CENTER FOR
Brains
Minds+
Machines

CBMM Memo No. 116 **February 18, 2021**

The Effects of Image Distribution and Task on Adversarial Robustness

Owen Kunhardt, Arturo Deza, Tomaso Poggio

Abstract

In this paper, we propose an adaptation to the area under the curve (AUC) metric to measure the adversarial robustness of a model over a particular ϵ-interval $[\epsilon_0, \epsilon_1]$ (interval of adversarial perturbation strengths) that facilitates unbiased comparisons across models when they have different initial ϵ_0 performance. This can be used to determine how adversarially robust a model is to different image distributions or task (or some other variable); and/or to measure how robust a model is comparatively to other models. We used this adversarial robustness metric on models of an MNIST, CIFAR-10, and a Fusion dataset (CIFAR-10 + MNIST) where trained models performed either a digit or object recognition task using a LeNet, ResNet50, or a fully connected network (FullyConnectedNet) architecture and found the following: 1) CIFAR-10 models are inherently less adversarially robust than MNIST models; 2) Both the image distribution and task that a model is trained on can affect the adversarial robustness of the resultant model. 3) Pretraining with a different image distribution and task sometimes carries over the adversarial robustness induced by that image distribution and task in the resultant model; Collectively, our results imply non-trivial differences of the learned representation space of one perceptual system over another given its exposure to different image statistics or tasks (mainly objects vs digits). Moreover, these results hold even when model systems are equalized to have the same level of performance, or when exposed to approximately matched image statistics of fusion images but with different tasks.

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

The Effects of Image Distribution and Task on Adversarial Robustness

Owen Kunhardt [1] **Arturo Deza** [*1] **Tomaso Poggio** [*1]

Abstract

In this paper, we propose an adaptation to the area under the curve (AUC) metric to measure the adversarial robustness of a model over a particular ϵ-interval $[\epsilon_0, \epsilon_1]$ (interval of adversarial perturbation strengths) that facilitates unbiased comparisons across models when they have different initial ϵ_0 performance. This can be used to determine how adversarially robust a model is to different image distributions or task (or some other variable); and/or to measure how robust a model is comparatively to other models. We used this adversarial robustness metric on models of an MNIST, CIFAR-10, and a Fusion dataset (CIFAR-10 + MNIST) where trained models performed either a digit or object recognition task using a LeNet, ResNet50, or a fully connected network (FullyConnectedNet) architecture and found the following: 1) CIFAR-10 models are inherently less adversarially robust than MNIST models; 2) Both the image distribution and task that a model is trained on can affect the adversarial robustness of the resultant model. 3) Pretraining with a different image distribution and task sometimes carries over the adversarial robustness induced by that image distribution and task in the resultant model; Collectively, our results imply non-trivial differences of the learned representation space of one perceptual system over another given its exposure to different image statistics or tasks (mainly objects vs digits). Moreover, these results hold even when model systems are equalized to have the same level of performance, or when exposed to approximately matched image statistics of fusion images but with different tasks.

[*] Denotes joint senior authorship [1]Center for Brains, Minds and Machines (CBMM), Massachusetts Institute of Technology, Cambridge, MA, USA.. Correspondence to: Owen Kunhardt <okunhardt@owenkunhardt.com>.

Under Review at the 38^{th} International Conference on Machine Learning, 2021.

1. Introduction

Adversarial images are perturbed visual stimuli that can fool a high performing image classifier with carefully chosen noise that is often imperceptible to humans (Szegedy et al., 2013; Goodfellow et al., 2014). These images are synthesized using an optimization procedure that maximizes the wrong output class of a model observer, while minimizing any noticeable differences in the image for a reference observer. Understanding why adversarial images exist has been studied extensively in machine learning as a way to explore gaps in generalization (Gilmer et al., 2018; Yuan et al., 2019; Ilyas et al., 2019), computer vision with applications to real-world robustness (Dubey et al., 2019; Yin et al., 2019; Richardson & Weiss, 2020), and recently in vision science to understand similar and divergent visual representations with humans (Zhou & Firestone, 2019; Feather et al., 2019; Golan et al., 2019; Reddy et al., 2020; Dapello et al., 2020). Thus far there have been gaps in the literature on how natural image distributions and classification task impact the robustness of a model to adversarial images.

This paper addresses whether training on a specific natural image distribution or task plays a role in the adversarial robustness of a model. *Natural images* are images that are representative of the real world. MNIST, CIFAR-10, and ImageNet are examples of natural image datasets. The goal is to understand what it means for a model to inherently be more adversarially robust to objects vs scenes or objects vs digits, where the latter is addressed in this paper. The thesis of this paper is that both the natural image distribution and task (independently and jointly) play a role in the adversarial robustness of a model trained on them.

Answering questions about the role the image distribution and task in adversarial robustness could be critical for applications where an adversarial image can be detrimental (e.g. self-driving cars (Lu et al., 2017), radiology (Hirano et al., 2020) and military surveillance (Ortiz et al., 2018; Deza et al., 2019)). These applications are often models of natural image distributions. Understanding if and how the dataset and task play a role in the adversarial robustness of a model could lead to better adversarial defenses for the aforementioned applications and a better understanding of the existence of adversarial images.

The works most closely related to the thesis of this paper

are the following: Ilyas et al. (2019) found that adversarial vulnerability is not necessarily tied to the training scheme, but rather is a property of the dataset. Similarly, Ding et al. (2019) finds that semantic-preserving shifts on the image distribution could result in drastically different adversarial robustness even for adversarially trained models.

All work in this paper studying the role of natural image distribution and classification task in the adversarial robustness of a model is empirical. The experiments presented require important performance comparisons. Therefore, we propose an unbiased metric to measure the adversarial robustness of a model for a particular set of images over an interval of perturbation strengths. Using this metric, we compare MNIST and CIFAR-10 models and find that MNIST models are inherently more adversarially robust than CIFAR-10 models. We then create a Fusion dataset (α-blend of MNIST and CIFAR-10 images) to determine whether the image distribution or task is causing this difference in adversarial robustness and discover that both play a role. Finally, we examine whether pretraining on one dataset (CIFAR-10 or MNIST), then training on the other results in a more adversarially robust learned representation of the dataset the model is trained on and find that this impacts robustness in unexpected ways.

2. Proposed Adversarial Robustness Metric

In order to make comparisons of how robust a model is to adversarial perturbations, a proper metric for adversarial robustness must be defined. We define the adversarial robustness R as a measure of the rate at which accuracy of a model changes as ϵ (adversarial perturbation strength) increases over a particular ϵ-interval of interest. The faster the accuracy of a model decreases as ϵ increases, the lower the adversarial robustness is for that model. We propose an adaptation of area under the curve (AUC) to measure adversarial robustness. A good measure of how much change of accuracy is occurring in an ϵ-interval for a model is the AUC of a function that outputs the accuracy for an adversarial attack given an ϵ for a model. This AUC provides a total measure of model performance for an ϵ-interval. If the accuracy decreases quickly as ϵ increases, then the AUC will be smaller.

Despite how intuitive the previous notion may sound, we immediately run into a problem: Some datasets are more discriminable than others independent of model observers, as shown in Figure 1(A). This must be taken into account when computing the area under the curve. It could be possible that under unequal initial performances, one model seems more 'adversarially robust' over the other by virtue purely of the initial offset in the better performance.

Figure 1(B) shows that one simple solution to solve the differences in accuracy between two model systems is by normalizing them with respect to their accuracy under non-adversarial ($\epsilon_0 = 0$) inputs. This yields the following expression:

$$R = \frac{1}{f(\epsilon_0)(\epsilon_1 - \epsilon_0)} \int_{\epsilon_0}^{\epsilon_1} f(\epsilon) d\epsilon \tag{1}$$

which can be interpreted as the normalized area under the curve of a function $f(\epsilon)$ that outputs the accuracy of a model for an adversarial attack of strength ϵ over an ϵ-interval (i.e. $[\epsilon_0, \epsilon_1]$). Note that $f(\epsilon_0) > 0$ and $\epsilon_1 > \epsilon_0$. Computing R is the same as integrating relative change (shown in supplementary material). Therefore, R is an aggregate measure of relative change in accuracy over an ϵ-interval. The division by $f(\epsilon_0)$ normalizes the function because the function now represents the change in accuracy with respect to no adversarial perturbations (i.e. it is now a relative change). Further, the accuracy at $f(\epsilon_0)$ can be considered an *'oracle'* for the adversarial attacks of the model (i.e. the likely optimal or best performance for that ϵ-interval). The term $\frac{1}{\epsilon_1 - \epsilon_0}$ of Eq. 1 puts the area under the curve of the normalized accuracy between $(0, 1]$. This is so that it is easier to interpret and so that the metric is normalized for different ϵ-intervals (i.e. the maximum value is not $\epsilon_1 - \epsilon_0$, but instead is 1). Note that the metric is valid independent of the adversarial attack method.

