34}$$

Depending on the application, the first term can be replaced by its mean value $\mathbb{E}_{p(z_h|z_v)}[-z_h^* x]$ since getting an unbiased estimate of this statistic is easy. Approximations to the bias gradients are made similarly. Often, setting $k = 1$ is sufficient for simple tasks.

For deeper models, we will acquire data-dependent statistics by running mean-field updates until convergence. Model-dependent statistics will still be acquired from K-step contrastive divergence.

4.3 Experiments

Historically, energy-based models have been used in at least three important applications. The first is of course the modeling of a target distribution with the intention of drawing novel samples from the distribution and understanding its latent structure as manifest in the model's learned features. The second is data completion, whereby a

noisy or otherwise incomplete sample from the true distribution is used as the starting point for a Gibbs sampling chain that travels to a nearby "clean" or "complete" sample which is considered less of a statistical outlier than the noisy input (Geman and Geman, 1984). A final application is to use the features learned by the model as unsupervised pretraining for some downstream process, often classification (Hinton and Salakhutdinov, 2006).

In this section, we will demonstrate all three of these applications using a KM, emphasizing in the latter two demonstrations the relevance to phase-coding. Throughout, we will focus on the BKM, and all data sets have been recreated from Reichert and Serre, 2013. First, we will show how a BKM can be made to fit data sets of simple scenes. Progress in learning will be measured by approximating the log partition function with annealed importance sampling (AIS) (Neal, 2001). The increased expressivity of a deep BKM will then be demonstrated on a slightly more sophisticated data set. Next, we will show how the BKM's "denoising" ability can be used to transform an image with random phases into one in which the distribution of phases come to respect to the statistical structure observed in the training data. That is, the images will come to "synchronize" into perceptually meaningful groups. In this way, we will have transformed the "binding by synchrony" found by Reichert and Serre, 2013 into bona fide conditional inference. Finally, we will show how a KM can be used to model "communication through coherence" in the sense of Fries, 2005, whereby a downstream population selects a single neural assembly from complex overlapping populations by "entraining" to that assembly's phase. In our model, this selection very naturally takes the form of Bayesian inference.

4.3.1 Fitting a shallow BKM

First, we measured the ability of a two-layer BKM to fit a simple 32×32 images of crossing bars. Six data sets were created according to whether or not there were

1-6 crossing vertical and horizontal bars in an image (Fig. 4.12). These data are an appropriate test bed for the system since their generative model is easy to specify. For each horizontal bar, a uniform random phase is chosen. Then, that bar is placed at a given column in the image that is not occupied by another bar. For any crossing with another bar, a Bernoulli random variable determines which bar will be "on top". This simple generative model gives us a rough idea of how many latent units we need to express the data exactly. Each data set contained 20000 images split evenly into training and testing sets. We used a BKM with 1024 visible units and 1024 hidden

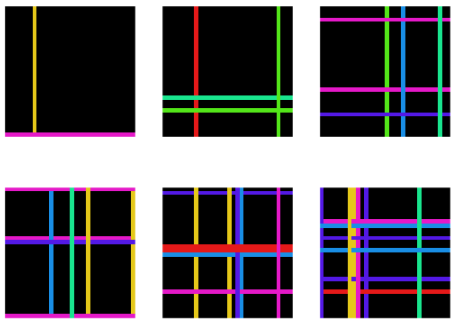
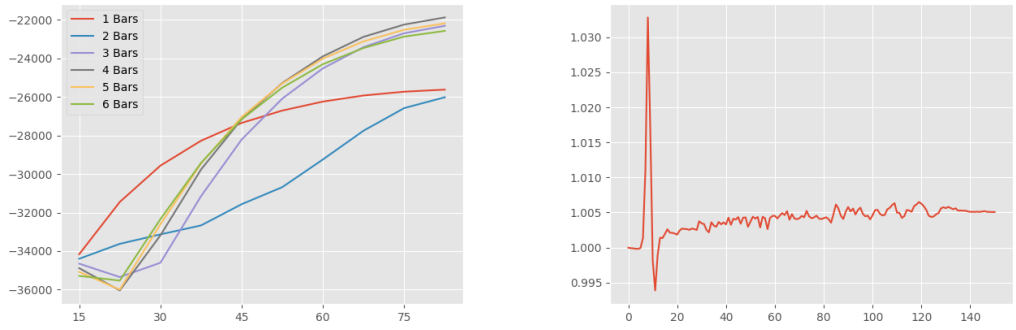


Figure 4.12 *A simple data set of crossing bars.* Images are generated by placing between 1 and 6 vertical and horizontal crossing bars. Each has a random phase and crossings belong to one or the other bar with equal probability.

units. Training lasted for 150 epochs and the data was split into mini-batches of size 64. We used persistent (Tieleman, 2008) CD-5 and a learning rate of .1.

Monitoring the progress of learning in energy-based models can be difficult. Although the goal of learning is to increase the log-likelihood of samples in the training set, actually measuring this log-likelihood is non-trivial since it technically requires calculating the partition function. The partition function can be estimated by AIS, but this is a costly process which involves getting good Gibbs samples from the model numerous times over the course of a slow annealing schedule. Therefore, we only estimate the log-likelihood every 15 epochs. We will supplement this direct



(a) **(b)**

Figure 4.13 *Monitoring learning in a BKM.. (a)* Log likelihood estimates plotted for the training data for each data set. Estimations were calculated by approximating the partition function with annealed importance sampling (Neal, 2001) and using this to normalize the marginal density on visible units given in Eq. 4.12 and evaluated on training data. *(b)* In parallel, the likelihood ratio of training data to testing data was computed. Note that the partition function cancels here so it can be calculated without numerical tricks. A value close to one indicates that the model is not overfitting. Figure here depicts the case of 4 bars.

measurement of learning with several indirect methods: plotting learned features, monitoring overfitting by calculating a likelihood ratio involve training and test sets and observing samples from the trained model.

Fig. 4.13a shows the AIS-estimated log-likelihood of the data over the course of training for each of the six training data sets. Calculating the ratio of log-likelihoods of a random training and testing batch over the course of training indicates that the model does not substantially overfit (4-bar case depicted in Fig. 4.13b). In short, the model has begun to approximate a density on the data.

Fig. 4.14 plots learned 25 random features for the 2-bar data set. We see immediately that the network has learned a sensible dictionary of features typically consisting of a few crossing bars to explain the data. Notice that the weights have all evolved to be real-valued. We can see this directly from the histogram of weight values (Fig.4.15). This is a sensible model of the data since a feature need only encode

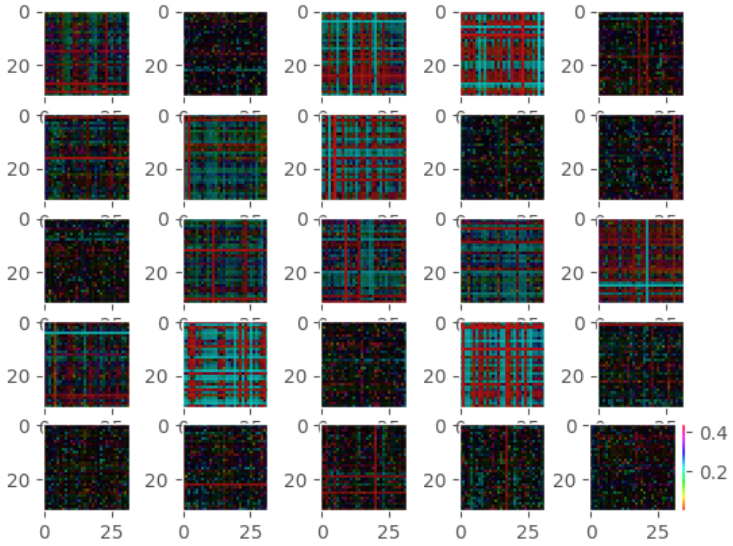
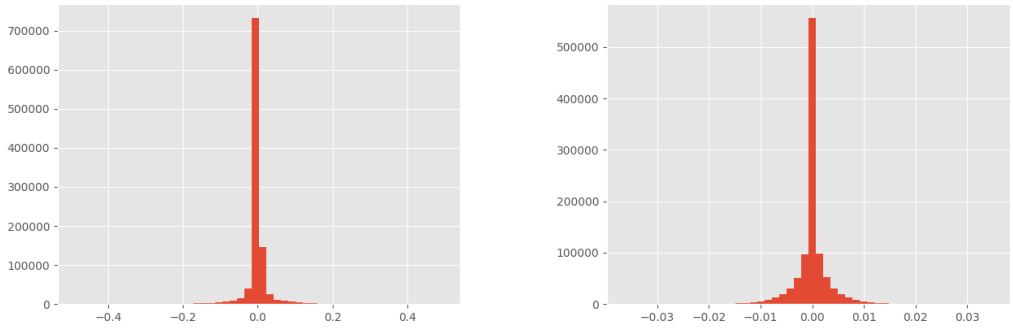


Figure 4.14 *Learned bar features.* 25 out of 1024 features are displayed here. Notice that features are overwhelmingly real-valued. Weights were penalized by an L_1 penalty during learning to encourage interpretable features. In some cases, features consisted of a single bar (e.g. row two, column one) and in other cases were a mix over overlapping bars (e.g. row one, column three).

the location of a bar and the phase of that bar can be decided by the phase of the hidden unit associated with that feature. Fig. 4.16 shows several samples from the 4-bar model acquired by running a Gibbs sampling chain for 1000 steps while the network is annealed.

4.3.2 Fitting a deep BKM

It is now well-known that the addition of extra layers to a Boltzmann machine results in a strictly better model of the data (Roux and Bengio, 2017; Salakhutdinov and Larochelle, 2010). Deep BKMs are no exception, and we will demonstrate this fact on a slightly more complicated data set involving 20×20 images of three overlapping squares and triangles (Fig. 4.17). Unlike the case of the bar images, where overlap among image components was minimal, these shape images cannot be easily described



(a) **(b)**

Figure 4.15 *BKM weights for bar images are nearly real-valued.* Plotted are the histograms of real *(a)* and imaginary *(b)* parts of the weights (note the scale of the x axis). For a possible explanation, see the main text.

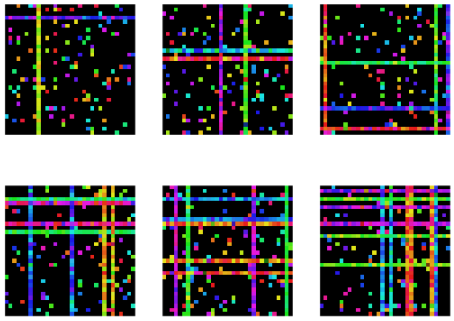


Figure 4.16 *Samples from a trained 2-layer BKM.*

as a linear superposition of features. We should expect that an additional layer of latent factors will help disambiguate the numerous overlaps in this new data set. For this demonstration, we trained both a two-layer BKM and a three-layer BKM and compared their performance over the course of training. The first two layers of the three-layer architecture were identical to the simpler two-layer system, as were all of the other learning parameters, so any improvement resulted from the additional layer. In both cases, the first two layers contained 400 and 1024 units respectively. The deep model had an additional layer of 512 units and was trained by greedy layerwise pretraining followed by joint training of all units (for details, see Salakhutdinov and

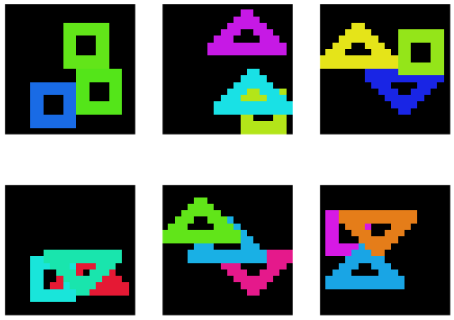


Figure 4.17 *Overlapping shapes for a deep BKM.* This data set was recreated from Reichert and Serre, 2013 and consists of images with three overlapping shapes (squares and triangles, the latter of which can be flipped upside down). The data set is significantly harder to learn than the bars data set because of the substantial, fundamentally ambiguous overlap among shapes.

Larochelle, 2010).

As before, we approximated log-likelihood during training using AIS for both models (Fig. 4.18). Observe that the log-likelihood of the deep model is significantly higher than the shallow counterpart. Since it has the pre-trained weights of the 2-layer system, it begins cross-layer training with quite good likelihood. The addition of the third layer evidently leads to a much better model of the data. Samples from the shallow vs the deep model are plotted in Fig. 4.19.

4.3.3 Perceptual grouping as conditional inference

One limitation of the model of Reichert and Serre, 2013 is that its ability to spontaneously group image features into perceptually meaningful segments had no tangible statistical interpretation. Consequently, this ability could not be encouraged by learning and remained somewhat unstable and inconsistent. Our current case is rather different, since the model has been explicitly tasked with learning a joint distribution on modulus and phase. Conditioned on one, we should expect the model to "paint in" the other, as in the case of other energy-based models (Lu, 2018).

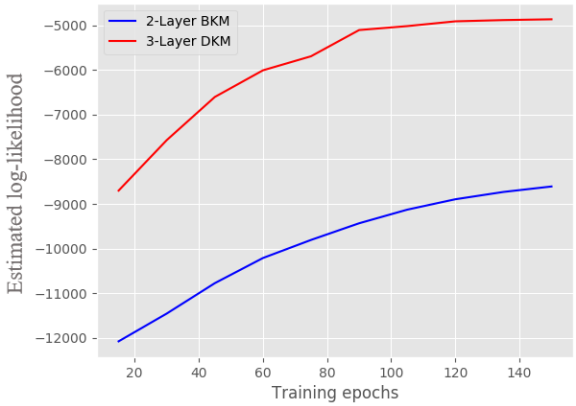


Figure 4.18 *Monitoring learning of a deep BKM.* As before, we plotted the approximate log-likelihood of data during training, now for both a shallow and deep model. The deep model was trained by first training each pair of layers on their own and then training the joint system. The 3-layer network achieves much better log-likelihood than the shallow system.

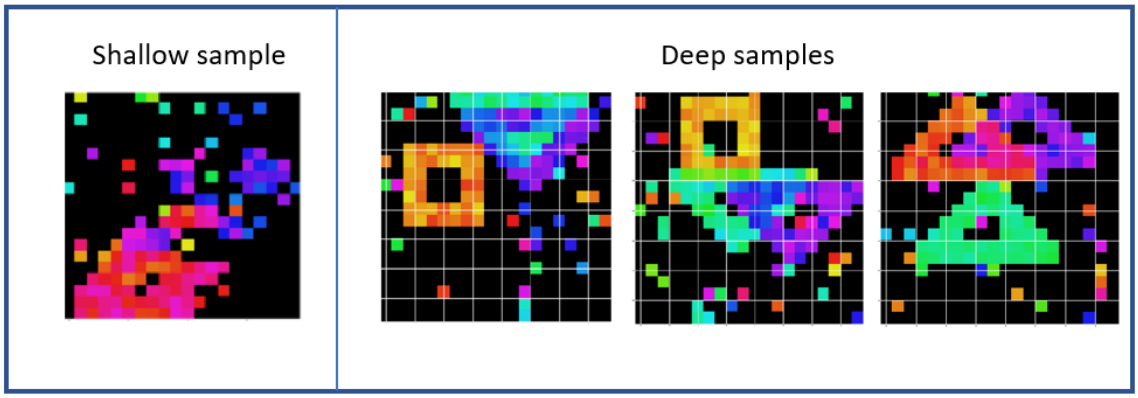


Figure 4.19 *Samples from a DKM.* The DKM produces much more faithful samples of the shape data set (right three images) than the shallow model (left image).

We can observe this directly by conditioning the model from Sec. 4.3.1 trained on 4-bar images on the moduli of a given sample and then sampling the rest of the model configuration as temperature is cooled according to an ad hoc annealing schedule. Specifically, we take a given sample and force the magnitudes of the units in the visible layer to assume the magnitudes of that sample and while stochastically updating both the visible phases and the phases and moduli of the hidden layer over the course of a 1000-step Gibbs sampling chain. Temperature is decreased linearly from 100 to 0

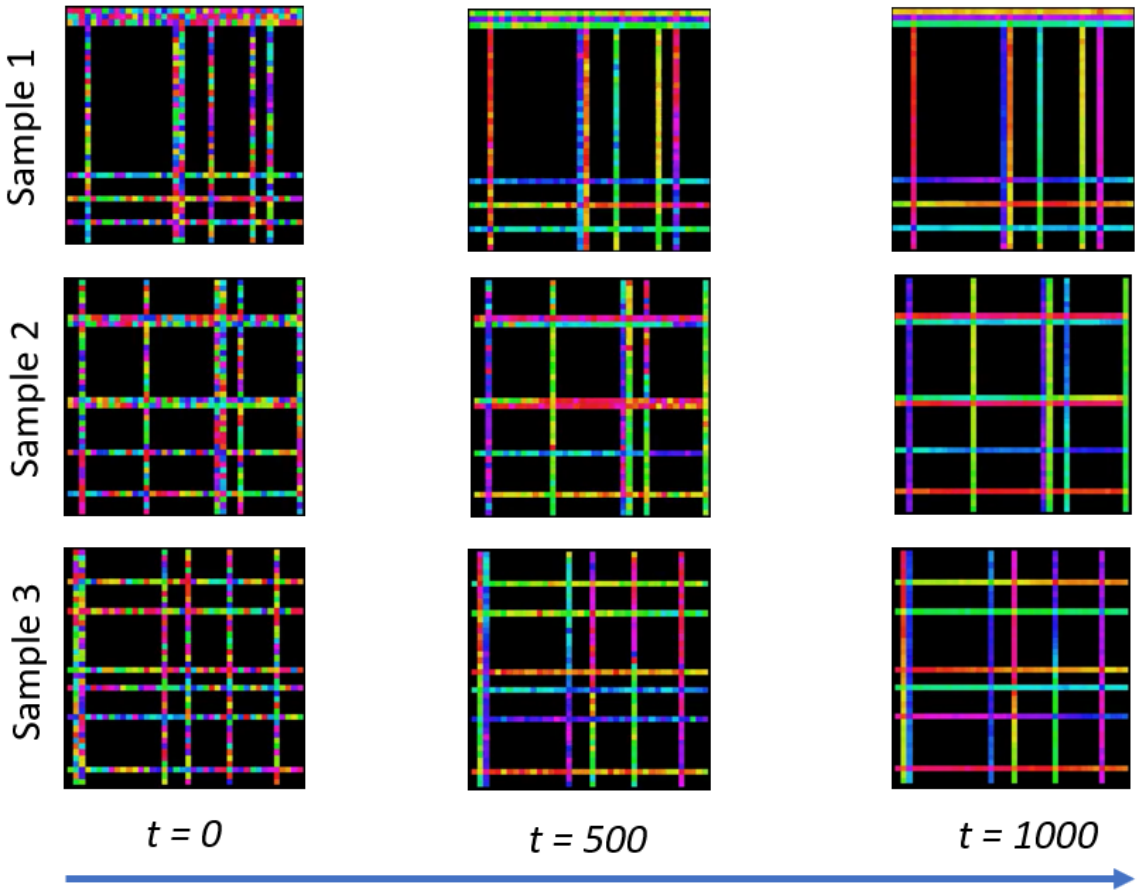


Figure 4.20 *Perceptual grouping as conditional inference.* A Gibbs sampling chain is run conditioned on the magnitudes of hidden units being fixed to values representing the luminances of a given image. As the system is annealed, the initially random input phases give way to structure learned from the data. This conceptually similar to binding-by-synchrony, although the KM framework ensures that the image at convergence is a maximum a posteriori estimate under the learned density.

in order to approximate a maximum a posteriori estimate of the phases of the given data sample at the end of the chain. The stages of this annealing process are depicted in Fig. 4.20. In the left column, we see the clamped images with random phases used to initialize the sampling chain. By step 500, the image has clustered into coherent segments and by step 1000 these segments extend along the whole of each bar. Notice that bar intersections are often ambiguous. In some cases, the model "commits" to a sensible interpretation of this crossing (e.g. the green bar crossing the red bar at the

upper left of the final image in the first column) and in other cases the phase at the crossing is simply the mean of the two crossing bars. This is likely a limitation of a two-layer generative model which can only account for occlusions by memorizing the two possible orders of any two bars at any two locations (as opposed to deciding the order by another step in the generative process; i.e. with another layer.)

When the generative model of the data is well-approximated by a network with real weights (as in the case of the bar images), a notable benefit of this manner perceptual grouping is its rotational symmetry. It is easy to demonstrate that the likelihood of a given datum under the visible marginal is invariant up to a constant phase shift. This differentiates the "grouping" abilities of this model from the "segmentation" abilities of similar models, since the latter case typically involves precise segment labels as opposed to the relationships among labels. In other words, the model proceeds by recognizing familiar objects and then separating them into coherent clusters without regard to the precise value of phase used for each cluster.

One obvious flaw in the conception of conditional inference over phase as a form of perceptual grouping is that the model must be trained on *pre-grouped* images. We address this flaw in Ch. 4 (at the cost of introducing a supervision signal), but for now we can simply imagine that the veridical grouping signal contained in the data comes from some source not made explicit here. For instance, perhaps the true phase information comes in the form of optic flow which transforms the image into a 2D vector field. Indeed, it is not uncommon to train complex-valued neural networks using optic flow or other image gradients (Guberman, 2016). As we will discuss in the final section of this chapter, future iterations of this model will make this interpretation of the data explicit, perhaps by training on moving images.

4.3.4 "Communication by coherence" with a phase prior

Perceptual groups are only valuable inasmuch as they can be used to easily select features being to this or that object in a computationally useful way. Fries, 2005's "communication by coherence" (CTC) theory, which we reviewed in Sec. 3.2.2, is one such group-selection mechanism. According to Fries, a single neural population can efficiently represent multiple objects simultaneously by segregating their representations in *time* rather than across anatomical space. If object A is represented by neurons spiking at time θ_A with respect to some macroscopic oscillation (for example in LFP) and object B is similarity represented at time θ_B, then a downstream population can selectively attend to object A by entraining to the afferent population at phase θ_A (see Sec. 3.2.2 for details).

We can simulate CTC in a Kosterlitz machine, and the resulting process has the flavor of Bayesian inference. For example, imagine we perceptually group a stimulus in the manner described in the previous section. After annealing the model to 0 temperature, the final layer approximates a maximum a posteriori estimate on image $z_h^* = \arg\max_{z_h} p(z_h|z_v)$. In particular, $\theta_h^* = \arg(z_h^*)$ is the maximum a posteriori estimate on latent phases. After this process as terminated, we imagine a downstream population imposes by some means a hard prior distribution

$$p(z_k = r_k e^{i\theta_k}) \propto \delta(r_k^* e^{i\theta'}) \tag{4.35}$$

on latent units encoding what phase θ' the downstream population wants to attend to. Flipping the model on its head, we can now conceive of the posterior $p(z_h|z_v)$ as a likelihood and compute the visible "posterior"

$$p(z_v|z_h) \propto p(z_h|z_v)p(z_h) \tag{4.36}$$

which can be easily calculated by zeroing-out all latent units not having phase θ' and then performing a single backwards pass onto the visible layer.

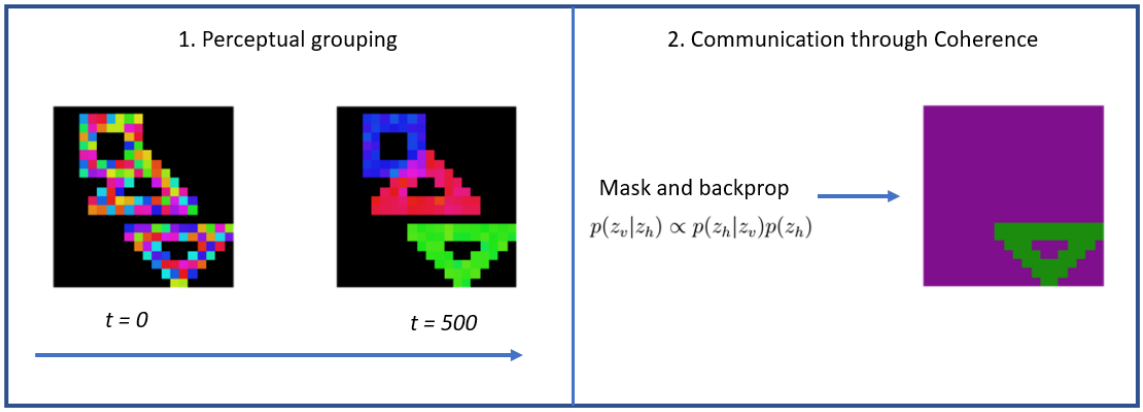


Figure 4.21 *Communication by coherence in a Kosterlitz Machine.* First, the perceptual grouping procedure from the previous section is run. After convergence, we imagine a downstream population multiplies the posterior estimate on the data encoded in the hidden layer with a prior which masks all but those units with a given phase. We then calculate the mean activities of the visible layer conditioned on this masked posterior. The result is that all shapes not represented in the hidden layer by the target phase are eliminated, allowing the selected shape to be routed to the downstream population.

We carry out this process in Fig. 4.21, using the case of highly overlapping objects and the deep model trained in Sec. 4.3.2 of this chapter. We see that this sort of "Bayesian CTC", can be used to select particular neural assemblies at the data layer. We could presumably add a third step to this algorithm in which the attended object is then routed to the downstream population for this or that purpose (e.g. storage in working memory). A similar method was described in Reichert and Serre, 2013, but, again, since the model had no known joint distribution, an interpretation of the attentional process as Bayesian inference was not possible.

