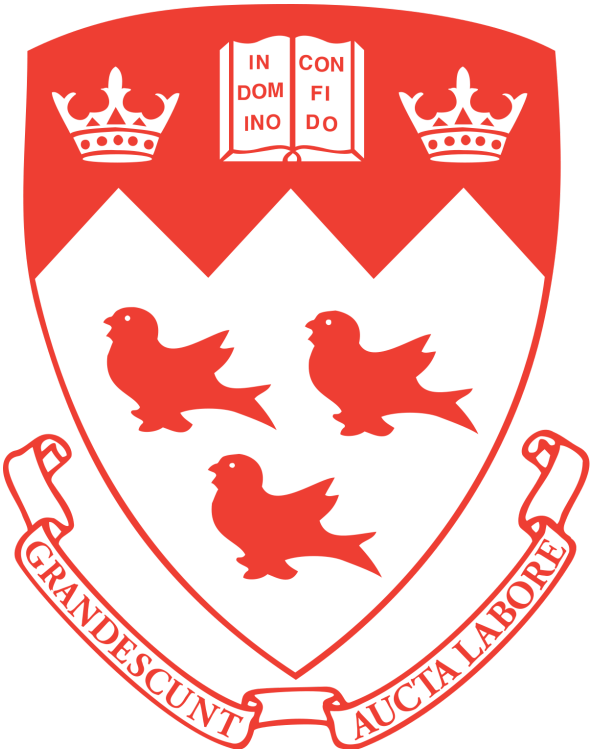
COMP 551 - Project 4



Viet Tran - 260924954
Oliver Stappas - 260930067
Lynn Cherif - 260822727

Abstract

In this project, the main claims of the paper titled "Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification" are investigated on KMNIST and CIFAR-10 datasets, as well as the CelebA dataset. The most important findings include that $\| \delta \|_2 \leq 3$ constrained robust models outperform their natural counterparts, particularly when using fewer training samples and on datasets most different from the source dataset, using 1-3 fine-tuning blocks. It is also found that robust models train faster and present shape bias, rather than texture bias, which is demonstrated through their better performance on low-resolution images. Additional findings include that target datasets more similar to the source set require fewer fine-tuning blocks in order to achieve good performance (0-1 fine-tuning blocks).

Introduction

The paper titled "Adversarially-Trained Deep Nets Transfer Better: Illustration on Image Classification" suggests that adversarially trained deep models transfer better to new datasets (i.e., non-source datasets) than natural models, particularly when fine-tuning on fewer training instances and when applied on less similar target datasets to the source dataset. Performance is measured through both test accuracy and number of training epochs to convergence. It is claimed that the better performance of robust models is attributed to its shape bias and its ability to capture class-level semantic properties, similar to human learning. In addition, the similarity between the source and target dataset decreases the optimal number of fine-tuned blocks and the robustness constraint required. As such, experiments supporting robust models' better performance claims were done as in the paper, using their naturally and adversially pre-trained ResNet-50 on ImageNet on the KMNIST and CIFAR-10 datasets for different training subsets, number of fine-tuning blocks, number of training epochs, and adversarial constraint types. Experiments using influence functions to support robust models' ability to capture class-level semantic data were only reproduced with 100 training samples, due to limited time and compute resources. Better performance claims of the robust models and their shape bias were also investigated on a new dataset, CelebA. Robust models' better transferability than natural models on new datasets and shape bias were indeed confirmed through both the reproduced and additional experiments.

Paper

In order to support the claims about adversarially-trained deep network's better and faster transferability to new domains, the authors propose experiments done on six different target domains, with 0, 1, 3 or 9 fine-tuned convolutional blocks, training subsets ranging from 100 to 25600 instances with 2-fold increments, and 5 to 20 random seeds, using a natural model and three adversarially pre-trained models using $\| \delta \|_2 \leq 3$, $\| \delta \|_\infty \leq 4/255$ and $\| \delta \|_\infty \leq 8/255$ constraints all based on Resnet50. These experiments are divided into four main experiments supporting different parts of the claim, where the authors report the mean and 95% confidence interval of the results over the random seeds.