If for a particular model, $R = 1$, this implies that $f(\epsilon)$ is constant over $[\epsilon_0, \epsilon_1]$. If for a model, $R \approx 0$, that means that for all ϵ in the interval, the model classifies nearly all the perturbed images of a given set incorrectly. R can be arbitrarily close to 0.

To guarantee that $R \leq 1$, the following constraint must be satisfied:

$$\int_{\epsilon_0}^{\epsilon_1} f(\epsilon) d\epsilon \leq f(\epsilon_0)(\epsilon_1 - \epsilon_0) \tag{2}$$

This is a reasonable constraint to make. An interpretation of $f(\epsilon_0)(\epsilon_1 - \epsilon_0)$ is a possible AUC for $f(\epsilon)$. This AUC occurs when $f(\epsilon) = f(\epsilon_0)$ for all $\epsilon \in [\epsilon_0, \epsilon_1]$. In other words, as ϵ increases, the classification performance of the adversarial images does not change. An AUC greater than $f(\epsilon_0)(\epsilon_1 - \epsilon_0)$ would imply that the accuracy increases above the starting accuracy (i.e. $f(\epsilon_0)$). This behavior would contradict what it means to perform an adversarial attack.

To measure the impact C, that adversarial attacks have on a model between two specific ϵ points instead of an interval, the following can be used:

$$C = \frac{f(\epsilon) - f(\epsilon_0)}{f(\epsilon_0)} \tag{3}$$

where C is the relative change between the performance of a model for two different ϵ's of adversarial attacks. Normalizing to compute R by taking the relative change in error with

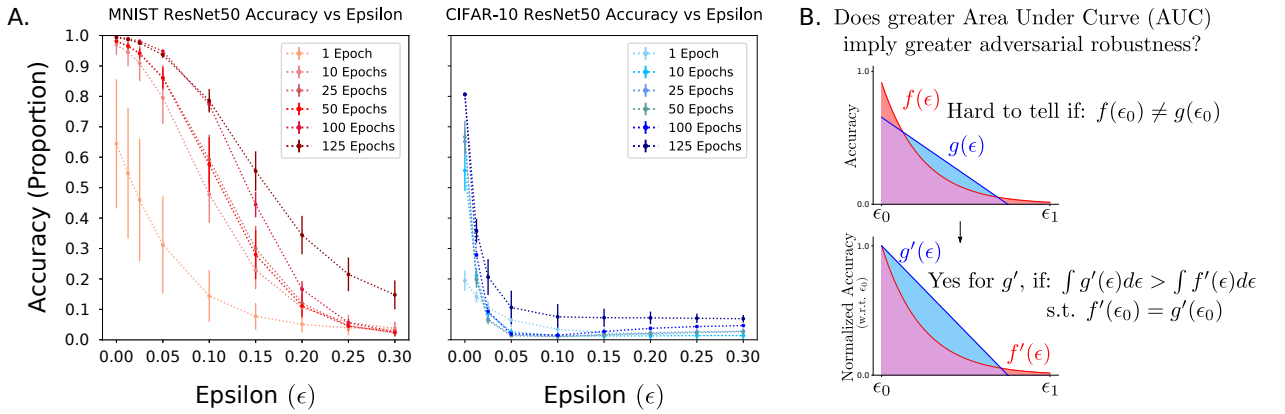

Figure 1. (A) After using the same hyperparameters and training scheme (SGD) for both models, MNIST achieves around 99% accuracy, while CIFAR-10 peaks around 80% with ResNet50 (both without data-augmentation). In cases like these it may be obvious to say that better performing models will be more adversarially robust – but this is not always the case, in some cases it is the opposite when fixing the image distribution (Zhang et al., 2019); (B) One solution: example graphs showing the area under the curve of $f(\epsilon)$ and $g(\epsilon)$, functions outputting the accuracy of an adversarial attack for a given ϵ of two models before (top) and after (bottom) accuracy normalization. This shows how at ϵ_0, models go from an unmatched accuracy to a matched upper-bounded score of 1, allowing an unbiased computation of area under the curve.

respect to a reference or optimal value $f(\epsilon_0)$ (i.e. Eq. 3) results in a less biased measure for adversarial robustness than other normalization schemes, such as taking the difference (i.e. $f(\epsilon) - f(\epsilon_0)$). This is because the other schemes are unable to properly account for differences in performance of models on a particular dataset or task. Broadly, we are not interested in how much the performance differs overall, but how much it differs relative from where it started.

There are two methods to find $f(\epsilon)$: 1) to empirically compute multiple values of ϵ and estimate the area under the curve using integral approximations, such as the trapezoid method; 2) to find the closed form expression of $f(\circ)$ as one would do for psychometric functions (Wichmann & Hill, 2001) and integrate. In this paper, we do the former (compute multiple values of ϵ and estimate the integral using the trapezoid method), although this method is extendable to the latter.

Picking ϵ_0 and ϵ_1 is an experimental choice. Choosing $\epsilon_0 = 0$ allows measures the adversarial robustness starting from no perturbations, yet $\epsilon_0 > 0$ can also be used. For too high a choice of ϵ_1, the image can saturate and the performance will likely approach chance. This rebounding effect can be seen in some of the CIFAR-10 curves in our experiments.

There are certain assumptions for this normalization scheme to hold. For example, in both of our experiments MNIST and CIFAR-10 are equalized to have 10 classes and we assume an independent and identically distributed testing distribution such that chance performance for any model observers is the same at 10%. One could see how the normalization scheme would give a misleading result if one dataset has 2 i.i.d classes that yield 50% chance and another dataset yields 10 i.i.d classes that yield 10% chance. In this case, proportions correct are not comparable and a more principled way of equalizing performance – likely using d' (a generalized form of Proportion Correct used in Signal Detection Theory) would be required (Green et al., 1966).

Overall, this robustness metric can be used to get a sense of whether a model is adversarially robust over a particular ϵ-interval or to measure how adversarially robust a model is comparatively to other models over that interval for a particular set of inputs. Note that this metric is not intended to be used to certify the adversarial robustness of an artificial neural network since it is an approximation of the change of accuracy of a model over an ϵ-interval for, in this paper, a specific set of images.

3. Experimental Design

Figure 2 visualizes the general experimental design, where models are trained on either MNIST or CIFAR-10 images, and later Fusion images. The architecture, optimization and learning scheme, and initial random weights between each MNIST and CIFAR-10 model is the same, allowing us to draw comparisons between the adversarial robustness of the models after attacking the trained models.

3.1. Architectures

All experiments used 3 networks: LeNet (LeCun et al., 1989), ResNet50 (He et al., 2015), and a fully connected network (FullyConnectedNet) where we explored adversarial robustness over 20 paired network runs and their learning

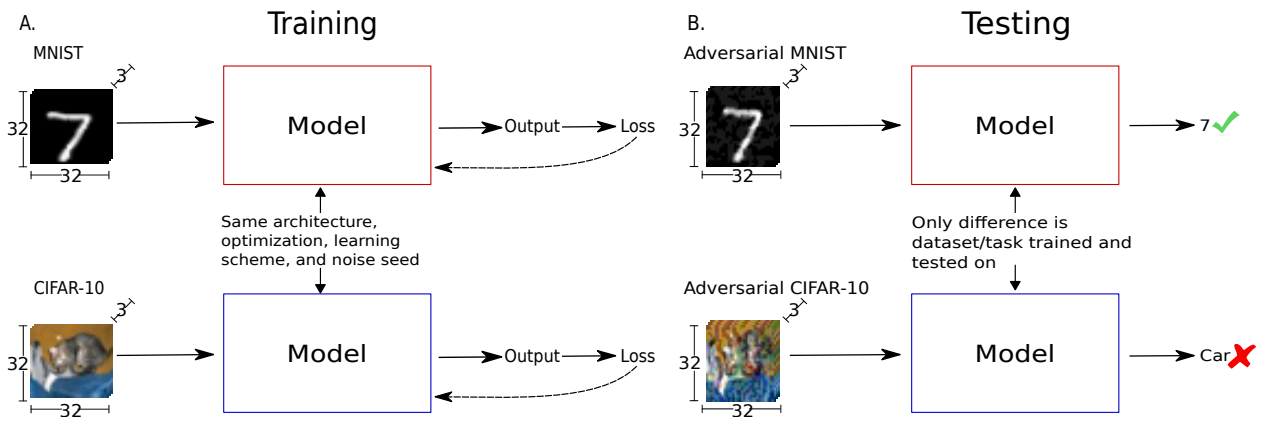

Figure 2. (A) 20 models are trained for each dataset/task (MNIST, CIFAR-10, and later a Fusion Dataset) and network architecture (LeNet, ResNet50, FullyConnectedNet), using a different set of randomly initialized weights (i.e. 60 models per dataset); (B) The models are then tested on adversarial images generated using FGSM (Goodfellow et al., 2014) of various perturbation strengths. The results from testing result in graphs similar to 1(A). Using these results, the adversarial robustness is computed using Eq. 1. The average adversarial robustness across the set of two models is compared to determine which model is more adversarially robust and analyze these results.

dynamics. FullyConnectedNet has 1 hidden layer with 7500 hidden units. This number of hidden units was chosen so the number of parameters for the FullyConnectedNet has the same order of magnitude as the number of parameters for ResNet50. FullyConnectedNet only have 1 hidden layer so that the network is not biased to approximate a hierarchical function as a convolutional neural network (See (Mhaskar & Poggio, 2016; Poggio et al., 2017) and recently (Neyshabur, 2020; Deza et al., 2020)).