4.4 Future work

In this chapter, we exploited the equilibrium nature of oscillatory systems lacking intrinsic activity to develop a hierarchical probabilistic graphical model on phase data. Through several simple simulations, we showed how shallow and deep versions of this system could be used to learn a density on phase data, how the trained model gave

rise to a sort of binding by synchrony, and how the system could hypothetically be joined with other modules to carry out a CTC attentional selection. Much of what we outlined in this chapter was simply a rigorization of Reichert and Serre, 2013, and we noted several instances where our statistical reformulation shed insight on model dynamics, including the emergence of fast synapse dynamics. Though early in development, the framework presented here could be scaled to richer data without terrible difficulty.

Before being scaled to more interesting problems, a careful examination of KMs' practical and theoretical limitations is warranted. First among practical limitations is the difficulty of learning, largely inherited from the Boltzmann machine literature although common to many generative models. For instance, we saw earlier how even the simple monitoring of learning requires costly numerical methods like annealed importance sampling. Acquiring gradients for learning also requires numerical approximations which are rather slow. Should the Kosterlitz Machine be pursued as an object of study in the future, these learning techniques should be modernized, perhaps by exploiting Hamiltonian MCMC (Neal, 2001) for sampling or score matching (Hyvarinen, 2005) for learning. We have no particular attachment to the Boltzmann Machine framework either: other generative modeling approaches, like adversarial training (Goodfellow et al., 2014) or Bayesian compressed sensing (Ji and Carin, 2010), could also be adapted to the case of phase-coding units.

The clearest area for future work, however, is theoretical. Currently, the Kosterlitz machine is designed to fit a density on 2D vector fields and so the system must be fed these "pre-grouped" images during training. Yet, this is a far cry from the intuitions of Malsburg, 1994, Fries, 2005 and others, who conceived of phase as a purely representational mechanism not inherent in the data itself. A system which could learn to "bind by synchrony" after having been trained only on typical real-valued images would be much closer to this original spirit. It is difficult to imagine, however, how this

would work for a generative model like the KM. How, for example, would phase play a role in generative model of images themselves lacking phase? One could imagine a generative model in which phase is ultimately killed by a pointwise modulus, but this trick calls into question the role of phase in the first place. In light of this conceptual difficulty, it is likely any future role for the KM as a model of perceptual processing will be at the post-grouping level, for instance, as a module for "filling in" the phase left out by prior mechanisms (as in Sec. 4.3.3) and switching between grouped assemblies (see 4.3.4) which are themselves the output of earlier processes which carry out bona fide phase-based grouping on traditional luminance-coded images.

In such a combined system, the statistical framework of the KM could presumably still be brought to bear in order to understand this full processing pipeline. This naturally leads us to the question of what type of model would fill the role of the earlier module, that using phase to group luminance-coded images. We provided one answer to this question in the next chapter.

Chapter Five

Kura-Net

Non-equilibrium[1] oscillatory systems for perceptual modeling are rather more numerous than their equilibrium counterparts, owing to the generally non-equilibrium nature of model neurons prevalent in computational neuroscience. Though it is rarely made explicit by modelers, one way to summarize the theoretical benefit of non-equilibrium systems is that they have the ability adapt without descending into chaos. This intuition is made more rigorous in the study of complex systems where a distinction is made between systems with rigid order (e.g. a ferromagnet) and chaotic systems. Systems far from equilibrium, like spin glasses, in some sense straddle the boundary between rigidity and order (Langton, 1990; Kauffman, 1992; Mitchell, Hraber, and Crutchfield, 1993), and are therefore considered a good starting point for the modeling of biological systems where regularity is just as important as the ability to adapt and change (see Stein and Newman, 2013 for a good review).

In this chapter, we will develop a differentiable learning procedure for non-equilibrium oscillatory systems. The result will be a neural network which we call "Kura-Net". Our principal application will continue to be perceptual grouping of image segments. Our focus will remain be the Kuramoto model, though we will now allow for non-constant intrinsic frequencies. This is at once a more biophysically

[1]Material from this chapter was presented as Ricci et al., 2020 and is the subject of a manuscript in publication with co-authors Yuwei Zhang, Minju Jung, Mathieu Chalvidal and Aneri Soni.

realistic scenario and a more difficult one from a learning perspective. On the one hand, biological neurons are governed by all sorts of intrinsic electrochemical processes (Llinás, 2014) and these processes may very well play a role in computation. On the other, the distribution on these intrinsically active neurons is hard to characterize at a macroscopic level, so we must go without the simple statistical interpretations of the system we enjoyed in the previous chapter. What's more, the types of synchrony speculated to underlie binding and selective attention will be harder to achieve in this regime: heterogeneous intrinsic activities are difficult to homogenize.

Luckily, we have modern machine learning at our disposal. These techniques will allow us to backpropagate directly through the flow of the Kuramoto model and thereby estimate the dependence of synchronous behavior on model parameters. Learning in this system differs significantly from the typical training scenario of neural networks, since we will not learn Kuramoto couplings/intrinsic frequencies themselves but rather a parametrized mapping from stimuli to Kuramoto couplings/intrinsic frequencies. The actual locus of learning is this parameterized mapping, and the Kuramoto parameters are merely outputs of this mapping. This makes learning in Kura-Net closer in spirit to optimal control: couplings/intrinsic frequencies are parameterized control variables which are optimized to guide the flow of phases.

The resulting model brings us closer to a goal stipulated in the previous chapter: phase-based grouping directly on natural images. The Kosterlitz Machine, the reader will recall, must be trained on pre-grouped images since it is difficult to construe phase as a bona fide generative variable. Kura-Net circumvents this problem essentially by learning a collection of synapses *per image*. These synaptic weights then link up oscillators associated to every image pixel which then participate in a Kuramoto dynamics. In a sense, these per-image weights are a continuation on the theme of fast synapses. The learned function undergirding Kura-Net is then a stand-in for whatever mysterious sub-synaptic biophysics transforms an impinging stimulus into an effective

conductance.

The chapter will proceed as follows. First, we will review the existing literature on non-equilibrium oscillator systems in image processing. Most of these systems use hardwired dynamical parameters, and although recent work by Meier, Haschke, and Ritter, 2014 and Finger and König, 2014 introduces a sort of learning, it does not depend on oscillatory dynamics whatsoever. We will then introduce Kura-Net more formally, noting its similarities to so-called "hypernetworks (Ha, Dai, and Le, 2014) and the interpretation of its learning algorithm as a form of optimal control. Finally, we will demonstrate Kura-Net in three experiments. The first is a recapitulation of earlier work on the synchronizability of complex graphs (Brede, 2008a) in which we learn a parametrized mapping from intrinsic frequencies to those couplings which will synchronize them. The second will reveal some qualitative similarities between Kura-Net and work by Brede, 2008a and others reviewed above. The third will demonstrate a quantitative advantage of Kura-Net over a competing CNN on a sort of same-different task. This experiment follows up on intuition established earlier in this thesis and elsewhere by Fries, 2005, McLelland and VanRullen, 2016 that phase-grouping can overcome systematicity deficits in CNNs during visual reasoning (Alamia et al., 2020).

5.1 Non-equilibrium oscillatory systems in image processing

Since the early speculations of Malsburg and Schneider, 1986, it has been generally recognized that the benefit of oscillatory systems for perceptual and particularly visual processing has been their ability to spontaneously partition into several synchronized groups. Synchrony is typically measured by coherence in a phase variable, though some researchers have investigated frequency-based grouping as well (Meier, Haschke,

and Ritter, 2014). The main computational problem in this line of work has been the understanding of network conditions which give rise to perceptually meaningful synchronous groupings, tying this literature to the research on multi-polar synchrony in oscillatory systems discussed in Sec. 3.3.1. These synchronous groups are interpreted as being "tagged" by their uniform phase/frequency so that constituent features of the group are bound into a coherent whole which can be selected en masse by downstream attentional and mnemonic processes in the manner of Fries, 2005 and McLelland and VanRullen, 2016. A hitherto secondary concern, though indeed the one that concerns us most, is how to make these systems learn. What few trainable systems of coupled oscillators exist typically involve ad hoc perceptual features acquired independent of oscillatory dynamics, and the problem of generalizing learned oscillatory grouping ability to new data is rarely considered.

Work on phase-based grouping in systems of coupled oscillators begins in earnest with (Malsburg and Schneider, 1986) who were also the first to hypothesize a general principle about network conditions giving rise to multiple synchronous groups. The authors showed that networks of spiking neurons (Fig. 5.1a) representing perceptual features (their application was auditory but the principle is modality independent) would gradually separate into synchronized groups when the units were connected by local excitatory and global inhibitory couplings. Units which were simultaneously activated, presumably by a single component (e.g. word, object, etc.) of the input stimulus, tended to mutually excite one another and gradually phase-locked at a common spiking frequency. The total activity at any time is limited by global inhibition so that synchronous groups representing different input components would burst sequentially at non-overlapping intervals (Fig. 5.1b). These groups would gradually decouple from one another according to fast synapses dynamics, which we will recall form a pillar of von der Malsburg's "correlation theory" Malsburg, 1994.

The notion of Malsburg and Schneider, 1986 that perceptual features only generate

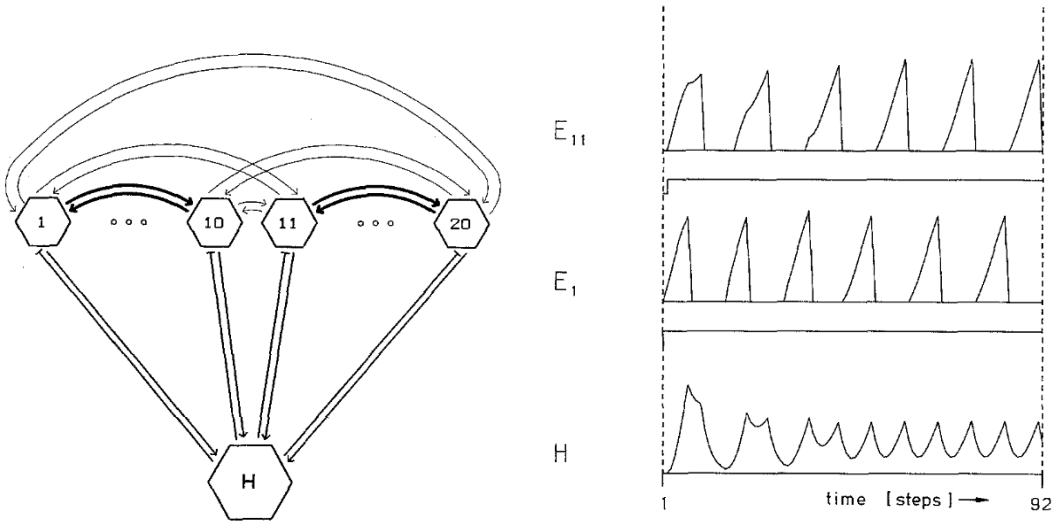


Figure 5.1 *Global inhibition and local excitation for phase grouping. (a)* A group of neural populations (numbered hexagons) were tuned to respond to different perceptual features and then linked by local excitatory couplings (thick black arrows). All of these units were innervated by a single inhibitory unit (hexagon H) limiting the amount of total activity in the network at any time. *(b)* Two of the representing populations, E_{11} and E_1, begin model dynamics by responding simultaneously to an input stimulus. This initiates feedback inhibition from the H unit (bottom trace) which gradually forces the E_{11} and E_1 bursts into anti-phase. These relative timings serve as a tag for the objects represented by each population.

synchronous activity in cortex when they belong to a single object physical object is essentially the principle of binding-by-synchrony made experimentally famous by Gray and Singer, 1989; Gray et al., 1989, as we discussed above. Gray's and colleague's work was in fact explicitly modeled by Wang, 1994 in a system similar to that of von der Malsburg. Similar to Gray and Singer, 1989's findings from cat striate cortex, Wang, 1994's results indicated that Wilson-Cowan oscillators (Wilson and Cowan, 1972) connected with local excitatory and global inhibitory connections would phase-lock when those oscillators took input from two regions of an elongated bar. When the bar was broken and the parts sufficiently separated, phase locking would disappear, much like the result of Gray and Singer, 1989, but relying explicitly on Malsburg and Schneider, 1986's principle of balanced excitation and inhibition (Fig.

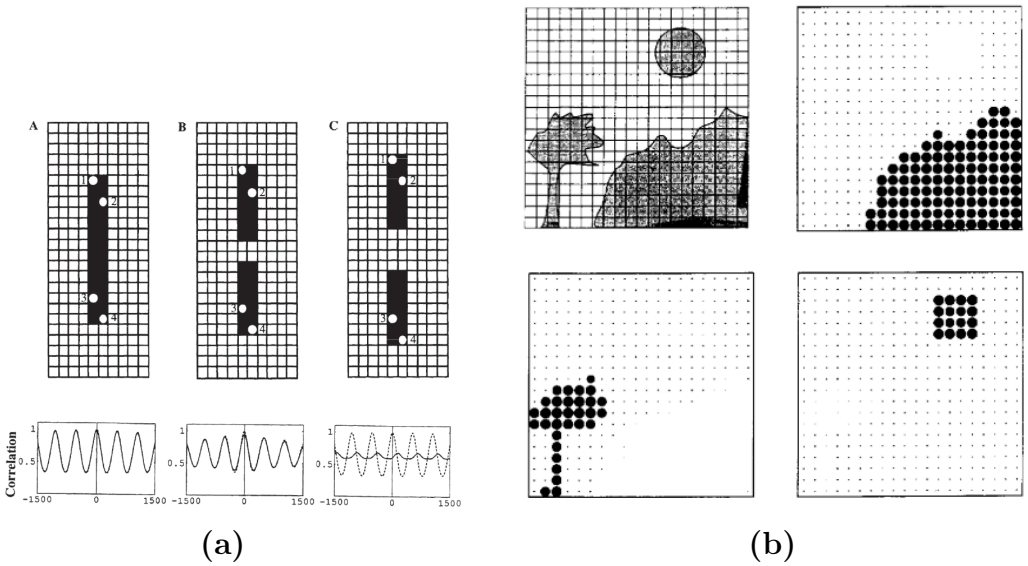


Figure 5.2 *A computational model of Gray and Singer. (a)* Using a similar network to Malsburg and Schneider, 1986, Wang recreated the cross-correlogram results of Gray et al., 1989. Top panels are input stimuli, and white dots indicating locations of units between which cross-correlations are calculated. As in the results of Gray and Singer, correlations display a marked rhythmic undulation only in the case that bar parts are sufficiently connected. If they are too far apart (panel C), this periodicity disappears (dotted line for comparison). *(b)* Later, Terman and Wang, 1995 extended this principle to simple images and showed how phase grouping à la Malsburg and Schneider, 1986 could be used to select single objects, prefiguring a kind of communication by coherence.

5.2a). Terman and Wang, 1995 extended this result to more complicated images (Fig. 5.2b) and provided mathematical guarantees for phase-locking. More recent work has investigated segmentation in systems of coupled oscillators with couplings determined by gestalt grouping rules (Yu and Slotine, 2009) and with sparse random connections (Li and Li, 2011; Raiko and Valpola, 2011). In parallel, Meier, Haschke, and Ritter, 2013; Meier, Haschke, and Ritter, 2014 specifically investigated grouping capability of the Kuramoto model.

To the author's knowledge, only one paper Quiles et al., 2011 investigates the ability of these non-equilibrium oscillatory systems to adapt "on the fly", in the manner emphasized as the beginning of the chapter. There, the authors show how

a downstream neural population can be used to selectively attend to individual synchronized groups, much like in McLelland and VanRullen, 2016 and our own demonstration in Sec. 4.3.4.

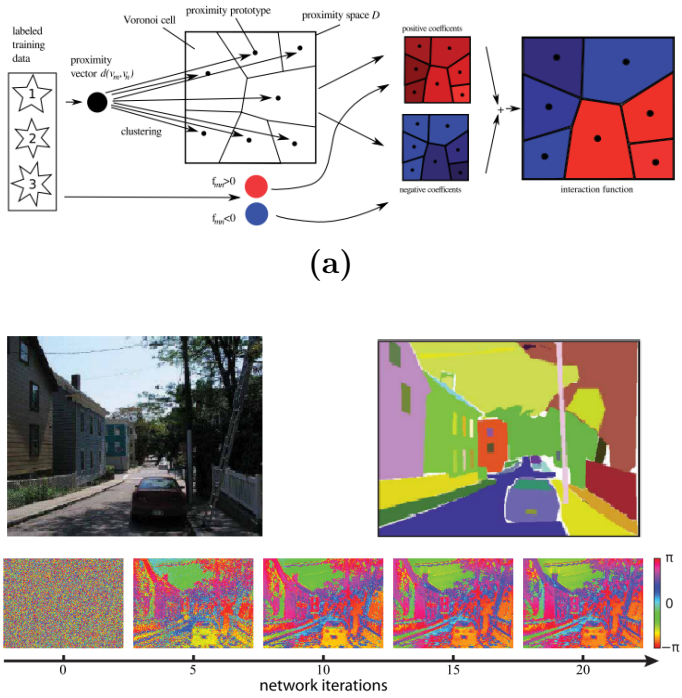


Figure 5.3 *Learning to synchronize: earlier attempts. (a)* Meier, Haschke, and Ritter, 2013 learned Kuramoto couplings which would synchronize images by computing coefficients of a training set in a Gabor basis and then constructing a metric in this Gabor space according to labels of image regions (see Meier, Haschke, and Ritter, 2013) for details. *(b)* Finger and König, 2014, on the other hand, constructed coupling matrices by computing a Grammian matrix from autoencoder features. The lower panel shows the evolution of their system. Notice that phase is not a unique signature of image objects. Further, neither of these examples learns how to synchronize by actually exploiting the Kuramoto dynamics. Any resulting synchrony is purely incidental.

So far, none of the systems we have described can learn in any meaningful sense. The typical locus of learning in neural models is the synapse, but, in the case of most coupled oscillator systems, the synapse must be hand-tuned so that the system synchronizes in a given manner (e.g. according to the objects in a *particular* image.). A new image requires, in effect, new synapses. Hence, any attempt to model learning in one of these systems must consider synapses themselves to be the output of some

parametrized *control function* whose parameters are the true locus of learning. For instance, Meier, Haschke, and Ritter, 2014 used a bank of Gabor filters as a control function (Fig. 5.3a). Image locations were coupled according to the similarity of their representation in this basis as measured by a metric learned by supervision. More recently, Finger and König, 2014 replaced Gabor filters with features learned by a simple auto-encoder and achieved fairly convincing image segmentation results (Fig. 5.3b). One neurobiological interpretation of these parameters, in line with the dynamical synapses of Malsburg and Schneider, 1986 and others, is that they control synaptic conductance at a fast timescale as a function of an input stimulus. That is, a stimulus arrives in the system and effects some dynamical process giving rise to the effective synapses used in later computation. A second interpretation is that the coupling function is itself a neural network whose outputs directly modulate the gain of the synapses of a second network, similar to the model of feedback modulation of V1 in Piëch et al., 2013.

Note, however, the difference between these learning algorithms and the optimization procedures we discussed in Sec. 3.3.3. The latter line of research concerns the optimization of network synchrony as a direct function of network dynamics. For instance, we reviewed the method of Brede, 2008a in which network couplings were adjusted depending on whether or not the current connectivity gave rise to a target behavior over the course of numerical simulation of the dynamics. Indeed, this is the gist learning in artificial neural networks as a whole. Instead, the nascent literature on learning in oscillator systems for specifically computer vision purposes separates the learning from the dynamics. Finger and König, 2014, for instance, generated couplings based on an auto-encoder's reconstruction loss and simply guessed that these couplings would give rise to the grouping behavior they wished for. Here, the insight is not learning through the dynamics, as it was for physicists like Brede, 2008a, but rather the off-loading of learning onto a background process with couplings as its

output.

In the next section, we propose a model which synthesizes these two ways of thinking. Like Meier, Haschke, and Ritter, 2014 and Finger and König, 2014, it has a trainable function that outputs the control variables determining network dynamics. As with Brede, 2008a, it is trained according to a loss function posed on the outputs of the dynamical evolution of the model. The whole system is end-to-end differentiable and can be trained with standard machine learning software. As always, our focus will be the Kuramoto model, for its generality, simplicity and neuroscientific relevance. Interestingly, the resulting system, "Kura-Net" has a simple physical interpretation as a Kuramoto model with quenched random parameters from a distribution which we will shape by backpropagation. Kura-Net also helps us realize a goal we stated in Ch. 3, the construction of an oscillatory system for perceptual grouping on real-valued images (as opposed to the pre-segmented training data required for the Kosterlitz machine). The cost is that Kura-Net must be trained by supervision in the form of pixel-wise masks, but the relevance of the resulting model to both computer vision and computational physics makes this cost bearable.

5.2 Kura-Net

Let X be a random variable with sample space $(\Psi, \mathcal{F})$. Let $\mathcal{K} = \mathbb{R}^{n\times n}$ be the space of (not necessarily symmetric) $n \times n$ coupling matrices and $\Omega = \mathbb{R}^n$ be the space of $n-$long intrinsic frequency vectors. We replace the quenched random parameters of Daido, 1987 with the parametrized *control function*

$$\begin{aligned} \Gamma_\Lambda &: \Psi \to \mathcal{K} \times \Omega \\ \Gamma_\Lambda(x) &= (K, \omega) \end{aligned} \tag{5.1}$$

where Λ is a collection of parameters. Here, we conceive of x as a latent representation of the Kuramoto parameters (K, ω) so that Γ_Λ functions like the reparameterization

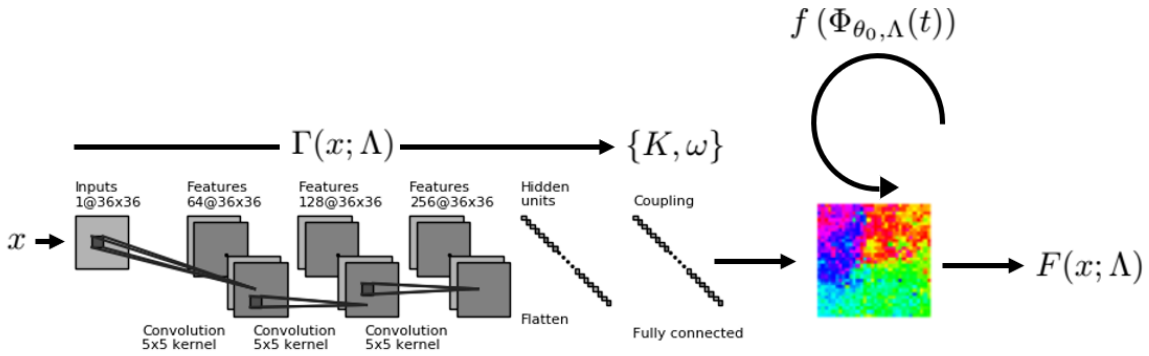


Figure 5.4 *Kura-Net.* A stimulus, x, is passed through the parametrized control function, Γ_Λ. The output is reshaped into a coupling matrix, $K(x)$, and a vector of intrinsic frequencies, $\omega(x)$. These parameters than determine a $T-$ long Kuramoto flow starting from a fixed initial phase. Each step of these dynamics is evaluated by an instantaneous loss function f and the integrated loss F is differentiated with respect to Λ in order to encourage a target behavior.

trick used in the construction of gradient estimators (Maddison, Mnih, and Teh, 2019). Depending on the application, we can also interpret x as a stimulus impinging on the system and influencing the flow the dynamics. We model the control function Γ_Λ as a hierarchical neural network (Fig. 5.4) whose weight and bias parameters comprise Λ. Notice that Γ_Λ is not necessarily a probability density on K, ω. For each sample x of the random variable X, we may the define equation of motion for the Kuramoto model

$$\dot{\theta}_i = \omega_i(x;\Lambda) + \sum_j K_{ij}(x;\Lambda)\sin(\theta_j - \theta_i), \tag{5.2}$$

where ω_i is the i^{th} component of $\omega(x)$ and so on. We have also made the dependence of the dynamics on Λ explicit. Given an initial condition θ_0, let $\Phi_{\theta_0,\Lambda} : \mathbb{R} \times [0, 2\pi]^n \to [0, 2\pi]^n$ denote the flow of the Kuramoto dynamics so that $\Phi_{\theta_0,\Lambda}(t) = \theta(t)$ is the state of the oscillator system at time t, having evolved from initial condition θ_0. We will evaluate each instantaneous $\theta(t)$ according to a loss function $f(\theta(t)) \geq 0$. Examples of f include *circular variance*, $1 - r(t)$, where r is the global order parameter (Eq. 3.14), and more complicated functions encouraging specific numbers of partially synchronous clusters (see Sec. 5.3.2). We refer to f as the "instantaneous loss". Details of relevant

loss functions are provided below for each application. Gradients with respect to control parameters Λ are acquired by differentiating the average value of the loss function on the dynamics after a t'-long burn-in period until some time T:

$$F(x;\Lambda,\theta_0) = \frac{1}{T-t'}\int_{t'}^{T} f(\Phi_{\theta_0,\Lambda}(t))\ dt \tag{5.3}$$

Hence, learning in Kura-Net is a gradient descent version of Brede, 2008a. We refer to F as the "integrated loss". Gradients $\nabla_\Lambda F$ propagate in two steps. First, gradients propagate backwards in time through the Kuramoto flow as a function of the initial condition θ_0 which is typically kept fixed during training. Once the sensitivity of the integrated loss with respect to all time steps has been calculated in this way, gradients are propagated down the layers of the control network Γ_Λ.

