Adversarial and Real World Image Robustness of Computer Vision Models With Better Human Visual Strategy Alignment

A thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science with Honors in the Neuroscience concentration of Brown University.

By Stephanie Olaiya

Class of 2023

Advised by Dr. Thomas Serre, PhD

April 18th, 2023

ACKNOWLEDGMENTS:

First, I would like to thank my thesis advisor, Dr. Thomas Serre, for helping me deepen my interest and understanding of computer vision models and their applications in Neuroscience. Working on projects in his lab has allowed me to explore my interest in both Neuroscience and Computer Science and how both are able to positively impact the other in various ways. Thank you for welcoming me into your lab and helping me develop my research and coding skills.

I would also like to thank Drew Linsley for providing me with mentorship throughout the duration of the project. Thank you for helping me get familiar with the analytical methods that I needed and guiding me through the steps I needed to take to complete this research.

Thank you to Ivan Rodriguez for always being so open to answering all of my questions and assisting me with the code as I worked on the project. You were an invaluable source of information and I would not have been able to get through the analysis without your support.

Thank you to all the other members of the Serre Lab who patiently offered me advice on how to move forward with the analysis of the project when I was stuck.

Finally, I would like to thank my parents, Dapo and Sade Olaiya, my siblings and the rest of my family for always supporting me and believing in me throughout my time at Brown and as I worked on this project. You all gave me the strength to keep persevering when mine failed me.

TABLE OF CONTENTS:

Acknowledgements.....2

Abstract.....4

Introduction 5

Overview of Artificial Neural Networks in Object

Recognition.....5

Limitations of Artificial Neural Networks: 6

Harmonized Models 8

ObjectNet Dataset…….....12

Adversarial Attacks and Robustness

.....14

Goals and Hypotheses 19

Methods.....20

Results 24

Discussion.....34

Limitations.....38

Conclusion and Future Directions.....39

References 41

ABSTRACT: *Analyzing the Adversarial and Real World Image Robustness of Computer Vision Models With Human Visual Feature Maps*

Deep neural networks (DNNs), which are based loosely on the ventral stream pathway of the primate visual system, are good models of the visual system because they fit neural data of the ventral steam (Yamins et al., 2013). Serre (2019) details that DNNs are also able to mimic functions of the human visual system such as image categorization, where they have achieved high accuracy. However, these models are vulnerable to adversarial attacks–small changes in image pixel values which have no effect on the accuracy of human vision. They have decreased performance when exposed to images from the ObjectNet dataset that contains objects of different rotations, viewpoints and backgrounds. The low adversarial and ObjectNet robustness of models mark a processing difference between models and the human visual system.

Work by Fel et al (2022) has resulted in harmonized models, which are models that are trained on both categorization accuracy loss and a loss in alignment of human feature importance maps that capture the features in an image important for classification. Harmonized models are able to achieve high categorization accuracy as well as alignment to human visual strategies. To further explore the capabilities of harmonized models, we compare the adversarial robustness of harmonized models compared to their corresponding baseline models (models without neural harmonizer) by comparing model accuracy on the Projected Gradient Descent (LinfPGD) adversarial attack from the Foolbox toolbox. Overall, we found that although accuracy drops with increasing image perturbation, harmonized models show greater adversarial robustness compared to baseline models. Furthermore, we compare the accuracy of the models on the ObjectNet dataset and find that harmonization does not increase real world image robustness which we attribute to model limitations due to training on ImageNet.

INTRODUCTION:

Overview of Artificial Neural Networks in Object Recognition:

In recent years, there has been a huge advancement of computer models in the form of artificial neural networks in emulating human vision functions. One of these functions is rapid image recognition where the models are able to achieve human level accuracy on image classification tasks after they have been trained, with some models even exceeding human performance on these tasks (Russakovsky et al., 2015). In humans, image categorization is believed to be performed by the ventral stream of the visual system. The primate ventral stream, commonly known as the "what" pathway of visual processing, has been found to be recurrent in nature in addition to having a hierarchical architecture. In this hierarchical architecture, visual input is passed from the primary visual cortex, to V2 and finally to the area IT. Some research shows that at each step of the pathway, a series of pooling and normalization operations are performed such that there is a tradeoff of selectivity and invariance. This theory is supported by research which shows that there is abstract level category information encoded in area IT which was not found in lower level visual areas such as V1 and V2 (DiCarlo et al., 2012). Thus, invariant object recognition needed for image classification is done rapidly and accurately in humans due to the highly complex hierarchical, recurrent architecture of the ventral visual pathway.

The leading type of models in image classification are deep neural networks (DNNs) which are artificial neural networks consisting of many layers composed of units activated by inputs that feed into them. Similar to how the neural system operates in humans, these layer units send on a weighted output. These weights are learned by the models during training through backpropagation in order to minimize the loss of the model. Loosely modeled after the human

visual system, these models are mainly feedforward hierarchical models in which the input are re-represented throughout the network as it is passed from the first layer (the input layer) to subsequent layers.

The success of these deep neural networks have had an impact on not only the computer vision field but also in the computational neuroscience field. The high performance of these models give an insight into the computational processes occurring in the ventral visual stream. Research shows that deep neural networks develop internal representational spaces that are similar to those seen in the higher level visual areas of human and non-human primates (Majaj et al., 2012; Yamins et al., 2013).

Limitations of Artificial Neural Networks:

DNN models of computer vision face many limitations in generalization and robustness. The performance of deep neural networks on images distorted through blur and noise is much lower than that of humans (Dodge & Karam, 2017). In addition, compared to humans DNNs have more problems generalizing to weaker changes in a variety of other distortions such as rotation, contrast, white and pepper noise and color (Geirhos et al., 2018). While the human visual system may have been exposed to more distortions which results in a good ability to generalize, it has also been shown that training models on distorted images (data augmentation) is not enough to close the gap between humans and the models. These imply that humans have more robust internal visual representations than DNNs and an important aspect of this internal representation is lacking in DNNs. One possible cause for these differences in visual representations may be differences in learning strategies or visual strategies.

Some studies have provided evidence of the difference in visual strategies of models and

humans used in object recognition. Ullman et al (2016) show that there are minimal recognizable configurations (MIRCs) which are crucial to object recognition in humans. A MIRC is defined as an image patch that can be reliably recognized by human observers and which is minimal in that further reduction in either size or resolution makes the patch unrecognizable. The low accuracy of minimal patches in models hints that these models and the human visual system use different processes in object recognition.

Linsley et al (2022) also demonstrated that deep neural networks utilize different visual strategies than those used in humans for image classification through investigating the visual features that both consider important for classification. Models and humans tend to attend to different diagnostic features for classification with models attending to background features. It has been suggested that the visual strategy that humans utilize in object detection combines the use of global and local feature detection. Local features characterize a localized region of the image and are involved in visual saliency (bottom-up attention). Global features characterize the entire image and provide scene contextual information about past search experiences in similar environments and strategies that were successful in finding the target object. Therefore, when detecting and recognizing a target object in a scene or an image, a combination of bottom-up, scene context and top-down attention help predict image regions to fixate on in humans (Lindsay, 2020; Torralba et al., 2006).

Surprisingly, although models show increasing image classification on ImageNet, as the performance of deep neural networks increases on ImageNet, there is a decrease in the alignment to human visual strategies. ImageNet is a collection of labeled images built upon the WordNet structure. The accuracy tradeoff is attributed to the difference in the visual features that humans and DNNs believe are important for recognizing an object. Consequently even though models

have high classification accuracies on ImageNet, there is a drop in accuracy when there is a minor change in these images that still contain the features that the human visual system employs for classification. If the attended feature for a model is grass, then removing the grass from an image of a cow would likely affect the model's ability to recognize the image. However, for a human observer, the recognition of the image as a cow would likely remain unchanged. This is because humans employ attention mechanisms including top down attention and bottom up attention that is lacking in models. Models are learning the statistical associations between labels and image pixels (Linsley et al., 2022) while humans have more of an all-or-nothing dependence on key image features in order to classify an image (Serre, 2019). As a result, during the learning process, models tend to take advantage of dataset biases using these statistical associations such that they achieve high accuracy on datasets similar to those that they are trained with but are not truly 'robust'.

Harmonized Models:

A lot of effort has been put into integrating feature attentional mechanisms or visual strategies similar to humans into models of computer vision. A product of one of said efforts is the development of "harmonized" models, which are models that are trained to both minimize categorical loss and the loss calculated through the difference in the feature importance maps of humans and a DNN. Feature importance maps are the areas that humans were found to be crucial to the categorical identification of an image. Humans tend to attend to similar features during object recognition (Tatler et al., 2010) so previous studies were able to collect data on human feature maps from the *ClickMe* and *Clicktionary* experiments.

There are two human feature importance datasets – *ClickMe* and *Clicktionary*– which can be used to compare the visual strategies of humans and DNNs. The *ClickMe* dataset contains feature importance maps for a non-overlapping set of 196,499 images from ImageNet training and validation sets obtained from a game of two players: a human and a model. Here, the two players are playing with the aim of locating the features in an image that are important for its classification. The human player selects the pixels of these important features one after the other and a blank canvas is filled with the selected pixels until the second player (the VGG16 model) correctly categorizes the image as fast as possible. The *Clicktionary g*ame follows the same steps as *ClickMe* except that the second player is another human, which provides a more reliable idea of what features are important for object recognition in humans. The *Clicktionary* dataset consists of 200 images. For both games, feature importance maps depicting the average object category diagnosticity of every pixel (i.e. the importance of that pixel in classifying the image) was computed as the probability of it being clicked by a participant (Linsley et al., 2022).

During training to develop optimal weights for the object recognition task, instead of the traditional training of models in order to minimize loss on categorization (matching model output to target outputs), models are trained using soft gradient descent optimization process to minimize both categorization loss on the ImageNet training set and feature map loss for images with importance maps from *ClickMe*. This training regimen allows models to be trained with human perceptual judgements. Thus, the training regimen is such that the categorization performance of harmonized models is not worse than their baselines which are versions of the models that are trained only for categorization.

To explain the alignment of model and human feature importance maps mathematically, let us take X as an input space, $Y \subseteq R^c$ as the output space and $f_\theta : X \rightarrow Y$ as the predictor

function that maps an input vector, x, to its output, $f_\theta(x)$. For a deep neural network model, a feature importance map (φ) for a given input x, is defined by the function g such that $\varphi = g(f_\theta, x)$. Thus during training, in addition to the categorical loss, harmonization introduces a loss that enforces feature importance map alignment on all N levels of a Guassian pyramid of an image. With $i \in \{1, ..., N\}$, $P_i(\varphi)$ is the function that maps a feature importance map to its representation on the given layer of the Guassian pyramid. $P_i(\varphi)$ is computed by downsampling $P_{i-1}(\varphi)$ using a Gaussian kernel, with $P_1(\varphi) = \varphi$. Thus, in order to align human and DNN feature importance maps, the feature map loss that is being minimized is $\sum_i^N || P_i(g(f_\theta, x)) - P_i(\varphi)||^2$. See (Fel et al., 2022; Linsley et al., 2022) for more details pertaining to the process of training the neural harmonizer for models.