3.2. Datasets

The datasets used were MNIST, CIFAR-10, and a Fusion Dataset. To use the exact same architectures with the datasets, MNIST was upscaled to 32×32 and converted to 3 channels to match the dimensions of CIFAR-10 (i.e. $32 \times 32 \times 3$). MNIST was changed instead of CIFAR-10 because given the low image complexity of MNIST images – mainly their low spatial frequency structure, that lends itself to upscaling – and thus it would be less likely to change the accuracy of models trained on that dataset. Preliminary results showed that there is a difference (insignificant in comparison to the other differences in results in this paper) in the adversarial robustness of models trained on the scaled and 3 color channel version and the regular version, with the scaled and 3 color channel version being less robust. Whether the changes to MNIST entirely caused the difference was not determined due to the differences between architectures that were used for each version. No other changes to the datasets were made (such as color normalization, which is typically used for CIFAR-10) in order to preserve the natural image distribution.

The Fusion dataset that is used in the experiments is not a natural image distribution. It was created with the purpose of better understanding the inherit adversarial robustness proprieties of natural image distribution models. Each fusion image in the dataset is generated with the following α-blending procedure:

$$F = 0.5M + 0.5C, \tag{4}$$

where F is a new fusion image, M is an MNIST image modified to 32x32x3 (by upscaling and increasing number of color channels), and C is a CIFAR-10 image. Example fusion images can be found in Figure 3. This dataset is similar to *Texture shiftMNIST* from Jacobsen et al. (2018).

The Fusion dataset was created online during training or testing during each mini-batch by formula 4. The fusion image training set was constructed using the MNIST and CIFAR-10 training set and the fusion image test set was constructed using the MNIST and CIFAR-10 test set. During training, the MNIST and CIFAR-10 datasets are shuffled at the start of every epoch. Therefore, it is likely that no fusion images are shown to the model more than once. This was done to ensure that the model cannot learn any correlation between any CIFAR-10 object and any MNIST digit, as well as, improve generalization of the model. Additionally, it is important to note that no two models were trained on the exact same set of fusion images, but were evaluated on the same test images. Since we train 20 random models, it should average out any possible noise to a certain degree, but strictly speaking the images were different but the statistics were approximately matched.

3.3. Hyperparameters, Optimization Scheme, and Initialization

It is important to note that all hyperparameters are held constant, including the ϵ-interval. The only difference between

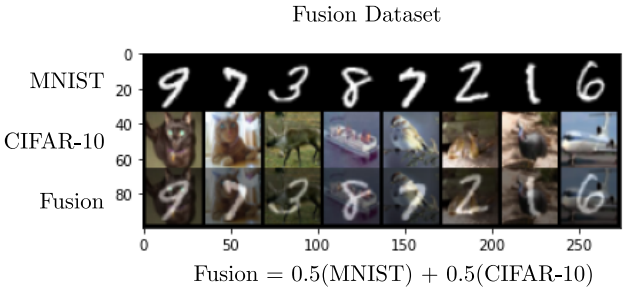

Figure 3. The Fusion dataset was created to tease apart the cause and effects of the inherit adversarial robustness of models trained on natural image distributions. Here, we show sample images from the Fusion dataset consisting of alpha-blended MNIST + CIFAR-10 stimuli.

the models using a certain architecture is the dataset/task they are trained and tested on (just the task in the case of the Fusion dataset). In the experiments presented, the independent variables are the dataset and task, while the dependent variable being measured is the adversarial robustness of the model. Since all other variables are held fixed, if the adversarial robustness of the models trained on the different datasets/tasks are different, then this change is due to the dataset/task itself (i.e. the image distribution and classification task). If the ϵ-interval used to attack the two models is different we could not directly conclude that any differences are due to the image distribution and task because the difference could also be due to the differences in the strengths of the adversarial attacks on each model. Experiments using the Fusion dataset are presented in this paper to investigate which of the independent variables (i.e. whether image distribution or task) is playing a role in the differences in adversarial robustness.

The loss function used for all models was cross-entropy loss and the optimizer used was stochastic gradient descent (SGD) with weight decay 5×10^{-4}, momentum 0.9, and with an initial learning rate 0.01 for the FullyConnectedNet and LeNet models and an initial learning rate 0.1 for the ResNet50 models. The learning rate was divided by 10 at 50% of the training. The FullyConnectedNet and LeNet models were trained to 300 epochs and the ResNet50 models were trained to 125 epochs. ResNet50 models required less epochs during training because those models reached high levels of performance sooner than the other architectures. A batch size of 125 was used. The batch size was 125 since this is the closest number to a more typical batch size of 128 that divides both the number of CIFAR-10 images and MNIST images. This was needed to ensure that the batches align properly when creating the fusion images. These hyperparameters and optimization scheme were chosen since they resulted in the best performance of those tested in preliminary experiments.

For all experiments, each model was trained 20 times with matched initial random weights across different datasets. For example in the case of LeNet, 20 different LeNet models all with different initial random weights: $\{w_1, w_2, ..., w_{20}\}$ were used to train for CIFAR-10 in our first experiment, and these same initial random weights were used to train for MNIST. This removed the variance induced by a particular initialization (*e.g.* a lucky/unlucky noise seed) that could bias the comparisons by arriving to a better solution via SGD. This procedure was possible because our MNIST dataset was resized to a 3-channeled version with a new size of $32 \times 32 \times 3$ instead of $28 \times 28 \times 1$ (original MNIST).

3.4. Adversarial Attacks

The method used for generating adversarial images in the experiments presented in this paper is the Fast Gradient Sign Method (FGSM) presented in Goodfellow et al. (2014). The focus of the attacks was to create images that cause the model to misclassify in general, rather than misclassifying an image to a particular class. FGSM was chosen over other optimization based attacks such as Projected Gradient Descent (PGD) (Madry et al., 2019) based on preliminary results as FGSM was sufficient to successfully adversarially attack the model. FGSM also has a lower computational cost than PGD allowing us to run more experiments and train more models. Adversarial training or other data-augmentations schemes that may bias the outcome were not performed. Importantly, given that an adversarial defense mechanism is not being proposed or used, strong adversarial attack methods, such as PGD, are not necessary in this first work – contrary, but justified to the advice from Carlini et al. (2019).

The ϵ-interval used in the experiments is $[0, 0.3]$ (i.e. $\epsilon_0 = 0, \epsilon_1 = 0.3$). The upper bound of 0.3 was chosen because adversarial images at that magnitude are difficult for many undefended classifiers to classify correctly. The trained models were adversarially attacked with $\epsilon \in \{0, 0.0125, 0.025, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ to approximate $f(\epsilon)$. For the models using LeNet and FullyConnectedNet architectures, they were adversarially attacked at 1, 10, 25, 50, 150, and 300 epochs. Models using the ResNet50 architecture were adversarially attacked at 1, 10, 25, 50, 100, and 125 epochs. Different epochs were adverarially attacked to determine whether the results differed at different stages of learning.

4. Experimental Results

The following experiments provide a glimpse into the role of classification task and image distribution in the adversarial robustness of models.