Note Λ comprises Kura-Net's only learnable parameters. In this sense, Kura-Net belongs to the class of hypernetworks (Ha, Dai, and Le, 2014), models whose outputs form the parameters for a second network such that the joint system is end-to-end differentiable. It is also similar in spirit to so-called neural optimal control (Chalvidal et al., 2020), in which a learned distribution on control functions is used to guide the flow of a stimulus-dependent dynamical system.

5.3 Experiments

The following three experiments demonstrate the ability of Kura-Net to control Kuramoto dynamics, particularly for use in the perceptual grouping relevant to visual reasoning. We first replicate the results of Brede, 2008a, Tanaka and Aoyagi, 2008 and others by showing that complex networks can be optimized to give rise to synchronous Kuramoto dynamics for much weaker coupling strengths than expected in either the all-to-all case (Eq. 3.15) or a random control. Our demonstration differs from this earlier work since we directly learn the function from intrinsic frequencies to optimal couplings. This demonstrates the utility of Kura-Net for the empirical study of general

coupled oscillator systems.

We then turn to two perceptual grouping experiments. In the first, we only report qualitative segmentation results and the in the second case we measure performance against a competing feedforward baseline. Both cases used the same control network Γ_Λ which assumes the form of a convolutional neural network with input images x. In the parlance of the previous section, we construe x as a random sample from some image distribution D and ultimately giving rise to the quenched random Kuramoto parameters $\Gamma_\Lambda(x) = (K, \omega)$. The network consisted of three convolutional layers with kernels of size 5×5 and no pooling (Fig. 5.4). Convolutional layers were interleaved with hyperbolic tangent activation layers. The first layer had 32 features and the number of features doubled each layer. The output of the final convolutional layer was subjected to a dropout rate of 50% during training. These features were passed through a single linear layer with an output of dimension $cn+n$, where n is the number pixels in the input and c is the number of neighbors of an oscillator. Even when $c \ll n$, the parameters of the final linear layer can be too large to hold in memory. To reduce memory load, we scale the dimensionality of the output down by a factor s and apply the resulting linear map to $\frac{cn+n}{s}$-sized sets of convolutional features s times in parallel. The output is then reshaped appropriately. This saves memory at the cost of reducing the expressivity of the network. The first cn dimensions of the CNN output are reshaped into a kernelized coupling matrix $K(x; \Lambda)$ with kernel width c and a vector of intrinsic frequencies $\omega(x; \Lambda)$

These parameters determined a Kuramoto flow which was run for T time-steps from a random initial condition θ_0 which was kept fixed throughout the experiment (Simulations in which θ_0 was randomized each trial were also run, but learning was much more difficult. Since the transients associated with initial conditions dissipate during the dynamics, use of randomized initial phases should be possible (Brede, 2008a; Brede, 2008b) if difficult to achieve in practice). Several masks were created

for each image indicating the group belonging of each pixel. For instance, an image with three target groups resulted in three binary masks acting as an indicator for each object (e.g. bottom row Fig. 5.8). We denote the set of pixel indices j in group g by G_g. These masks acted as a supervision signal for an instantaneous loss f which was the difference of two terms, $f_{\text{synch}}(\theta)$ and $f_{\text{desynch}}(\theta)$, designed to encourage synchrony within groups and desynchrony between groups respectively. Specifically, an $M-$group image was evaluated by the loss instantaneous loss function

$$\begin{aligned} f(\theta) &= -(f_{\text{synch}}(\theta) - f_{\text{desynch}}(\theta)) \\ &= -\left(\frac{1}{M} \sum_{g=1,\ldots,M} \overbrace{\frac{1}{|G_g|} \left| \sum_{j \in G_g} e^{i\theta_j} \right|}^{\text{Global order}} - \frac{1}{\prod_{g=1}^{M} |G_g|} \sum_{j_1 \in G_1, \ldots j_M \in G_M} \underbrace{\sum_{(p,q) \in (j_1,\ldots,j_M)} |\cos(\theta_p - \theta_q)|^2}_{\text{Frame potential}} \right), \end{aligned} \tag{5.4}$$

where second-to-last sum is taken over all M-tuples of locations and the last sum is taken over all pairs of indices from a given M-tuple (see Fig 5.8, right). The first bracketed term is simply the global order parameter r for the oscillators in group g (see Eq. 3.14). This term is maximized exactly when the oscillators in group g are synchronized. The second bracketed term is the so-called *frame potential* and is provably minimal when the M-tuple of phases $(\theta_{j_1}, \ldots, \theta_{j_M})$ equidistribute the unit circle (Benedetto and Fickus, 2003); i.e., for some permutation of $(j_1, \ldots, j_M)$, $\theta_{j_k} - \theta_{j_{k+1}} = \frac{2\pi}{M}$ for each k. Put another way, the frame potential is minimized when the complex exponentials $z_{j_k} = e^{i\theta_{j_k}}$ are a rotation of the M-th roots of unity[2]. The

[2]The frame potential gets its name from that fact that a collection of vectors $\{v_i \in \mathbb{R}^d\}_i^n$ is a minimizer of frame potential if and only if it forms a finite normalized tight frame of $\mathbb{R}^d$. When $d > 2$, the cosine difference is replaced by a general inner product. A frame is a generalization of a basis whose elements are not necessarily linearly independent. A signal $x \in \mathbb{R}^d$ can be represented by coefficients in the frame via the mapping $x \to \{\langle x, v_i \rangle\}_{i=1}^n$. The frame in some sense captures the energy of the signal since by definition there exist scalars A, B such that $A\|x\| \leq \sum_{i=1}^n |\langle x, v_i \rangle| \leq B\|x\|$, which

negative sign outside of Eq. 5.4 means that f is minimized when both the synchrony term f_{synch} and desynchrony term $-f_{\text{desynch}}$ are maximized. This is precisely the case that groups are internally synchronized but as mutually orthogonal as possible. Since we discretized the dynamics into T steps, the ultimate integrated loss was

$$F(x; \Lambda, \theta_0) = \frac{1}{T} \sum_{0}^{T} f(\Phi_{\theta_0, \Lambda}(t)) \tag{5.5}$$

where Φ once again denotes a solution to the Kuramoto dynamics with initial (fixed but random) condition θ_0 and f is the instantaneous loss just described in Eq. 5.4.

5.3.1 Learning to synchronize on complex networks.

Understanding the emergence of synchrony on graphs with constant coupling is no simple matter, but understanding synchrony on complex networks is harder still. By complex networks, we mean graphs whose properties deviate significantly from both regularity (e.g. lattices) and uniform randomness (e.g. an Erdös Rényi graph). These mysterious structures are an active area of exciting research and are speculated to be good models of real-world biological (Bashan et al., 2012) and social (Watts and Strogatz, 1998) networks.

In Sec. 3.3.3, we saw how earlier researchers, like Gleiser and Zanette, 2006; Brede, 2008a; Brede, 2008b; Tanaka and Aoyagi, 2008; Brede, Stella, and Kalloniatis, 2018, used numerical techniques to understand when synchrony emerges in these networks. For instance, Brede, 2008a gradually transformed a random network to optimize network synchrony by randomly switching links and keeping the new connectivity if it led to improved global order (Eq. 3.14)). This procedure was run for a single intrinsic frequency vector. Brede and others found that this sort of optimization procedure

guarantees that x can be reconstructed from the coefficients $\langle x, v_i \rangle$ up to a given error. A frame is tight when $A = B$ and normalized when $A = B = 1$. When $n = d$, the frames of $\mathbb{R}^d$ are precisely the orthonormal bases of $\mathbb{R}^d$. The frame potential therefore measures the "degree" of orthogonality of a collection of vectors in $\mathbb{R}^d$.

always resulted in increased connectivity between oscillators of disparate intrinsic frequencies. This makes intuitive sense, since oscillators with similar frequencies will synchronize anyway, and, since the number of links cannot increase, this limited resource should be spent enforcing connectivity between units which would otherwise desynchronize. This intuition was corroborated by theoretical work from Gottwald, 2015 and Kelly and Gottwald, 2011 who proved that increasing the correlation between disparate oscillators, under mild conditions, would always increase synchrony.

Note that this type of optimized connectivity is not only relevant to the particular case of Kuramoto dynamics. Recall from Sec. 3.3.2 how Honey and Sporns, 2008, Gómez-Gardeñes et al., 2010 and others used the Kuramoto model merely as a tool for understanding the more general issue of distributed communication in cortical networks. In that sense, maximized global order simply implies that units are still capable of communicating at long anatomical distances. This type of optimization procedure could therefore be used to make predictions about patterns of connectivity in the brain, particularly those optimized for task performance.

In this section, we replicate the experiment of Brede, 2008a, but now learn a *distribution* on connectivity conditioned on intrinsic frequencies instead of a single connectivity for a given intrinsic frequency vector. This distribution will be constrained during training to have a fixed mean, $\bar{p}$, so that the probability of two units' being connected does not change during optimization. Otherwise, an obvious optimal solution would be a fully-connected network. We will then compare the synchronizability of the optimized network to random (Erdös Rényi (Erdös and Rényi, 1959)) networks with the same $\bar{p} = .035$. Any improvement in synchronizability will therefore arise from non-random factors beyond first order graph statistics.

To that end, we define a Kura-Net instance

$$\Gamma_\Lambda : [-1,1]^{100} \to \mathbb{R}^{100\times 100}$$
$$\Gamma_\Lambda : \omega \mapsto \ell, \tag{5.6}$$

where ω is a vector of intrinsic frequencies uniformly distributed on $[-1,1]^{100}$ for a graph of 100 oscillators and ℓ is a symmetric matrix. We model Γ_Λ as a three-layer multi-layer perceptron (MLP) whose layers are interleaved with rectified non-linearities. Hidden layers each had 256 units. Λ are the weights and biases of this network.

We conceive of ℓ as a symmetric array of logits which are then centered by

$$\ell_{ij} \leftarrow \ell_{ij} - \left(\langle \ell \rangle - \log\left(\frac{\bar{p}}{1-\bar{p}}\right)\right) \tag{5.7}$$

where $\langle \ell \rangle$ is the sample mean of ℓ, to maintain a constant coupling probability. Logits are then transformed into probabilities of connectivity via $p_{ij} = \frac{1}{1+e^{-\ell_{ij}}}$ from which is sampled[3] a symmetric adjacency matrix A. This adjacency matrix then participates in the Kuramoto dynamics defined by

$$\dot{\theta}_j = \omega_j + \sum_i^{100} \sigma^* A_{ij}(\omega;\Lambda)\sin(\theta_i - \theta_j), \tag{5.8}$$

where σ^* is a constant coupling strength set to .7, as in Brede, 2008a. We have made the dependence of A on the Kura-Net input, ω, and the Kura-Net parameters, Λ, explicit. We solved this equation with Euler's method for $T = 100$ steps and a step size of $\alpha = .1$.

The result was a sequence of phase vectors $\{\theta(t)\}_{t=1}^{100}$ each of which was evaluated according to its circular variance

$$1 - r(t) = 1 - \frac{1}{100}\left|\sum_{j=1}^{100} e^{i\theta_j(t)}\right| \tag{5.9}$$

These circular variances were integrated via

$$F(\omega;\Lambda) = \frac{1}{T-t'}\sum_{t=t'}^{T}(1-r(t)) \tag{5.10}$$

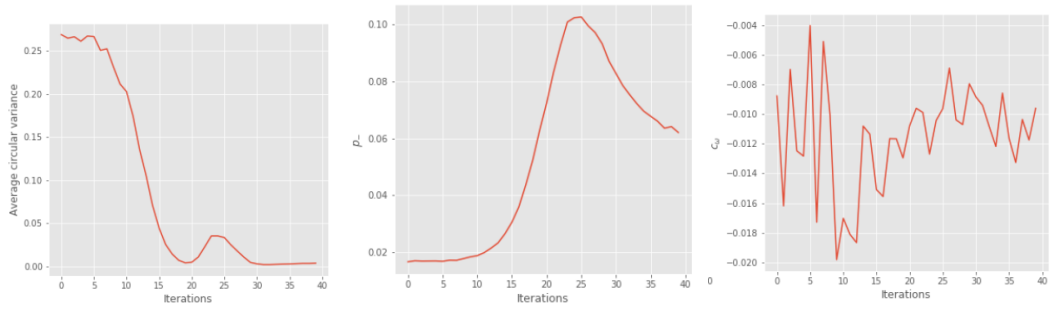
[3]We used the straight-through estimator (Yin et al., 2019) to maintain automatic differentiability.

where t' was a relaxation time set to 50. Batches of 256 intrinsic frequencies were fed to Kura-Net for 40 iterations. Gradients were calculated using Pytorch autograd software (Paszke et al., 2019).

We find that the loss decreases rapidly to 0 in around 20 batches at which point there is a ricochet which quickly decays to 0 once again (Fig. 5.5a). Brede, 2008a and subsequent authors claimed that optimized networks featured strong connectivity between oscillators of differing intrinsic frequencies and measured this effect by calculating the probability, p_-, of connectivity in the optimized graph between units whose intrinsic frequencies had opposite signs and by computing

$$c_\omega = \frac{\sum_{i,j} A_{ij}(\omega_i - \langle\omega\rangle)(\omega_j - \langle\omega\rangle)}{\sum_{i,j} A_{ij}(\omega_i - \langle\omega\rangle)^2} \tag{5.11}$$

where $\langle\omega\rangle$ the sample mean of intrinsic frequencies. Contrary to this earlier work, we find no correlation between these values (Fig. 5.5b, 5.5c). There was a gradual increase of p_- over training until a peak at around 25 iterations (shortly after the bump in the loss curve), but it decayed steadily after that. Compare this to the result of Brede, 2008a, who found that p_- tended to 1 as synchronizability increased. Further, we find no trend in c_ω whatsoever. Whatever strategy Kura-Net has found to increase synchronizability, it is different from that proposed by Brede, 2008a.



(a) **(b)** **(c)**

Figure 5.5 *Learning to synchronize on complex networks. (a)* The loss function, Eq. 5.10 decreasing over the course of 40 iterations. During optimization, neither p_- *(b)* nor c_ω *(c)* were found to increase, contra Brede.

We also find that Kura-Net learns synchronizable graph structures that generalize

beyond the coupling strength, σ^*, used in training. We varied coupling strength σ from 0 to 1.6 in 25 increments and measured the average synchrony during dynamics of Kura-Net and a random control, again with the same probability of connectivity, $\bar{p} = .035$. We averaged over 10 random ω and 10 initial phases for the dynamics. Results are shown in Fig. 5.6. Brede, 2008a found much the same (compare to Fig. 3.15), though we note that our network synchronizes to much higher levels at much lower coupling strengths at many fewer iterations than in his case.

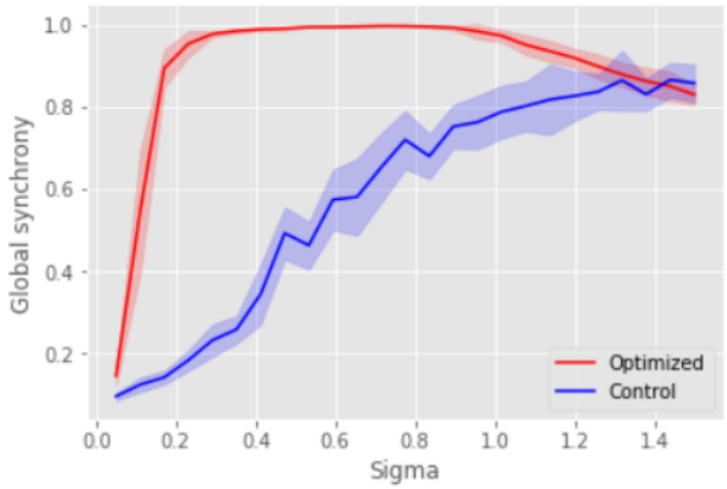


Figure 5.6 *Optimized synchrony across a range of coupling strengths.* Here, we reproduced the result of Brede, 2008a by showing Kura-Net optimized or coupling strengths at $\sigma^* = .7$ also produced better synchrony across a range of coupling strengths than a random Erdös Rényi control. Average global synchrony (Eq. 5.10) was calculated for 10 initial phase configurations and intrinsic frequencies. Shaded regions are $\pm$ one standard deviation over initial phase and intrinsic frequency samples.

We can get a little intuition about Kura-Net's solution by averaging some statistics of its output adjacencies. First, we see that the degree distribution of optimized connectivities are approximately Gaussian (red bars, Fig. 5.7a). This differentiates the optimized graphs from both the binomially distributed random control (blue bars) and the fat-tailed distributions which characterize many naturally occurring networks. Gaussian degree distributions have been reported in cortex, however (Ivković, Kuceyeski, and Raj, 2012). Next, we plot the observed distribution of

coupling probabilities (red, Fig. 5.7b) with the mean, $\bar{p}$ indicated by the dashed blue line. These probabilities seem to follow a power-law distribution with a large mode close to 0 and a fat tail, suggesting the existence of a few units, on average, with high connectivity. Finally, we depict the eigenspectrum of the symmetric, normalized graph Laplacian (Gallier, 2019) for both optimized and control graphs. The random controls tend to have eigenvalues, λ, close to 0, indicating weak connectivity and the presence of numerous connected components in the case of the 0 eigenvalue. The optimized graphs have a spectrum which is shifted rightward, as is characteristic for highly synchronizable graphs (see Brede, 2008b).

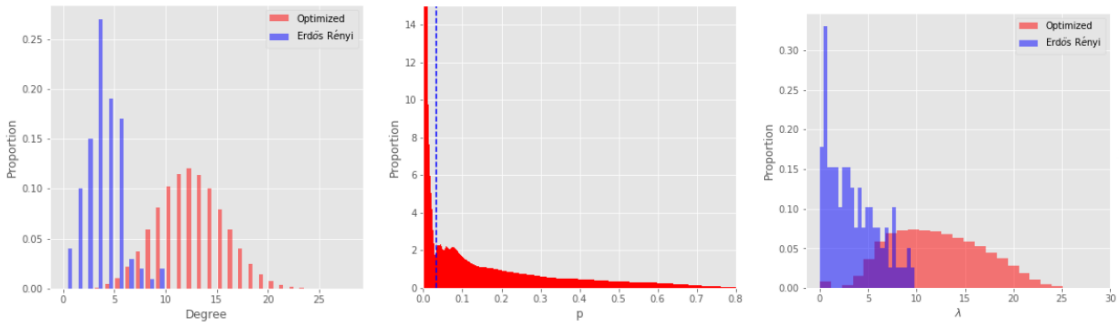


Figure 5.7 *Optimized graph statistics. (a)* Degree distribution of optimized graphs (red) and random control with same probability of connectivity. The optimized degree distribution is markedly Gaussian with a mean at about 12. The control distribution is binomial. *(b)* Here, we plot the distribution of p_{ij} output by Kura-Net. The mode is overwhelmingly at 0 (it extends much higher than shown), indicating that the graphs are very sparsely connected. The distribution has a flat region centered around $\bar{p}$ given by the dashed blue line. This is followed by a long decaying tail, suggesting the presence of a few highly connected oscillators. *(c)* The graph Laplacian operator is given by $L = D - A$, where D is the degree matrix and A is the adjacency matrix. The symmetric, normalized Laplacian is given by $D^{-1/2}LD^{1/2}$. The spectrum of this operator is known to contain interesting information about graph structure (Gallier, 2019). The lack of 0 eigenvalues compared to the control network indicates the graph has become highly connected. Together with the fact that the graph is sparsely connected indicates that the graphs have evolved a small-world structure.

In this section, we chose to replicate and generalize the result of Brede, 2008a,

along the way demonstrating that connectivity between disparate intrinsic frequencies is not the only path to synchrony. However, we could very well have used Kura-Net to study other aspects of synchrony on graphs. For instance, we could have examined the opposite case of optimized intrinsic frequencies for random input couplings. We might have also looked at directed graphs or negative couplings. Future work will address these questions.

5.3.2 Texture Segmentation

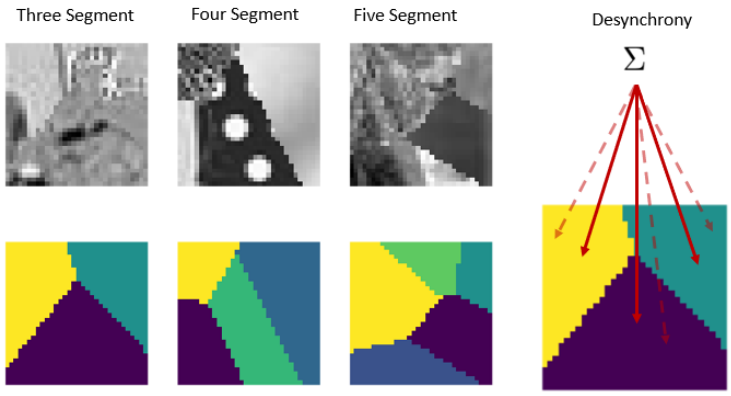


Figure 5.8 *Composite textures.* Composite textures are shown in the first three upper panels and their masks are shown directly below. The desynchrony part of the loss is calculated by exhaustively comparing all k-tuples of locations where k is the number of segments.

To qualitatively evaluate the performance of Kura-Net on a perceptual grouping task, we trained it on a data set of 32×32 images with composite textures (Fig. 5.8) similar to the stimuli of Meier, Haschke, and Ritter, 2014. Images were divided into $k = 3, \ldots, 5$ randomly arranged Voronoi cells which were populated with textures from the Describable Textures Database (Cimpoi, Maji, and Kokkinos, 2014). 10000 training and 10000 testing images were created for each setting of k. An instance of Kura-Net was then trained on each of the k training sets. T was set to 8. The dynamics were solved using Euler's method with step size 6. The step size is large in an absolute sense, since otherwise the coupling-dependent term in Eq. 3.13 is so small

the model only evolves very little in 8 time steps. Fig. (5.9) shows an example of the

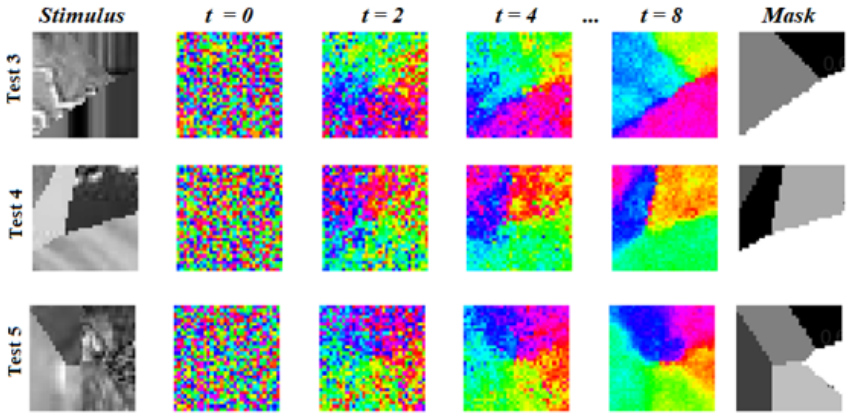


Figure 5.9 *Texture segmentation in Kura-Net.* Three, four and five text stimuli are shown in the first column. Starting from a random initial condition, the system evolves for 8 timesteps towards a phase segmentation of the stimulus which closely matches the ground truth in the final column.

evolution of the trained model on test images for each k. By the 2nd time step, we can observe that the system has self-organized into three or four coarse clusters. As the system evolves, these clusters resolve into sharpened regions which closely match the target groups depicted in the sixth column. By design, phases at convergence tend to equidistribute the unit circle (Fig. 5.10).

An inspection of the intrinsic frequency vectors learned by the model is revealing. Fig. (5.11) depicts a collection of input images x together with the intrinsic frequencies output by the trained model $\omega(x)$. Notice that intrinsic frequencies on either side of a texture boundary tend to have opposite sign. One interpretation of this result is that the model has learned to place anti-correlated frequencies at these locations to encourage desynchrony at texture edges. Connectivity in the model is local by design, so there is only pressure to desynchronize phases at boundaries. It is tempting to interpret this result as a corroboration of Tanaka and Aoyagi, 2008; Brede, 2008a; Brede, 2008b who found that synchrony on complex graphs was encouraged by connectivity between units of anti-correlated intrinsic frequencies. However, recall that those authors investigated the optimal couplings for *fixed* intrinsic frequencies,

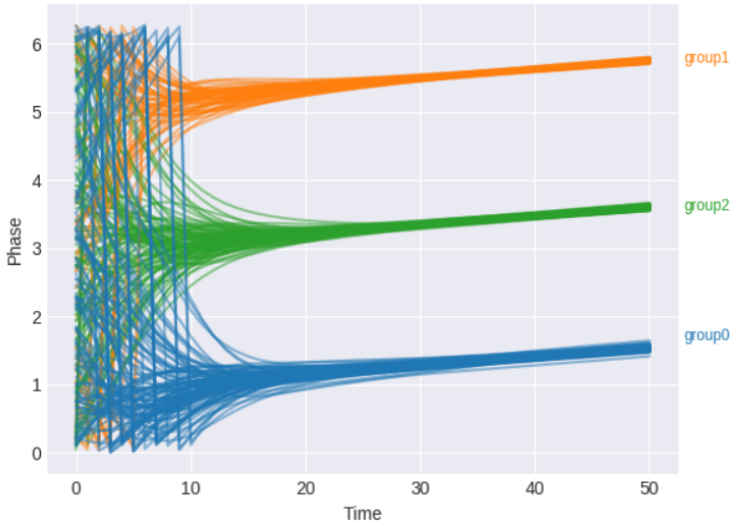


Figure 5.10 *Phase equidistribution in Kura-Net.* Here, we run the trained model on a three-segment image for 50 timesteps to allow more time for convergence. The mean phases within a group come to equidistribute the unit circle. Vertical lines are plotting artifacts from where oscillators crossed 2π.

whereas both are optimized in our case, and indeed for multi-modal synchrony and not global synchrony. Of course, these results are mostly interesting for the perceptual quality of the segmentation rather than for the quantitative segmentation performance compared to a baseline. We turn to this type of experiment next.