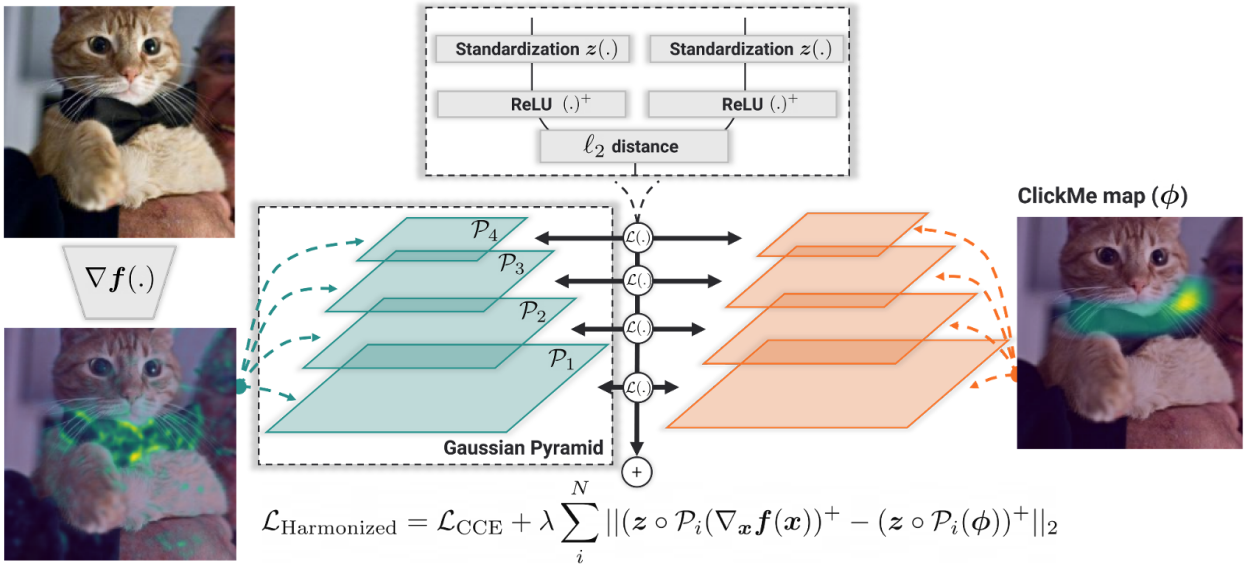


Figure 1: Explanation of the method for aligning human and DNN visual strategies using the neural harmonizer during training. (From Fel et al presentation).

Using Spearman's rank-correlation as a metric to measure the similarity between the feature importance maps of humans obtained from the *ClickMe* dataset and DNNs, harmonized models were found to be more accurate and aligned to human perception than their baselines

(Linsley et al., 2022). This is in contrast to the tradeoff in accuracy and human alignment that is shown in conventional baseline models. It was also reported that the feature importance maps of the harmonized are also less reliant on contexts such as background which bears closer

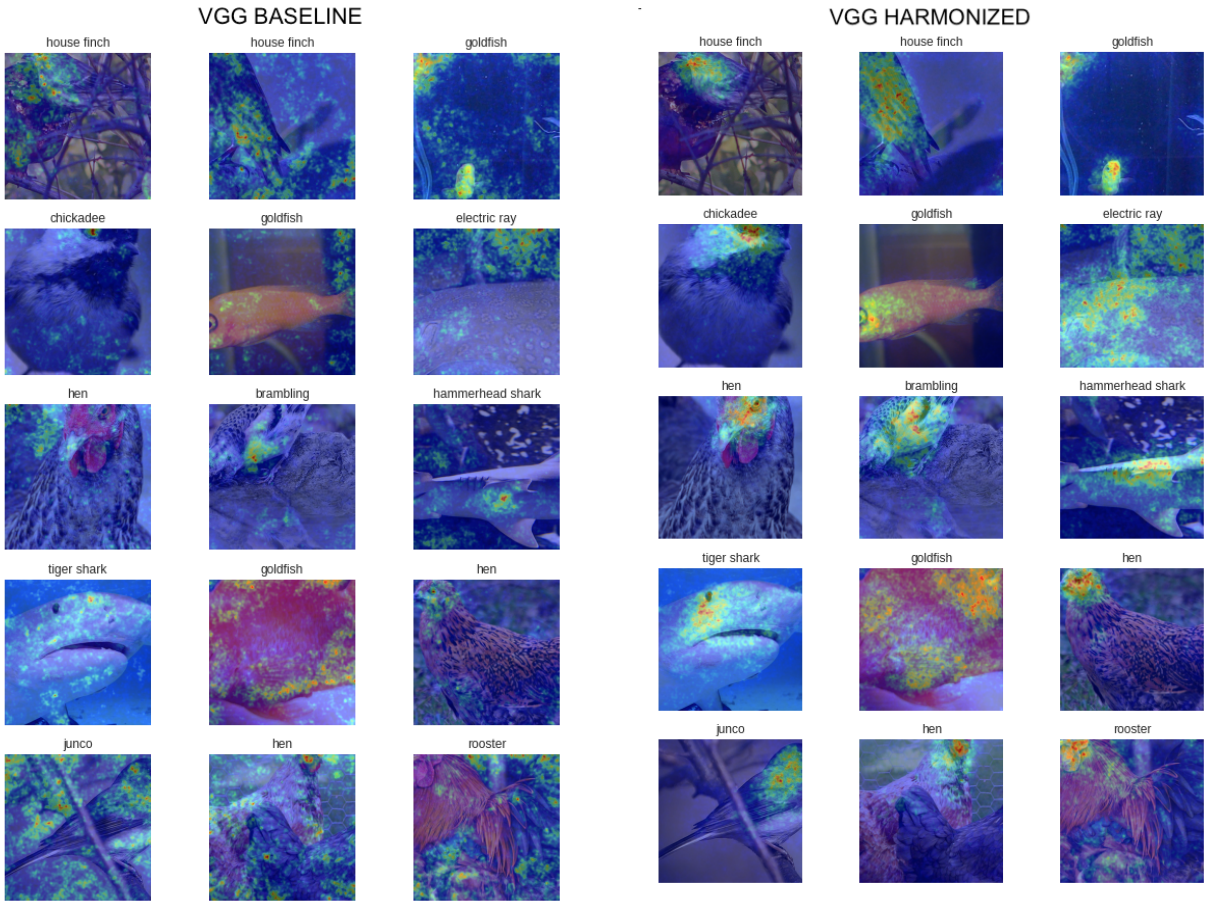


Figure 2: Example ImageNet validation dataset with overlaid saliency (feature importance maps) from the VGG baseline model (left) and VGG harmonized model (right). Image label for each image located at the top.

resemblance to the feature importance maps humans. In addition, the absence of the tradeoff between ImageNet accuracy and human alignment is also found when using the feature importance maps from the *Clicktionary* game. This provides evidence that harmonized models may be better models of human vision, preserving their accuracy on image classification while also possessing the feature maps that are similar to those humans use in object recognition.

Furthermore, using a classification psychophysics experiment, we see that there is evidence that unharmonized models integrate visual information into accurate decisions less efficiently than humans or harmonized models. In the psychophysics experiment, both models and humans use the feature importance maps collected from *Clicktionary* to classify whether an image is an animal or non-animal. This tests whether there is also an alignment between how models and humans integrate these feature importance maps into their decision-making (how they use these features in recognizing objects). Here, it was also found that there is a similar trade-off between ImageNet accuracy and alignment with human visual decision making in baseline models but this tradeoff is absent in the harmonized models.

ObjectNet Dataset:

One of the major benchmarks of current models in the computer vision field is accuracy on the ImageNet test datasets. As of today, many models are able to achieve high accuracy in ImageNet, meaning they are able to match the class of their output to the target ImageNet label corresponding to an image (for many images in the ImageNet validation and test datasets). However, ImageNet and other popular image datasets have been found to be biased in their image conformations, such as having images where objects correlate with the background e.g. a snake in the grass, where there is little occlusion and where objects are in stereotypical rotations. Therefore, there are questions challenging the merit of accuracy on these datasets as benchmarks for the relationship between the performance of models of computational vision and that of humans.

Many efforts have been made to create datasets for object recognition consisting of images that are more representative of how objects are viewed in real life by humans. One of

these datasets is ObjectNet, which is a dataset of 50000 crowdsourced images of real world objects in natural settings where there are controls for object rotations, viewpoints and backgrounds. Due to the nature of the data collection, the images in the dataset are of household object which are captured in 50 rotation angles, 4 backgrounds (kitchens, living rooms, bedrooms, washrooms), from 3 viewpoints (top, angled at 45 degrees, and side) (Barbu et al., 2019). The dataset contains images of 313 classes, 113 of which are overlapping with ImageNet.

Researchers found that the performance of popular deep neural networks dropped 40-45% on the ObjectNet dataset compared to ImageNet, regardless of looking at the Top-1 or Top-5 predictions, indicating that models are biased towards the nature of images on ImageNet. It also reveals that these models are not robust enough to correctly identify objects in real-life/natural environments which the human visual system is able to accomplish. Therefore, a higher performance on ObjectNet may be indicative that a model is better aligned with the human visual system as it is able to better recognize objects in real-life, unstaged settings.