All differences in robustness that are mentioned are statis-

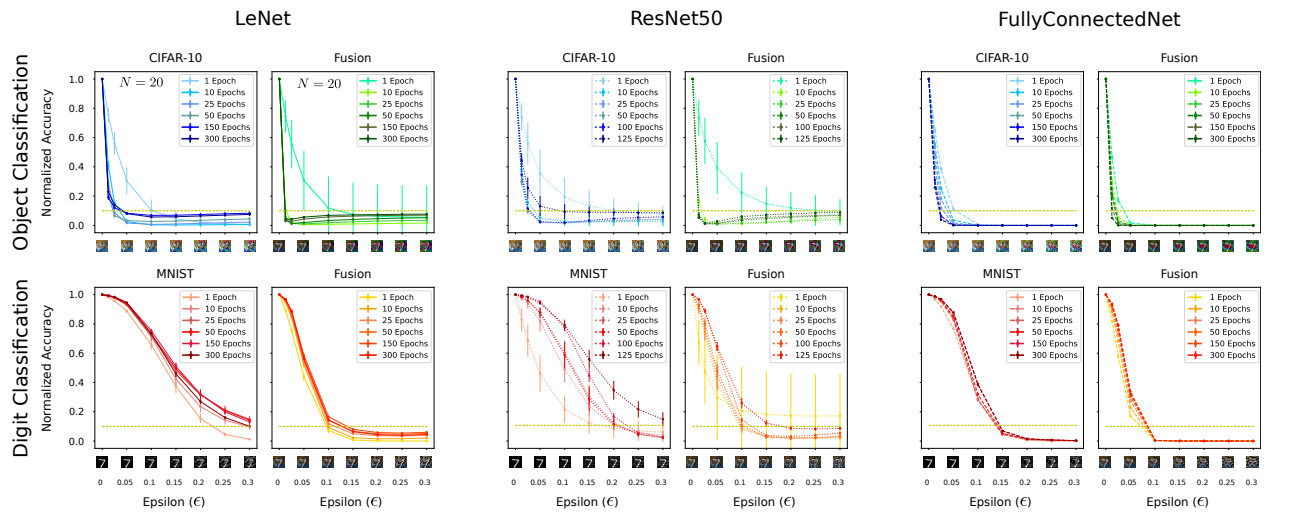

Figure 4. MNIST-trained networks (bottom left) across all architectures show greater adversarial robustness after accuracy normalization than CIFAR-10 trained networks (top left for each architecture). Notice too that ResNet50 appears to be the more adversarially robust network across network architectures (LeNet and FullyConnectedNet) independent of learning dynamics. Graphs of the normalized accuracy of the Fusion dataset on the object recognition task (top right) and digit recognition task (bottom right) for LeNet, ResNet50, and FullyConnectedNet. Generally, models trained on the digit task were more adversarially robust than those trained on the object task, showing the role that task plays in the adversarial robustness of a model. Additionally, these models were generally less adversarially robust than their MNIST and CIFAR-10 model counterparts. In combination, these results imply that both task and image distribution play distinct roles in the adversarial robustness of a model. The gold lines represent chance performance in the graphs.

tically significant using a Welch's t-test with significance level $\alpha = 0.05$. This test was used because the models are unpaired and do not have equal variance since the models are trained on different datasets.

4.1. Comparing MNIST vs CIFAR-10 Adversarial Robustness

This experiment investigates whether MNIST models are inherently more adversarially robust than CIFAR-10 models. This was investigated by comparing the adversarial robustness of CIFAR-10 models and the MNIST models for the three architectures. Figure 4 (top left for each architecture) shows normalized accuracy graphs for the CIFAR-10 trained models and Figure 4 (bottom left) shows graphs of normalized accuracy for MNIST trained models. Both LeNet and FullyConnectedNet, the MNIST models were more adversarially robust than CIFAR-10 models, for each epoch we examined. The same pattern of results held for ResNet50 models except for the first epoch where there was no difference between the MNIST and CIFAR-10 models.

Result 1: For the three network architectures tested (that all vary in approximation power and architectural constraints), MNIST trained models are inherently more adversarially robust than CIFAR-10 models. This implies that the task and/or image distribution play a role in the adversarial robustness of the model.

4.2. Comparing Object vs Digit Classification in the Fusion (MNIST + CIFAR-10) dataset

The previous results suggested that after taking into account different measures of accuracy normalization, MNIST (both dataset and digit recognition task) models are intrinsically more adversarially robust than CIFAR-10 models. This implies that it is harder to fool an MNIST model, than a CIFAR-10 model, likely, in part, due to the fact that number digits are highly selective to shape, and show less perceptual variance than objects.

Naturally, the next question that arises is if the task itself is somehow making each perceptual system less adversarially robust. To test this hypothesis the Fusion dataset was used. Models were trained to perform either digit recognition or object recognition on these fusion images – thus we have approximately fixed the image distribution but varied the approximation task (Deza et al., 2020). They are approximately matched because no model is trained on the exact same images, the image distribution is approximately the same on average given the random sampling procedure. The goal with this new hybrid dataset is to re-run the same set of previous experiments and test adversarial robustness for both the digit recognition task and the object recognition task and probe the role of the type of classification task when fixing the dataset to test how adversarial robustness varies when all other variables remain constant.

Observation: When examining the first epoch for the fusion

trained models, the standard deviation of the curves in 3(B) are generally high. This is likely due to design choice of avoiding to show the same fusion image twice. This does not occur in later stages of training.

Result 2a: Task plays a critical role in the adversarial robustness of a model. Figure 4 contains the normalized curves of the results for the digit and object recognition tasks on the fusion dataset for each of the architectures. The models were evaluated on fusion images constructed from the MNIST and CIFAR-10 test sets. The FullyConnectedNet (all epochs), ResNet50 and LeNet fusion image models were more adversarially robust on the digit recognition task than the object recognition task for all epochs examined excluding the first epoch. This suggests that even if the image distribution is approximately equalized at training, the representation learned varies given the task, and impacts adversarial robustness differently.

Result 2b: Image distribution also plays a role in the adversarial robustness of a model. Comparing the three architectures trained on the Fusion Dataset vs their regular image-distribution trained models show that increasing the image complexity (by adding a conflicting image with the hope of increasing invariance) in fact decreases adversarial robustness when compared to regularly trained networks. Comparing fusion image models trained on the digit task and MNIST models: for the FullyConnectedNet and LeNet architecture, the MNIST models were more robust. The same holds for the ResNet50 MNIST models except at the first epoch, where there was no difference. CIFAR-10 models using the FullyConnectedNet architecture were more adversarially robust than the fusion image models trained on the object recognition task for all epochs tested. The same was true for the LeNet and ResNet50 architectures except there were no differences between CIFAR-10 models and fusion images with object task in adversarial robustness for 1 and 50 epochs.

4.3. Impact of Pretraining on Out-Of-Distribution (o.o.d) image datasets

This experiment investigates whether pretraining on one dataset (CIFAR-10 or MNIST), then training on the other results in a more adversarially robust learned representation of the dataset the model is trained on.

The pretraining procedure was done by using the existing fully trained CIFAR-10 or MNIST FullyConnectedNet, LeNet, and ResNet50 models as bases and then training/fine-tuning them using the same training scheme but with MNIST or CIFAR-10 respectively. These models were then tested using the test sets of the datasets the models were fine-tuned on.

For the FullyConnectedNet, the MNIST models were more adversarially robust than the MNIST pretrained on CIFAR-10 model during early stages of learning, but the pretrained models were more robust when examined at 150 and 300 epochs of fine-tuning. The MNIST LeNet models were more adversarially robust for all stages of learning than the pretrained model. The pretrained ResNet50 models had no differences in robustness compared to the MNIST ResNet50 models, except for the first epoch where the pretrained models were more robust. This result is unexpected as this does not occur for the other architectures. These results would seem to suggest that architecture plays a role in the adversarial robustness of the learned representation contingent on the given dataset/task and potentially compositional nature.

Pretraining on CIFAR-10 and then training on MNIST generally does not lead to more adversarially robust models. Next we investigate whether pretraining on MNIST and then training on CIFAR-10 has this same effect. We find that this is not always the case. Pretraining on MNIST then training on CIFAR-10 led to marginal improvements in adversarial robustness for LeNet, except for the first epoch (Figure 5). For ResNet50, pretraining resulted in less adversarially robust models at the start and end of training (1 and 125 epochs), otherwise there was no difference compared to not pretraining. The FullyConnectedNet pretrained models were more adversarially robust in earlier stages of learning, but were less robust in later stages. Tables of the robustness metrics for the CIFAR-10 models pretrained on MNIST (as well as for other experiments) can be found in the supplementary material. These findings requires further investigation.

For the ResNet50, LeNet, and FullyConnectedNet architectures, the models pretrained on CIFAR-10 then trained on MNIST were statistically significantly more adversarially robust than models pretrained on MNIST then trained on CIFAR-10 for all epochs examined.