5.3.3 A quasi-Same-Different problem

The moral of Ch. 1 was visual relation problems, like the same-different relation, reveal certain deficiencies in CNNs. The results presented in and literature reviewed in that chapter highlight the need for various dynamical mechanisms not found in CNNs, notably attention, working memory, and perceptual grouping, the last of which has occupied us for many pages now. Recall also that brute force implementations of these mechanisms, like the relation net, suffer from their reliance on rigid, grid-like arrays of convolutional features. When objects in the relation tend to be closer or farther away than expected, these types of "attention" algorithms fail (see Sec. 2.3.3). As noted by Fries, 2005, McLelland and VanRullen, 2016 and others, the case of

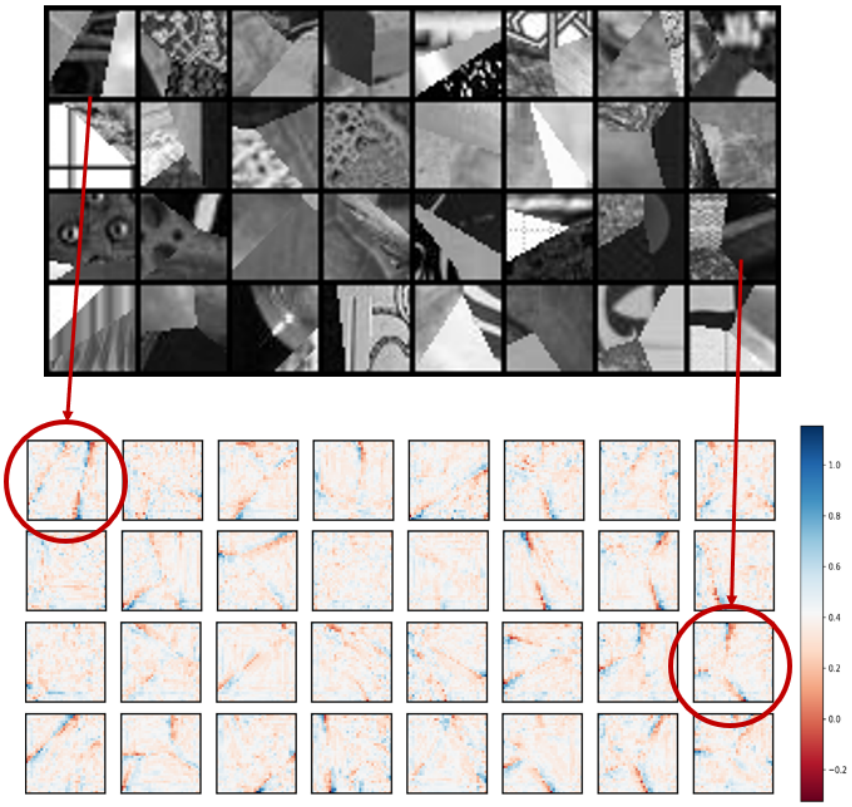


Figure 5.11 *Learned intrinsic frequencies.* For various input images (upper panels), intrinsic frequencies output by the control network are displayed (lower panels). Observe how pairs of clusters with opposite signs congregate at texture boundaries.

mutually occluding and highly overlapping objects is exactly where a neural grouping mechanism based on phase coding could prove useful.

In this experiment, we take a first step towards modeling the types of oscillatory mechanisms speculated to underlie visual reasoning (see Ch. 3, Sec. 1). To that end, we generated a data set of 36×36 images each with two overlapping MNIST (LeCun, Cortes, and Burges, 2010) digits. In half of the images, the two digits belonged to the same class; in the other half, the digits belonged to different classes. We sought to compare the performance of Kura-Net on a "quasi-" same-different task versus a CNN baseline. Instead of prompting these models to classify the images directly, we rather tasked the systems with segmenting the images as in the previous experiment and then determined a "same" or "different" label based on the number of predicted

segments. A “same" prediction was chosen if the model predicted two segments (a background segment plus a single segment comprising the two same-class digits), and a “different" prediction was chosen if the model predicted three segments (a background segment and one for each digit). Digits were allowed to appear anywhere in the image, and no restriction on digit overlap was enforced. This makes segmentation inherently ambiguous.

Kura-Net produced a phase segmentation in the manner described above, although we found increasing T to 15 produced better results. The CNN baseline output phases directly and had almost exactly the same architecture as that used to produce couplings and intrinsic frequencies in Kura-Net. The only architectural difference between the two models was in the final linear layer, since it was shaped to produce an output that was quadratic in image size for Kura-Net and linear in size for the baseline. To compensate for Kura-Net's increased number of parameters, we increased the parameter s until Kura-Net's linear layer was small enough so that both models had approximately the same number of parameters (Kura-Net actually had fewer). Both models were trained with the loss function of Eq. 5.4.

The continuous-valued output of both models was quantized by K-means for both $K = 2$ and $K = 3$. Then, a silhouette score (Rousseeuw, 1987) was produced for each K to measure the confidence of the segmentation. The more confident clustering was chosen. The quasi-same-different accuracy was determined by calculating the proportion of the data assigned the correct K value. To ensure that each model was actually segmenting the data correctly (as opposed to putting the correct number of clusters in the wrong place), we also measured the perceptual quality of the segmentation using mean intersections-over-unions (mean IoU; Tan, 2005) and panoptic quality (PQ) metric, an IoU metric normalized by the sum of false positive and false negative segmentation predictions (see Kirillov et al., 2018 for details).

The qualitative performance of trained model is depicted in Fig. 5.12. As before,

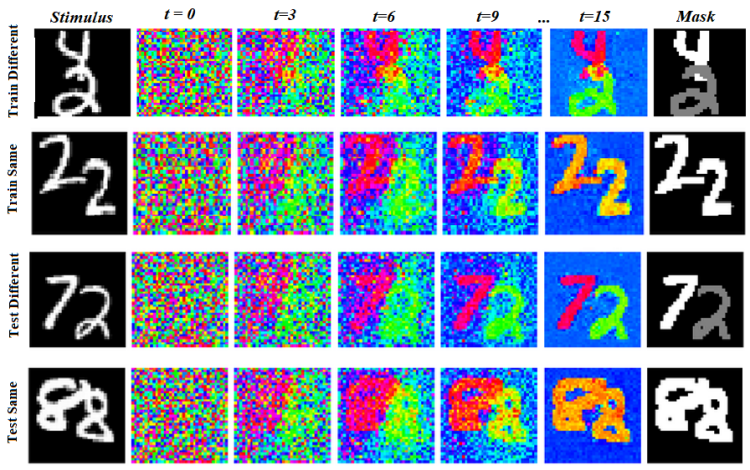


Figure 5.12 *Kura-Net segmenting digits by class.* We plot once again the evolution of a trained Kura-Net, though this time we include both training and testing examples. Further, the system must learn digit class features in order to correctly segment.

the model evolves from a random state to fairly sharp segments, now respecting the syntax of the sameness rule. Interestingly, we found that the model tended to incorrectly desynchronize "same" digits only to synchronize them later in the dynamics (e.g. Fig. 5.12, rows 2 and 4), concordant with theoretical evidence that partial synchrony in numerous groups is often easier than in few groups (Pham and Slotine, 2007).

Loss curves for the CNN baseline and Kura-Net are shown in Fig. 5.13. Kura-Net outperforms the baseline in both the training and testing regimes. Note however that Kura-Net substantially overfits to the training set, indicating that Kura-Net has a much greater effective capacity than the baseline. Improved regularization (by the introduction of weight decay or properly optimized dropout rate) might eliminate overfitting.

The comparative segmentation and quasi-same-different accuracies are collected Table 5.1. Both models achieve comparable mean IoU and PQ scores, though Kura-Net succeeds marginally in both cases. Kura-Net's margin of improvement on the final quasi-same-different measure is quite a bit larger, at about 10%. Digits often

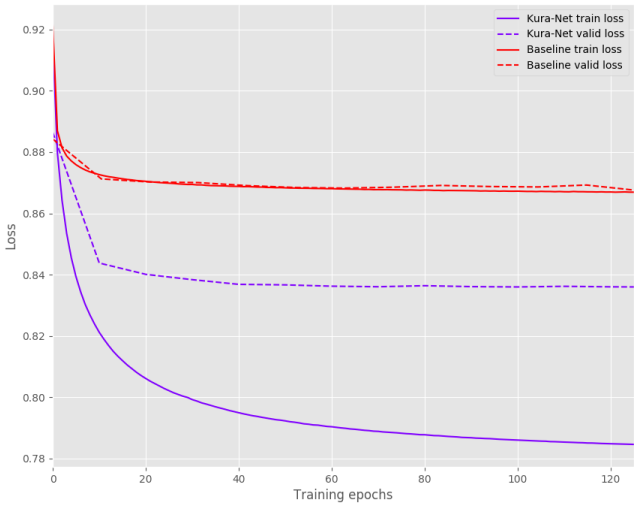


Figure 5.13 *Kura-Net training.* Loss curves on training (solid) and testing (dashed) data for Kura-Net (purple) and a baseline (red). The baseline is the same as Kura-Net with the Kuramoto dynamics removed.

Model/Data	Mean IoU	PQ	qSD
Kura-Net	**0.8405**	**0.9139**	**0.7176**
Baseline	0.8012	0.9060	0.6125

Table 5.1

substantially overlapped (Fig. 5.14, first panel), and, according to the results of Sec. 2.3.3, this is precisely the scenario where a CNN should struggle. Note that the substantial overlap of digits places a cap on the maximal possible segmentation scores, since overlapping regions are inherently ambiguous. Both models consistently incurred more false same detections than false different detections (25.21% vs 3.03% for Kura-Net; 37.30% vs 1.45% for the baseline). This is either because the substantial overlap of digits encourages the model to incorrectly synchronize the digits in impossibly ambiguous cases or because of a problem in the loss of Eq. 5.4. We discuss this possibility in the next section.

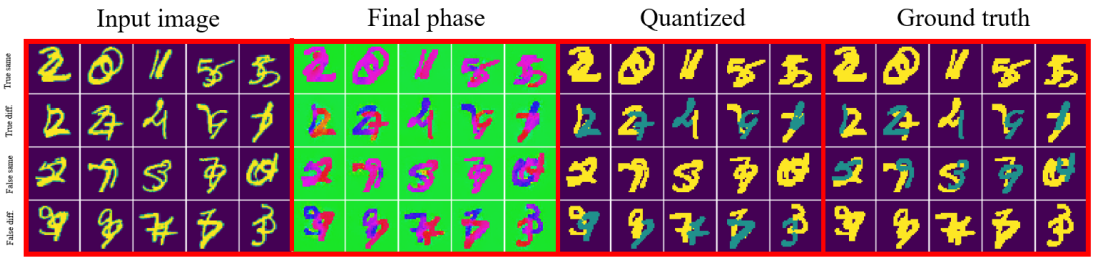


Figure 5.14 *Kura-Net quasi-Same-Different examples.* From left to right, successive panels show the input to the system, the phase after 15 timesteps of evolution in the trained model, quantization according to the method described in the main text, and ground truth masks. Note that the correspondence between colors in the quantized and ground truth images do not matter, only the degree of overlap between segments. Rows index true same, true different, false same, and false different images.

5.4 Future work

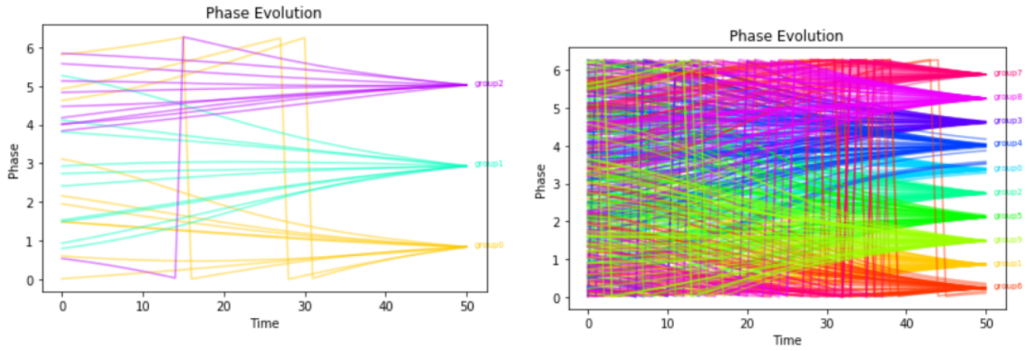
In this chapter, we proposed an end-to-end differentiable learning method for the Kuramoto model. The abilities of the resulting Kura-Net system were demonstrated both on the problem of learning to synchronize complex networks and on various perceptual grouping tasks, including one which we analogized to same-different reasoning. The results presented here are promising but obviously early in development.

A clear area of future work is the refinement of perceptual grouping quality. Here, there have been some recent developments. We can see from the second panel of Fig. 5.14 that different digits are often very close in phase, even if the quantization eventually comes to discriminate between them. That is, spurious synchronization is a problem. Future work will investigate other loss functions for desynchrony. One promising example comes from Cohn, 1960, who proposed the following potential energy for maximizing the distances between electrical charges on a circle:

$$U(\theta_1, \theta_2) = -\log(|2\sin(.5(\theta_1 - \theta_2))|). \tag{5.12}$$

Cohn noted that such a potential is even, 2π-periodic, concave up on intervals $[2\pi k, 2\pi(k+1)]$ for $k \in \mathbb{Z}$ and, crucially, obeys $U'(\epsilon)/\epsilon \to -\infty$ as $\epsilon \to 0$. In other words, the repulsive force felt between two oscillators increases super-linearly as

the oscillators approach one another. Some simple tests with this desynchrony loss have indicated that this strong repulsion does indeed promote well-spaced groupings (Fig. 5.15)



(a) **(b)**

Figure 5.15 *A logarithmic desynchrony potential.* A single coupling matrix was evolved for multimodal synchrony using the desynchrony potential of Eq. 5.12 in two cases: *(b)* 30 oscillators to be synchronized into 3 groups and *(b)* 500 oscillators to be synchronized into 10 groups. Again, only a single coupling matrix is being optimized, so there is no control network. The selection of an appropriate desynchrony loss is a topic for further research.

A second area of future work is the transformation of our current quasi-same-different task into a genuine visual reasoning problem. So far, we have only measured a sort of same-different-based clustering. This is of course rather different than the visual relation tasks explored in Ch. 2, in which CNNs and related networks were trained by a simple supervision loss on image labels. At present, it is unclear how Kura-Net could be trained in this way. In experiments not reported here, we found that simply passing phases through a classifier is insufficient, since the system learns to encode the predicted label in seemingly random, desynchronized phases, essentially eliminating the value of a synchrony-based system. Although one motivation for investigating oscillatory systems was their hypothesized role in visual reasoning (e.g. Alamia et al., 2019), it could very well be that their inclusion in future neural networks will be only as a module meant to support perceptual organization as opposed to a

complete system. Particularly intriguing would be an implementation of Kura-Net for these purposes that made good on the intuition that non-equilibrium dynamics are good for adaptation in the sense of Stein and Newman, 2013: could these adaptive dynamics be used to model transient grouping and selective attention as in the spiking model of McLelland and VanRullen, 2016 or the oscillatory system of Quiles et al., 2011?

There are branches of science, however, where oscillatory systems are an end unto themselves, and it is here that Kura-Net might have its most direct application. As we mentioned above, Kura-Net could be used to study numerous aspects of synchrony on graphs beyond the ones discussed in Sec. 5.3.1. Development of the method proposed here could therefore be useful to understanding the emergence of synchrony in areas as diverse as social networks, power grids, and laser arrays (see Arenas and Albert, 2008, Sec. 5 for a summary). Further, there is an alternative body of literature concerning the estimation of coupling parameters from observed phases in oscillatory systems (Kralemann, Pikovsky, and Rosenblum, 2011; Kralemann, Pivovsky, and Rosenblum, 2014), for instance, central pattern generators in the spinal cord. A solution to this problem would be of use to neuroscientists seeking to infer patterns of functional connectivity among oscillating neurons. Existing methods depend on a computationally taxing estimation of the probability densities on observed phases. If, however, a Kura-Net-like system uses a control network to model these couplings and the resulting coupling distribution is constrained by empirically relevant prior knowledge (e.g. couplings should be sparse), then the earlier brute force methods could be avoided.

Chapter Six

Conclusion

This dissertation has two parts, a first part on systematicity in Ch. 2 and a long second part on oscillatory neural networks in Chs. 4 and 5. The dissertation's cohesion depends on the connective tissue of Ch. 3, in which I describe the purported role of cortical oscillations in systematic cognition. In this concluding section, we will take inventory both of the thesis' two main parts and the fulcrum of Ch. 3. Where weaknesses are found, we will offer potential future directions.

In Ch. 2, we argued that the state-of-the-art model of image classification, the CNN, is pathologically sensitive to non-syntactic image properties when attempting to learn a visual relation discrimination task. We analogized this deficiency to a lack of systematic cognition, inasmuch as systematicity is the ability to understand structured expressions by virtue of their syntax alone. We used this deficiency to motivate a search for neural mechanisms, namely attention, grouping and working memory, which would improve the systematicity of neural machines. It is no coincidence that these mechanisms are features of neural dynamics, that is, neural phenomena operating on a timescale faster than long-term synaptic learning. Indeed, the whole crux of Fodor and Pylyshyn, 1988 and of the theory of automata that clearly inspired them, is that computation is efficient when it is done "on-the-fly", without storing solutions in advance (for instance in the long-term storage medium of synapses). Since systematicity ideally requires the understanding of an *infinite* number of expressions,

long-term storage solutions are untenable, and the only real solution would seem to be a set of flexible mechanisms for building new representations out of a dictionary of fixed ones. Attention, grouping, and memory are three of these mechanisms.

In practice, it is very difficult to make this argument in the current machine learning milieu. First of all, one must combat the folk science that the ability of deep networks to learn highly selective and invariant object categories should support systematic reasoning. Disregarding the fact that these networks do not even generalize on image classification tasks the way they should (Szegedy, Zaremba, and Sutskever, 2013), this folk science confuses template matching with reasoning, an error already pointed out by Fodor and Pylyshyn, 1988. Only the latter involves complex expressions with real constituents which are computationally accessible. For example, the complex representation "The red object has the same shape as the green object" has "red object" as an actual part so that the true proposition "There is a red object" can be automatically deduced by mechanically separating the first part of the expression from the second. This type of flexible, mechanical reasoning is exactly what one needs for visually-informed behavior instead of just image classification. Next, there is the vexing claim that "more layers" are all that is needed to solve the problem. There is little explanatory merit to these claims in a day and age where it is recognized that deep networks are merely a case of more general, infinite-layer systems (Knauf, 2018). Far more interesting are the practical cases of small networks with known expressivity bounds. How do we get these systems to perform well on a huge range of recognition and reasoning problems at least as well as humans and animals? The results of Ch. 2 show we are far from solving this problem.

After the arguments of Ch. 3 (which, again, we will address below), we proposed two new models. The first, the Kosterlitz Machine, is a probabilistic graphical model with phases. After a mathematical exposition, we showed how it could be used to formally mimic certain oscillatory processes in cortex. Its relevance to future modeling,

however, is uncertain. Its primary practical limitation is the manner in which is must be trained on data with phase, which avoids the main problem of using oscillations as a purely representational mechanism. In the main text, we noted some ways of getting around this problem by using optic flow from moving images. The actual novelty of this system from a purely physics perspective is its use of unit magnitudes. To the knowledge of the author, this is a novel development and should be explored elsewhere.

The second proposed model was Kura-Net, more or less a neural network parametrization of the Kuramoto model. We demonstrated its utility for phase-based perceptual grouping as well as for the study of functional connectivity in complex networks. This system is still early in development and the various experiments shown in the main text represent subtly different interpretations of the model. The greatest practical limitation of this system is its prohibitive memory costs, imposed by the large coupling matrix output of the control network. The sparser this matrix, the lesser the memory cost but the worse the performance (or so it seems from early experiments). Future work should examine whether there is a cheap and easy way to learn small-world graph structures which are sparse and known to support synchrony in complex networks. This practical limitation is also related to the somewhat modest performance gains exhibited by Kura-Net over baseline CNNs. It is likely that the future alleviation of technical limitations in training the system will result in better overall performance compared to feedforward models. An important limitation from the perspective of theoretical neuroscience is the relevance of this system to actual neural processes. First, as noted above, weakly coupled neural networks are not representative of general neurophysiology. Second, the interpretation of the control network as an abstraction of the biophysics underlying fast synapses has not been worked out in detail. Future work will address all of these problems but will also endeavor to align Kura-Net more closely to work in which the Kuramoto model is unquestionably relevant, particularly

in the study of functional connectivity.

The whole weight of the dissertation's argument rests on Ch. 3, where the supposed link between systematicity and neural oscillations is provided. Let me defuse the tension surrounding this link immediately by admitting that I am not ideologically committed to any role of cortical oscillations in cognition whatsoever. There are, of course, convincing studies, some of them very recent, but our computational and physiological understanding of cortical oscillations and their role in mental processing is still early in development. Further, there are numerous interesting proposals for making flexible cognitive machines out of neurally plausible hardware not involving oscillatory dynamics. For instance, working memory models like that of O'Reilly and Frank, 2006 are purely rate-based, as are the grouping models of Roelfsema and Houtkamp, 2011. I will leave the adjudication between rate-based and oscillations-based models to the neurophysiologists. This thesis only attempts to even the playing field by giving scholars of neural oscillations machine learning methods which have hitherto been exclusively enjoyed by the rate modeling camp. Whether or not these systems reveal anything meaningful about neural processing is a separate question.

Overall, I hope that this dissertation is viewed as a continuation of the tradition in computational neuroscience which attempted to link "strong" cognition (in the sense of Fodor) and neural dynamics. In my view, this tradition, associated with Bienenstock, Geman, von der Malsburg, and others is ripe for a new era of exploration in the age of modern parallel computing. So far, this era has seen the ascendance of deep rate-based models, and discussions of systematicity and other cognitive properties have mostly been forgotten. It is a great irony that modeling in the 1980s and 90s, when technical limitations were so numerous, would give rise to such diverse and interesting formalisms as "neural gas" models (Martinetz and Schulten, 1991) and "dynamic link architectures" (Lades et al., 1993), but that the modern age of computational abundance would be accompanied by a stark conformity of thought. I hope that this

thesis represents a first step in the rejuvenation of this older way of thinking.

REFERENCES

Abadi, Martın et al. (2016). "TensorFlow: A System for Large-scale Machine Learning". In: *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*. OSDI'16. Berkeley, CA, USA: USENIX Association, pp. 265–283.

Acebrón, Juan A. et al. (Apr. 2005). "The Kuramoto model: A simple paradigm for synchronization phenomena". In: *Rev. Mod. Phys.* 77 (1), pp. 137–185. DOI: 10.1103/RevModPhys.77.137. URL: https://link.aps.org/doi/10.1103/RevModPhys.77.137.

Ackley, D, G Hinton, and T Sejnowski (Mar. 1985). "A learning algorithm for boltzmann machines". In: *Cognitive Science* 9.1, pp. 147–169. ISSN: 03640213. DOI: 10.1016/S0364-0213(85)80012-4. URL: http://www.sciencedirect.com/science/article/pii/S0364021385800124.

Aizawa, Kenneth (1997). "Exhibiting versus explaining systematicity: A reply to Hadley and Hayward". In: *Minds and Machines* 7.1, pp. 39–55. ISSN: 09246495. DOI: 10.1023/A:1008203312152.

Alamia, Andrea et al. (Jan. 2019). "Differential involvement of EEG oscillatory components in sameness vs. spatial-relation visual reasoning tasks". In: *bioRxiv*, p. 2019.12.16.877829. DOI: 10.1101/2019.12.16.877829. URL: http://biorxiv.org/content/early/2019/12/16/2019.12.16.877829.abstract.

Alamia, Andrea et al. (2020). "Differential involvement of EEG oscillations in identity vs. spatial-relation reasoning tasks". In: *(Submitted)*. ISSN: 1534-7362. DOI: 10.1167/19.10.44b.

Alvarez, G a and P Cavanagh (2004). "The capacity of visual short-term memory is set both by visual information load and by number of objects." In: *Psychological science : a journal of the American Psychological Society / APS* 15.2, pp. 106–111. ISSN: 0956-7976. DOI: 10.1167/2.7.273.

Amorapanth, Px, P Widick, and a Chatterjee (2010). "The neural basis for spatial relations". In: *Journal of Cognitive ...* 22.8, pp. 1739–1753. DOI: 10.1162/jocn.2009.21322.The. URL: http://www.mitpressjournals.org/doi/abs/10.1162/jocn.2009.21322.

Andersen, H.H. et al. (1995). *Linear and Graphical Models for the Multivariate Complex Normal Distribution*. Ed. by P. Diggle et al. New York: Springer. ISBN: 1359-7345. DOI: 10.1016/S0733-8619(03)00096-3. arXiv: 1707.04192. URL: http://linkinghub.elsevier.com/retrieve/pii/S0733861903000963.