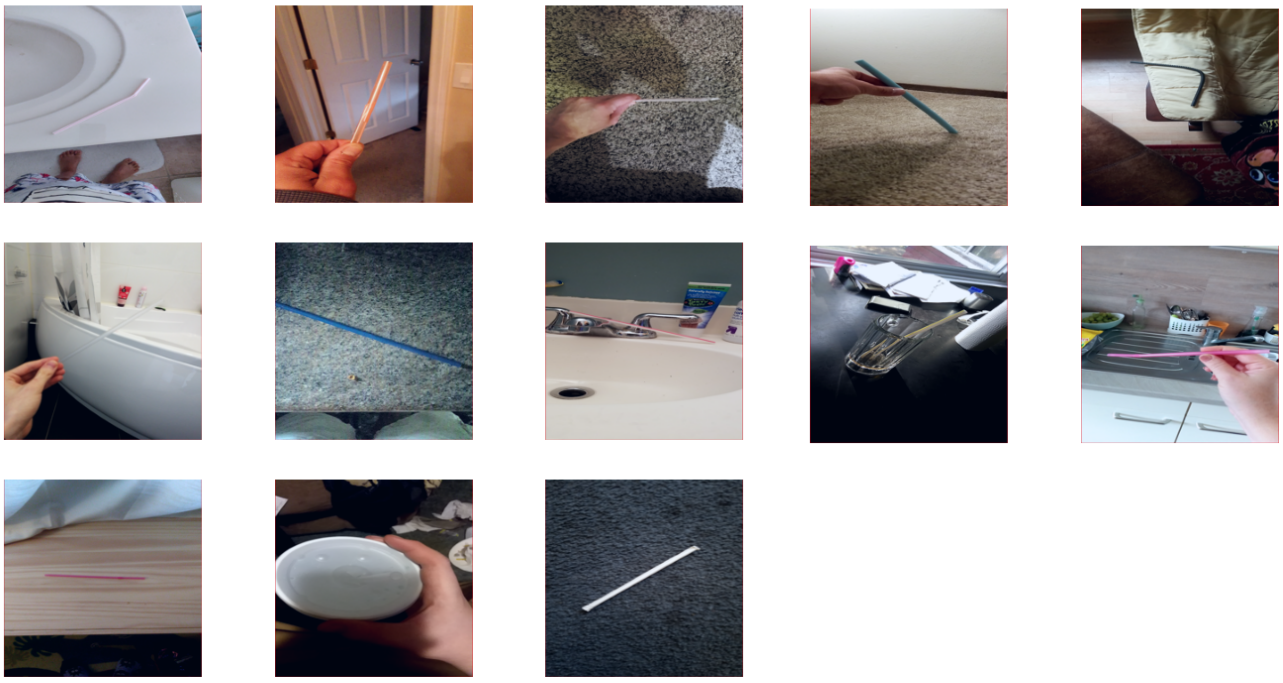


Figure 3: Example images of the "drinking straw" label from the ObjectNet dataset

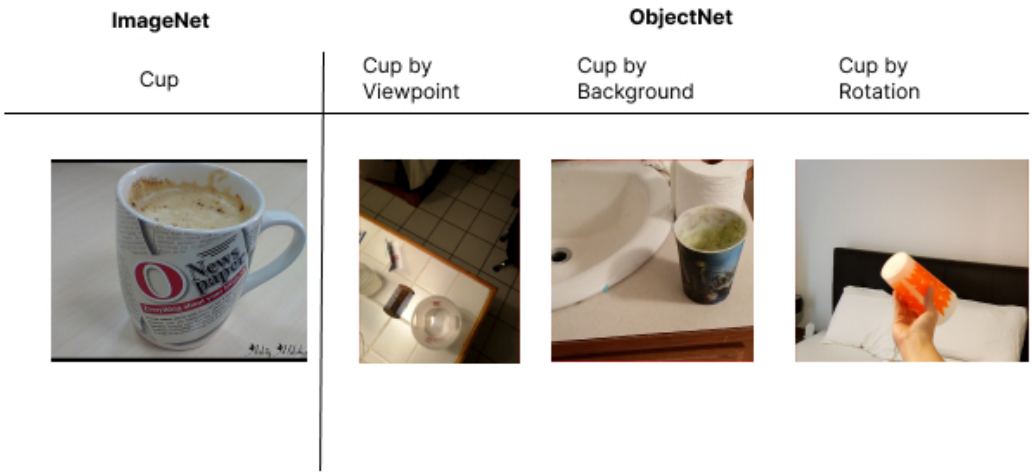


Figure 4: ImageNet (left column) shows objects on a typical background, not rotated and from a front facing viewpoint. Typical ObjectNet objects (right column) are imaged in rotated, on different backgrounds, from multiple viewpoints. The first three columns show a cup varying by the three properties that are being controlled for: rotation, background, and viewpoint.

***Adversarial Attacks and Robustness**:*

Adversarial examples are images that are able to fool a model into misclassifying. There are broadly two types, "fooling images" and perturbed images. Fooling images are meaningless images that are classified as familiar objects by models but mean nothing to humans such as a pattern of noise being classified as a robin or a cheetah (Nguyen et al., 2015; Zhou & Firestone, 2019). Nyugen et al found that these unrecognizable images are incorrectly classified as correct images by model with high confidence, a striking difference to the human visual system. Perturbed images are images where an intentional small perturbation has been applied such that a model that formerly correctly classifies said image, now incorrectly classifies the image with high confidence (Zhou & Firestone, 2019). These images are another evidence of the difference between human and machine perception.

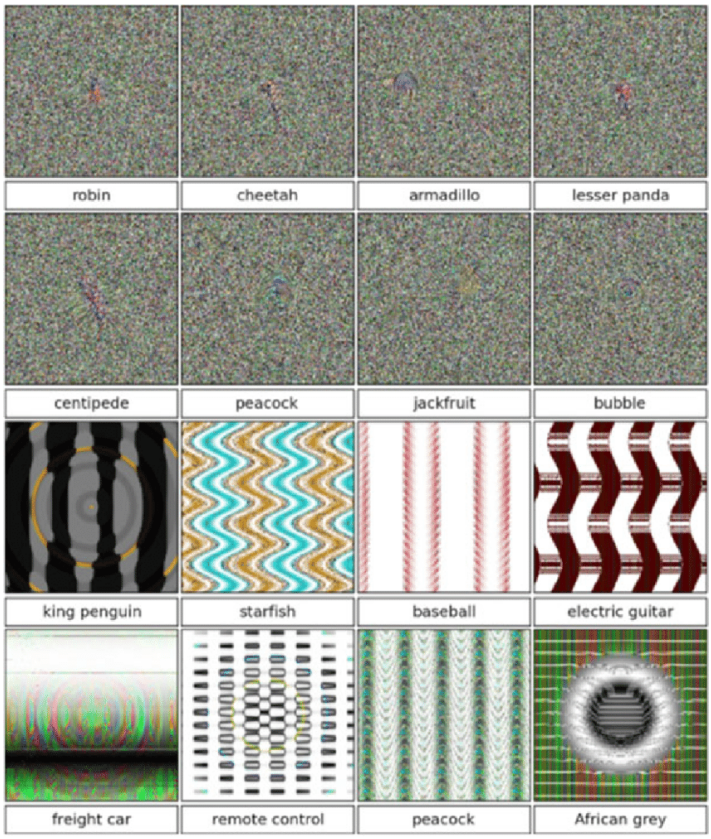


Figure 5: Examples of adversarial images with their incorrectly predicted model output label. Images are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with ≥ 99.6% certainty to be a familiar object. Images are either directly (top) or indirectly (bottom) encoded (from Nyugen et al paper)

The adversarial image type focused on in this study are perturbed images. There are many ways to generate these classes of adversarial images. These approaches can be broadly classified into gradient based attacks, score based attacks, transfer based attacks and decision-based attacks that take advantage of the loss/gradients, the predicted scores (e.g. class probabilities or logits), training data information and the top-1 prediction of a model respectively in order to fool it (Brendel et al., 2017; Ciolino et al., 2021).

I. Gradient based attacks are attacks that linearize the loss around an input x to find directions ρ to which the model predictions for class ℓ are most sensitive to:

$$L(x + \rho, \ell) \approx L(x, \ell) + \rho^{\top} \nabla_x L(x, \ell).$$

where the gradient of the loss with regards to input x is $g(x) = \nabla_x L(x, \ell)$, x is a model input , ℓ a class label, x the reference input and ℓ the reference label such that $L(x, \ell)$ is the loss.

II. Score-based attacks are attacks that use predicted scores (e.g. class probabilities or logits) of the model to numerically estimate the gradient of the loss. An example is the Local Search Attack which applies perturbations to some pixels to construct a new perturbed image until the probability of the correct class is below a given kth place. It does this by searching for additional critical pixels in the neighborhood of previously found ones (Brendel et al., 2017; Narodytska & Kasiviswanathan, 2016; Rauber et al., 2018)

III. Transfer-based attacks are attacks that rely on the data that is used to train a DNN in order to train a substitute model that can be used to generate adversarial images. These attacks are based on the fact that adversarial examples for one model are transferable to another.

IV. Decision-based attacks are attacks that rely on the top-1 prediction (class-decision) of a model i.e. its classification 'decision' in order to generate adversarial examples. They do not require gradients or probabilities. An example is the Boundary which starts from a large adversarial perturbation and then seeks to reduce the perturbation while staying adversarial, i.e. while the decision is still incorrect (decision-based).

Research shows that models are highly susceptible to adversarial attacks such that they are easily influenced by these changes while the image changes are insusceptible to humans

(Goodfellow et al., 2014). There have been many suspected causes for this vulnerability with researchers considering overfitting and extreme nonlinearity and others considering too much linearity as possible explanations for the effect. This susceptibility is also further emphasized by the inability of these state of the art models to generalize to the same degree as humans. Goodfellow et al found that an adversarial example generated for one model is also misclassified by other models. Furthermore, the models tend to misclassify these images to the same categories. They hypothesized that adversarial examples for one model have been seen to be transferable to others due to the alignment between the adversarial perturbations and the weights of a model (Goodfellow et al., 2014). During training, models are learning parameters for an arbitrary function that maps input presented to the model to outputs. Therefore, when models are trained on the same task, they develop similar functions, so adversarial examples that affect one model are likely to affect the other that was trained similarly on the same task.

There are many possible reasons as to why humans are far less susceptible to adversarial attacks than models are. One such possible reason is the early vision limitations of the human visual system. It is possible that due to the limitations of low level image processing in the human visual system such as human visual acuity, resolution, and sensitivity to contrast, the perturbations on images during adversarial attacks are not visible to humans (Zhou & Firestone, 2019). Adversarial attacks may be less about the high level visual processing differences between humans and models and more about the early vision processing differences. Another related reason for “adversarial robustness” seen in humans is that if a perturbation requires high acuity in the periphery of an image, due to high foveal resolution in the fovea which decreases with increasing eccentricity, this perturbation would be undetectable by the eye, and thus would have no impact on human perception (Elsayed et al., 2018). Also, these perturbations might also

not be visible as pixel changes when shown to humans in a monitor due to image display limitations as well, making it impossible for humans to notice a change in the images. However, models are able to detect the slightest change in pixel value and factor these input changes into subsequent predictions.

In addition to adversarial attacks, there has also been efforts to create datasets of adversarial images. An example of a dataset is the natural adversarial dataset (ImageNet-A) which consists of clean natural images that are difficult for models but easy for humans. Images in this dataset do not undergo synthetic perturbations like those previously explained but are images that take advantage of the context clues that are usually present in images to fool models, thus making them "harder" to classify (Hendrycks et al., 2021). The images contained in this dataset belong to the same 1000 classes as those in the ImageNet training and validation dataset but the images that are easily classified by Resnet50 are removed. Research has shown that these adversarial images are also transferable across models as they cause consistent classification mistakes in different models with a DenseNet-121 model obtaining around 2% accuracy, an accuracy drop of approximately 90% (using ImageNet validation as a reference).