Result 3: Pretraining on CIFAR-10 followed by training on MNIST does not generally produce a more adversarially robust model than training on MNIST alone, with any of the tested architectures. This is counter intuitive given that humans typically base their learned representations on objects rather than figures (Janini & Konkle, 2019). On the other hand, pretraining on MNIST, then training on CIFAR-10 only aided LeNet; for FullyConnectedNet it helped in earlier stages of learning, while decreased robustness later. Generally, however, ResNet50 models were not affected in terms of carried-over robustness at any intermediate stages of learning. Investigating the origins of this visual hysteresis (an asymmetry in learned representation visible through robustness given the pretraining scheme) (Sadr & Sinha, 2004) and how it may relate to shape/texture bias (Geirhos et al., 2018; Hermann & Kornblith, 2019), spatial frequency sensitivity (Dapello et al., 2020; Deza & Konkle, 2020), or

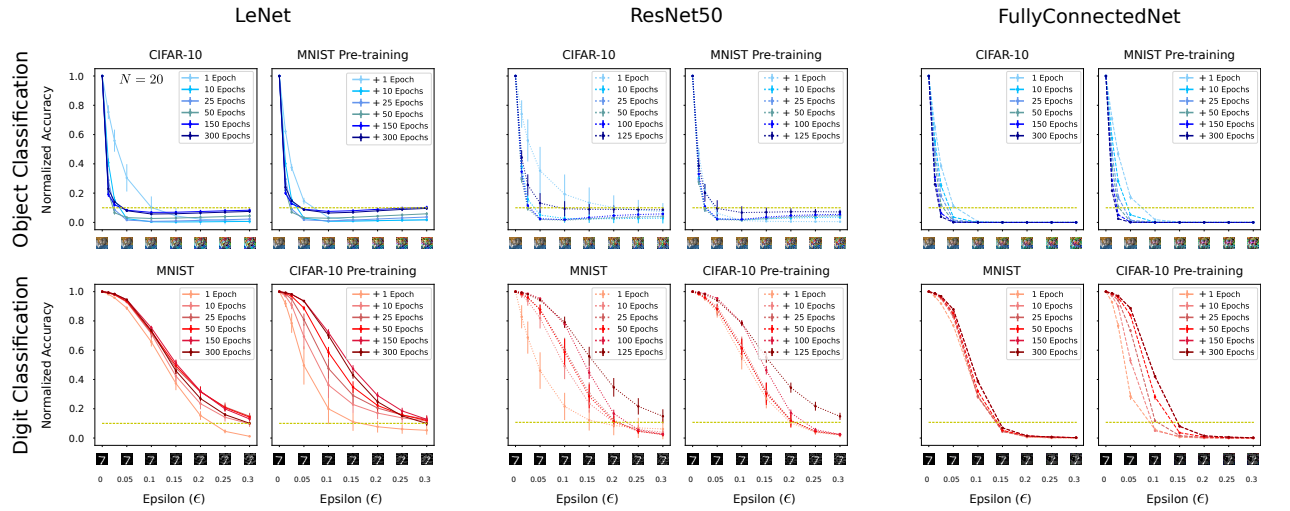

Figure 5. Visual Hysteresis: FullyConnectedNet and LeNet networks seems to carry over the learned representation and adversarial vulnerability from the pretrained system. However, only LeNet experiences a clear visual hysteresis where pretraining on CIFAR-10 for MNIST is worse (less adversarially robust) than only training on MNIST, yet pretraining on MNIST for CIFAR-10 is better (more adversarially robust) than only training on CIFAR-10 (See supplementary material). The gold lines represent chance performance in the graphs.

common perturbations (Hendrycks & Dietterich, 2019) is a subject of on-going work.

5. Discussion

This work verified that both the image distribution and task (independently or jointly) can impact the adversarial robustness of a model under FGSM. The next step is to investigate why, and what specific factors of the image statistics and task play a role. It is likely that MNIST trained networks are intrinsically more adversarially robust than CIFAR-10 trained networks in part due to the lower-dimensional subspace in which they live in given their image structure (Hénaff et al., 2014) compared to CIFAR-10 (*i.e.* MNIST has less non-zero singular values than CIFAR-10 allowing for greater compression for a fixed number of principal components). Additionally, in future work we want to know whether these observations hold with other optimization based attacks and gradient-free attacks, such as PGD (Madry et al., 2019) and NES (Ilyas et al., 2018) respectively. Given that FGSM is not considered a strong attack, would a stronger attack exacerbate these results? Based on the noticeable differences in adversarial robustness between the models testing only using FGSM, this is a promising direction.

Indeed, this paper has only scratched the surface of the role of natural image distribution and task in the adversarial robustness of a model by comparing two well known candidate datasets over their learning dynamics: MNIST and CIFAR-10. Continuing this line of work onto exploring the role of the image distribution on adversarial robustness for other natural image distributions such as textures or scenes is another promising next step. Finally, future experiments should continue to investigate the effect of the learning objective on the learned representation induced from the image distribution. We have already seen how the task affects the adversarial robustness of a model even when image statistics are approximately matched under a supervised training paradigm. With the advent of self-supervised (Konkle & Alvarez, 2020; Geirhos et al., 2020; Purushwalkam & Gupta, 2020) and unsupervised (Zhuang et al., 2020) objectives that may be predictive of human visual coding, it may be relevant to investigate the changes in adversarial robustness for the current (objects, digits) and new (texture, scenes) image distributions with the proposed adversarial robustness metric for these new learning objectives.

6. Acknowledgements

The authors thank Dr. Christian Bueno and Dr. Susan Epstein for their helpful feedback on this paper. This work was supported in part by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF – 1231216.

References

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. On evaluating adversarial robustness, 2019.

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., and DiCarlo, J. J. Simulating a primary visual

cortex at the front of cnns improves robustness to image perturbations. *BioRxiv*, 2020.

Deza, A. and Konkle, T. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020.

Deza, A., Surana, A., and Eckstein, M. P. Assessment of faster r-cnn in man-machine collaborative search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3185–3194, 2019.

Deza, A., Liao, Q., Banburski, A., and Poggio, T. Hierarchically local tasks and deep convolutional networks. *arXiv preprint arXiv:2006.13915*, 2020.

Ding, G. W., Lui, K. Y.-C., Jin, X., Wang, L., and Huang, R. On the sensitivity of adversarial robustness to input data distributions. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1xNEhR9KX.

Dubey, A., Maaten, L. v. d., Yalniz, Z., Li, Y., and Mahajan, D. Defense against adversarial images using web-scale nearest-neighbor search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8767–8776, 2019.

Feather, J., Durango, A., Gonzalez, R., and McDermott, J. Metamers of neural networks reveal divergence from human perceptual systems. In *Advances in Neural Information Processing Systems*, pp. 10078–10089, 2019.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Geirhos, R., Narayanappa, K., Mitzkus, B., Bethge, M., Wichmann, F. A., and Brendel, W. On the surprising similarities between supervised and self-supervised models. In *NeurIPS 2020 Workshop SVRHM*, 2020. URL https://openreview.net/forum?id=q2ml4CJMHAx.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

Golan, T., Raju, P. C., and Kriegeskorte, N. Controversial stimuli: pitting neural networks against each other as models of human recognition. *arXiv preprint arXiv:1911.09288*, 2019.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Green, D. M., Swets, J. A., et al. *Signal detection theory and psychophysics*, volume 1. Wiley New York, 1966.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Hénaff, O. J., Ballé, J., Rabinowitz, N. C., and Simoncelli, E. P. The local low-dimensionality of natural images. *arXiv preprint arXiv:1412.6626*, 2014.

Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Hermann, K. L. and Kornblith, S. Exploring the origins and prevalence of texture bias in convolutional neural networks. *arXiv preprint arXiv:1911.09071*, 2019.

Hirano, H., Koga, K., and Takemoto, K. Vulnerability of deep neural networks for detecting covid-19 cases from chest x-ray images to universal adversarial attacks. *arXiv preprint arXiv:2005.11061*, 2020.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information, 2018.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, pp. 125–136, 2019.

Jacobsen, J.-H., Behrmann, J., Zemel, R., and Bethge, M. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

Janini, D. and Konkle, T. Shape features learned for object classification can predict behavioral discrimination of written symbols. *Journal of Vision*, 19(10):32d–32d, 2019.

Konkle, T. and Alvarez, G. A. Instance-level contrastive learning yields human brain-like representation without category-supervision. *bioRxiv*, 2020.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. doi: 10.1162/neco.1989.1.4.541. URL https://doi.org/10.1162/neco.1989.1.4.541.

Lu, J., Sibai, H., Fabry, E., and Forsyth, D. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv preprint arXiv:1707.03501*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks, 2019.

Mhaskar, H. N. and Poggio, T. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

Neyshabur, B. Towards learning convolutions from scratch. *arXiv preprint arXiv:2007.13657*, 2020.

Ortiz, A., Fuentes, O., Rosario, D., and Kiekintveld, C. On the defense against adversarial examples beyond the visible spectrum. In *MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM)*, pp. 1–5. IEEE, 2018.

Poggio, T., Mhaskar, H., Rosasco, L., Miranda, B., and Liao, Q. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

Purushwalkam, S. and Gupta, A. Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases. *arXiv preprint arXiv:2007.13916*, 2020.

Reddy, M. V., Banburski, A., Pant, N., and Poggio, T. Biologically inspired mechanisms for adversarial robustness, 2020.

Richardson, E. and Weiss, Y. A bayes-optimal view on adversarial examples. *arXiv preprint arXiv:2002.08859*, 2020.