Angelucci, Alessandra and S. Shushruth (2013). "Beyond the Classical Receptive Field: Surround Modulation in Primary Visual Cortex ." In: *The New Visual Neurosciences*. Ed. by John S. Werner and Leo M. Chalupa. Cambridge: MIT Press, pp. 425–444.

Arenas, Alex and D Albert (2008). "Synchronization in complex networks". In: pp. 1–80. arXiv: arXiv:0805.2976v3.

Awh, Edward, Brian Barton, and Edward K Vogel (2007). "Visual Working Memory Represents a Fixed Number of Items Regardless of Complexity". In: *Psychological Science* 18.7, pp. 622–628. ISSN: 0956-7976. DOI: 10.1111/j.1467-9280.2007.01949.x.

Bamber, D. (1972). "Reaction times and error rates for judging nominal identity of letter strings". In: *Perception & Psychophysics* 12, pp. 321–326.

Bamber, D. and S. Paine (1973). "Information retrieval processes in same-different judgments of letter strings". In: *Attention and Performance IV*. Ed. by S. Kornblum. New York, NY: Academic Press.

Bashan, Amir et al. (2012). "Network physiology reveals relations between network topology and physiological function". In: *Nature Communications* 3. ISSN: 20411723. DOI: 10.1038/ncomms1705. arXiv: 1203.0242.

Benedetto, John J and Matthew Fickus (2003). "Finite normalized tight frames". In: *Advances in Computational Mathematics* 18, pp. 357–385. DOI: 10.1023/A.

Bever, Thomas G. (1988). "The psychological reality of grammar: A student's eye view of cognitive science". In: *Giving Birth to Cognitive Science: A Festschrift for George A. Miller*. Ed. by W. Hirst. Cambridge, UK: Cambridge University Press.

Biederman, I (1987). "Recognition-by-components: a theory of human image understanding." In: *Psychol. Rev.* 94.2, pp. 115–147. ISSN: 0033-295X. DOI: 10.1037/0033-295X.94.2.115.

Bienenstock, E et al. (July 1987). "A Neural Network for Invariant Pattern Recognition". en. In: *Europhysics Letters (EPL)* 4.1, pp. 121–126. ISSN: 0295-5075. DOI: 10.1209/0295-5075/4/1/020. URL: http://iopscience.iop.org/article/10.1209/0295-5075/4/1/020.

Blasius, B, A Huppert, and L Stone (1999). "Complex dynamics and phase synchronisation in spatially extended systems". In: *Nature* 397.May, pp. 354–359.

Bonilla, L. L., C. J. Pérez Vicente, and J. M. Rubí (1993). "Glassy synchronization in a population of coupled oscillators". In: *Journal of Statistical Physics* 70.3-4, pp. 921–937. ISSN: 00224715. DOI: 10.1007/BF01053600.

Brady, Timothy F (2011). "Structured representations in visual working memory". In: 2006, pp. 1–195. URL: http://dspace.mit.edu/handle/1721.1/68420%7B%5C%%7D5Cnpapers3://publication/uuid/46F14103-B182-4D3B-A3F0-94500E729582.

Brady, Timothy F. and Joshua B. Tenenbaum (2013). "A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates." In: *Psychological Review* 120.1, pp. 85–109. ISSN: 1939-1471. DOI: 10.1037/a0030779. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0030779.

Brady, Timothy F and George A Alvarez (2015). "Contextual effects in visual working memory reveal hierarchically structured memory representations". In: *Journal of Vision* 15.2015, pp. 1–69. ISSN: 15347362. DOI: 10.1167/15.15.6.doi.

Breakspear, Michael, Stewart Heitmann, and Andreas Daffertshofer (2010). "Generative Models of Cortical Oscillations: Neurobiological Implications of the Kuramoto Model". In: *Frontiers in Human Neuroscience* 4.November, pp. 1–14. ISSN: 1662-5161. DOI: 10.3389/fnhum.2010.00190. URL: http://journal.frontiersin.org/article/10.3389/fnhum.2010.00190/abstract.

Brede, M. (2008a). "Locals vs. global synchronization in networks of non-identical Kuramoto oscillators". In: *European Physical Journal B* 62.1, pp. 87–94. ISSN: 14346028. DOI: 10.1140/epjb/e2008-00126-9.

Brede, Markus (2008b). "Synchrony-optimized networks of non-identical Kuramoto oscillators". In: *Physics Letters, Section A: General, Atomic and Solid State Physics* 372.15, pp. 2618–2622. ISSN: 03759601. DOI: 10.1016/j.physleta.2007.11.069. arXiv: 0809.4531.

Brede, Markus, Massimo Stella, and Alexander Kalloniatis (2018). "Competitive influence maximization and enhancement of synchronization in populations of non-identical Kuramoto oscillators". In: *Scientific Reports* 8.1, pp. 1–9. ISSN: 20452322. DOI: 10.1038/s41598-017-18961-z. URL: http://dx.doi.org/10.1038/s41598-017-18961-z.

Bromley, Jane et al. (1994). "Signature verification using a "siamese" time delay neural network". In: *Advances in Neural Information Processing Systems*, pp. 737–744.

Bronski, Jared C et al. (2016). "The stability of fixed points for a Kuramoto model with Hebbian interactions". In: pp. 1–11. arXiv: arXiv:1611.09941v1.

Brown, Eric, Jeff Moehlis, and Philip Holmes (2004). "On the Phase Reduction and Response Dynamics of Neural Oscillator Populations". In: *Neural Computation* 16.4, pp. 673–715. ISSN: 08997667. DOI: 10.1162/089976604322860668.

Brzezicka, Aneta, Adam N Mamelak, and Ueli Rutishauser (2020). "Combined Phase-Rate Coding by Persistently Active Neurons as a Mechanism for Maintaining Multiple Items in Working Memory in Humans Report Combined Phase-Rate Coding by Persistently Active Neurons as a Mechanism for Maintaining Multiple Items in Working M". In: pp. 1–9. DOI: 10.1016/j.neuron.2020.01.032.

Burkell, Jacquelyn A. and Zenon W. Pylyshyn (1997). "Searching through subsets: A test of the visual indexing hypothesis". In: *Spatial Vision* 11.2, pp. 225–258. ISSN: 01691015. DOI: 10.1163/156856897X00203.

Burmann, Britta, Guido Dehnhardt, and Björn Mauck (2005). "Visual information processing in the lion-tailed macaque (Macaca silenus): Mental rotation or rotational invariance?" In: *Brain, Behavior and Evolution* 65.3, pp. 168–176. ISSN: 00068977. DOI: 10.1159/000083626.

Burrows, D. (1972). "Modality effects in retrieval of information from short-erm memory". In: *Perception and Psychophysics* 11, pp. 365–372.

Buzsaki, Gyorgi and Xiao-jing Wang (2012). "Mechanisms of Gamma Oscillations". In: *Annual review of neuroscience* 6.9, pp. 2166–2171. ISSN: 00092665. DOI: 10.1021/nl061786n.Core-Shell. arXiv: NIHMS150003.

Cabral, Joana et al. (2012). "Modeling the outcome of structural disconnection on resting-state functional connectivity". In: *NeuroImage* 62.3, pp. 1342–1353. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2012.06.007. URL: http://dx.doi.org/10.1016/j.neuroimage.2012.06.007.

Cabral, Joana et al. (2014). "Exploring mechanisms of spontaneous functional connectivity in MEG: How delayed network interactions lead to structured amplitude envelopes of band-pass filtered oscillations". In: *NeuroImage* 90, pp. 423–435. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2013.11.047. URL: http://dx.doi.org/10.1016/j.neuroimage.2013.11.047.

Calvo, Paco and John Symons (2014). *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*. Cambridge: MIT Press.

Cardin, Jessica A. et al. (2009). "Driving fast-spiking cells induces gamma rhythm and controls sensory responses". In: *Nature* 459.7247, pp. 663–667. ISSN: 00280836. DOI: 10.1038/nature08002.

Carlson, Laura A. and Gordan D. Logan (2001). "Using spatial terms to select an object." In: *Memory & cognition* 29.6, pp. 883–892. ISSN: 0090-502X. DOI: 10.3758/BF03196417.

Carpenter, P a and P Eisenberg (1978). "Mental rotation and the frame of reference in blind and sighted individuals." In: *Perception & psychophysics* 23.2, pp. 117–124. ISSN: 0031-5117. DOI: 10.3758/BF03208291.

Chalmers, David J (1993). "Why Fodor and Pylyshyn Were Wrong : The Simplest Refutation". In: *Cognitive Science*, pp. 340–347.

Chalvidal, Mathieu et al. (2020). "Neural Optimal Control for Representation Learning". In: *Neural Information Processing Systems (NeurIPS)*. Vancouver: Curran Associates, Inc.

Chavanis, Pierre-Henri (May 2014). "The Brownian mean field model". In: *The European Physical Journal B* 87.5. ISSN: 1434-6036. DOI: 10.1140/epjb/e2014-40586-6. URL: http://dx.doi.org/10.1140/epjb/e2014-40586-6.

Chen, Minggui et al. (2014). "Incremental Integration of Global Contours through Interplay between Visual Cortical Areas". In: *Neuron* 82.3, pp. 682–694. ISSN: 10974199. DOI: 10.1016/j.neuron.2014.03.023.

Chikkerur, Sharat et al. (2010). "What and where: A Bayesian inference theory of attention". In: *Vision Research* 50.22, pp. 2233–2247. ISSN: 00426989. DOI: 10.1016/j.visres.2010.05.013. URL: http://dx.doi.org/10.1016/j.visres.2010.05.013.

Chollet, François (2019). "On the Measure of Intelligence". In: pp. 1–64. arXiv: 1911.01547. URL: http://arxiv.org/abs/1911.01547.

Chomsky, Noam (2009). *Cartesian Linguistics*. Ed. by J. McGilvray. 3rd. Cambridge, UK: Cambridge University Press, pp. 1–158. ISBN: 9780511803116. DOI: 10.1017/CBO9780511803116. URL: http://www.scopus.com/inward/record.url?eid=2-s2.0-84925072526%7B%5C&%7DpartnerID=tZOtx3y1.

Cimpoi, M, S. Maji, and I. Kokkinos (2014). "Describing textures in the wild". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ciobotaru, Corina et al. (2018). "Mean field repulsive Kuramoto models : Phase locking and spatial signs". In: 1.1, pp. 1–18. arXiv: arXiv:1803.02647v1.

Clark, Herbert H (1972). "On the process of comparing sentence agaist pictures". In: *Cognitive Psychology* 3, pp. 472–517.

— (1973). "Space, time, semantics, and the child." In: *Cognitive development and the acquisition of language*. Ed. by T.E Moore. New York, NY: Academic Press, pp. 27–63.

Clevenger, Pamela E and John E Hummel (2014). "Working memory for relations among objects". In: *Attention, Perception, & Psychophysics* 76.December 2013, pp. 1933–1953. ISSN: 1943-393X. DOI: 10.3758/s13414-013-0601-3.

Cohn, Harvey (1960). "Global Equilibrium Theory of Charges on a Circle". In: *The American Mathematical Monthly* 67.4, pp. 338–343.

Constantinidis, Christos and Torkel Klingberg (2016). "The neuroscience of working memory capacity and training". In: *Nature Publishing Group* 17.7, pp. 438–449. ISSN: 1471-0048. DOI: 10.1038/nrn.2016.43. URL: http://dx.doi.org/10.1038/nrn.2016.43%7B%5C%%7D5Cnpapers3://publication/doi/10.1038/nrn.2016.43.

Cooper, Lynn A. (1975). "Mental Rotation of Random Two-Dimensional Shapes". In: *Cognitive Psychology* 7, pp. 20–43.

Cooper, Lynn a. (1976). "Demonstration of a mental analog of an external rotation". In: *Perception & Psychophysics* 19.4, pp. 296–302. ISSN: 0031-5117. DOI: 10.3758/BF03204234.

Cooper, Lynn A. and Roger N. Shepard (1973). "Chronometric studies of the rotation of mental images". In: *Visual information processing*. Ed. by W. G. Chase. New York, NY.

Courtney, S. M. et al. (1996). "Object and Spatial Visual Working Memory Activate Separate Neural Systems in Human Cortex". In: *Cereb. Cortex* 6.1, pp. 39–49. ISSN: 1047-3211. DOI: 10.1093/cercor/6.1.39. URL: http://cercor.oxfordjournals.org/content/6/1/39.abstract.

Cowan, N. (2001). "The magical number 4 in short term memory. A reconsideration of storage capacity". In: *Behavioral and Brain Sciences* 24.4, pp. 87–186. ISSN: 0140525X. DOI: 10.1017/S0140525X01003922. arXiv: 0140-525X.

Cybenko, G. (1989). "Approximation by Superpositions of a Sigmoidal Function". In: *Mathematics of control, signals and systems* 9.3, pp. 303–314. ISSN: 10009221. DOI: 10.1007/BF02836480.

Daido, Hiroaki (1987). "Population Dynamics of Randomly Interacting Self-Oscillators". In: *Progress of Theoretical Physics* 77.3, p. 622634.

— (1992). "Quasientrainment and slow relaxation in a population of oscillators with random and frustrated interactions". In: *Physical Review Letters* 68.7, pp. 1073–1076. ISSN: 00319007. DOI: 10.1103/PhysRevLett.68.1073.

— (2000). "Self-averaging of an order parameter in randomly coupled limit-cycle oscillators". In: *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* 61.2, pp. 2148–2149. ISSN: 1063651X. DOI: 10.1103/PhysRevE.61.2148.

Daniel, Thomas A., Anthony A. Wright, and Jeffrey S. Katz (2015). "Abstract-concept learning of difference in pigeons". In: *Animal Cognition* 18.4, pp. 831–837. ISSN: 14359448. DOI: 10.1007/s10071-015-0849-1.

Dean, Jeff, David Patterson, and Cliff Young (2018). "A New Golden Age in Computer Architecture: Empowering the Machine-Learning Revolution". In: *IEEE Micro* 38.2, pp. 21–29. ISSN: 02721732. DOI: 10.1109/MM.2018.112130030.

Delis, Dean C., Lynn C. Robertson, and Robert Efron (1986). "Hemispheric specialization of memory for visual hierarchical stimuli". In: *Neuropsychologia* 24.2, pp. 205–214. ISSN: 00283932. DOI: 10.1016/0028-3932(86)90053-9.

Delius, Juan D. (1994). "Comparative Cognition of Identity". In: *XXV International Congress of Psychology*. Ed. by Paul Bertselson, Paaul Eelen, and Gery D'Ydewalle. Vol. 1. January 1994. Burssels, pp. 25–40. ISBN: 0863772986.

Delius, Juan D. and Valerie D. Hollard (1987). "Rotational Invariance in Visual Pattern Recognition by Pigeons and Humans". In: *Science* 218, pp. 804–806.

Deng, J. et al. (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *Conference on Computer Vision and Pattern Recognition*.

Dessalegn, B. and B. Landau (2008). "More than meets the eye: the role of language in binding and maintaining feature conjunctions". In: *Psychological Science* 19, pp. 189–195.

Diaz-Guilera, Albert et al. (Feb. 2009). "SYNCHRONIZATION IN RANDOM GEOMETRIC GRAPHS". In: *International Journal of Bifurcation and Chaos* 19.02, pp. 687–693. ISSN: 1793-6551. DOI: 10.1142/s0218127409023044. URL: http://dx.doi.org/10.1142/S0218127409023044.

DiCarlo, James J. and John H. R. Maunsell (2003). "Anterior Inferotemporal Neurons of Monkeys Engaged in Object Recognition Can be Highly Sensitive to Object Retinal Position". In: *Journal of Neurophysiology* 89.6, pp. 3264–78. ISSN: 0022-3077. DOI: 10.1152/jn.00358.2002. URL: http://jn.physiology.org/cgi/doi/10.1152/jn.00358.2002%7B%5C%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/12783959.

DiCarlo, James J and David D Cox (2007). "Untangling invariant object recognition". In: *Trends Cogn. Sci.* 11.8, pp. 333–341. ISSN: 13646613. DOI: 10.1016/j.tics.2007.06.010.

Donderi, D C (1983). "Acquisition and decision in visual same–different search of letter displays." In: *Perception and Psychophysics* 33.3, pp. 271–282. ISSN: 0031-5117. DOI: 10.3758/BF03202865.

Donderi, D. and Dorothy Zelnicker (1969). "Parallel processing in visual same-different". In: *Perception & psychophysics* 5.4, pp. 197–200.

Doumas, Leonidas A. A., John E Hummel, and Catherine M Sandhofer (2008). "A theory of the discovery and predication of relational concepts." In: *Psychological review* 115.1, pp. 1–43. ISSN: 0033-295X. DOI: 10.1037/0033-295X.115.1.1. URL: http://babytalk.psych.ucla.edu/documents/Doumas%7B%5C_%7Det%7B%5C_%7Dal%7B%5C_%7D2008.pdf.

Dubey, Rachit et al. (2015). "What makes an object memorable?" In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1089–1097.

Eberhardt, Sven, Jonah Cader, and Thomas Serre (2016). "How Deep is the Feature Analysis Underlying Rapid Visual Categorization ?" In: *Neural Information Processing Systems*. Ed. by DD Lee et al. Red Hook, NY: Curran Associates, pp. 1100–8. arXiv: arXiv:1409.1556.

Edelman, G.M. (1978). *The Mindful Brain*. Cambridge, MA: MIT Press.

Edelman, Shimon and Nathan Intrator (2000). "Coarse coding of shape fragments plus retinotopy equals representation of structure". In: *Spat. Vis.* 13.2, pp. 255–264. ISSN: 0169-1015. DOI: 10.1163/156856800741072. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Ddb=PubMed%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Ddopt=Citation%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Dlist%7B%5C%%7D7B%7B%5C_%7D%7B%5C%%7D7Duids=11198236%7B%5C$%7D%7B%5C%%7D5C%7B%5C$%7Dnhttp://www.ncbi.nlm.nih.gov/pubmed/11198236%7B%5C$%7D%7B%5C%%7D5C%7B%5C$%7Dnhttp://booksandjournals.brillonline.com/content/journals/10.1163/1568568007.

Egeth, H.E. (1966). "Parallel versus serial processes in multidimensional stimulus discrimination". In: *Perception & Psychophysics* 1, pp. 245–252.

Ellis, Kevin, Armando Solar-lezama, and Joshua B Tenenbaum (2015). "Unsupervised Learning by Program Synthesis". In: *NIPS*, pp. 1–9. ISSN: 10495258.

Erdogan, Goker and Robert A Jacobs (Nov. 2017). "Visual shape perception as Bayesian inference of 3D object-centered shape representations". en. In: *Psychol. Rev.* 124.6, pp. 740–761.

Erdös, Paul and Alfréd Rényi (1959). "On random graphs I." In: *Publicationes Mathematicae* 6, pp. 290–297.

Eriksen, C W, W P O'Hara, and B Eriksen (1982). "Response competition effects in same-different judgments." In: *Perception & psychophysics* 32.3, pp. 261–270. ISSN: 0031-5117. DOI: 10.3758/BF03206230.

Eriksen, C W and D W Schultz (1979). "Information processing in visual search: a continuous flow conception and experimental results." In: *Perception & psychophysics* 25.4, pp. 249–263. ISSN: 0031-5117. DOI: 10.3758/BF03198804.

— (1977). "Retinal locus and acuity in visual information processing". In: *Bulletin of the Psychonomic Society* 9.2, pp. 81–84.

Ermentrout, G. B. and N. Kopell (1991). "Multiple pulse interactions and averaging in systems of coupled neural oscillators". In: *Journal of Mathematical Biology* 29.3, pp. 195–217. ISSN: 14321416. DOI: 10.1007/BF00160535.

Ermentrout, George Bard (1996). "Type I Membranes, Phase Resetting Curves, and Synchrony". In: *Neural Computation* 8.979.

Ermentrout, George Bard and Nancy Kopell (1984). "Frequency Plateaus in a Chain of Weakly Coupled Oscillators, I." In: *SIAM Journal on Mathematical Analysis* 15.2, pp. 215–237. ISSN: 0036-1410. DOI: 10.1137/0515019.

Eurich, Christian W and Helmut Schwegler (1997). "Coarse coding: calculation of the resolution achieved by a population of large receptive field neurons". In: *Biol. Cybern* 76, pp. 357–363.

Evans, Karla K and Anne Treisman (2005). "Perception of objects in natural scenes: is it really attention free?" In: *Journal of experimental psychology. Human perception and performance* 31.6, pp. 1476–1492. ISSN: 0096-1523. DOI: 10.1037/0096-1523.31.6.1476.

Eviatar, Z, E Zaidel, and T Wickens (1994). "Nominal and physical decision criteria in same-different judgments." In: *Perception & psychophysics* 56.1, pp. 62–72. ISSN: 0031-5117. DOI: 10.3758/BF03211691.

Fabiani, Miguel, Demetrios Karis, and Emanuel Donchin (1986). "P300 and recall in an incidental memory paradigm." In: *Psychophysiology* 23 3, pp. 298–308.

Farid, Hany (2002). "Temporal synchrony in perceptual grouping: A critique". In: *Trends in Cognitive Sciences* 6.7, pp. 284–288. ISSN: 13646613. DOI: 10.1016/S1364-6613(02)01927-7.

Feldman, Jacob and Patrice D. Tremoulet (2006). "Individuation of visual objects over time". In: *Cognition* 99.2, pp. 131–165. ISSN: 00100277. DOI: 10.1016/j.cognition.2004.12.008.

Fell, Juergen et al. (2003). "Is synchronized neuronal gamma activity relevant for selective attention?" In: *Brain Research Reviews* 42.3, pp. 265–272. ISSN: 01650173. DOI: doi:DOI:10.1016/S0165-0173(03)00178-4. URL: http://www.sciencedirect.com/science/article/B6SYS-48JSR90-1/2/999000e842097afdf4d6c4242dca436a.

Fiebig, Florian, Pawel Herman, and Anders Lansner (2020). "An Indexing Theory for Working Memory Based on Fast Hebbian Plasticity". In: *Eneuro* 7.2, ENEURO.0374–19.2020. ISSN: 23732822. DOI: 10.1523/eneuro.0374-19.2020.

Filatrella, Giovanni, Arne Hejde Nielsen, and Niels Falsig Pedersen (2008). "Analysis of a power grid using a Kuramoto-like model". In: *The European Physical Journal B* 61, pp. 485–491.

Finger, Holger and Peter König (2014). "Phase synchrony facilitates binding and segmentation of natural images in a coupled neural oscillator network". In: *Frontiers in Computational Neuroscience* 7.January, pp. 1–21. ISSN: 1662-5188. DOI: 10.3389/

fncom.2013.00195. URL: http://journal.frontiersin.org/article/10.3389/fncom.2013.00195/abstract.

Fink, G.R. et al. (1996). "Where in the brain does visual attention select the forest and the trees?" In: *Nature* 382, pp. 626–628.

Fink, Gereon R. et al. (1997). "Neural mechanisms involved in the processing of global and local aspects of hierarchically organized visual stimuli". In: *Brain* 120.10, pp. 1779–1791. ISSN: 00068950. DOI: 10.1093/brain/120.10.1779.

Fischer, Asja and Christian Igel (2014). "Training restricted Boltzmann machines: An introduction". In: *Pattern Recognition* 47.1, pp. 25–39. ISSN: 00313203. DOI: 10.1016/j.patcog.2013.05.025.

Fleuret, François et al. (2011). "Comparing machines and humans on a visual categorization test." In: *Proceedings of the National Academy of Sciences of the United States of America* 108.43, pp. 17621–5. ISSN: 1091-6490. DOI: 10.1073/pnas.1109168108. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3203755%7B%5C&%7Dtool=pmcentrez%7B%5C&%7Drendertype=abstract.

Fodor, Jerry and Brian P. McLaughlin (1990). "Connectionism and the problem of systematicity (continued): Why Smolensky's solution still doesn't work". In: *Cognition* 35.1, pp. 183–204. ISSN: 00100277. DOI: 10.1016/S0010-0277(96)00780-9.

Fodor, Jerry and Zenon Pylyshyn (1988). "Connectionism and cognitive architecture: A critical analysis". In: *Cognition* 28.1-2, pp. 3–71. ISSN: 00100277. DOI: 10.1016/0010-0277(88)90031-5. arXiv: arXiv:1011.1669v3. URL: http://linkinghub.elsevier.com/retrieve/pii/0010027788900315%7B%5C%%7D5Cnpapers2://publication/uuid/EF730E3F-DF39-4397-BA0F-D68A9F483A0F.

— (2014). *Minds without Meaning*. Cambridge, MA: MIT Press.

Fougnie, Daryl, Christopher L. Asplund, and Rene Marois (2010). "What are the units of storage in visual working memory?" In: *Journal of Vision* 10.12, p. 27. DOI: 10.1167/10.12.27.What.

Franconeri, Steven L. et al. (2012). "Flexible visual processing of spatial relationships". In: *Cognition* 122.2, pp. 210–227. ISSN: 00100277. DOI: 10.1016/j.cognition.2011.11.002. URL: http://dx.doi.org/10.1016/j.cognition.2011.11.002.

Franconeri, Steven L, George a Alvarez, and James T Enns (2007). "How many locations can be selected at once?" In: *Journal of experimental psychology. Human perception and performance* 33.5, pp. 1003–12. ISSN: 0096-1523. DOI: 10.1037/0096-1523.33.5.1003. URL: http://www.ncbi.nlm.nih.gov/pubmed/17924803.