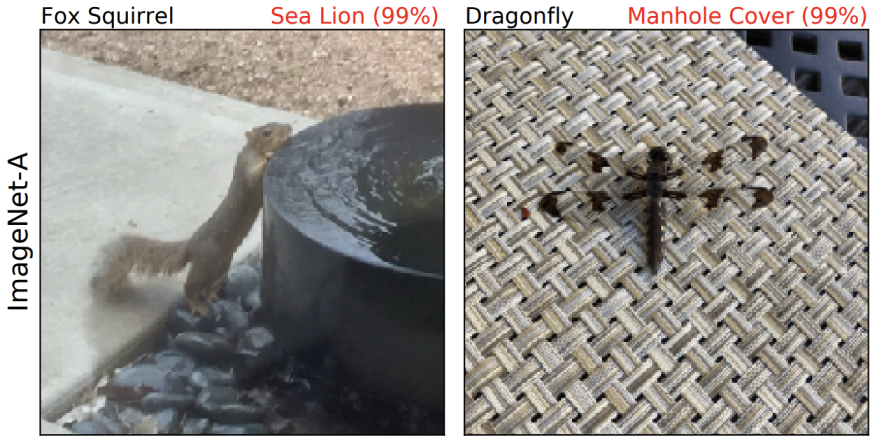


Figure 6: Example images from ImageNet-A dataset. The black text is the actual class, and the red text is a ResNet-50 prediction and its confidence (from Hendrycks et al paper).

Goals and Hypotheses:

Previous studies have shown that there is a tradeoff in adversarial robustness and ImageNet accuracy of models. We see a similar tradeoff in accuracy with the ObjectNet dataset, which indicates that current models in the field are missing a key component present in the human visual system that renders humans insusceptible to these effects on images. One reason for this could be that in general, humans are exposed to many more images than deep neural network models could possibly be trained with. Including adversarial images in the training datasets for models has been found to reduce the susceptibility of the models to similar adversarial examples of those they are trained with. However, while this improves the performance of these models, it is still not a method that closes the knowledge gap in understanding human and machine perception as it is a data driven method. In addition, unlike other datasets, showing models on images similar to those found in ObjectNet during training ie finetuning does not result in a large increase in performance in the dataset (Barbu et al., 2019). Thus, looking away from the data driven approaches and towards architectural ones that align models with the human visual system to increase the robustness of these models may be key in producing models that perform better on these benchmarks of vision. Work done by Linsley et al have attempted an architectural approach by aligning the visual strategies of models and humans to create “harmonized models”. Therefore, in this study we will be evaluating the adversarial robustness and objectness performance of harmonized models compared to their baselines. This would potentially evidence that incorporating visual strategies similar to those found in humans in DNNs would allow these models to attend to similar features as humans, making them to be less vulnerable to adversarial attacks which target the statistical associations that models utilize. We also hypothesize that the alignment of model feature important maps with that of humans

would result in better performance on the ObjectNet dataset. This would show that creating models of vision that better incorporate aspects of human vision would also lead to better performance and robustness of these models with conventional computer vision benchmarks as well.

METHODS:

There are two methods of evaluation used to analyze the robustness of the harmonized models compared to their baselines. Firstly, we will be evaluating the adversarial robustness of both harmonized and baseline models through an implementation of the Linf Projected Gradient Descent Attack in the Foolbox adversarial attack toolkit. Secondly, we compare the robustness of the harmonized to real world images in the ObjectNet dataset to that of the baseline models. We then look at the relationship between human alignment and the adversarial robustness of the models. We will be running these analyses on seven harmonized models: VGG16, resnet_sgd, EfficientNet, LeViT, ViT, ConvNeXT and MaxViT and the corresponding baseline model for each of the above-mentioned harmonized models.

Foolbox Adversarial Attack Toolbox:

The Foolbox adversarial toolbox is a python toolkit which allows the implementation of the popular adversarial attacks that are used to test models today. In general, in order to create an adversarial image, five elements are needed (Rauber et al., 2018):

- A model to run the attack on, i.e., an image classification model that takes in an input and returns an output as a prediction

- A criterion that defines if an image is adversarial or not. The default criterion in the toolkit is misclassification i.e. whether or not the model correctly classifies an image.
- A distance measure that quantifies the size of a perturbation that is applied to an image
- An attack algorithm that combines the other four elements by taking in an image and its corresponding label, the model, the criterion and the distance measure in order to generate a perturbed image to run against the model.

To evaluate the adversarial robustness of models, we implemented the Linf Projected Gradient Descent (LinfPGD) attack with 1000 images from the imagenet validation dataset with perturbation epsilons of 0, 0.0002, 0.0005, 0.0008, 0.001, 0.0015, 0.002, 0.003, 0.01, 0.1, 0.3, 0.5, and 1.0.

To explain how LinfPGD works, let x represent a model input , ℓ a class label, x_0 the reference input, ℓ_0 the reference label such that L(x, ℓ) is the loss (e.g. cross-entropy).

Fast-Gradient Sign attack is a type of gradient based attack where the gradient of the loss with regards to input x, $g(x_0) = \nabla_x L(x_0, \ell_0)$ is computed once and then computes the minimum step size ϵ such that $x_0 + \epsilon \text{ sign}(g(x_0))$ is adversarial (Goodfellow et al., 2014). Basic Iterative Method (BIM) is an iterative form of FGSM where the FGSM attack is performed multiple times, iteratively (Kurakin et al., 2017). At each iteration, the generated pixel value is projected back on the ε-norm ball around the original input at each iteration of the attack (Nicolae et al., 2019; Wang et al., 2020). This projection or "clipping" of the pixel value ensures that the pixel values are within a certain distance (ε-neighborhood) of the original image:

$$\boldsymbol{X}_0^{adv} = \boldsymbol{X}, \quad \boldsymbol{X}_{N+1}^{adv} = Clip_{X,\epsilon}\left\{\boldsymbol{X}_N^{adv} + \alpha \operatorname{sign}\left(\nabla_X J(\boldsymbol{X}_N^{adv}, y_{true})\right)\right\}$$

Projected Gradient Descent (PGD) is an iterative extension of the FGSM attack method that is very similar to BIM, except that this algorithm adds randomization such that a random point in the ε-norm ball of interest around the original image is chosen as the starting point where BIM starts at the original image (Feinman et al., 2017). Linf Projected Gradient Descent is a PGD attack with order = Linf. (Rauber et al., 2018).

Using the LinfPGD attack, we compute the Top-1 accuracy of each model with the perturbation at each epsilon step. We then compare the accuracies of each harmonized model to its corresponding baseline model. In addition, we compute the accuracy difference between the harmonized and baseline model at each epsilon in order to analyze how the difference in adversarial robustness for all model pairs (a harmonized model and its corresponding baseline) changes with increasing epsilons.

A two-way ANOVA is performed to analyze the effect of the epsilon of perturbation and model type (either baseline or harmonized) on the accuracy of the model.

Evaluating Models on ObjectNet Dataset:

We evaluate the robustness of harmonized and baseline models by computing the accuracy of the harmonized and the baseline models on the images in the ObjectNet dataset. However, because only 113 out of the 313 of the categories in the dataset overlap with the 1000 imagenet classes that the models are trained with, we only tested the accuracy of the models to the images belonging to the 113 overlapping categories. We used the top-1 and the top-5 prediction accuracy as evaluation metrics for robustness in this dataset. Top-1 accuracy computes the accuracy of the model based on if the top prediction of the model matches the true label for an image. Top-5 accuracy computes the accuracy of the model based on if the true label for a

given image is in the top 5 predictions of the model (i.e. the five classes that have the highest probability in the output of the model).

A one way ANOVA is performed to compare the effect of the addition of the neural harmonizer to Top-1 and Top-5 accuracy on the ObjectNet dataset.

Correlation of Adversarial Robustness with Human Alignment Metrics:

We highlighted above that due to their training, harmonized models consider both the categorization loss and also the loss on the feature importance maps alignment to humans. This allows harmonized models to have saliency maps that closer resemble human feature importance maps. Using the feature importance maps obtained from the *ClickMe* dataset, we compute the human alignment score of a given model as the mean Spearman correlation between human and the model's feature importance maps, normalized by the average inter-rater alignment of humans (Fel et al., 2022). Thus, we evaluate the correlation between the difference in human alignment scores and the difference in adversarial robustness of the model pairs using the Spearman correlation coefficient as a metric where the adversarial robustness of a given model is measured by the mean Top-1 accuracy of the model over all the epsilons of the LinfPGD attack.

Correlation of Real World Image Robustness with Human Alignment Metrics:

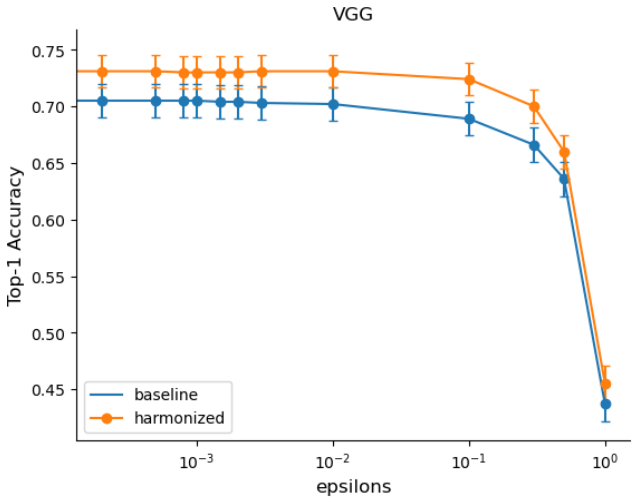
In order to analyze if harmonization truly leads to more robustness of the ObjectNet dataset, we evaluate the correlation between the difference in human alignment scores and the Top 1 accuracy difference of the harmonized-baseline model pairs using the Spearman correlation coefficient as a metric. We also compute the Spearman correlation between the difference in human alignment scores and the Top-5 accuracy difference.