Sadr, J. and Sinha, P. Object recognition and random image structure evolution. *Cognitive Science*, 28(2):259–287, 2004.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Wichmann, F. A. and Hill, N. J. The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, 63(8):1293–1313, 2001.

Yin, D., Lopes, R. G., Shlens, J., Cubuk, E. D., and Gilmer, J. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, pp. 13276–13286, 2019.

Yuan, X., He, P., Zhu, Q., and Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9): 2805–2824, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.

Zhou, Z. and Firestone, C. Humans can decipher adversarial images. *Nature communications*, 10(1):1–9, 2019.

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M., DiCarlo, J., and Yamins, D. Unsupervised neural network models of the ventral visual stream. *bioRxiv*, 2020.

A. Appendix

A.1. Metric

The relative change between the accuracy of a model for two different ϵ's is Eq. 3. We will now show that the robustness R (Eq. 1) is indeed a measure relative change.

$$\frac{1}{\epsilon_1-\epsilon_0}\int_{\epsilon_0}^{\epsilon_1}\frac{f(\epsilon)-f(\epsilon_0)}{f(\epsilon_0)}d\epsilon = \frac{1}{\epsilon_1-\epsilon_0}\int_{\epsilon_0}^{\epsilon_1}\left(\frac{f(\epsilon)}{f(\epsilon_0)}-\frac{f(\epsilon_0)}{f(\epsilon_0)}\right)d\epsilon$$

$$= \frac{1}{\epsilon_1-\epsilon_0}\left(\int_{\epsilon_0}^{\epsilon_1}\frac{f(\epsilon)}{f(\epsilon_0)}d\epsilon - \int_{\epsilon_0}^{\epsilon_1}\frac{f(\epsilon_0)}{f(\epsilon_0)}d\epsilon\right)$$

$$= \frac{1}{\epsilon_1-\epsilon_0}\left(\frac{1}{f(\epsilon_0)}\int_{\epsilon_0}^{\epsilon_1}f(\epsilon)d\epsilon - \int_{\epsilon_0}^{\epsilon_1}1d\epsilon\right)$$

$$= \frac{1}{\epsilon_1-\epsilon_0}\left(\frac{1}{f(\epsilon_0)}\int_{\epsilon_0}^{\epsilon_1}f(\epsilon)d\epsilon - (\epsilon_1-\epsilon_0)\right)$$

$$= \frac{1}{f(\epsilon_0)(\epsilon_1-\epsilon_0)}\int_{\epsilon_0}^{\epsilon_1}f(\epsilon)d\epsilon - \frac{\epsilon_1-\epsilon_0}{\epsilon_1-\epsilon_0}$$

$$= \frac{1}{f(\epsilon_0)(\epsilon_1-\epsilon_0)}\int_{\epsilon_0}^{\epsilon_1}f(\epsilon)d\epsilon - 1$$

Notice that the output of this is in $(-1, 0]$, where values close to -1 imply low adversarial robustness and 0 implies a model did not change in classification accuracy at all as ϵ increases. Therefore, ignoring the shift of -1 so that the measure is easier to interpret and more intuitive, we get R.

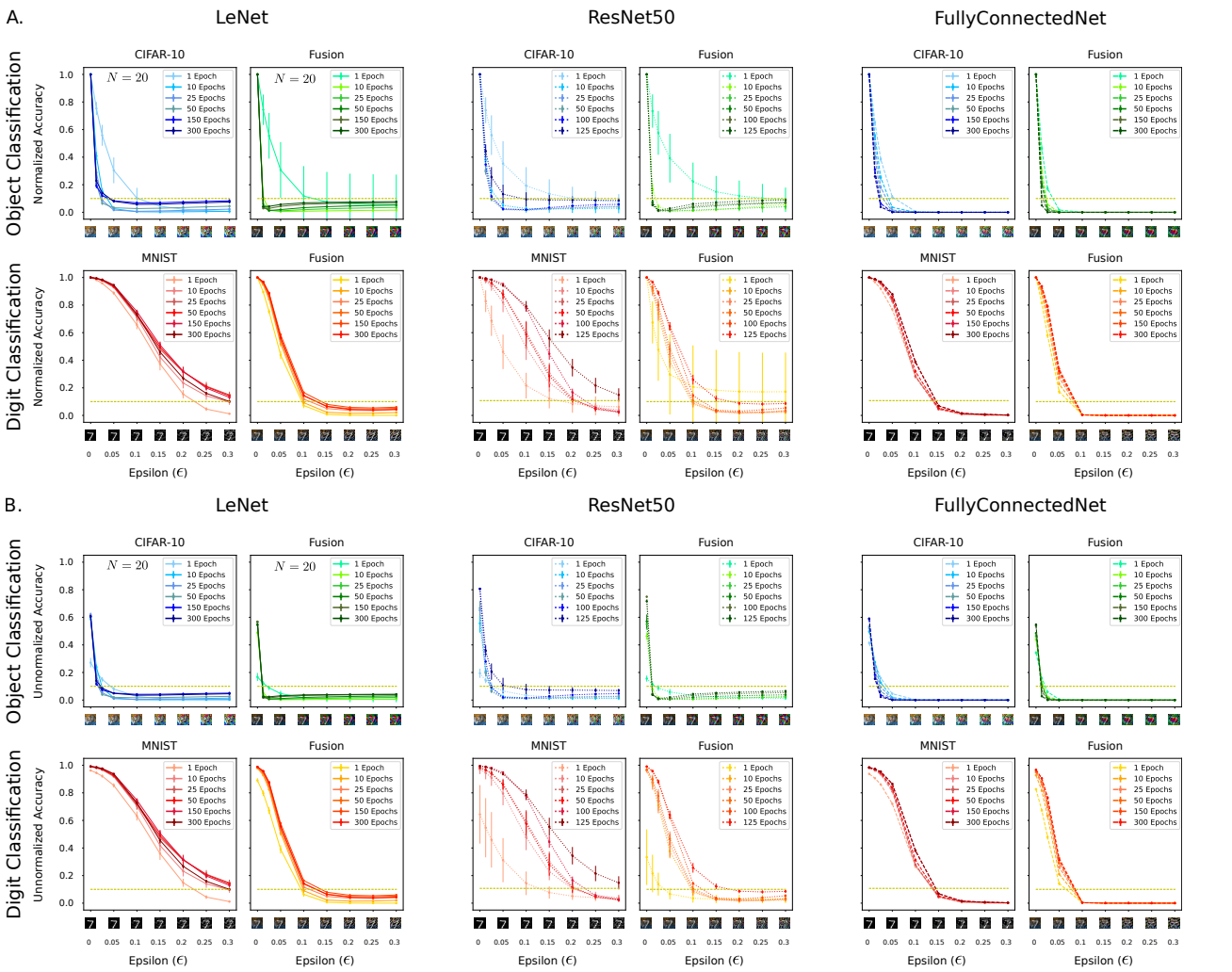

Figure 6. (A): Redrawn graphs from Figure 3 as a reference; (B): The un-normalized adversarial robustness trade-off curves for each network (LeNet, ResNet50, FullyConnectedNet) and dataset (MNIST and CIFAR-10) for the FGSM-based Attack (Goodfellow et al., 2014).

A.2. Unnormalized Adversarial Robustness Curves (Raw-Data)

Figure 6A are redrawn graphs from Figure 3 as a reference. Figure 6B are the unnormalized versions of those graphs. In other words, the points in the normalized graphs are not a proportion out of $f(\epsilon_0)$ and are the raw accuracies for the models at each ϵ.

Case study: In Figure 6B, notice that for LeNet, ResNet50, and FullyConnectedNet (with the exception of epoch 1 for the ResNet50 models), the models for the Fusion dataset with the digit recognition task and MNIST have approximately the same average initial performance for each epoch examined. Observe that although starting at similar points, Fusion digit recognition models accuracies decreases faster as ϵ increases than the MNIST models. We find that indeed MNIST models are more adversarially robust than Fusion digit recognition models. Here, it is clear to see without the normalization.

Figure 7A are redrawn graphs from Figure 5 as a reference. Figure 7B are the unnormalized versions of those graphs.