The first experiment plots the test accuracy results of the $\| \delta \|_2 \leq 3$ adversarially trained model and naturally trained model, fine-tuned on 3 blocks, and applied on the 6 respective target datasets: CIFAR-100, CIFAR-10, SVHN, FMNIST, KMNIST, MNIST. The results consistently indicate the robust model's superior performance. The second experiment shows the faster transferability of the robust models by plotting the test accuracy against the number of training epochs for each respective dataset, using both the $\| \delta \|_2 \leq 3$ adversarially and naturally pretrained models, with 3 fine-tuning blocks on 3200 images (~ 5% of the target dataset). As for the third experiment, the $\| \delta \|_2 \leq 3$ adversarially pre-trained model is applied on the respective datasets using 0,1,3, and 9 fine-tuning blocks across training subsets from 100-25600. This experiment shows the superior performance of fine-tuning 1-3 blocks in general, and the need for less fine-tuning blocks (0-1) on datasets that are more similar to source dataset such as CIFAR-10 and CIFAR-100. The fourth experiment shows the test accuracy of the three robust models ($\| \delta \|_2 \leq 3$, $\| \delta \|_\infty \leq 4/255$ and $\| \delta \|_\infty \leq 8/255$) applied on the most similar dataset to the source dataset (CIFAR-10), and the most different (SVHN), with 3 fine-tuning blocks and across the multiple training subsets. The $\| \delta \|_2 \leq 3$ constrained model outperformed the two others.

With regards to the shape bias, the most notable experiment was downscaling the Caltech101 dataset from 224x224 to 32x32 and observing the $\| \delta \|_2 \leq 3$ robust model's better performance over the natural model using 3 fine-tuning blocks. Finally, to the robust model learning class-level semantic information, the authors use influence functions in order to output the most influential image and label on the test prediction.

Some other smaller experiments were also done throughout the paper, such as using Gaussian perturbation to improve or replace adversarial training. However, these experiments were not considered as they required to retrain the model from scratch which was not an option for us given the time and compute constraints.

Experiments & Results

The first experiment attempted to verify the claim that robust models transfer better to new domains. The setup is as follows, first the target dataset was chosen to be KMNIST as it is the most different when compared to the original training set (ImageNet). The fine-tuning parameters used were taken directly from the paper's Tables 3 and 4 (see appendix Table A1 & A2 for reference). However, only one seed was chosen for time purposes. The transforms applied to a dataset varied, grayscale was applied to MNIST, KMNIST, while horizontal flips, rotations, jitter, and random cropping were applied to every other dataset. Pretrained models provided by the authors of the paper were used. Given the focus was to prove that the robust model performs better, the $\| \delta \|_2 \leq 3$ constrained model was chosen considering its superior results out of the three mentioned in the paper. Using 3 fine-tuning blocks, the $\| \delta \|_2 \leq 3$ robust model and natural model's test accuracy were compared for several training subset sizes on KMNIST and CIFAR-10. As can be seen in Figures A1 and A2, the robust model outperforms the natural model, particularly using smaller training subsets, as expected from the paper.

Further, the experiments to verify that robust models train faster than their natural counterparts were done also using the $\| \delta \|_2 \leq 3$ constrained and natural models. Although the author's only performed the experiment on 3 fine-tuning blocks and 3200 training samples, we reproduced the experiments on 1,3, and 9 blocks in order to also assess the influence of fine-tuning blocks on training speed (Figures A3-A6). The results obtained show that the robust model achieves good results with a small number of epochs, while the natural model needs to be trained longer to achieve good accuracy. This falls in line with the findings in the paper, as the robust model requires less epochs than the natural model.

The third experiment was done to verify the target datasets that are more similar to the source dataset require fewer fine-tuning blocks to perform well. CIFAR-10, amongst the most similar to ImageNet, was chosen for the experiment. The training method followed the same process as in experiment 1. As claimed and can be seen in Figure A7, it is verified that training on 0-1 blocks on this dataset outperforms the 1-3 blocks required for KMNIST.