Freund, Y and Re Schapire (1995). "A desicion-theoretic generalization of on-line learning and an application to boosting". In: *Computational learning theory* 55,

pp. 119–139. ISSN: 00220000. DOI: 10.1006/jcss.1997.1504. URL: http://link.springer.com/chapter/10.1007/3-540-59119-2%7B%5C_%7D166.

Freyer, Frank et al. (2009). "Bistability and non-Gaussian fluctuations in spontaneous cortical activity". In: *Journal of Neuroscience* 29.26, pp. 8512–8524. ISSN: 02706474. DOI: 10.1523/JNEUROSCI.0754-09.2009.

Fries, Pascal (2005). "A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence". In: *Trends in Cognitive Sciences* 9.10, pp. 474–480. ISSN: 13646613. DOI: 10.1016/j.tics.2005.08.011. arXiv: 1111.7219.

— (2015). "Rhythm for Cognition: Communication Through Coherence". In: *Neuron* 88.1, pp. 220–235. DOI: 10.1016/j.neuron.2015.09.034.Rhythms.

Fukuda, Hirokazu et al. (2007). "Synchronization of plant circadian oscillators with a phase delay effect of the vein network". In: *Physical Review Letters* 99.9, pp. 1–4. ISSN: 00319007. DOI: 10.1103/PhysRevLett.99.098102.

Fukushima, Kunihiko (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological Cybernetics* 36.4, pp. 193–202. ISSN: 03401200. DOI: 10.1007/BF00344251.

Fuster, J. M. (1997). *The Prefrontal Cortex: Anatomy, Physiology, and Neurophysiology of the Frontal Lobe*. New York, NY: Lippincott-Raven.

Gallier, Jean (2019). "Spectral Theory of Unsigned and Signed Graphs. Applications to Graph Clustering: a Survey". In: arXiv: 1601.04692. URL: http://arxiv.org/abs/1601.04692.

Gallistel, CR. and Adam King (2009). *Memory and the Computational Brain: Why Cognitive Science will Transform Neuroscience*. 1st ed. Wiley-Blackwell. ISBN: 1405122889.

Garcia-Ojalvo, Jordi, Michael B. Elowitz, and Steven H. Strogatz (2004). "Modeling a synthetic multicellular clock: Repressilators coupled by quorum sensing". In: *Proceedings of the National Academy of Sciences* 101.30, pp. 10955–10960. ISSN: 0027-8424. DOI: 10.1073/pnas.0307095101. eprint: https://www.pnas.org/content/101/30/10955.full.pdf. URL: https://www.pnas.org/content/101/30/10955.

Garlaschelli, D. et al. (May 2007). "Interplay between topology and dynamics in the World Trade Web". In: *The European Physical Journal B* 57.2, pp. 159–164. ISSN: 1434-6036. DOI: 10.1140/epjb/e2007-00131-6. URL: http://dx.doi.org/10.1140/epjb/e2007-00131-6.

Garnham, A. (1989). "A unified theory of the meaning of some spatial relational terms". In: *Cognition* 31, pp. 45–60.

Geman, Donald et al. (2015). "Visual Turing test for computer vision systems." In: *Proc. Natl. Acad. Sci. U. S. A.* 112.12, pp. 3618–3623. ISSN: 1091-6490. DOI: 10.1073/pnas.1422953112. arXiv: arXiv:1406.2375. URL: http://www.pnas.org/content/112/12/3618.short.

Geman, Stuart and Donald Geman (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6, pp. 721–741. ISSN: 01628828. DOI: 10.1109/TPAMI.1984.4767596.

Gentner, Dedre (2010). "Bootstrapping the mind: Analogical processes and symbol systems". In: *Cognitive Science* 34.5, pp. 752–775. ISSN: 03640213. DOI: 10.1111/j.1551-6709.2010.01114.x.

Giurfa, M et al. (2001). "The concepts of 'sameness' and 'difference' in an insect". In: *Nature* 410.6831, pp. 930–933. ISSN: 0028-0836. DOI: 10.1038/35073582.

Gleiser, P. M. and D. H. Zanette (2006). "Synchronization and structure in an adaptive oscillator network". In: *European Physical Journal B* 53.2, pp. 233–238. ISSN: 14346028. DOI: 10.1140/epjb/e2006-00362-y.

Glorot, Xavier and Yoshua Bengio (13–15 May 2010). "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, pp. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html.

Goel, Vinod and Raymond J. Dolan (2004). "Differential involvement of left prefrontal cortexin inductive and deductive reasoning". In: *Cognition* 93.3. ISSN: 00100277. DOI: 10.1016/j.cognition.2004.03.001.

Golde, Maria, D. Yves von Cramon, and Ricarda I. Schubotz (2010). "Differential role of anterior prefrontal and premotor cortex in the processing of relational information". In: *NeuroImage* 49.3, pp. 2890–2900. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2009.09.009. URL: http://dx.doi.org/10.1016/j.neuroimage.2009.09.009.

Gómez-Gardeñes, Jesús et al. (2010). "From modular to centralized organization of synchronization in functional areas of the cat cerebral cortex". In: *PLoS ONE* 5.8. ISSN: 19326203. DOI: 10.1371/journal.pone.0012313.

Goodfellow, Ian J. et al. (2014). "Generative Adversarial Nets". In: pp. 1–9. ISSN: 10495258. DOI: 10.1001/jamainternmed.2016.8245. arXiv: 1406.2661. URL: http://arxiv.org/abs/1406.2661.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2015). "Explaining and Harnessing Adversarial Examples". In: *Iternational Conference on Learning Representations*.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.

Gottwald, Georg A. (2015). "Model reduction for networks of coupled oscillators". In: *Chaos* 25.5. ISSN: 10541500. DOI: 10.1063/1.4921295. arXiv: 1505.05243. URL: http://dx.doi.org/10.1063/1.4921295.

Gray, Charles M and Wolf Singer (1989). "Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex". In: 86.March, pp. 1698–1702.

Gray, Charles M et al. (1989). "Oscillatory responses in cat visual cortex exhibit intercolumnar synchronization which reflects global stimulus properties". In: 338.March.

Griffiths, D.J. (1999). *Introduction to Electrodynamics*. 3rd. New York: Pearson Education Inc.

Grossberg, Stephen, Ennio Mingolla, and William D. Ross (1997). "Visual brain and visual perception: How does the cortex do perceptual grouping?" In: *Trends in Neurosciences* 20.3, pp. 106–111. ISSN: 01662236. DOI: 10.1016/S0166-2236(96)01002-8.

Guberman, Nitzan (2016). "On Complex Valued Convolutional Neural Networks". In: arXiv: 1602.09046. URL: http://arxiv.org/abs/1602.09046.

Guckenheimer, J. and P. Holmes (1991). *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. New York: Springer.

Guclu, U and M A J van Gerven (July 2015). "Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream". In: *Journal of Neuroscience* 35.27, pp. 10005–10014.

Gülçehre, Çağlar and Yoshua Bengio (2013). "Knowledge Matters : Importance of Prior Information for Optimization". In: *arXiv preprint arXiv:1301.4083*, pp. 1–12. arXiv: 1301.4083. URL: http://arxiv.org/abs/1301.4083.

Gupta, Shamik, Alessandro Campa, and Stefano Ruffo (Aug. 2014a). "Kuramoto model of synchronization: equilibrium and nonequilibrium aspects". In: *Journal of Statistical Mechanics: Theory and Experiment* 2014.8, R08001. DOI: 10.1088/1742-5468/14/08/r08001. URL: https://doi.org/10.1088%2F1742-5468%2F14%2F08%2Fr08001.

— (2014b). "Nonequilibrium first-order phase transition in coupled oscillator systems with inertia and noise". In: pp. 1–13. arXiv: arXiv:1309.0035v2.

Ha, David, Andrew Dai, and Quoc V. Le (2014). "Hypernetworks". In: *Hypernetworks in the Science of Complex Systems*, pp. 151–176. DOI: 10.1142/9781860949739_0006. arXiv: 1609.09106.

Ha, Seung Yeal, Zhuchun Li, and Xiaoping Xue (2013). "Formation of phase-locked states in a population of locally interacting Kuramoto oscillators". In: *Journal of Differential Equations* 255.10, pp. 3053–3070. ISSN: 00220396. DOI: 10.1016/j.jde.2013.07.013. URL: http://dx.doi.org/10.1016/j.jde.2013.07.013.

Ha, Seung-yeal, Jaeseung Lee, and Zhuchun Li (2018). "Synchronous harmony in an ensemble of Hamiltonian mean-field oscillators and inertial Kuramoto oscillators". In: 113112.July. DOI: 10.1063/1.5047392. URL: http://dx.doi.org/10.1063/1.5047392.

Ha, Seung-yeal, Se Eun Noh, and Jinyeong Park (2016). "Synchronization of Kuramoto Oscillators with Adaptive Couplings ". In: *SIAM Journal on Applied Dynamical Systems* 15.1, pp. 162–194.

Halford, G S, W H Wilson, and S Phillips (1998). "Processing capacity defined by relational complexity: implications for comparative, developmental, and cognitive psychology." In: *The Behavioral and brain sciences* 21.6, pp. 803–864. ISSN: 0140-525X. DOI: 10.1017/S0140525X98001769. URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed%7B%5C&%7Did=10191879%7B%5C&%7Dretmode=ref%7B%5C&%7Dcmd=prlinks.

Halford, Graeme S., William H. Wilson, and Steven Phillips (2010). "Relational knowledge: The foundation of higher cognition". In: *Trends in Cognitive Sciences* 14.11, pp. 497–505. ISSN: 13646613. DOI: 10.1016/j.tics.2010.08.005. URL: http://dx.doi.org/10.1016/j.tics.2010.08.005.

Ham, Ineke J M van der et al. (2012). "Retinotopic mapping of categorical and coordinate spatial relation processing in early visual cortex". In: *PLoS ONE* 7.6, pp. 1–8. ISSN: 19326203. DOI: 10.1371/journal.pone.0038644.

Hannay, H J, N R Varney, and A L Benton (1976). "Visual localization in patients with unilateral brain disease." In: *Journal of neurology, neurosurgery, and psychiatry* 39.4, pp. 307–313. ISSN: 0022-3050. URL: http://www.ncbi.nlm.nih.gov/pubmed/932747.

Hansel, David, G Mato, and C. Meunier (1993). "Phase Dynamics for Weakly Coupled Hodgkin-Huxley Neurons". In: *Europhysics Letters* 23.5, pp. 367–72. DOI: 10.1209/0295-5075/23/5/011.

Harris, K D and A Thiele (2011). "Cortical state and attention". In: *Nat Rev Neurosci* 12.9, pp. 509–523. ISSN: 1471-0048. DOI: 10.1038/nrn3084. URL: http://www.ncbi.nlm.nih.gov/pubmed/21829219%7B%5C%%7D5Cnhttp://www.nature.com/nrn/journal/v12/n9/pdf/nrn3084.pdf.

Hawkins, H. L., G. J. McDonald, and A. K. Cox (1973). "Effects of irrelevant information in speeded discrimination". In: *Journal of Experimental Pscyhology* 98, pp. 435–437.

Hawkins, H. L. and R.H. Shigley (1972). "Irrelevant information and processing mode in speeded discrimination". In: *Journal of Experimental Psychology* 96, p. 389.

Hayworth, K J, M D Lescroart, and I Biederman (2011). "Neural encoding of relative position". In: *J Exp Psychol Hum Percept Perform* 37.4, pp. 1032–1050. ISSN: 1939-1277. DOI: 10.1037/a0022338. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve%7B%5C&%7Ddb=PubMed%7B%5C&%7Ddopt=Citation%7B%5C&%7Dlist%7B%5C_%7Duids=21517211%20http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Ddb=PubMed%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Ddopt=Citation%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Dlist%7B%5C%%7D7B%7B%5C_%7D%7B%5C%%7D7Duids=21517211.

He, Kaiming et al. (Feb. 2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". In: *arXiv Prepr. arXiv1502.01852*, pp. 1–11.

Hécaen, H and J de Ajuriaguerra (1954). "Balint's Syndrome (Psychic Paralysis of Visual Fixation) and Its Minor Forms". In: *Brain* 77.3, pp. 373–400. DOI: 10.1093/brain/77.3.373. URL: http://brain.oxfordjournals.org/content/77/3/373.full.pdf+html?sid=4477773f-8e83-47f9-9153-238deddadbfd.

Hemmen, J.L. van and W.F. Wreszinski (1993). "Lyapunov Function for the Kuramoto Model of Nonlinearly Coupled Oscillators". In: 72, pp. 145–166.

Herskovits, A. (1986). *Language and spatial cognition: An interdisciplinary study of the prepositions in English*. Cambridge, UK: Cambridge University Press.

Hinton, G. E., James L. McClelland, and David E Rumelhart (1986). "Distributed Representations". In: *Parallel Distributed Processing: Explorations in the Microsctructure of Cognition*. Ed. by G. E. Hinton, James L. McClelland, and David E Rumelhart. Cambridge, MA: MIT Press, pp. 77–109.

Hinton, Geoffrey E and Ruslan Salakhutdinov (2006). "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.July, pp. 504–507. ISSN: 0036-8075. DOI: 10.1126/science.1127647. arXiv: 20.

Holcombe, Alex O., Daniel Linares, and Maryam Vaziri-Pashkam (2011). "Perceiving spatial relations via attentional tracking and shifting". In: *Current Biology* 21.13, pp. 1135–1139. ISSN: 09609822. DOI: 10.1016/j.cub.2011.05.031. URL: http://dx.doi.org/10.1016/j.cub.2011.05.031.

Hollard, Valerie D. and Juan D. Delius (1982). "Rotational invariance in visual pattern recognition by pigeons and humans". In: *Science* 218.4574, pp. 804–806. ISSN: 00368075. DOI: 10.1126/science.7134976.

Honey, Christopher J. and Olaf Sporns (2008). "Dynamical consequences of lesions in cortical networks". In: *Human Brain Mapping* 29.7, pp. 802–809. ISSN: 10659471. DOI: 10.1002/hbm.20579.

Hong, Ha et al. (2016). "Explicit information for category-orthogonal object properties increases along the ventral stream". In: *Nature Neuroscience* 19.4, pp. 613–622. ISSN: 1097-6256. DOI: 10.1038/nn.4247. URL: http://www.nature.com/doifinder/10.1038/nn.4247%7B%5C%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26900926.

Hong, Hyunsuk and Steven H. Strogatz (2011). "Kuramoto model of coupled oscillators with positive and negative coupling parameters: An example of conformist and contrarian oscillators". In: *Physical Review Letters* 106.5, pp. 1–4. ISSN: 00319007. DOI: 10.1103/PhysRevLett.106.054102. arXiv: arXiv:1011.1669v3.

Hopcroft, J.E., Rajeev Motwani, and Jeffrey Ullman (2008). *Introduction to Automata Theory, Languages, and Computation*. 3rd Editio. London: Pearson.

Hoppensteadt, Frank C. and Eugene M. Izhikevich (1997). *Weakly Connected Neural Networks*. Vol. 126. ISBN: 978-1-4612-7302-8. DOI: 10.1007/978-1-4612-1828-9. URL: http://link.springer.com/10.1007/978-1-4612-1828-9.

Hübner, Ronald (1998). "Hemispheric Differences in Global / Local Processing Revealed by Same- Different Judgements". In: *Visual Cognition* 5.4, pp. 457–468. DOI: 10.1080/713756793.

Hübner, Ronald and Tobias Studer (2009). "Functional hemispheric differences for the categorization of global and local information in naturalistic stimuli". In: *Brain and Cognition* 69.1, pp. 11–18. ISSN: 02782626. DOI: 10.1016/j.bandc.2008.04.009. URL: http://dx.doi.org/10.1016/j.bandc.2008.04.009.

Hummel, J E and I Biederman (1992). "Dynamic binding in a neural network for shape recognition." In: *Psychol. Rev.* 99.3, pp. 480–517. ISSN: 0033-295X. DOI: 10.1037/0033-295X.99.3.480.

Hummel, J E and Keith J. Holyoak (2003). "A symbolic-connectionist theory of relational inference and generalization". In: *Psychological Review* 110.2, pp. 220–64.

Hummel, John E and Brian J Stankiewicz (1998). "Hummel, J. E., & Stankiewicz, B. J. (1998). Two roles for attention in shape perception: A structural description model of visual scrutiny. Visual Cognition, 5, 49-79. Please note: I am not confident this document is the absolutely final (i.e., in print) ". In: *Psychology*.

Hyötyniemi, Heikki (1996). "Turing Machines are Recurrent Neural Networks". In: *STeP'96 Genes, Nets and Symbols*. Ed. by Jarmo Alander, Timo Honkela, and Matti Jakobsson. Vaasa: The Finnish Artificial Intelligence Society, pp. 13–24. URL: http://lipas.uwasa.fi/stes/step96/step96/hyotyniemi1/.

Hyvarinen, Aapo (2005). "Estimation of Non-Normalized Statistical Models by Score Matching". In: *Journal of Machine* 6, pp. 695–709.

Iatsenko, D., P. V.E. McClintock, and A. Stefanovska (2014). "Oscillator glass in the generalized Kuramoto model: synchronous disorder and two-step relaxation". In: *Nature Communications* 5, pp. 1–5. ISSN: 20411723. DOI: 10.1038/ncomms5118. arXiv: 1303.4453.

Ivković, Miloš, Amy Kuceyeski, and Ashish Raj (2012). "Statistics of weighted brain networks reveal hierarchical organization and gaussian degree distribution". In: *PLoS ONE* 7.6. ISSN: 19326203. DOI: 10.1371/journal.pone.0035029.

Ivry, R.B. and L.C. Roberston (1998). *The two sides of perception*. Cambridge, MA: MIT Press.

Izhikevich, Eugene M. (2007). *Dynamical Systems in Neuroscience*. Vol. 25. 1. Cambridge, MA: MIT Press, pp. 435–489. ISBN: 9780262090438. DOI: 10.1017/S0143385704000173. arXiv: 0310317 [math]. URL: http://books.google.com/books?hl=en%7B%5C&%7Dlr=%7B%5C&%7Did=kVjM6DFk-twC%7B%5C&%7Doi=fnd%7B%5C&%7Dpg=PR15%7B%5C&%7Ddq=Dynamical+Systems+in+Neuroscience%7B%5C&%7Dots=KRExnXb9si%7B%5C&%7Dsig=eN72JzIWk6-LfvNDSFETcxn0vyo.

James, William (1890). *The Principles of Psychology*, p. 688. ISBN: 0674705599. DOI: 10.1037/10538-000.

Jansen, Peter and Scott Watter (Mar. 2012). "Strong systematicity through sensorimotor conceptual grounding: An unsupervised, developmental approach to connectionist sentence processing". In: *Connection Science - CONNECTION* 24, pp. 1–31. DOI: 10.1080/09540091.2012.664121.

Ji, Shihao and Lawrence Carin (2010). "Bayesian compressive sensing". In: *IEEE Transactions on Signal Processing* 19.1, pp. 53–63. ISSN: 1941-0042. DOI: 10.1109/TIP.2009.2032894. URL: http://www.ncbi.nlm.nih.gov/pubmed/21245010.

Jia, Xiaoxuan, Seiji Tanabe, and Adam Kohn (2013). "Gamma and the Coordination of Spiking Activity in Early Visual Cortex". In: *Neuron* 77.4, pp. 762–774. ISSN: 08966273. DOI: 10.1016/j.neuron.2012.12.036. URL: http://dx.doi.org/10.1016/j.neuron.2012.12.036.

Jin, Y. and S. Geman (2006). "Context and Hierarchy in a Probabilistic Image Model". English. In: *2006 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. - Vol. 2*. Vol. 2. IEEE, pp. 2145–2152. ISBN: 0-7695-2597-0. DOI: 10.1109/CVPR.2006.86. URL: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=1641016.

Johnson, Justin et al. (2017). "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning". In: *Computer Vision and Pattern Recognition (CVPR)*.

Jolicoeur, P., S. Ullman, and M. Mackay (1984). "Boundary Tracing: a possible elementary operation in the perception of spatial relations". In: *Unpublished*.

Jonides, J et al. (1993). "Spatial Working-Memory in Humans As Revealed By Pet". In: *Nature* 363.6430, pp. 623–625. ISSN: 0028-0836. DOI: 10.1038/363623a0.

Kahneman, D and A Treisman (1984). *Changing views of attention and automaticity*.

Karras, Tero, Samuli Laine, and Timo Aila (2019). "A style-based generator architecture for generative adversarial networks". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* 2019-June, pp. 4396–4405. ISSN: 10636919. DOI: 10.1109/CVPR.2019.00453. arXiv: 1812.04948.

Katz, J. S. and A. A. Wirght (2006). "Same/different abstract-concept learning by pigeons". In: *Journal of Experimental Psychology: Animal Behavior Processes* 32.1, pp. 80–86.

Kauffman, Stuart A. (1992). *The Origins of Order: Self-Organization and Selection in Evolution*, pp. 61–100. ISBN: 0195058119. DOI: 10.1142/9789814415743_0003.

Kelly, David and Georg A. Gottwald (2011). "On the topology of synchrony optimized networks of a Kuramoto-model with non-identical oscillators". In: *Chaos* 21.2. ISSN: 10541500. DOI: 10.1063/1.3590855.

Kemelmacher-Shlizerman, Ira et al. (2016). "The megaface benchmark: 1 million faces for recognition at scale". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4873–4882.

Khaligh-Razavi, Seyed-Mahdi and Nikolaus Kriegeskorte (Nov. 2014). "Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation". en. In: *PLoS Comput. Biol.* 10.11. Ed. by Jörn Diedrichsen, e1003915.

Kheradpisheh, Saeed Reza et al. (Sept. 2016). "Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition". en. In: *Sci. Rep.* 6, p. 32672.

Kim, Junkyung et al. (2018). "Not-So-CLEVR : learning same – different relations strains feedforward neural networks". In: *Royal Society Interface* 8.2018011. DOI: http://dx.doi.org/10.1098/rsfs.2018.0011.

Kingma, D P and J L Ba (2015). "Adam: a Method for Stochastic Optimization". In: *International Conference on Learning Representations*.

Kirillov, Alexander et al. (2018). *Panoptic Segmentation*. arXiv: 1801.00868 [cs.CV].

Kitzbichler, Manfred G. et al. (2009). "Broadband criticality of human brain network synchronization". In: *PLoS Computational Biology* 5.3. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000314.

Knauf, Andreas (2018). "Neural Ordinary Differential Equations". In: *UNITEXT - La Matematica per il 3 piu 2* 109.NeurIPS, pp. 31–60. ISSN: 20385757. DOI: 10.1007/978-3-662-55774-7_3. arXiv: arXiv:1806.07366v5.

Knauff, Markus et al. (2003). "Reasoning, models, and images: behavioral measures and cortical activity." In: *Journal of cognitive neuroscience* 15.4, pp. 559–73. ISSN: 0898-929X. DOI: 10.1162/089892903321662949. URL: http://www.ncbi.nlm.nih.gov/pubmed/12803967.

Kok, Albert (May 2001). "On the utility of P3 amplitude as a measure of processing capacity". In: *Psychophysiology* 38, pp. 557–577. DOI: 10.1017/S0048577201990559.

Kopp, Lars (1994). *A Neural Network for Spatial Relations : Connecting Visual Scenes To Linguistic Descriptions*. Tech. rep.

Korniss, G. et al. (Feb. 2000). "From Massively Parallel Algorithms and Fluctuating Time Horizons to Nonequilibrium Surface Growth". In: *Physical Review Letters* 84.6, pp. 1351–1354. ISSN: 1079-7114. DOI: 10.1103/physrevlett.84.1351. URL: http://dx.doi.org/10.1103/PhysRevLett.84.1351.

Kosslyn, Stephen M. et al. (1989). "Evidence for two types of spatial representations: hemispheric specialization for categorical and coordinate relations." In: *Journal of Experimental Psychology: Human perception and performance* 14.4, pp. 723–35.

Kralemann, B., Arkady Pivovsky, and M Rosenblum (2014). "Reconstructing effective phase connectivity of oscillator networks from observations feature ranking Reconstructing effective phase connectivity of oscillator networks from observations". In: *New Journal of Physics* 16.085013. DOI: 10.1088/1367-2630/16/8/085013.

Kralemann, Björn, Arkady Pikovsky, and Michael Rosenblum (2011). "Reconstructing phase dynamics of oscillator networks". In: *Chaos* 21.2. ISSN: 10541500. DOI: 10.1063/1.3597647. arXiv: 1102.3064.

Kriegeskorte, Nikolaus (Nov. 2015). "Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing". en. In: *Annu Rev Vis Sci* 1, pp. 417–446.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Adv. Neural Inf. Process. Syst.* ISSN: 10495258. arXiv: 1102.0183.

Kroger, James K. et al. (2002). "Recruitment of Anterior Dorsolateral Prefrontal Cortex in Human Reasoning: a Parametric Study of Relational Complexity". In: *Cerebral Cortex* 12.5, pp. 477–485. ISSN: 1047-3211, 1460-2199. DOI: 10.1093/cercor/12.5.477. URL: http://cercor.oxfordjournals.org/content/12/5/477%7B%5C%%7D5Cnhttp://cercor.oxfordjournals.org/content/12/5/477.full.pdf%7B%5C%%7D5Cnhttp://cercor.oxfordjournals.org/content/12/5/477.short%7B%5C%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/11950765.

Krueger, David et al. (2020). "Out-of-Distribution Generalization via Risk Extrapolation (REx)". In: arXiv: 2003.00688. URL: http://arxiv.org/abs/2003.00688.