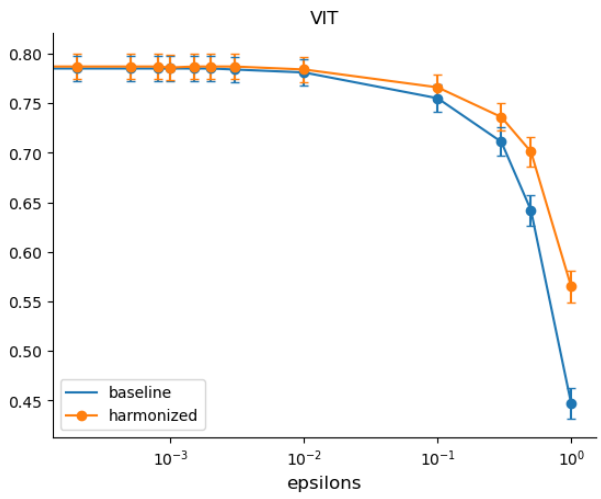
RESULTS:

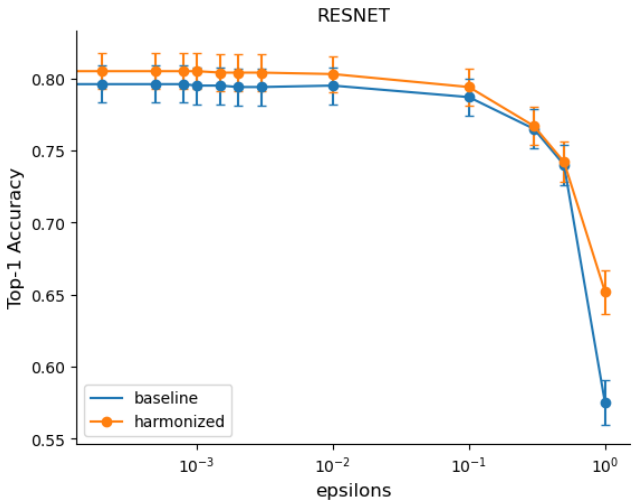
Accuracy to Linf Projected Gradient Descent Adversarial Attack:

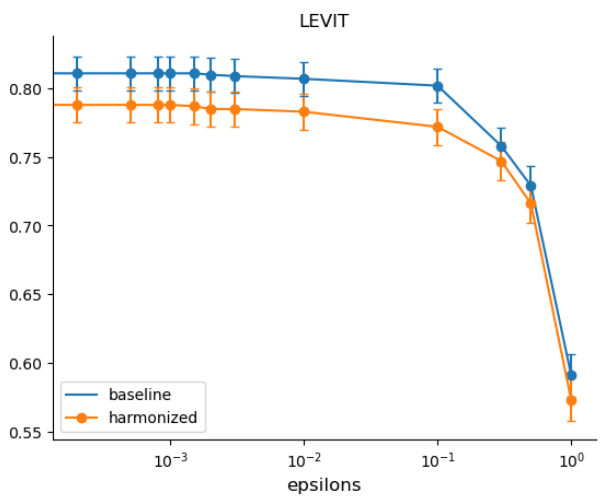
We obtained the Top-1 ImageNet accuracies of seven models at different epsilons: 0, 0.0002, 0.0005, 0.0008, 0.001, 0.0015, 0.002, 0.003, 0.01, 0.1, 0.3, 0.5, and 1.0. Consistent with previous results, we see that for all models except the LeViT models, the harmonized models have a higher accuracy on the unperturbed images than the baseline models. We also see that at epsilon 0 (unperturbed images), that all models have an accuracy greater than 70% (with the lowest accuracy being the VGG16 baseline model with an accuracy of 71.2%).

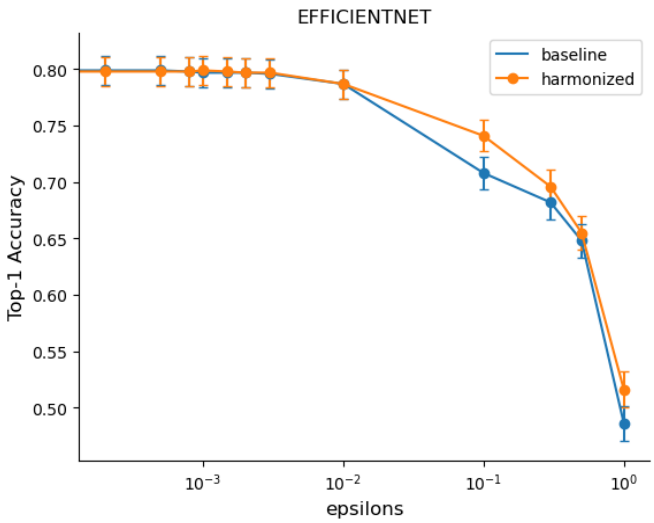
We also see that with increasing epsilon of perturbation, all models show a decrease in accuracy with the epsilon of 1.0 having the lowest accuracies for all models. In general, the slope of decline for harmonized models is less steep compared to their baselines. We see that for all model pairs except the LeViT model, the harmonized models have a higher accuracy than their corresponding baselines for all epsilons.

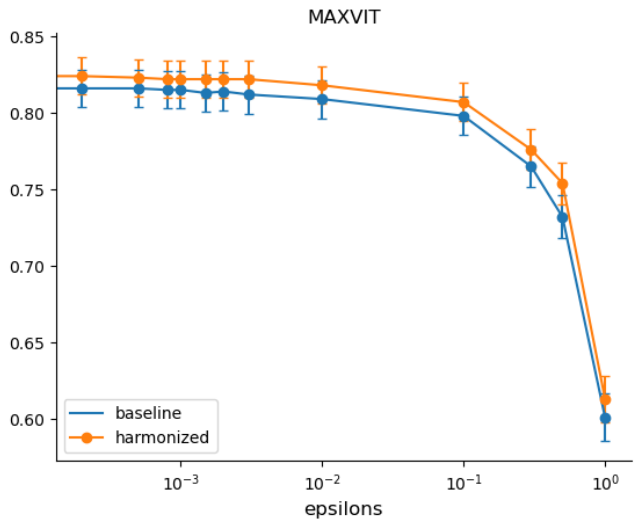

VGG
Top-1 Accuracy
epsilons
baseline
harmonized


VIT
epsilons
baseline
harmonized


RESNET
Top-1 Accuracy
epsilons
baseline
harmonized


LEVIT
epsilons
baseline
harmonized


EFFICIENTNET
Top-1 Accuracy
epsilons
baseline
harmonized


MAXVIT
epsilons
baseline
harmonized

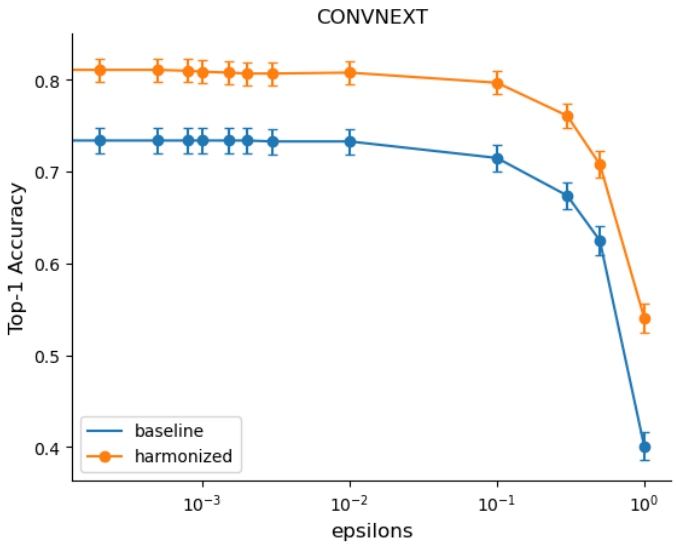


Figure 7: Top-1 ImageNet accuracy of Harmonized (orange) and corresponding baseline (blue) models at each epsilon of perturbation value.

In agreement with previous results, all model pairs excluding the LeViT models have a harmonized-baseline accuracy difference greater than 0 at all epsilons. We also see that looking at the difference in accuracy between the models at each epsilon, for four out of the seven models, with increasing epsilon magnitude, there is a trend of an increasing gap between the accuracy of the harmonized and baseline models.

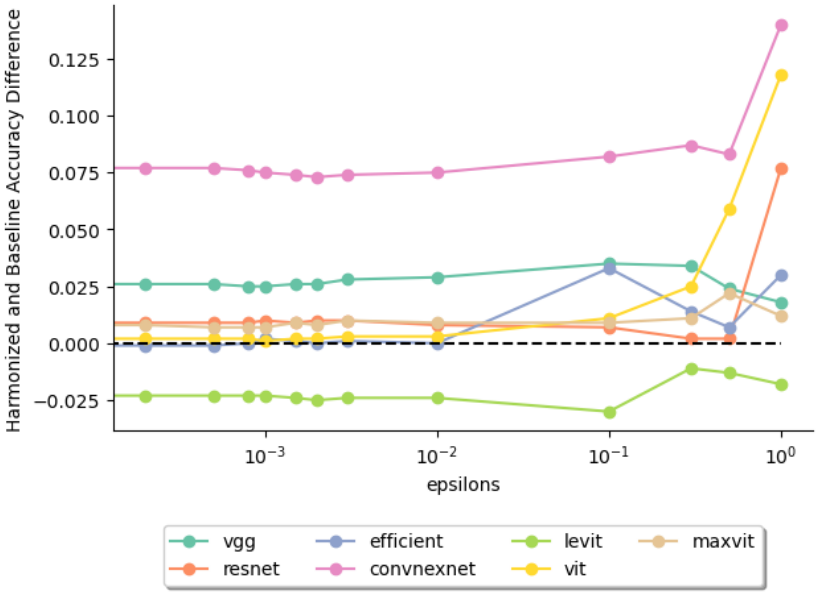


Figure 8: Difference in Top-1 accuracy between seven harmonized models and their corresponding baseline at each epsilon of perturbation value. Top-1 ImageNet accuracy of Harmonized (orange) and corresponding baseline (blue) models at each epsilon of perturbation value.

Furthermore, we found a statistically-significant difference in top-1 accuracy by both the epsilon in adversarial perturbation ($F(12)=42.22$, $p=7.79e\text{-}43$) and by model type ($F(1)=9.92$, $p=1.96e\text{-}03$), though the interaction between these terms was not significant. Therefore, we see that it is very likely that there is a significant difference in the top-1 accuracy of the harmonized and baseline models using adversarially perturbed images.

	df	sum_sq	mean_sq	F	PR(>F)
epsilons	12.0	0.866632	0.072219	42.221409	7.793775e-43
model_type	1.0	0.016962	0.016962	9.916339	1.963540e-03
epsilons:model_type	12.0	0.005278	0.000440	0.257152	9.943660e-01

Residual	156.0	0.266837	0.001710	NaN	NaN

Table 1: Descriptive statistics for Top-1 accuracy by model type (harmonized vs baseline and epsilon including results for the interaction of model type and epsilon

How does adversarial robustness change as a function of human alignment?