However, it can also be seen that in general, as claimed in the paper, 1-3 blocks perform best out of the 0,1,3, or 9 fine-tuning blocks see Figure A5-A6, particularly when using smaller training subsets. It is also worth pointing out that our results experienced some outliers given we did not train using multiple random seeds as in the paper.

In a new experiment, the performance of the different models on a dataset outside of the ones mentioned in the paper was considered. Indeed, in the "future works" section of the paper, it is mentioned that considering robust models' low-resolution and low-frequencies bias, it could be possible that the robust model would perform worse in this case. With this in mind, a high resolution (178x218) face dataset, CelebA, was used. The dataset was set up so that the models would perform a simple classification on whether or not an image was a male. A 32x32 transform provided by the authors was used on the dataset, as done on the Caltech101 dataset in the paper. The same training parameters as the other experiments were used, and the best performing robust model ($\| \delta \|_2 \leq 3$) with 3 unfrozen blocks, considering it had the best performance on new domains different from ImageNet.

As hypothesized in the paper, the natural model performs better on the high-resolution face dataset, although by slight margins when classifying an image of a person by gender. The natural model is particularly superior when training on a small number of images. Figure A11 shows that the robust model performs much worse when only training on 100 images, although the difference in accuracy when increasing the size of the training set between the two models is minimal.

Moreover, in order to further prove that robust models are biased toward low-resolution, i.e., have a shape bias rather than texture bias, a downscaled 32x32 CelebA dataset was used. The rest of the training methodology was exactly the same as what was done in experiment 1. It can be seen in Figure A12 that reducing the resolution indeed positively affects the robust model as it is capable of achieving better or close to the same accuracy as the natural model, confirming robust models' shape bias.

In another experiment, the best performing robustness constraint out of the $\| \delta \|_2 \leq 3$, $\| \delta \|_\infty \leq 4/255$ and $\| \delta \|_\infty \leq 8/255$ constraints. Three datasets were considered in this experiment: the downscaled CelebA, KMNIST and CIFAR-10. Each of the three constrained models were trained from 100 to 6400, with two-fold increments on the CelebA, while only 100 to 1600 were considered for the other two datasets considering time and compute constraints. Valuable results were still achieved, as the required trends can still be observed in this range. The training parameters were also reused as in Tables A1 and A2. Figures A8, A9 and A10 in the appendix show the different constrained models are very close in terms of performance. Although the $\| \delta \|_2 \leq 3$ performs the best beating out the other two models for accuracy on KMNIST and CIFAR-10. For the CelebA dataset, $\| \delta \|_2 \leq 3$ model performs poorly with a small number of training images (100) but surpasses the other two afterwards. Which is similar to what is described in figure 5c of the paper as this constraint performs worse at the start on unsimilar datasets but catches up afterwards.

The final experiment was done to find the most influential images for the robust and natural models on a subset of CIFAR-10 in order to confirm the use of class-level semantic information of the robust model in the learning process, as suggested in the paper. The original paper used a subset of 3200 training images, but due to considerably high computational costs, we attempted to reach the same conclusions as the paper with a 100 training image subset. Another issue faced was that the supplied code for this part was not well documented and presented many bugs. After computing the influence functions for the $\| \delta \|_2 \leq 3$ constrained model and the natural model, an output of test images with the most influential images for both the natural and robust model was generated (see figure A13), similar to Figure 6 in the paper. Unfortunately, it is evident that we could not demonstrate that the robust model's influential images and their labels were more similar to the test images and their labels than those of the natural model. Considering that both the robust and natural models only achieved 57% and 52% test accuracy respectively on a subset of 100 CIFAR-10 training images, this is indeed expected.