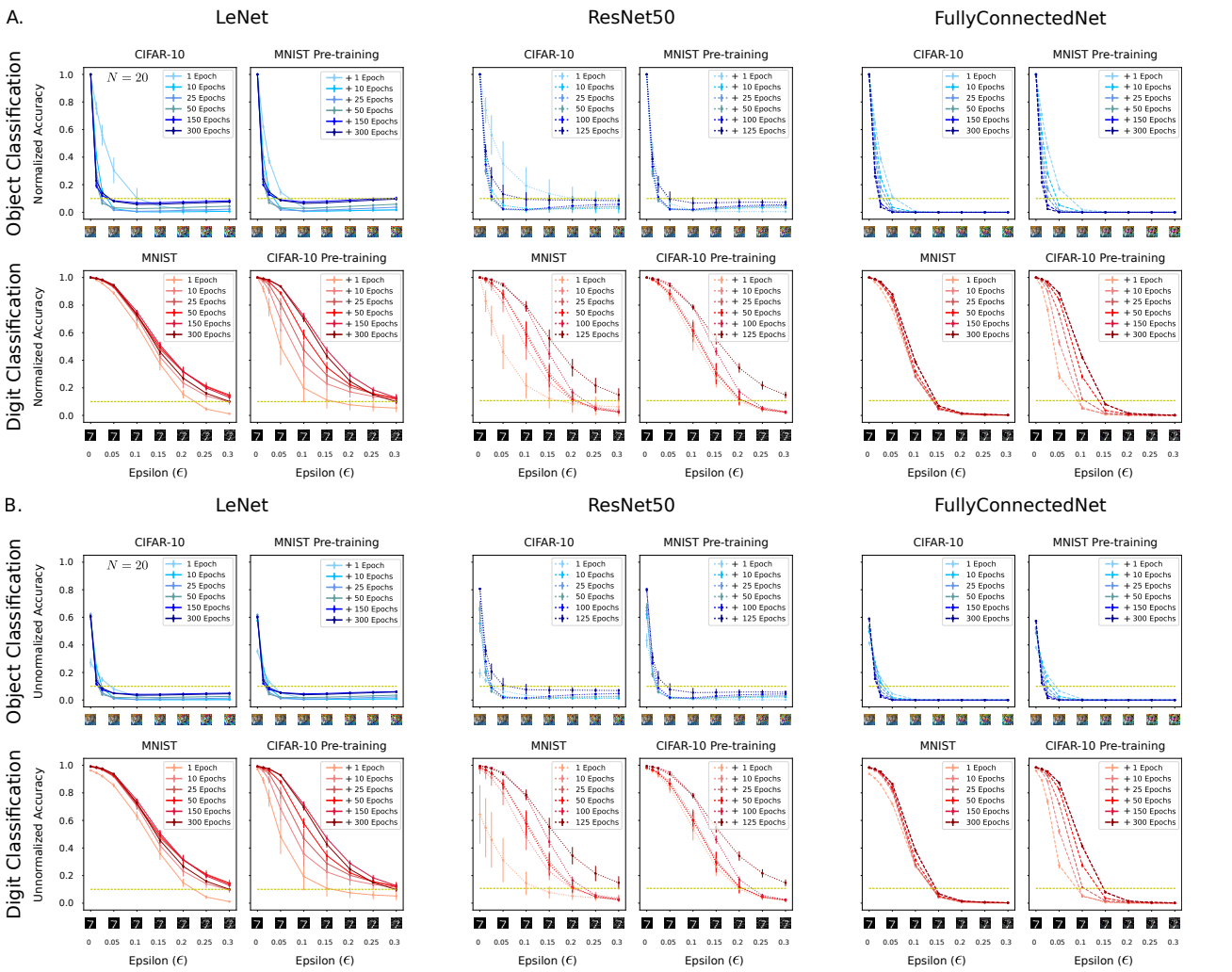

Figure 7. (A): Redrawn graphs from Figure 5; (B): The un-normalized adversarial robustness trade-off curves for each network (LeNet, ResNet50, FullyConnectedNet) and Fusion dataset for the FGSM-based Attack (Goodfellow et al., 2014).

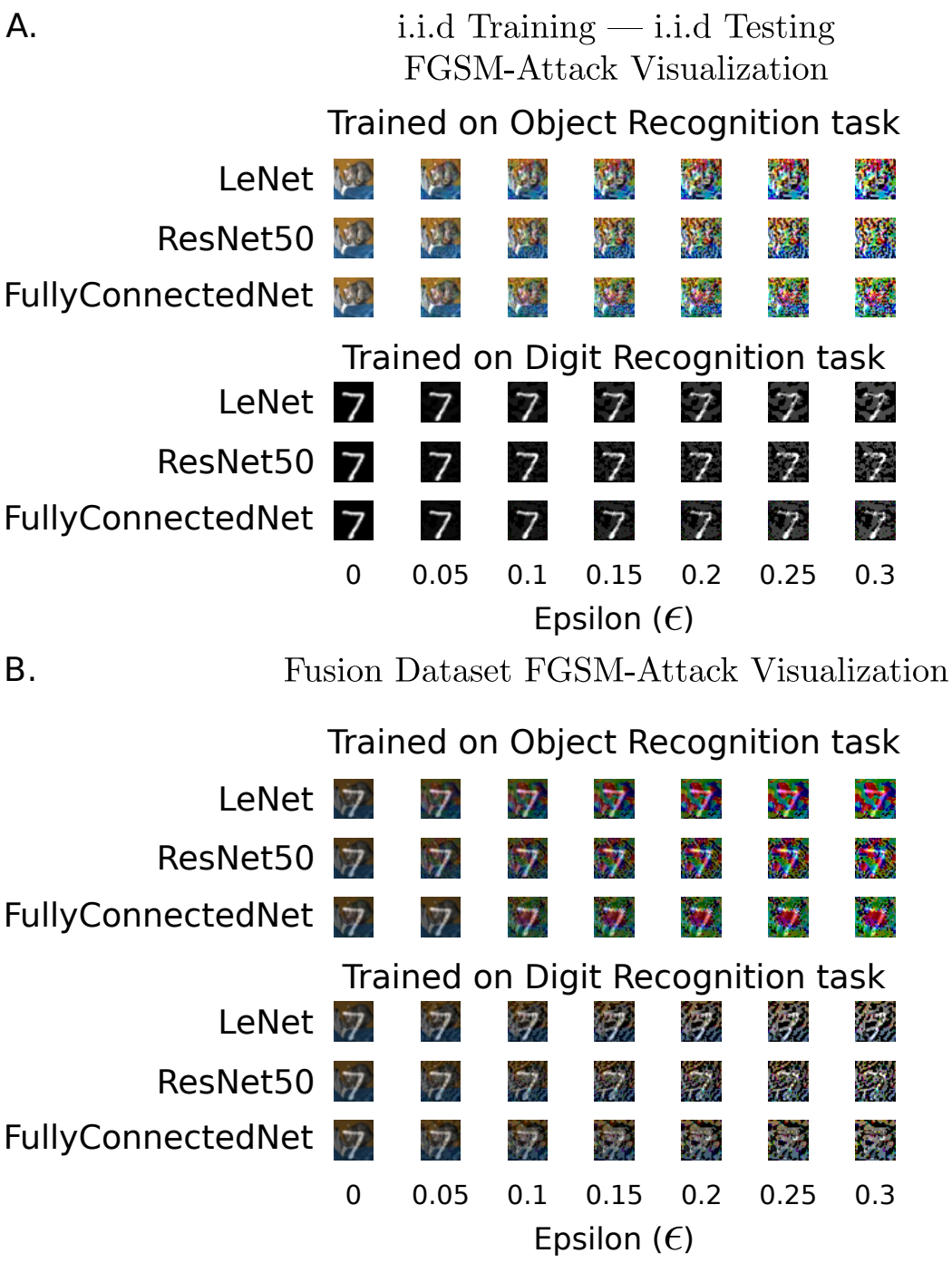

Figure 8. Zoomed in versions of the adversarial patches created after an FGSM-Attack. Shown images are the perturbed stimuli for the networks at the 300th,125th,300th epoch for LeNet, ResNet50 and FullyConnectedNet respectively. A and B show sample stimuli from our first experiment. Interestingly, the differences in the adversarial noise pattern are more salient across architectures for the Fusion Dataset.