We then consider the relationship between the difference in the mean top-1 accuracy score across all epsilons and the human alignment score difference for all the harmonized-baseline model pairs. We see that there is a strong positive correlation between adversarial robustness and the human alignment score according to the ClickMe data (R^2 score = 0.7946). Results of the Spearman correlation indicated that there is a positive association between ClickMe human alignment scores and mean Top-1 accuracy scores but this association is not significant, (rs[7] = 0.7500, n.s.)

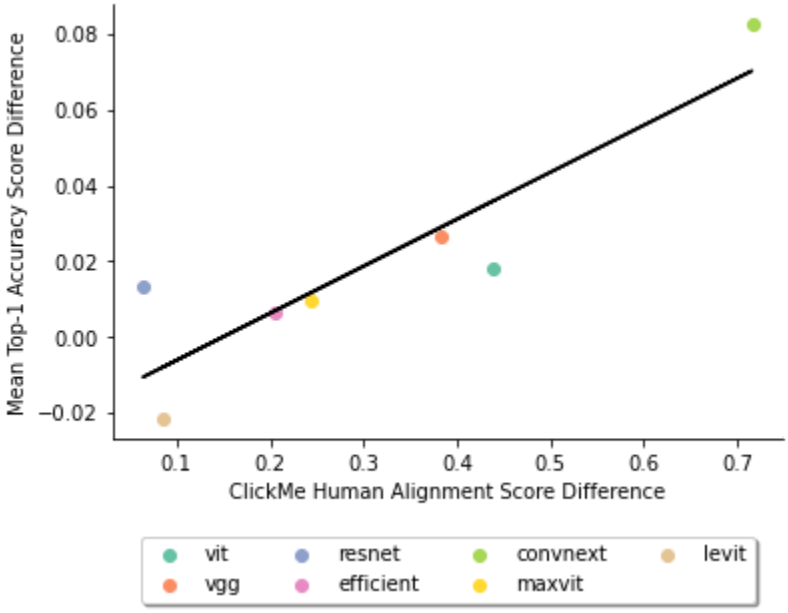


Figure 9: Difference in Top-1 accuracy on adversarial images between harmonized and baseline models against difference in human alignment scores.

Accuracy on Objectnet Dataset:

We obtained the Top-1 and Top-5 accuracies of the harmonized models and their corresponding baselines to the overlapping imagenet classes in the ObjectNet test dataset. Using a one-way ANOVA, we find that there is a significant difference between the Top-1 and Top-5 accuracy Objectnet accuracy of models ($F(12) = 95.9535$, $p < 0.001$).

We see that looking at these accuracy scores of the fourteen models, for five pairs of the models, the model with the neural harmonizer is slightly more accurate than its corresponding baseline model.

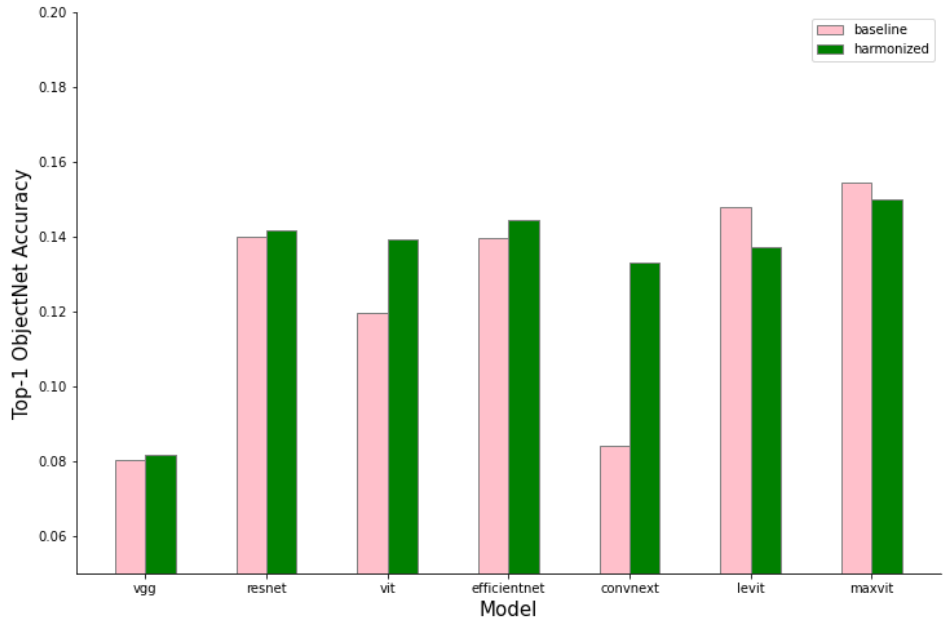


Figure 10: Top-1 ObjectNet accuracy results for seven harmonized models (green) and their corresponding baselines (pink)

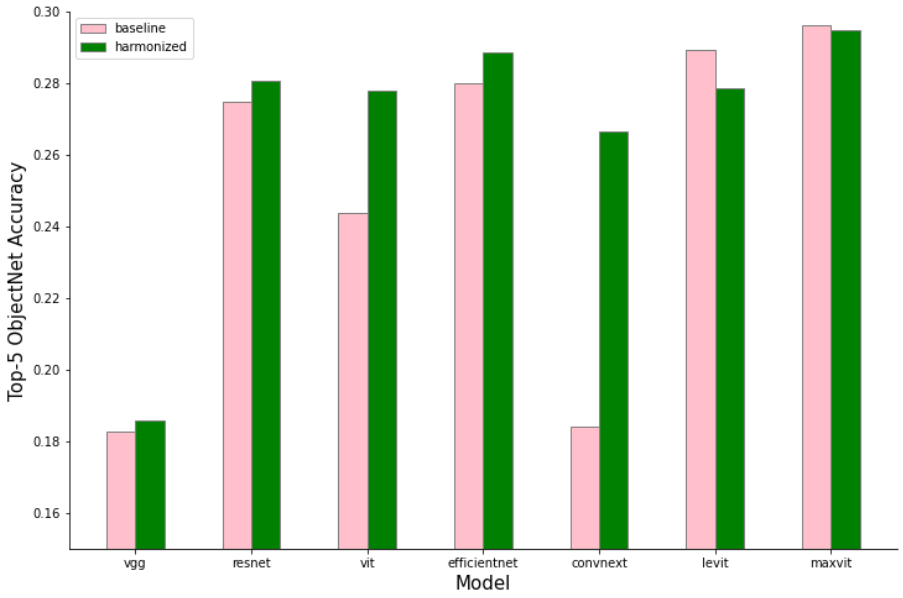


Figure 11: Top-5 ObjectNet accuracy results for seven harmonized models (green) and their corresponding baselines (pink)

Model	Harmonized ObjectNet Top-1 Accuracy	Baseline ObjectNet Top-1 Accuracy
VGG16	8.16 %	8.02 %
Resnet	14.16 %	14.00 %
ConvNeXT	13.29 %	8.40 %
LeViT	13.71 %	14.77 %
ViT	13.94 %	11.96 %
MaxViT	14. 99 %	15.45 %
EfficientNet	14.44 %	13.95 %

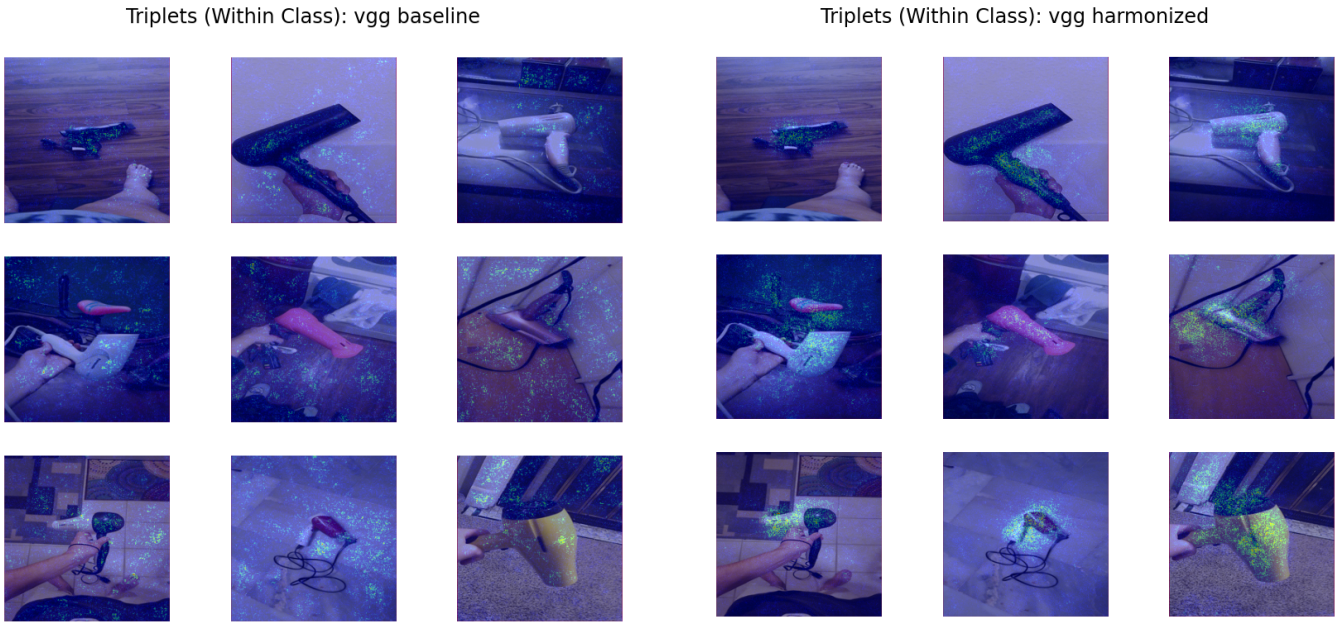
Table 2: Top-1 ObjectNet accuracy results for seven harmonized models and their corresponding baselines (higher values highlighted)

Model	Harmonized ObjectNet Top-5 Accuracy	Baseline ObjectNet Top-5 Accuracy
VGG16	18.59 %	18.27 %

Resnet	28.06 %	27.47 %
ConvNeXT	26.64 %	18.40 %
LeViT	27.86 %	28.93 %
ViT	27.80 %	24.37 %
MaxViT	29.48 %	29.61 %
EfficientNet	28.86 %	28.01 %

Table 3: Top-5 ObjectNet accuracy results for seven harmonized models and their corresponding baselines (higher values highlighted)

However, using a one-way ANOVA, we find that the differences in top-1 accuracy of the harmonized and baseline models are not significant ($F(12) = 0.3710$, n.s). The differences in the Top-5 accuracy of harmonized and baseline models are also not significant ($F(12) = 0.5730$, n.s). Thus, based on the models studied, the harmonized models do not perform significantly better than the baseline models on real world images used in the ObjectNet dataset.