Krueger, Lester E (1973a). "Effect of irrelevant surrounding material on speed of same-different judgment of two adjancent letters". In: *Journal of Experimental Psychology* 98, pp. 252–259.

— (1973b). "Effect of stimulus frequency on speed of "same"-"different" judgments." In: *Attention and Performance IV*. Ed. by S. Kornblum. New York, NY: Academic Press.

— (1978). "A Theory of Perceptual Matching". In: *Psychological review* 85.4, pp. 278–304. ISSN: 0033-295X. DOI: 10.1037/0033-295X.85.4.278.

Krusemark, Elizabeth, Kent Kiehl, and Joseph Newman (May 2016). "Endogenous attention modulates early selective attention in psychopathy: An ERP investigation". In: *Cognitive, Affective, and Behavioral Neuroscience* 16. DOI: 10.3758/s13415-016-0430-7.

Kubilius, Jonas, Stefania Bracci, and Hans P Op de Beeck (Apr. 2016). "Deep Neural Networks as a Computational Model for Human Shape Sensitivity". en. In: *PLoS Comput. Biol.* 12.4, e1004896.

Kundu, Prosenjit et al. (2017). "Perfect synchronization in networks of phase-frustrated oscillators". In: *Europhysics Letters* 120.4, pp. 1–11. ISSN: 12864854. DOI: 10.1209/0295-5075/120/40002. arXiv: arXiv:1801.05660v2.

Kuramoto, Yoshiki (1975). "Lecture Notes in Physics". In: *39th International Symposium on Mathematical Problems in Theoretical Physics*. Ed. by H Araki. New York: Springer-Verlag, p. 420.

Lades, Martin et al. (1993). "Distortion invariant object recognition in the Dynamic Link Architecture". In: *IEEE Transactions on Computers* 42.3, pp. 300–311. ISSN: 00189340. DOI: 10.1109/12.210173.

Laeng, Bruno (1994). "Lateralization of categorical and coordinate spatial functions: a study of unilateral stroke patients." In: *Journal of cognitive neuroscience* 6.3, pp. 189–203. ISSN: 0898-929X. DOI: 10.1162/jocn.1994.6.3.189. URL: http://www.ncbi.nlm.nih.gov/pubmed/23964971.

Laeng, Bruno et al. (2011). "Processing Spatial Relations With Different Apertures of Attention". In: *Cognitive Science* 35.2, pp. 297–329. ISSN: 03640213. DOI: 10.1111/j.1551-6709.2010.01139.x.

Lake, B., R. Salakhutdinov, and J. Tenenbaum (2015a). "Human-level concept learning through probabilistic program induction". In: *Science (80-.).* 350.6266, pp. 1332–1338. DOI: 10.1126/science.aab3050.

Lake, B, R Salakhutdinov, and J Tenenbaum (2015b). "Human-level concept learning through probabilistic program induction". In: *Science* 350.6266, pp. 1332–1338.

Lake, Brenden M. et al. (2016). "Building Machines That Learn and Think Like People". In: 2012, pp. 1–58. ISSN: 0140-525X. DOI: 10.1017/S0140525X16001837. arXiv: 1604.00289. URL: http://arxiv.org/abs/1604.00289.

Laming, D.R.J. (1968). *Information theory of choice reaction times*. London: Academic Press.

Lamme, Victor A F and Henk Spekreijse (1998). "Neuronal synchrony does not represent texture segregation". In: *Nature* 396.November, pp. 362–366.

Lämmer, Stefan et al. (Apr. 2006). "Decentralised control of material or traffic flows in networks using phase-synchronisation". In: *Physica A: Statistical Mechanics and its Applications* 363.1, pp. 39–47. ISSN: 0378-4371. DOI: 10.1016/j.physa.2006.01.047. URL: http://dx.doi.org/10.1016/j.physa.2006.01.047.

Langton, Chris G. (1990). "Computation at the edge of chaos: Phase transitions and emergent computation". In: *Physica D: Nonlinear Phenomena* 42.1-3, pp. 12–37. ISSN: 01672789. DOI: 10.1016/0167-2789(90)90064-V.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444.

LeCun, Yann, Corinna Cortes, and CJ Burges (2010). "MNIST handwritten digit database". In: *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2.

LeCun, Yann et al. (1998). "Gradient-based learning applied to document recognition". In: *Proc. IEEE* 86.11, pp. 2278–2323. ISSN: 00189219. arXiv: 1102.0183.

Leeuwen, C. van (2007). "Synchrony, Binding, and Consciousness: How Are They Related?" In: *Theory & Psychology* 17.6, pp. 779–790. ISSN: 0959-3543. DOI: 10.1177/0959354307083493. URL: http://tap.sagepub.com/cgi/doi/10.1177/0959354307083493.

Levelt, W.J.M (1984). "Limits on perception". In: *Some perceptual limitations in talking about space*. Ed. by A.J. van Doorn, W.A. de Grind, and J.J. Koenderink. Utrecht: VNU Science Press, pp. 323–358.

Li, Chunguang and Yuke Li (2011). "Fast and robust image segmentation by small-world neural oscillator networks". In: *Cognitive Neurodynamics* 5, pp. 209–220. DOI: 10.1007/s11571-011-9152-2.

Linsley, Drew et al. (2017). "What are the visual features underlying human versus machine vision?" In: *IEEE ICCV Workshop on the Mutual Benefit of Cognitive and Computer Vision*.

Linsley, Drew et al. (2018). "Learning long-range spatial dependencies with horizontal gated recurrent units". In: *Neural Information Processing Systems*. Ed. by S. Bengio et al. Red Hook, NY: Curran Associates, pp. 152–64.

Lisman, John E. and Marco a. P. Idart (1995). "Storage of 7 +/- 2 Short-Term Memories in Oscillatory Subcycles". In: *Science* 267.March, pp. 1512–1515.

Llinás, Rodolfo R. (2014). "Intrinsic electrical properties of mammalian neurons and CNS function: A historical perspective". In: *Frontiers in Cellular Neuroscience* 8.November, pp. 1–14. ISSN: 16625102. DOI: 10.3389/fncel.2014.00320.

Loewenstein, J. and Dedre Gentner (2005). "Relational language and the development of relational mapping". In: *Cognitive Psychology* 50, pp. 315–353.

Logan, Gordan D. (1995). *Linguistic and conceptual control of visual spatial attention*. DOI: 10.1006/cogp.1995.1004.

Logan, Gordon D. (1994). "Spatial attention and the apprehension of spatial relations." In: *Journal of experimental psychology. Human perception and performance* 20.5, pp. 1015–36. ISSN: 0096-1523. DOI: 10.1037/0096-1523.20.5.1015. URL: http://www.ncbi.nlm.nih.gov/pubmed/7964527.

Logan, Gordon D. and Daniel D. Sadler (1996). *A computational analysis of the apprehension of spatial relations*.

Lopes, M A et al. (2016). "Synchronization in the random field Kuramoto model on complex networks". In: pp. 1–7. arXiv: arXiv:1605.04733v2.

Loula, João, Marco Baroni, and Brenden Lake (2019). "Rearranging the Familiar: Testing Compositional Generalization in Recurrent Networks". In: pp. 108–114. DOI: 10.18653/v1/w18-5413. arXiv: 1807.07545.

Lu, Hangwei (2018). "Missing area completion in facial images using maximum-correntropy-criterion regularized cascading autoencoder". In: *2017 International Conference on Security, Pattern Analysis, and Cybernetics, SPAC 2017* 2018-January.1, pp. 696–700. DOI: 10.1109/SPAC.2017.8304364.

Luck, S J and EK Vogel (1997). "The capacity of visual working memory for features and conjunctions." In: *Nature* 390.6657, pp. 279–281. ISSN: 0028-0836. DOI: 10.1038/36846. URL: http://www.ncbi.nlm.nih.gov/pubmed/9384378.

Luck, Steven J. (2012). *Electrophysiological Correlates of the Focusing of Attention within Complex Visual Scenes: N2pc and Related ERP Components*. January 2017, pp. 1–56. ISBN: 9780199940356. DOI: 10.1093/oxfordhb/9780195374148.013.0161.

Luck, Steven J. et al. (1997). "Bridging the Gap between Monkey Neurophysiology and Human Perception: An Ambiguity Resolution Theory of Visual Selective Attention". In: *Cognitive Psychology* 33.1, pp. 64–87. ISSN: 0010-0285. DOI: S0010-

0285(97)90660-5[pii]\r10.1006/cogp.1997.0660. URL: http://www.ncbi.nlm.nih.gov/pubmed/9212722.

Maddison, Chris J., Andriy Mnih, and Yee Whye Teh (2019). "The concrete distribution: A continuous relaxation of discrete random variables". In: *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, pp. 1–20. arXiv: 1611.00712.

Mallat, Stéphane (2016). "Understanding Deep Convolutional Networks". In: *Philosophical Transactions of the Royal Society A* 374.20150203, pp. 1–17. ISSN: 1364503X. DOI: 10.1098/rsta.2015.0203. arXiv: 1601.04920.

Malsburg, Christoph von der (1994). "The Correlation Theory of Brain". In: *Models of Neural Networks II*. Ed. by E. Domany, JL van Hemmen, and K Schulten. January 1994. Berlin: Springer. ISBN: 9781461243205. DOI: 10.1007/978-1-4612-4320-5.

— (1999). "The What and Why of Binding : The Modeler ' s Perspective". In: *Cell* 24, pp. 95–104. ISSN: 0023-5946. DOI: 10.1016/S0896-6273(00)80825-9.

Malsburg, Christoph von der and W. Schneider (1986). "A neural cocktail-party processor". In: 40, pp. 29–40.

Marcus, Gary (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.

— (2018). "Deep Learning: A Critical Appraisal". In: pp. 1–27. arXiv: 1801.00631. URL: http://arxiv.org/abs/1801.00631.

Marcus, Gary F., Keith J. Fernandes, and Scott P. Johnson (2007). "Infant rule learning facilitated by speech: Research report". In: *Psychological Science* 18.5, pp. 387–391. ISSN: 09567976. DOI: 10.1111/j.1467-9280.2007.01910.x.

Marmor, Gloria S. and Larry A Zaback (1976). "Mental Rotation by the Blind: Does Mental Rotation Depend on Visual Imagery?" In: *Journal of Experimental Psychology: Human Perception and Perjormance* 2.4, pp. 51–521. ISSN: 0096-1523. DOI: 10.1037/0096-1523.2.4.515.

Martin, A. B. and R. von der Heydt (2015). "Spike Synchrony Reveals Emergence of Proto-Objects in Visual Cortex". In: *Journal of Neuroscience* 35.17, pp. 6860–6870. ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3590-14.2015. URL: http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.3590-14.2015.

Martinetz, Thomas and Klaus Schulten (1991). *A "Neural-Gas" Network Learns Topologies*. URL: http://web.cs.swarthmore.edu/%7B~%7Dmeeden/DevelopmentalRobotics/fritzke95.pdf.

Martinho III, Antone and Alex Kacelnik (2016). "Ducklings imprint on the relational concept of "same or different"". In: *Science* 353.6296, pp. 286–288. ISSN: 0036-8075. DOI: 10.1126/science.aaf4247. arXiv: arXiv:1011.1669v3.

Masse, Nicolas Y. et al. (2019). "Circuit mechanisms for the maintenance and manipulation of information in working memory". In: *Nature Neuroscience* 22.7, pp. 1159–1167. ISSN: 15461726. DOI: 10.1038/s41593-019-0414-3. URL: http://dx.doi.org/10.1038/s41593-019-0414-3.

Matthews, Robert J. (1997). "Can Connectionists Explain Systematicity?" In: *Mind and Language* 12.2, pp. 154–177. ISSN: 0268-1064. DOI: 10.1111/1468-0017.00041.

Mauck, B and G Dehnhardt (1997). "Mental rotation in a California sea lion (Zalophus californianus)." In: *The Journal of experimental biology* 200.Pt 9, pp. 1309–1316. ISSN: 0022-0949.

Mayer, Eugène et al. (1999). "A pure case of Gerstmann syndrome with a subangular lesion". In: *Brain* 122.6, pp. 1107–1120. ISSN: 00068950. DOI: 10.1093/brain/122.6.1107.

McEvoy, Linda K., Michael E. Smith, and Alan S. Gevins (1998). "Dynamic cortical networks of verbal and spatial working memory: effects of memory load and task practice." In: *Cerebral cortex* 8 7, pp. 563–74.

McLelland, Douglas and Rufin VanRullen (2016). "Theta-Gamma Coding Meets Communication-through-Coherence: Neuronal Oscillatory Multiplexing Theories Reconciled". In: *PLoS Computational Biology* 12.10, pp. 4–10. ISSN: 15537358. DOI: 10.1371/journal.pcbi.1005162.

Meier, Martin, Robert Haschke, and Helge J Ritter (2013). "Perceptual Grouping through Competition in Coupled Oscillator Networks". In: April, pp. 24–26.

— (2014). "Perceptual grouping by entrainment in coupled kuramoto oscillator networks". In: *Network: Computation in Neural Systems* 25.June, pp. 72–84. DOI: 10.3109/0954898X.2014.882524.

Michal, A. L. and S. L. Franconeri (2014). "The order of attentional shifts determines what visual relations we extract". In: *Journal of Vision* 14.10, pp. 1033–1033. ISSN: 1534-7362. DOI: 10.1167/14.10.1033. URL: http://jov.arvojournals.org/article.aspx?articleid=2144911.

Michal, Audrey L. et al. (2016). "Visual routines for extracting magnitude relations". In: *Psychonomic Bulletin & Review* 23.6, pp. 1802–1809. ISSN: 1069-9384. DOI: 10.3758/s13423-016-1047-0. URL: http://link.springer.com/10.3758/s13423-016-1047-0.

Miller, George A. (1956). "The Magical Number 7, Plus or Minus 2 : Some Limits on Our Capacity for Processing Information". In: *Psychological Review-* 101.2,

pp. 343–352. ISSN: 0033-295X. DOI: 10.1037/h0043158. URL: papers3://publication/uuid/9D829293-04E4-4BB7-8B3B-1E868E1EC915.

Miller, George A. (1962). "Some psychological studies of grammar". In: *American Psychologist* 7, pp. 748–62.

Miller, George A. and Phil Johnson-Laird (1976). *Language and perception*. Cambridge, MA: Harvard University Press.

Mitchell, Melanie, Peter Hraber, and James P. Crutchfield (1993). "Revisiting the Edge of Chaos: Evolving Cellular Automata to Perform Computations". In: 7, pp. 89–130. arXiv: 9303003 [adap-org]. URL: http://arxiv.org/abs/adap-org/9303003.

Miyano, Takaya and Takako Tsutsui (2007). "Data synchronization in a network of coupled phase oscillators". In: *Physical Review Letters* 98.2, pp. 1–4. ISSN: 00319007. DOI: 10.1103/PhysRevLett.98.024102.

Moore, Cathleen M., Catherine L. Elsinger, and Alejandro Lleras (1994). "Visual attention and the apprehension of spatial relations: The case of depth". In: *J. Exp. Psychol. Hum. Percept. Perform.* 20.5, pp. 1015–1036. ISSN: 0096-1523. DOI: 10.1037/0096-1523.20.5.1015. URL: http://www.ncbi.nlm.nih.gov/pubmed/7964527.

Motter, B C and V B Mountcastle (1981). "The functional properties of the light-sensitive neurons of the posterior parietal cortex studied in waking monkeys: foveal sparing and opponent vector organization." In: *Journal of Neuroscience* 1.1, pp. 3–26. ISSN: 0270-6474. URL: http://www.jneurosci.org/content/1/1/3.short%7B%5C%%7D5Cnpapers3://publication/uuid/3844BF7A-E105-4232-89D8-A969CB617BAC.

Motter, B C et al. (1987). "Functional properties of parietal visual neurons: mechanisms of directionality along a single axis." In: *The Journal of Neuroscience* 7.1, pp. 154–76. ISSN: 0270-6474. URL: http://www.ncbi.nlm.nih.gov/pubmed/3806192.

Nakao, Hiroya (2016). "Phase reduction approach to synchronization of nonlinear oscillators". In: 214, pp. 188–214. arXiv: arXiv:1704.03293v1.

Navon, David (1977). "Forest before trees: The precedence of global features in visual perception". In: *Cognitive Psychology* 9.3, pp. 353–383. ISSN: 00100285. DOI: 10.1016/0010-0285(77)90012-3. arXiv: ISSN0010-0285.

Nayebi, Aran et al. (2018). "Task-Driven Convolutional Recurrent Models of the Visual System". In: *Neural Information Processing Systems*. Ed. by S. Bengio et al. Red Hook, NY. arXiv: 1807.00053. URL: http://arxiv.org/abs/1807.00053.

Neal, Radford M. (2001). "Annealed importance sampling". In: *Statistics and Computing* 11.2, pp. 125–139. ISSN: 09603174. DOI: 10.1023/A:1008923215028. arXiv: 9803008 [physics].

Nekovarova, Tereza et al. (2013). "Mental transformations of spatial stimuli in humans and in monkeys: Rotation vs. translocation". In: *Behavioural Brain Research* 240.1, pp. 182–191. ISSN: 01664328. DOI: 10.1016/j.bbr.2012.11.008. URL: http://dx.doi.org/10.1016/j.bbr.2012.11.008.

Nickerson, Raymond S. (1968). "Note on same-different response times". In: *Perceptual and Motor Skills* 27, pp. 565–566.

— (1972). "Frequency, recency, and repetition effects on same and different response times". In: *Journal of Experimental Psychology* 101.2, pp. 330–336.

— (1978). "On the time it takes to tell things apart". In: *Attention and Performance VII*. Ed. by J. Resquin. Hillsdale, NJ: Erlbaum.

Nishimori, Hidetoshi (2010). *Statistical Physics of Spin Glasses and Information Processing*. ISBN: 0198509413. DOI: 10.1093/acprof:oso/9780198509417.001.0001.

O'Reilly, Randall C. and Michael J. Frank (2006). "Making Working Memory Work: A Computational Model of Learning in the Prefrontal Cortex and Basal Ganglia". In: *Neural Computation* 18.2, pp. 283–328. ISSN: 0899-7667. DOI: 10.1162/089976606775093909. URL: http://www.mitpressjournals.org/doi/10.1162/089976606775093909.

Onnela, J.-P. et al. (Nov. 2003). "Dynamics of market correlations: Taxonomy and portfolio analysis". In: *Physical Review E* 68.5. ISSN: 1095-3787. DOI: 10.1103/physreve.68.056110. URL: http://dx.doi.org/10.1103/PhysRevE.68.056110.

Ostojic, Srdjan and Nicolas Brunel (2011). "From spiking neuron models to linear-nonlinear models". In: *PLoS Computational Biology* 7.1. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1001056.

Ott, Edward and Thomas M. Antonsen (2008). "Low dimensional behavior of large systems of globally coupled oscillators". In: *Chaos* 18.3, pp. 1–6. ISSN: 10541500. DOI: 10.1063/1.2930766. arXiv: 0806.0004.

Ottino-l, Bertrand and Steven H Strogatz (2018). "Volcano transition in a solvable model of oscillator glass". In: 14853, pp. 1–5. arXiv: arXiv:1712.05850v2.

Overmyer, S.P. and J.R. Simon (1985). "The effect of irrelevant cues on 'same-different' judgments in a sequential information processing task". In: *Acta Psychologica (Amsterdam)* 58, pp. 237–249.

Palanca, Ben J.A. and Gregory C. DeAngelis (2005). "Does neuronal synchrony underlie visual feature grouping?" In: *Neuron* 46.2, pp. 333–346. ISSN: 08966273. DOI: 10.1016/j.neuron.2005.03.002.

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*.

Ed. by H. Wallach et al. Curran Associates, Inc., pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pearson, J.C., L.H. Finkel, and G.M. Edelman (1987). "Plasticity in the Organizatino of Adult Cerebral Cortical Maps: A Computer Simulation Based on Neuronal Group Selection". In: *Journal of Neuroscience* 7.12.

Perez-Vicente, C.J. and F. Ritort (1997). "A moment-based approach to the dynamical solution of the Kuramoto model". In:

Petersen, S.E and M Posner (2012). "The attention system of the human brain: 20 years after". In: *Annual review of neuroscience* 21.35, pp. 73–89. ISSN: 1879-307X. DOI: 10.1146/annurev-neuro-062111-150525.The. arXiv: NIHMS150003.

Peterson, J, J Abbott, and T Griffiths (2016). "Adapting deep network features to capture psychological representations". In: *38th annual conference of the cognitive science society*. Ed. by D Grodner et al., pp. 2363–2368.

Pham, Quang Cuong and Jean-Jacques Slotine (2007). "Stable concurrent synchronization in dynamic system networks". In: *Neural Networks* 20.1, pp. 62–77. ISSN: 08936080. DOI: 10.1016/j.neunet.2006.07.008.

Picallo, Clara B. and Hermann Riecke (2011). "Adaptive oscillator networks with conserved overall coupling: Sequential firing and near-synchronized states". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 83.3. ISSN: 15393755. DOI: 10.1103/PhysRevE.83.036206. arXiv: 1008.0333.

Piëch, Valentin et al. (2013). "Network model of top-down influences on local gain and contextual interactions in visual cortex". In: *Proceedings of the National Academy of Sciences of the United States of America* 110.43. ISSN: 00278424. DOI: 10.1073/pnas.1317019110.

Pluchino, A., V. Latora, and A. Rapisarda (2004). *Changing Opinions in a Changing World: a New Perspective in Sociophysics*. arXiv: cond-mat/0410217 [cond-mat.other].

Posner, M.I., M.J. Nissen, and W.C. Ogden (1978). "Attended and unattended processing modes: the role of set for spatial location". In: *Modes of Perceiving and Processing Information*. Ed. by I.J Saltzman and H.L Pick. Hillsdale, NJ: Erlbaum.

Prabhakaran, V et al. (2000). "Integration of diverse information in working memory within the frontal lobe." In: *Nature neuroscience* 3.1, pp. 85–90. ISSN: 1097-6256. DOI: 10.1038/71156. URL: http://www.ncbi.nlm.nih.gov/pubmed/10607400.

Prinzmetal, William (2005). "Location perception: the X-Files parable." In: *Perception & psychophysics* 67.1, pp. 48–71. ISSN: 0031-5117.

Pylyshyn, Z W and R W Storm (1988). "Tracking multiple independent targets: evidence for a parallel tracking mechanism". In: *Spat. Vis.* 3.3, pp. 179–197.

Pylyshyn, Zenon (1989a). "The role of location indexes in spatial perception: A sketch of the FINST spatial-index model". In: *Cognition* 32, pp. 65–97.

— (1989b). "The role of location indexes in spatial perception: A sketch of the FINST spatial-index model". In: *Cognition* 32.1, pp. 65–97. ISSN: 00100277. DOI: 10.1016/0010-0277(89)90014-0.

— (1994). "Some primitive attention mechanisms of spatial". In: *Cognition* 50, pp. 363–384.

Pylyshyn, Zenon W. (2004). "Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities". In: *Visual Cognition* 11.7, pp. 801–822. ISSN: 13506285. DOI: 10.1080/13506280344000518.

Quiles, Marcos G. et al. (2011). "Selecting salient objects in real scenes: An oscillatory correlation model". In: *Neural Networks* 24.1, pp. 54–64. ISSN: 08936080. DOI: 10.1016/j.neunet.2010.09.002. URL: http://dx.doi.org/10.1016/j.neunet.2010.09.002.

Quiroga, R Quian et al. (2005). "Invariant visual representation by single neurons in the human brain." In: *Nature* 435.7045, pp. 1102–7. ISSN: 1476-4687. DOI: 10.1038/nature03687. URL: http://www.ncbi.nlm.nih.gov/pubmed/15973409.

Raffone, Antonino and Gezinus Wolters (2001). "A Cortical Mechanism for Binding in Visual Working Memory". In: *Journal of Cognitive Neuroscience* 13.6, pp. 766–785. ISSN: 0898-929X. DOI: 10.1162/08989290152541430. arXiv: 1511.04103. URL: http://www.mitpressjournals.org/doi/10.1162/08989290152541430.

Raiko, Tapani and Harri Valpola (2011). "Oscillatory Neural Network for Image Segmentation with Biased Competition for Attention". In: *Advances in Experimental Medicine and Biology: From Brains to Systems*. Ed. by C. Hernandez. Vol. 718. New York, NY: Springer, pp. 75–88.

Rainer, G, W F Asaad, and E K Miller (1998a). "Memory fields of neurons in the primate prefrontal cortex". In: *Proceedings of the National Academy of Sciences of the United States of America* 95.25, pp. 15008–13. ISSN: 0027-8424. DOI: 10.1073/pnas.95.25.15008. URL: http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?db=m%7B%5C&%7Dform=6%7B%5C&%7Ddopt=r%7B%5C&%7Duid=9844006%7B%5C%%7D5Cnhttp://www.pnas.org/cgi/content/full/95/25/15008.

— (1998b). "Selective representation of relevant information by neurons in the primate prefrontal cortex." In: *Nature* 393.6685, pp. 577–579. ISSN: 0028-0836. DOI: 10.1038/31235.

Rao, S, Gregor Rainer, and Earl Miller (1997). "Integration of what and where in the primate prefrontal cortex". In: *Science* 276.5313, pp. 821–824. ISSN: 0036-8075.

DOI: 10.1126/science.276.5313.821. URL: papers2://publication/uuid/644A1A12-8D93-406C-AD25-A5FA9059639E.