A.3. Zoomed in Sample Adversarial Stimuli

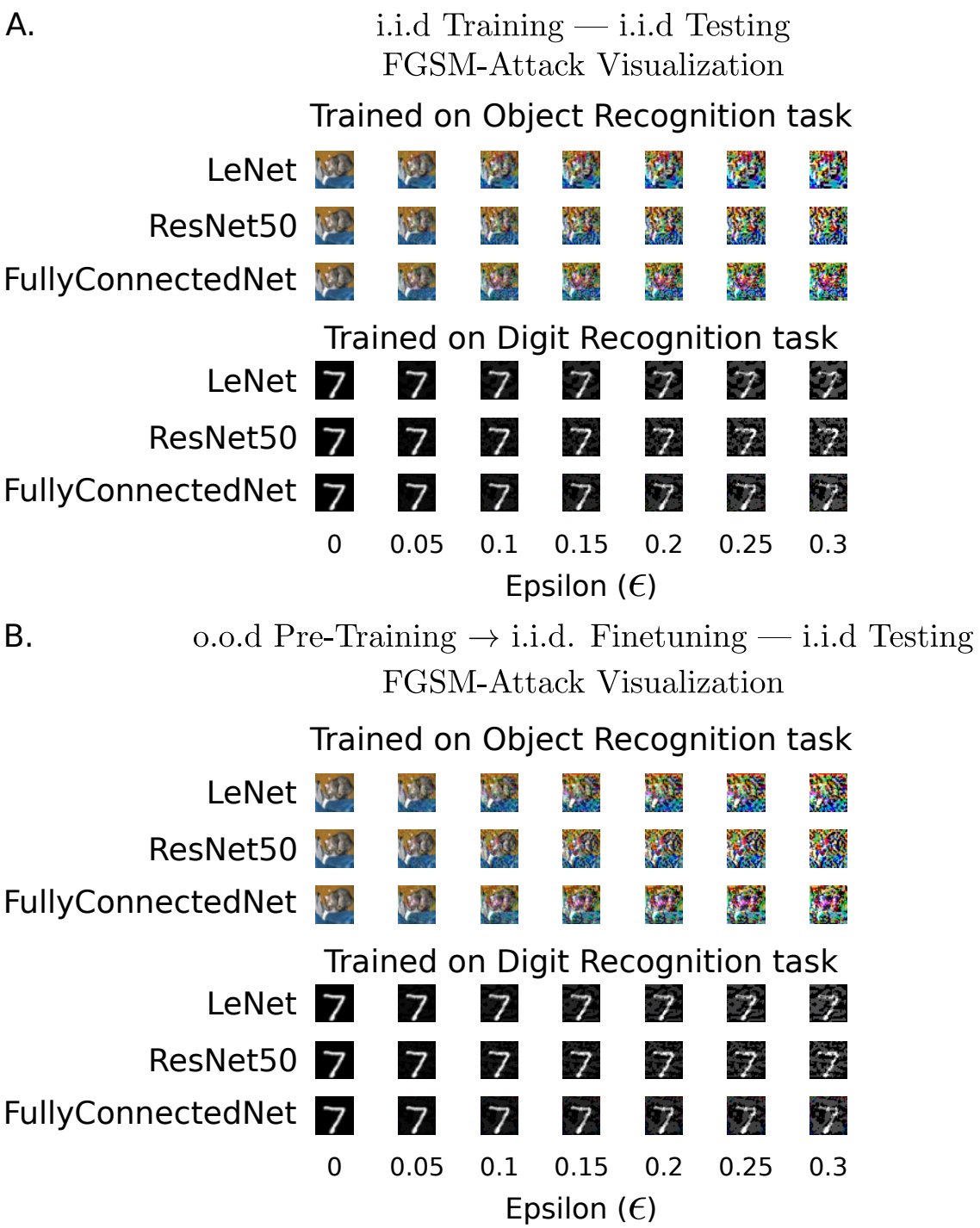

Figure 9. Zoomed in versions of the adversarial patches created after an FGSM-Attack. Shown images are the perturbed stimuli for the networks at the 300th,125th,300th epoch for LeNet, ResNet50 and FullyConnectedNet respectively. A and B show sample stimuli from our second experiment. Interestingly, the differences in the adversarial noise pattern are more salient across architectures for the Fusion Dataset.

A.4. Statistical Testing of Results (Extended)

The tables below contain the robustness R computed using Eq. 1, averaged over the 20 models for the specified dataset/task and architecture at various stages of learning.

Table Legend:
⋆: Denotes statistically significantly higher adversarial robustness for MNIST vs CIFAR-10
•: Denotes statistically significantly higher adversarial robustness for a pretrained MNIST model vs a pretrained CIFAR-10 model
◇: Denotes statistically significantly higher adversarial robustness for a pretrained model vs non-pretrained model both trained on same dataset
†: Denotes statistically significantly higher adversarial robustness for Fusion digit task vs Fusion object task
‡: Denotes statistically significantly higher adversarial robustness for MNIST vs Fusion digit task or CIFAR-10 vs Fusion object task
Results from fully trained models are bolded.

Table 1. MNIST LeNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.438451⋆ ◇ ‡	0.020373
10	0.499361⋆ ◇ ‡	0.023713
25	0.539882⋆ ◇ ‡	0.025632
50	0.543602⋆ ◇ ‡	0.019689
150	0.549709⋆ ◇ ‡	0.013361
300	**0.520547**⋆ ◇ ‡	**0.010607**

Table 2. CIFAR-10 LeNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.155815◇	0.048486
10	0.053812‡	0.003744
25	0.050288‡	0.00447
50	0.063126	0.009296
150	0.102256‡	0.009832
300	**0.097237**‡	**0.008319**

Table 3. Pretrained on CIFAR-10 trained on MNIST LeNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.249363•	0.05819
10	0.364303•	0.049268
25	0.420838•	0.039604
50	0.462459•	0.022959
150	0.529665•	0.013517
300	**0.503499**•	**0.013186**

Table 4. Pretrained on MNIST trained on CIFAR-10 LeNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.102636	0.008461
10	0.059423◇	0.005675
25	0.057364◇	0.008041
50	0.069428◇	0.006776
150	0.113739◇	0.007872
300	**0.109077**◇	**0.008987**

Table 5. Fusion digit task LeNet adversarial robustness

Epoch:	Mean Robustness:	STD:
1	0.173351	0.008095
10	0.204077†	0.006719
25	0.225057†	0.008112
50	0.238905†	0.0096
150	0.254536†	0.005374
300	**0.239281**†	**0.004036**

Table 6. Fusion object task LeNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.181474	0.1951
10	0.035735	0.007395
25	0.046072	0.005877
50	0.058456	0.006017
150	0.077806	0.005403
300	**0.087175**	**0.00396**

Table 7. MNIST ResNet50 adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.243206	0.064093
10	0.37482⋆‡	0.040255
25	0.413437⋆‡	0.037413
50	0.405367⋆‡	0.03901
100	0.486548⋆‡	0.015996
125	**0.572381**⋆‡	**0.03999**

Table 8. CIFAR-10 ResNet50 adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.222479◇	0.098118
10	0.072652‡	0.027665
25	0.06675‡	0.010177
50	0.062354	0.007254
100	0.075903 ‡	0.00379
125	**0.138936**◇‡	**0.038124**

Table 9. Pretrained on CIFAR-10 trained on MNIST ResNet50 adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.392197●◇	0.036562
10	0.422000●	0.023966
25	0.414428●	0.035434
50	0.413840●	0.034093
100	0.491886●	0.012084
125	**0.569465●**	**0.020692**

Table 10. Pretrained on MNIST trained on CIFAR-10 ResNet50 adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.067825	0.008714
10	0.063501	0.0099
25	0.068075	0.012437
50	0.064227	0.010571
100	0.074815	0.004281
125	**0.114783**	**0.03279**

Table 11. Fusion digit task ResNet50 adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.252273	0.267929
10	0.180888†	0.014562
25	0.190701†	0.022835
50	0.202564†	0.009819
100	0.242128†	0.006506
125	**0.296214**†	**0.018022**

Table 12. Fusion object task ResNet50 adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.243001	0.10571
10	0.053983	0.014983
25	0.045387	0.008189
50	0.065714	0.017707
100	0.070246	0.00476
125	**0.084441**	**0.007446**

Table 13. MNIST FullyConnectedNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.272342⋆ ◇ ‡	0.004514
10	0.280533⋆ ◇ ‡	0.002443
25	0.284781⋆ ◇ ‡	0.002756
50	0.29422⋆ ◇ ‡	0.002926
150	0.311588⋆‡	0.000868
300	**0.31289**6⋆‡	**0.00083**

Table 14. CIFAR-10 FullyConnectedNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.086983‡	0.003529
10	0.064134‡	0.002695
25	0.054232‡	0.00201
50	0.045977◇‡	0.001153
150	0.037929◇‡	0.000255
300	**0.034145**◇‡	**0.000274**

Table 15. Pretrained on CIFAR-10 trained on MNIST FullyConnectedNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.153764•	0.005824
10	0.194182•	0.00334
25	0.235387•	0.002786
50	0.281417•	0.004191
150	0.320011•◇	0.001317
300	**0.321303•◇**	**0.001126**

Table 16. Pretrained on MNIST trained on CIFAR-10 FullyConnectedNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.104627◇	0.004495
10	0.069091◇	0.002008
25	0.053113	0.002535
50	0.044058	0.001362
150	0.036256	0.00035
300	**0.031516**	**0.000222**

Table 17. Fusion digit task FullyConnectedNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.113213†	0.003298
10	0.129283†	0.001951
25	0.139871†	0.001997
50	0.145459†	0.001856
150	0.15254†	0.000776
300	**0.15016**†	**0.000758**

Table 18. Fusion object task FullyConnectedNet adversarial robustness

Epoch:	Mean Robustness:	SD:
1	0.052986	0.002406
10	0.036917	0.00154
25	0.03305	0.001239
50	0.030376	0.001014
150	0.026116	0.00027
300	**0.022955**	**0.000141**