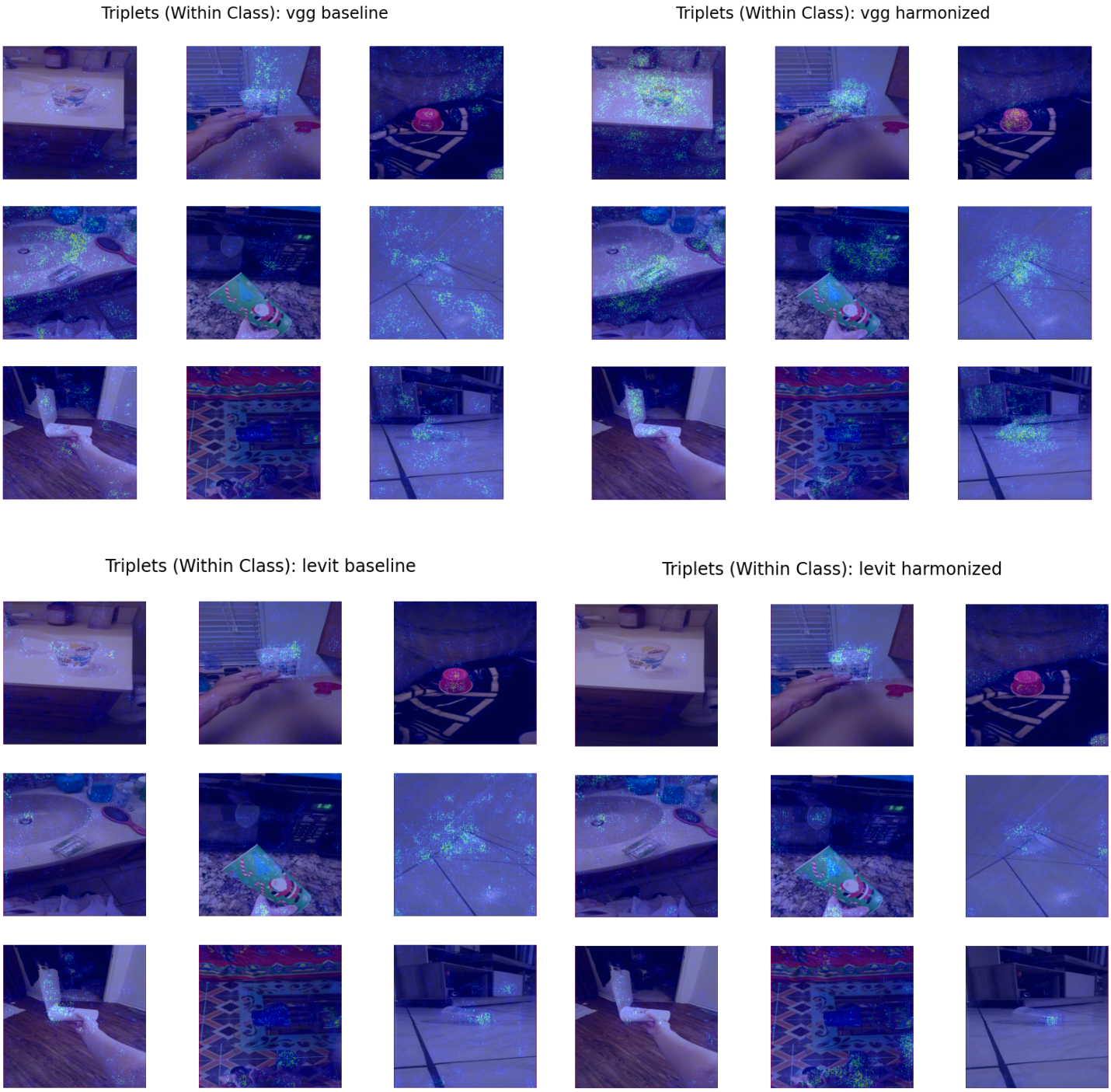


Figure 12: Example ObjectNet dataset images with overlaid saliency (feature importance maps) from the VGG baseline model (upper and middle left), VGG harmonized model (upper and middle right), LeViT baseline (lower left) and LeViT harmonized (lower right). Images of hairdryer (first three rows) and plastic cup (last nine rows)

How does real world image robustness change as a function of human alignment?

We then consider the relationship between the ObjectNet Top-1 accuracy and the human alignment score for all the models. We see that there is a strong positive correlation between the performance on the model performance on the ObjectNet dataset and increase in human alignment score according to the ClickMe data due to harmonization (R^2 score = 0.8029). Results of the Spearman correlation indicated that there is a positive association between ClickMe human alignment score difference and Top-1 ObjectNet accuracy difference but this association is not significant, (rs[7] = .6071, n.s.)

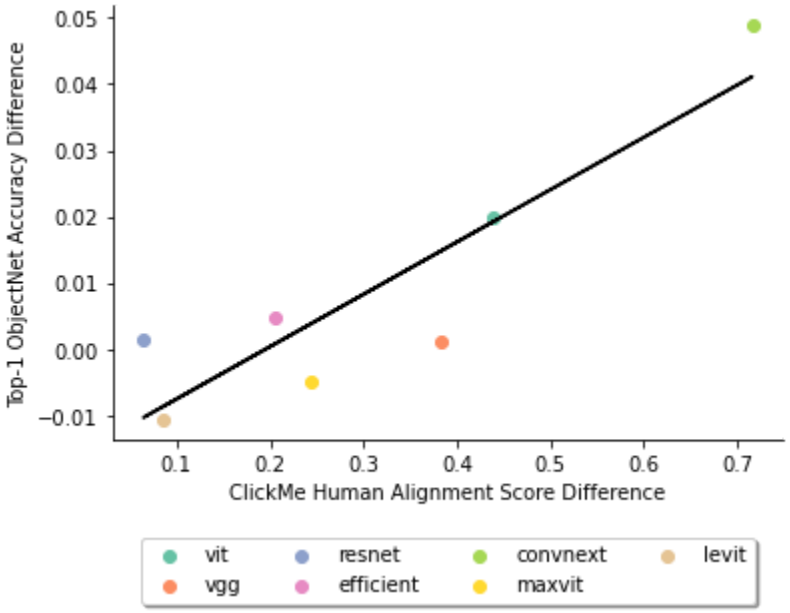


Figure 13: Difference in Top-1 ObjectNet accuracy between harmonized and baseline models against difference in human alignment scores.

In addition, results of the Spearman correlation indicated that there is also a positive association between the difference between difference in *ClickMe* human alignment scores and difference in Top-5 ObjectNet accuracy ($R^2 = 0.7965$) but this association was not significant, (rs[7] = 0.6071, n.s.).

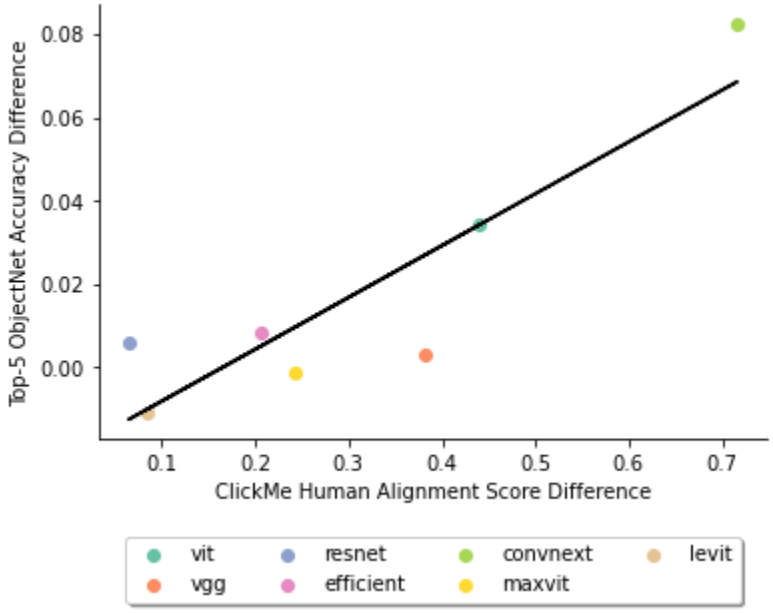


Figure 14: Difference in Top-5 ObjectNet accuracy between harmonized and baseline models against difference in human alignment scores.

DISCUSSION:

Adversarial Robustness

The purpose of this study was to evaluate the association between human alignment of models and their performance on adversarial and real world images. Previous research has shown that current models in the field that achieve high accuracy on ImageNet have very low performance on these images. This is in contrast to the human visual system being able to correctly classify these images that successfully fool models. This work is a continuation of previous work done by (Linsley et al., 2022) which shows that including a neural harmonizer in

the training paradigm of a model allows it to have both better aligned feature-importance maps and high ImageNet accuracy. The goal of this study is to evaluate if based on the more human-like feature importance maps and high ImageNet accuracy, harmonized models will be more robust to adversarial attacks and real world images. This work provides more evidence that models that are better aligned with humans are less susceptible to adversarial attacks.

Overall, the results indicate that harmonized models perform better than baseline models on one of popular adversarial attacks. We see that with increasing epsilon of perturbation in the adversarial attack used, the harmonized models continue to outperform their corresponding baseline models. One possible reason for the better adversarial robustness of harmonized models could be that due to the alignment of visual strategies, slight changes in pixel values in the non-important features of an image would have less of an effect on the classification decision of a given model. This is because the integration of a neural harmonizer results in a model developing stronger weights on the features of an image that are important for classification and less on context. As a result, we see that in general, harmonized models are able to perform better than baseline models without the need for fine-tuning models by training with more adversarial data.

The result also supports previous work done by Lindsay et al (2022) by showing that harmonized models in general have higher accuracy than baseline models on unperturbed images. In addition, also in support of previous works on harmonized models, we see that there is an absence of tradeoff between human alignment and accuracy on the unperturbed images.

ObjectNet Robustness

We see that there is no significant improvement in performance on the ObjectNet dataset with the integration of the neural harmonizer in models. One possible reason for this is because feature importance maps collected and used to calculate the loss during training are collected based on ImageNet images. The images in the ObjectNet dataset depict objects in different rotations, backgrounds and viewpoints, which is distinctly different from those seen in the ImageNet dataset where images are centered and backgrounds usually provide context in the images. However, our results also reveal that the saliency maps that harmonized models are trained to develop which are similar to humans transfer well to ObjectNet images. Previous studies theorized that one of the reasons for poor model performance on ObjectNet dataset is that models are unable to capture the same generalizable features that humans use. However, our results indicate that this does not hold true for harmonized models, with their saliency maps showing that they are capturing the important classification features in an image. This is unlike the baseline models where areas of saliency are sometimes the background or another contextual object such as a hand holding the plastic cup. Therefore, the continued low Top-1 accuracy of even harmonized models on the real world images may be because there are many objects in the same image which the model determines is important for classification. For example, in one of the images in Figure 12, both a cup and a hairbrush are placed on a sink and the saliency maps for both the VGG and LeViT harmonized models show that these models also considered the hairbrush a feature of importance in the image in addition to the cup, which was the target label.