Recht, Benjamin et al. (2019). "Do ImageNet classifiers generalize to ImageNet?" In: *36th International Conference on Machine Learning, ICML 2019* 2019-June, pp. 9413–9424. arXiv: 1902.10811.

Regier, Terry and Laura A. Carlson (2001). *Grounding spatial language in perception: an empirical and computational investigation*. DOI: 10.1037/0096-3445.130.2.273.

Reichert, David P (2014). "Deep Boltzmann Machines as Hierarchical Generative Models of Perceptual Inference in the Cortex Deep Boltzmann Machines as Hierarchical Generative Models of Perceptual Inference in the Cortex Doctor of Philosophy Institute for Adaptive and Neural Computat". In: May.

Reichert, David P. and Thomas Serre (Dec. 2013). "Neuronal Synchrony in Complex-Valued Deep Networks". In: *Int. Conf. Learn. Represent.* arXiv: 1312.6115. URL: http://arxiv.org/abs/1312.6115.

Rescorla, Michael (2020). "The Computational Theory of Mind". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2020. Metaphysics Research Lab, Stanford University.

Reynolds, John H., Leonardo Chelazzi, and Robert Desimone (1999). "Competitive mechanisms subserve attention in macaque areas V2 and V4". In: *Journal of Neuroscience* 19.5, pp. 1736–1753. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.211.8831.

Ricci, Matthew G., Junkyung Kim, and Thomas Serre (2018). "Same-different problems strain convolutional neural networks". In: *Proceedings of the 40th Annual Cognitive Science Society*. Madison, WI: Cognitive Science Society.

Ricci, Matthew, Remi Cadene, and Thomas Serre (Submitted). "Same-different relations in machine vision". In: *Current Opinion in Behavioral Sciences*.

Ricci, Matthew, Charles Windolf, and Thomas Serre (2019). "A formal neural synchrony model for unsupervised image grouping". In: *COSYNE*. Lisbon.

Ricci, Matthew et al. (2020). "Kura-net: Exploring systems of coupled oscillators with deep learning". In: *COSYNE*. Denver.

Rice, S.O. (1948). "Statistical Properties of a Sine Wave Plus Random Noise". In: *Bell Syst. Tech. J.*

Riesenhuber, M and T Poggio (1999). "Hierarchical models of object recognition in cortex". In: *Nat Neurosci* 2.11, pp. 1019–1025. ISSN: 1097-6256 (Print). DOI: 10.1038/14819. URL: http://www.ncbi.nlm.nih.gov/pubmed/10526343.

Roberts, David C. (2008). "Linear reformulation of the Kuramoto model of self-synchronizing coupled oscillators". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 77.3, pp. 1–12. ISSN: 15393755. DOI: 10.1103/PhysRevE.77.031114. arXiv: arXiv:0704.1166v5.

Robertson, Lynn C. (2003). "Binding, Spatial Attention and Perceptual Awareness". In: *Nature Reviews Neuroscience* 4.2, pp. 93–102. DOI: 10.1038/nrn1030.BINDING.

Robertson, Lynn C and Min-Shik Kim (1999). "Effects of Perceived Space on Spatial Attention". In: *Psychological Science* 10.1, pp. 76–79. ISSN: 0956-7976. DOI: 10.1111/1467-9280.00110. URL: http://pss.sagepub.com/lookup/doi/10.1111/1467-9280.00110.

Roelfsema, Pieter R and Roos Houtkamp (Nov. 2011). "Incremental grouping of image elements in vision". In: *Attention, Perception and Psychophysics* 73.8, pp. 2542–2572. ISSN: 1943-393X. DOI: 10.3758/s13414-011-0200-0. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3222807%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Dtool=pmcentrez%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Drendertype=abstract.

Roelfsema, Pieter R, Victor a F Lamme, and Henk Spekreijse (Sept. 2004). "Synchrony and covariation of firing rates in the primary visual cortex during contour grouping." In: *Nature neuroscience* 7.9, pp. 982–991. ISSN: 1097-6256. DOI: 10.1038/nn1304. URL: http://www.ncbi.nlm.nih.gov/pubmed/15322549.

Rogers, Timothy T. and James L. McClelland (2004). "Semantic cognition: A parallel distributed processing approach". In: *Attention And Performance* 425, p. 439. URL: http://books.google.com/books?hl=en%7B%5C&%7Dlr=%7B%5C&%7Did=AmB33Uz2MVAC%7B%5C&%7Doi=fnd%7B%5C&%7Dpg=PR9%7B%5C&%7Ddq=Semantic+Cognition+:+A+Parallel+Distributed+Processing+Approach%7B%5C&%7Dots=BtcuT%7B%5C_%7Dumzk%7B%5C&%7Dsig=bvPMpcNL77MzrcT4STKTha4cfJs%7B%5C%%7D5Cnhttp://mitpress.mit.edu/catalog/item/default.asp?ttype=2%7B%5C&%7Dtid=10117.

Rosielle, Luke J, Brian T Crabb, and Eric E Cooper (2002). "Attentional coding of categorical relations in scene perception: evidence from the flicker paradigm." In: *Psychonomic bulletin & review* 9.2, pp. 319–26. ISSN: 1069-9384. URL: http://www.ncbi.nlm.nih.gov/pubmed/12120795.

Roth, J. C. and Steven L. Franconeri (2012). "Asymmetric coding of categorical spatial relations in both language and vision". In: *Frontiers in Psychology* 3.NOV, pp. 1–22. ISSN: 19326203. DOI: 10.1371/journal.pone.0163141.

Rousseeuw, Peter J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20.C, pp. 53–65. ISSN: 03770427. DOI: 10.1016/0377-0427(87)90125-7.

Roux, Nicolas Le and Yoshua Bengio (2017). "Representational Power of Restricted Boltzmann Machines and Deep Belief Networks". In: pp. 1–18.

Ruelles, D. (2004). *Thermodynamic Formalism*. Second Edi. Cambridge, UK: Cambridge University Press.

Russakovsky, Olga et al. (2010). "Imagenet large scale visual recognition challenge 2010". In: arXiv: arXiv:1409.0575v1. URL: http://scholar.google.com/scholar?hl=en%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7DbtnG=Search%7B%5C%%7D7B%7B%5C&%7D%7B%5C%%7D7Dq=intitle:Large+Scale+Visual+Recognition+Challenge+2010%7B%5C%%7D7B%7B%5C#%7D%7B%5C%%7D7D2%20http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:Large+Scale+Visual+Recognition+Challenge+2010%7B%5C#%7D2.

Sakaguchi, Hidetsugu (1988). "Cooperative Phenomena in Coupled Oscillator Systems under External Fields". In: 79.1.

Sakaguchi, Hidetsugu, Shigeru Shinomoto, and Yoshiki Kuramoto (1988). "Phase Transitions and Their Bifurcation Analysis in a Large Population of Active Rotators with Mean-Field Coupling". In: 79.3, pp. 600–607.

Salakhutdinov, Ruslan and Hugo Larochelle (2010). "Efficient learning of Deep Boltzmann Machines". In: *Journal of Machine Learning Research* 9, pp. 693–700. ISSN: 1533-7928.

Saleh, Babak, Ahmed Elgammal, and Jacob Feldman (Feb. 2016). "The Role of Typicality in Object Classification: Improving The Generalization Capacity of Convolutional Neural Networks". In: arXiv: 1602.02865 [cs.CV].

Santoro, Adam et al. (2017). "A simple neural network module for relational reasoning". In: pp. 1–16. arXiv: 1706.01427. URL: http://arxiv.org/abs/1706.01427.

Schaub, Michael T. et al. (2016). "Graph partitions and cluster synchronization in networks of oscillators". In: *Chaos* 26.9. ISSN: 10541500. DOI: 10.1063/1.4961065.

Schmidt, Ruben et al. (2015). "Kuramoto model simulation of neural hubs and dynamic synchrony in the human cerebral connectome". In: *BMC Neuroscience* 16.1, pp. 1–13. ISSN: 14712202. DOI: 10.1186/s12868-015-0193-z.

Schneider, G. E. (1969). "Two Visual Systems". In: *Science* 163.3870, pp. 895–902.

Sergent, Justine (1991). "Judgments of relative position and distance on representations of spatial relations." In: *Journal of Experimental Psychology: Human Perception and Performance* 17.3, pp. 762–780.

Serre, Thomas (2015). "Hierarchical Models of the Visual System". en. In: *Encyclopedia of Computational Neuroscience*. Ed. by Dieter Jaeger and Ranu Jung. Springer New York, pp. 1309–1318.

— (May 2016). "Models of visual categorization". en. In: *Wiley Interdiscip. Rev. Cogn. Sci.* 7.3, pp. 197–213.

Serre, Thomas, Aude Oliva, and Tomaso Poggio (Apr. 2007). "A feedforward architecture accounts for rapid categorization". In: *Proceedings of the National Academy of Sciences* 104.15, pp. 6424–6429. ISSN: 0027-8424. DOI: 10.1073/pnas.0700622104. URL: http://www.pnas.org/content/104/15/6424.full.

Shadlen, Michael N, J Anthony Movshon, and Howard Hughes (1999). "Synchrony Unbound : A Critical Evaluation of the Temporal Binding Hypothesis". In: 24, pp. 67–77.

Shastri, Lokendra (1999). "Advances in Shruti—A neurally motivatedmodel of relational knowledge representation and rapid inference using temporal synchrony". In: *Applied Intelligence* 11, pp. 79–108. ISSN: 0924669X. DOI: 10.1023/A:1008380614985.

Shepard, Roger N. and Lynn A. Cooper (1982). *Mental images and their transformations*. Cambridge, MA: Bradford.

Shepard, Roger N. and Jacqueline Metzler (1971). "Mental Rotation of Three-Dimensional Objects". In: *Science* 171.3972, pp. 701–703. ISSN: 0036-8075. DOI: 10.1126/science.171.3972.701.

Shiffrin, R.M., D.P. McKay, and W.O. Shaffer (1976). "Attending to forty-nine spatial positions at once". In: *Journal of Experimental Pscyhology: Human Perception and Performance* 2.1, pp. 14–22.

Shulman, G.L., R.W. Remington, and I.P. Mclean (1979). "Moving attention through visual space". In: *Journal of Experimental Pscyhology: Human Perception and Performance* 5, pp. 522–526.

Siegle, Joshua H., Dominique L. Pritchett, and Christopher I. Moore (2014). "Gamma-range synchronization of fast-spiking interneurons can enhance detection of tactile stimuli". In: *Nature Neuroscience* 17.10, pp. 1371–1379. ISSN: 15461726. DOI: 10.1038/nn.3797.

Silverman, W.P. (1973). "The perception of identity in simultaneously presented complex visual displays". In: *Memory & Cognition* 1, pp. 459–466.

Slotnick, Scott D et al. (2001). "Hemispheric Asymmetry in Categorical Versus Coordinate Visuospatial Processing Revealed by Temporary Cortical Deactivation". In: *Journal of Cognitive Neuroscience* 13, pp. 1088–1096.

Smith, Lachlan D. and Georg A. Gottwald (2019). "Chaos in networks of coupled oscillators with multimodal natural frequency distributions". In: *Chaos* 29.9. ISSN: 10541500. DOI: 10.1063/1.5109130. arXiv: 1905.02859.

Snodgrass, J. G. (1972). "Matching patterns vs matching digits: The effect of memory dependence and complexity on same-different reaction times". In: *Perception & Psychophysics* 11, pp. 341–349.

Sougné, Jacques P. (1999). "Infernet: a Neurocomputational Model of Binding and Inference". PhD thesis. Universite de Liege.

Spelke, Elizabeth S. et al. (1994). "Early knowledge of object motion: continuity and inertia". In: *Cognition* 51.2, pp. 131–176. ISSN: 0010-0277. DOI: https://doi.org/10.1016/0010-0277(94)90013-2. URL: http://www.sciencedirect.com/science/article/pii/0010027794900132.

Sperling, G. (1960). "The information available in brief visual presentations". In: *Psychological Monographs* 74.11.

Stabinger, Sebastian and Antonio Rodriguez-Sanchez (Oct. 2017). "Evaluation of Deep Learning on an Abstract Image Classification Dataset". In: *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

Stabinger, Sebastian, Antonio Rodríguez-Sánchez, and Justus Piater (2016). "25 years of CNNS: Can we compare to human abstraction capabilities?" In: *ICANN*. Vol. 9887 LNCS, pp. 380–387. ISBN: 9783319447803. DOI: 10.1007/978-3-319-44781-0_45. arXiv: 1607.08366.

Stein, Daniel L. and Charles M. Newman (2013). *Spin Glasses and Complexity*. Princeton University Press. ISBN: 9780691147338. URL: http://www.jstor.org/stable/j.ctt12f4hf.

Sternberg, Saul (1966). "High-speed scanning in human memory". In: *Science* 153, pp. 652–654. ISSN: 0036-8075. DOI: 10.1126/science.153.3736.652.

Stich, Kai Petra, Guido Dehnhardt, and Björn Mauck (2003). "Mental rotation of perspective stimuli in a California sea lion (Zalophus californianus)". In: *Brain, Behavior and Evolution* 61.2, pp. 102–112. ISSN: 00068977. DOI: 10.1159/000069355.

Strogatz, Steven H (2000). "From Kuramoto to Crawford : exploring the onset of synchronization in populations of coupled oscillators". In: 143, pp. 1–20.

Strogatz, Steven H. and Renato E. Mirollo (1993). "Splay states in globally coupled Josephson arrays: Analytical prediction of Floquet multipliers". In: *Physical Review E* 47.1, pp. 220–227. ISSN: 1063651X. DOI: 10.1103/PhysRevE.47.220.

Szegedy, Christian, W Zaremba, and I Sutskever (2013). "Intriguing properties of neural networks". In: *arXiv preprint arXiv: ...*, pp. 1–10. arXiv: arXiv:1312.6199v4. URL: http://arxiv.org/abs/1312.6199.

Talmy, L. (1983). "How language structures space". In: *Spatial orientationL Theory, research, and application*. Ed. by H.L. Pick and L.P. Acredolo. New York, NY: Plenum Press, pp. 225–282.

Tan, Pang-Ning (May 2005). *Introduction to Data Mining*. Pearson. ISBN: 0321321367. URL: https://www.xarg.org/ref/a/0321321367/.

Tanaka, K (Jan. 2003). "Columns for Complex Visual Object Features in the Inferotemporal Cortex: Clustering of Cells with Similar but Slightly Different Stimulus Selectivities". In: *Cereb. Cortex* 13.1, pp. 90–99.

Tanaka, Takuma and Toshio Aoyagi (2008). "Optimal weighted networks of phase oscillators for synchronization". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 78.4, pp. 1–10. ISSN: 15393755. DOI: 10.1103/PhysRevE.78.046210.

Tasal, Y. (1983). "Movements of attention across the visual field". In: *Journal of Experimental Pscyhology: Human Perception and Performance*.

Tass, P.A. (20007). *Phase resetting in medicine and biology: stochastic modeling and data analysis*. Berlin: Springer.

Tenenbaum, Joshua B. et al. (2011). "How to Grow a Mind: Statistics, Structure, and Abstraction". In: *Science* 331.6022, pp. 1279–1285.

Terman, David and DeLiang Wang (1995). "Global competition and local cooperation in a network of neural oscillators". In: *Physica D* 81, pp. 148–176.

Thomas, E. A. C. (1974). "The selectivity of preparation". In: *Psychological Review* 81, pp. 442–464.

Tieleman, Tijmen (2008). "Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient". In: *Proceedings of the 25th International Conference on Machine Learning* 307, p. 7. ISSN: 21576904. DOI: 10.1145/1390156.1390290.

Timms, L and L Q English (2014). "Synchronization in phase-coupled Kuramoto oscillator networks with axonal delay and synaptic plasticity". In: *Physical Review E* 89.032906, pp. 1–9. DOI: 10.1103/PhysRevE.89.032906.

Tong, David (2012). "Statistical Physics". In:

Treisman, A., D. Kahneman, and J. Burkell (1983). "Perceptual objects and the cost of filtering". In: *Perception & Psychophysics* 33.6, pp. 527–532. ISSN: 00315117. DOI: 10.3758/BF03202934.

Treisman, A and S Gormican (1988). *Feature analysis in early vision - evidence from search asymmetries..pdf*. DOI: 10.1037/0033-295X.95.1.15. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-295X.95.1.15.

Treisman, Anne M. and R. Patterson (1984). "Emergent features, attention, and object perception". In: *Journal of Experimental Psychology: Human perception and performance* 10, pp. 12–31.

Treisman, Anne M. and Janet Souther (1985). "Search asymmetry: A diagnostic for preattentive processing of separable features". In: *Journal of Experimental Pscyhology: General* 114.3, pp. 285–310.

Treisman, Anne. M. (1996). "The binding problem". In: *Current Opinion in Neurobiology* 6, pp. 171–178. ISSN: 19395086. DOI: 10.1002/wcs.1279. arXiv: arXiv:1011.1669v3.

Treisman, Anne and Garry Gelade (1980). "A Feature-Integration Theory of Attention". In: *Cognitive Psychology* 12, pp. 97–136. ISSN: 00100285. DOI: 10.1016/0010-0285(80)90005-5.

Tsodyks, Misha V. and Henry Markram (1997). "The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability". In: *Proceedings of the National Academy of Sciences of the United States of America* 94.2, pp. 719–723. ISSN: 00278424. DOI: 10.1073/pnas.94.2.719.

Ullman, Shimon (1984). "Visual Routines". In: *Cognition* 18, pp. 97–159.

Ullman, Shimon et al. (2016). "Atoms of recognition in human and computer vision". In: pp. 1–6. DOI: 10.1073/pnas.1513198113.

Ungerleider, Leslie G and J V Haxby (1994). "'What' and 'where' in the human brain". In: *Curr. Opin. Neurobiol.* 4.2, pp. 157–165. ISSN: 0959-4388. DOI: 10.1016/0959-4388(94)90066-3. arXiv: 0959-4388. URL: http://www.ncbi.nlm.nih.gov/pubmed/8038571.

Van Voorhis, Steven and Steven A. Hillyard (1977). "Visual evoked potentials and selective attention to points in space". In: *Perception & Psychophysics* 22.1, pp. 54–62. ISSN: 00315117. DOI: 10.3758/BF03206080.

Vandaloise, C. (1991). *Spatial preposition: A case study from French*. Chicago: University of Chicago Press.

VanRullen, Rufin (2009). "Binding hardwired versus on-demand feature conjunctions". In: *Visual Cognition* 17.1-2, pp. 103–119. ISSN: 1350-6285. DOI: 10.1080/13506280802196451.

Villegas, Pablo, Paolo Moretti, and Miguel A. Muñoz (2014). "Frustrated hierarchical synchronization and emergent complexity in the human connectome network". In: *Scientific Reports* 4. ISSN: 20452322. DOI: 10.1038/srep05990.

Wainwright, Martin J and Michael I Jordan (2008). "Graphical Models , Exponential Families , and Variational Inference". In: 1, pp. 1–305. DOI: 10.1561/2200000001.

Wang, DeLiang (1994). "Modeling Global Synchrony in the Visual Cortex by Locally Coupled Neural Oscillators". In: *Computation in Neurons and Neural Systems*, pp. 109–114. ISBN: 9781461361695.

Wang, Wei and Jean-Jacques Slotine (2005). "On partial contraction analysis for coupled nonlinear oscillators". In: *Biological Cybernetics* 92.1, pp. 38–53. ISSN: 03401200. DOI: 10.1007/s00422-004-0527-x.

Wasserman, E. a., L. Castro, and J. H. Freeman (2012). "Same-different categorization in rats". In: *Learning & Memory* 19.4, pp. 142–145. ISSN: 1072-0502. DOI: 10.1101/lm.025437.111.

Watts, Duncan J and Steven H Strogatz (1998). "Collective dynamcics of "small-world" networks". In: *Nature* 393.June, pp. 440–442. URL: https://www.ncbi.nlm.nih.gov/pubmed/9623998.

Whitney, David and Dennis M Levi (2011). "Visual Crowding: a fundamental limit on conscious perception and object recognition". In: *Trends Cogn Sci* 15.4, pp. 160–168. ISSN: 1879-307X. DOI: 10.1016/j.tics.2011.02.005.Visual.

Wiley, Robert W, Colin Wilson, and Brenda Rapp (2016). "The Effects of Alphabet and Expertise on Letter Perception." In: *Journal of Experimental Psychology: Human Perception and Performance* 42.8, No Pagination Specified. ISSN: 1939-1277. DOI: 10.1037/xhp0000213. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/xhp0000213.

Wilken, Patrick and Wei Ji Ma (2004). "A detection theory account of change detection." In: *Journal of vision* 4.12, pp. 1120–35. ISSN: 1534-7362. DOI: 10.1167/4.12.11. URL: http://jov.arvojournals.org/article.aspx?articleid=2192647.

Williams, C. (1974). "The effect of an irrelevant dimension on "same-different" judgements of multi-dimensional stimuli". In: *Quarterly Journal of Experimental Psychology* 26, pp. 26–31.

Wilson, Hugh R. and Jack D. Cowan (1972). "Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons". In: *Biophysical Journal* 12.1, pp. 1–24. ISSN: 00063495. DOI: 10.1016/S0006-3495(72)86068-5. URL: http://dx.doi.org/10.1016/S0006-3495(72)86068-5.

Wright, Anthony A. and Debbie M. Kelly (2017). "Comparative approaches to same/different abstract concept-learning". In: *Learn. Behav.* 45, pp. 323–324. ISSN: 19383711.

Yamaguchi, S, S Yamagata, and S Kobayashi (2000). "Cerebral asymmetry of the "top-down" allocation of attention to global and local features." In: *The Journal of neuroscience : the official journal of the Society for Neuroscience* 20.9, RC72. ISSN: 1529-2401. DOI: 20/9/RC72[pii].

Yamins, D. L. K. et al. (2014). "Performance-optimized hierarchical models predict neural responses in higher visual cortex". In: *Proceedings of the National Academy of Sciences* 111.23, pp. 8619–8624. ISSN: 0027-8424. DOI: 10.1073/pnas.1403112111. arXiv: 0706.1062v1. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1403112111.

Yin, Penghang et al. (2019). "Understanding straight-through estimator in training activation quantized neural nets". In: *7th International Conference on Learning Representations, ICLR 2019*, pp. 1–30. arXiv: 1903.05662.

Yu, Guoshen and Jean-Jacques Slotine (2009). "Visual grouping by neural oscillator networks". In: *IEEE Transactions on Neural Networks* 20.12, pp. 1871–1884. ISSN: 10459227. DOI: 10.1109/TNN.2009.2031678.

Yuan, Lei, David Uttal, and Steven Franconeri (2016). "Are categorical spatial relations encoded by shifting visual attention between objects?" In: *PLoS ONE* 11.10, pp. 1–22. ISSN: 19326203. DOI: 10.1371/journal.pone.0163141.

Yuille, Alan L., Peter W. Hallinan, and David S. Cohen (1992). "Feature extraction from faces using deformable templates". In: *International Journal of Computer Vision* 8.2, pp. 99–111. ISSN: 09205691. DOI: 10.1007/BF00127169.

Zandvakili, Amin and Adam Kohn (2015). "Coordinated Neuronal Activity Enhances Corticocortical Communication". In: *Neuron* 87.4, pp. 827–839. ISSN: 10974199. DOI: 10.1016/j.neuron.2015.07.026. URL: http://dx.doi.org/10.1016/j.neuron.2015.07.026.

Zanette, D. H. (2005). "Synchronization and frustration in oscillator networks with attractive and repulsive interactions". In: *Europhysics Letters* 72.2, pp. 190–196. ISSN: 02955075. DOI: 10.1209/epl/i2005-10238-4.

Zeki, S and S Shipp (1988). *The functional logic of cortical connections*. DOI: 10.1038/335311a0. URL: http://dx.doi.org/10.1038/335311a0.

Zemel, Richard S., Christopher K. I. Williams, and Michael C. Mozer (1995). "Lending Direction to Neural Networks". In: *Neural Networks* 8.4, pp. 503–512. ISSN: 08936080. DOI: 10.1016/0893-6080(94)00094-3.

Zhang, Ruiling et al. (2013). "Neural processes underlying the"same"-"different" judgment of two simultaneously presented objects–an EEG study." In: *PloS one*

8.12, e81737. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0081737. URL: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0081737.

Zhang, W and S J Luck (2008). "Discrete fixed-resolution representations in visual working memory". In: *Nature* 453.7192, 233–U13. ISSN: 0028-0836. DOI: Doi10.1038/Nature06860. URL: http://links.isiglobalnet2.com/gateway/Gateway.cgi?GWVersion=2%7B%5C&%7DSrcAuth=mekentosj%7B%5C&%7DSrcApp=Papers%7B%5C&%7DDestLinkType=FullRecord%7B%5C&%7DDestApp=WOS%7B%5C&%7DKeyUT=000255592400042%7B%5C%%7D5Cn(null)%7B%5C%%7D5Cn%7B%5C%%7D3CGo%20to%20ISI%7B%5C%%7D3E://000255592400042.

Zhang, Ying et al. (May 2011). "Object decoding with attention in inferior temporal cortex". In: *Proc. Natl. Acad. Sci. U. S. A.* 108.21, pp. 8850–8855.

Zhu, Song-Chun and David Mumford (2006). "A Stochastic Grammar of Images". In: *Foundations and Trends in Computer Graphics and Vision* 2.4, pp. 259–362. ISSN: 1572-2740. DOI: 10.1561/0600000018.