Furthermore, the nature of the images in the ObjectNet dataset could explain the low performance of harmonized models. Because ObjectNet contains images images with different contexts (backgrounds), viewpoints and rotations, and the models evaluated are only trained on

ImageNet (which has object facing forward and centered), we would expect models, including harmonized models, to fail in classifying images in which the viewpoint and rotation differ from those traditionally seen in ImageNet images. This is because due to their training with ImageNet images, models have not been exposed to other angles and viewpoints of objects other than those presented in ImageNet. However, because harmonization aids in shifting model "attention" to the objects rather than the backgrounds, we hypothesize that it may help in correctly classifying images of objects with different backgrounds. Our study provides evidence that model feature importance maps obtained from ImageNet images are transferable to other datasets. Therefore, more research analyzing the accuracy of harmonized models in the different image groups (background, viewpoint and rotation) would show if harmonization (increased human visual strategy alignment) helps models better generalize to real world situations and what image group it helps increase classification accuracy.

Likewise, the presence of many objects in the same image frame may also be the reason why the Top-5 accuracy of all models were significantly higher than their Top-5 accuracy. Models may end up classifying an image based on another object present in said image. Therefore, we see that our results show that even when models know *where* to attend to, they still show a drop in accuracy in the ObjectNet dataset unlike in the ImageNet dataset. Therefore, more research is needed to identify if the reason for poor performance on the real world images contained in ObjectNet is the inability of deep neural networks to generalize to the same degree as humans or another problem involving incorporating features in decision making.

The lack of improvement of robustness of the ObjectNet dataset may also point to more differences between the internal representations of DNNs and humans. ObjectNet images are more difficult for models than other datasets because they contain objects in more variations of

background, rotation and viewpoint than conventional image datasets like ImageNet. Ghodrati et al (2014) demonstrated that models only perform on the same level as humans on image classification tasks on low-level image variations in size, position, view, etc. They also indicated that models might also lack the figure-ground segregation due to their lack of recurrent processing. Humans are capable of invariant object recognition to a higher degree than models i.e. a more invariant representation of objects than models. Learning sparse important features might not be enough to result in the same level of invariant representation like humans have in higher visual areas. Consequently, harmonization, which aids models focusing on these features might not have a significant effect if learning the right image features does not help for the invariant representations needed to classify more complex variations.

Limitations

There are some limitations to this study. One of the major limitations is that only a small number of harmonized-baseline model pairs were included in the analysis. Therefore, the results obtained from the analysis might change if more model pairs are considered. Also, stronger relationships between variables may be established with the evaluation of more harmonized and baseline models. In the analysis of the correlations between improvement in human alignment and improvement in adversarial and real world image robustness, the Spearman correlation statistics show that there is a positive correlation but the correlations are not significant. However, the high R^2 scores indicate that with the evaluation of more harmonized-baseline model pairs and the subsequent inclusion of more data points, clearer results can be obtained on whether the positive relationship between human alignment and adversarial and ObjectNet robustness is truly significant.

Another limitation is the relationship between adversarial decisions between models and humans was not further explored when evaluating adversarial robustness. In this study, after the perturbation method was used to generate adversarial examples, there was not a test to correlate model accuracy with human accuracy using psychophysics experiments. Previous research has shown sometimes, human and machine classification of adversarial images are robustly related in that human subjects correctly anticipated the machine's preferred label over relevant foils—even for images described as "totally unrecognizable to human eyes." Consequently, there may be a relationship between the wrongly chosen categories of models in adversarial attacks and human classifications for some adversarial images in this study as well. Thus, further research is needed to explore if there is a correlation between those images that the harmonized model fails at classifying correctly and human recognition of those images. There is a chance that the images that the models fail at with high epsilon values are also difficult for humans to correctly classify to the correct test labels which then does show that humans and machines are becoming more aligned in their functions with the integration of a neural harmonizer.

Conclusion and Future Directions

Overall, our results show that harmonized models, which have better human aligned feature importance maps show increased adversarial robustness compared to the models that do not. We see with the saliency maps of both models that harmonized models are better able to know where to "attend" during image classification. This research provides evidence that aligning model architecture and processing more with humans could potentially improve model robustness on a variety of benchmarks of vision.

However, we see that while the saliency maps obtained through training with ImageNet are transferable to images in other datasets, harmonization did not result in increased real world image robustness, i.e. an increased performance on the ObjectNet dataset. This reveals that there is still more work to be done integrating attention modules within computer vision models. For example, more efforts need to be done on collecting feature maps on images outside of the usual ImageNet configuration.

Increased adversarial robustness of harmonized models in our results indicates that aligning feature maps has potential in closing the gap in the object recognition processes of humans and models. One way to further study this would be to compare the performance of harmonized models, baseline models and humans on MIRCs.In addition, in order to confirm that the lack of increased performance of harmonized models is due to the structure of the objects in the dataset and not due to a difference in features used in classification, more studies could be done on analyzing the performance of model pairs on MIRCs obtained from ObjectNet images. This would remove the effect of detecting features of other objects present in image. A higher accuracy of harmonized models compared to baseline models on the minimal patches that humans are able to accurately classify would demonstrate that harmonization truly results in a closer likeness between the features and processes that the human visual system and models use in object recognition.

REFERENCES:

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., & Katz, B. (2019). ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf

Brendel, W., Rauber, J., & Bethge, M. (2017a). Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. *ArXiv E-Prints*, arXiv:1712.04248. https://doi.org/10.48550/arXiv.1712.04248

Ciolino, M., Kalin, J., & Noever, D. (2021). *Fortify Machine Learning Production Systems: Detect and Classify Adversarial Attacks* (arXiv:2102.09695). arXiv. https://doi.org/10.48550/arXiv.2102.09695

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Dodge, S., & Karam, L. (2017). *A Study and Comparison of Human and Deep Learning Recognition Performance Under Visual Distortions* (arXiv:1705.02498). arXiv. http://arxiv.org/abs/1705.02498

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). *Adversarial Examples that Fool both Computer Vision and Time-Limited Humans* (arXiv:1802.08195). arXiv. http://arxiv.org/abs/1802.08195

Feinman, R., Curtin, R. R., Shintre, S., & Gardner, A. B. (2017). *Detecting Adversarial Samples from Artifacts* (arXiv:1703.00410). arXiv. http://arxiv.org/abs/1703.00410

Fel, T., Felipe, I., Linsley, D., & Serre, T. (2022). *Harmonizing the object recognition strategies of deep neural networks with humans* (arXiv:2211.04533). arXiv. https://doi.org/10.48550/arXiv.2211.04533

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, *31*. https://papers.nips.cc/paper/2018/hash/0937fb5864ed06ffb59ae5f9b5ed67a9-Abstract.html

Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2014a). Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in Computational Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fncom.2014.00074

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *ArXiv E-Prints*, arXiv:1412.6572. https://doi.org/10.48550/arXiv.1412.6572

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., & Song, D. (2021). *Natural Adversarial Examples* (arXiv:1907.07174). arXiv. http://arxiv.org/abs/1907.07174

Kurakin, A., Goodfellow, I., & Bengio, S. (2017). *Adversarial Machine Learning at Scale* (arXiv:1611.01236). arXiv. http://arxiv.org/abs/1611.01236

Lindsay, G. W. (2020). Attention in Psychology, Neuroscience, and Machine Learning. *Frontiers in Computational Neuroscience*, *14*. https://www.frontiersin.org/articles/10.3389/fncom.2020.00029

Linsley, D., Shiebler, D., Eberhardt, S., & Serre, T. (2022, February 10). *Learning what and where to attend*. International Conference on Learning Representations. https://openreview.net/forum?id=BJgLg3R9KQ

Majaj, N. J., Hong, H., Solomon, E., & DiCarlo, J. J. (2012, February). A unified neuronal population code fully explains human object recognition. *Computation and Systems Neuroscience (COSYNE)*. http://www.cosyne.org/c/index.php?title=Cosyne_12

Narodytska, N., & Kasiviswanathan, S. P. (2016). *Simple Black-Box Adversarial Perturbations for Deep Networks* (arXiv:1612.06299). arXiv. https://doi.org/10.48550/arXiv.1612.06299

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., Zantedeschi, V., Baracaldo, N., Chen, B., Ludwig, H., Molloy, I. M., & Edwards, B. (2019). *Adversarial Robustness Toolbox v1.0.0* (arXiv:1807.01069). arXiv. https://doi.org/10.48550/arXiv.1807.01069

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 427–436. https://doi.org/10.1109/CVPR.2015.7298640

Rauber, J., Brendel, W., & Bethge, M. (2018). *Foolbox: A Python toolbox to benchmark the robustness of machine learning models* (arXiv:1707.04131). arXiv. https://doi.org/10.48550/arXiv.1707.04131

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). *ImageNet Large Scale*

Visual Recognition Challenge (arXiv:1409.0575). arXiv. https://doi.org/10.48550/arXiv.1409.0575

Serre, T. (2019). Deep Learning: The Good, the Bad, and the Ugly. *Annual Review of Vision Science*, *5*(1), 399–426. https://doi.org/10.1146/annurev-vision-091718-014951

Tatler, B. W., Wade, N. J., Kwan, H., Findlay, J. M., & Velichkovsky, B. M. (2010). Yarbus, eye movements, and vision. *I-Perception*, *1*(1), 7–27. https://doi.org/10.1068/i0382

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*, 766–786. https://doi.org/10.1037/0033-295X.113.4.766

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016a). Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, *113*(10), 2744–2749. https://doi.org/10.1073/pnas.1513198113

Wang, L., Zhang, C., Luo, Z., Liu, C., Liu, J., Zheng, X., & Vasilakos, A. (2020). *Progressive Defense Against Adversarial Attacks for Deep Learning as a Service in Internet of Things* (arXiv:2010.11143). arXiv. http://arxiv.org/abs/2010.11143

Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. *Advances in Neural Information Processing Systems*, *26*. https://proceedings.neurips.cc/paper/2013/hash/9a1756fd0c741126d7bbd4b692ccbd91-Abstract.html

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*(1), 1334. https://doi.org/10.1038/s41467-019-08931-6