Discussion & Conclusion

In conclusion, we validated the authors' claims that the robust model outperforms the natural model when transferring to a new domain particularly when using less training samples, and using less training to classify data accurately. Robust models' shape bias was also verified through its better performance on the downscaled resolution CelebA dataset than on the higher resolution where the natural model performed better. The $\| \delta \|_2 \leq 3$ constrained robust model transferred best out of three adversarially trained models, particularly when using 1-3 fine-tuning blocks. Moreover, tuning 3 block showed the best results when training on small image samples. Attempting to prove the robust models' ability to capture class-level semantic properties during the training process was insufficient for a reliable conclusion using only 100 training samples of CIFAR-10, as both the natural and robust models perform poorly using that subset size. Having more computational power and time, it would be worth further investigating the claims on the robust models' ability to capture class-level semantic properties using the downscaled CelebA dataset, and a larger training subset. Moreover, adversarial attacks' contribution to improving transferability could be evaluated.

References

[1] F. Utrera, E. Kravitz, N. B. Erichson, R. Khanna, and M. W. Mahoney, "Adversarially-trained deep nets transfer better," *arXiv preprint arXiv:2007.05869,* 2020.

Appendix

Table A1: Hyper-parameter summary for all fine-tuned source models, “Table 3” [1]

Learning rate	Batch size	Momentum	Weight decay	LR decay	LR decay schedule	Fine-tuned adversarially?
0.1	128	0.9	5×10^{-4}	10x	1/3, 2/3 epochs	No

Table A2: Batch summary for every target dataset and source model, “Table 4” [1]

Number of images	Fine-tuning epochs	Number of random seeds	Test accuracy frequency (epochs)	LR decay schedule
100	100	20	20	33/66
200	100	20	20	33/66
400	100	20	20	33/66
800	100	20	20	33/66
1,600	100	20	20	33/66
3,200	150	10	10	50/100
6,400	150	10	10	50/100
12,800	150	5	10	50/100
25,600	150	5	10	50/100
All	150	1	10	50/100

Experiment 1

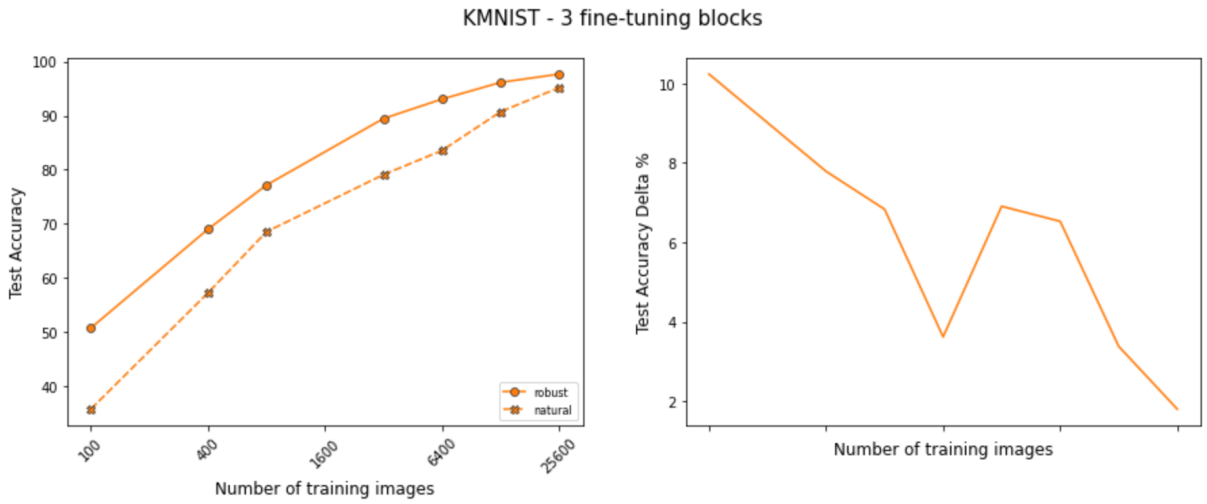


Figure A1: Test accuracy vs number of training images on KMNIST using $\| \delta \|_2 \leq 3$ constrained model and natural model with 3 fine-tuning blocks

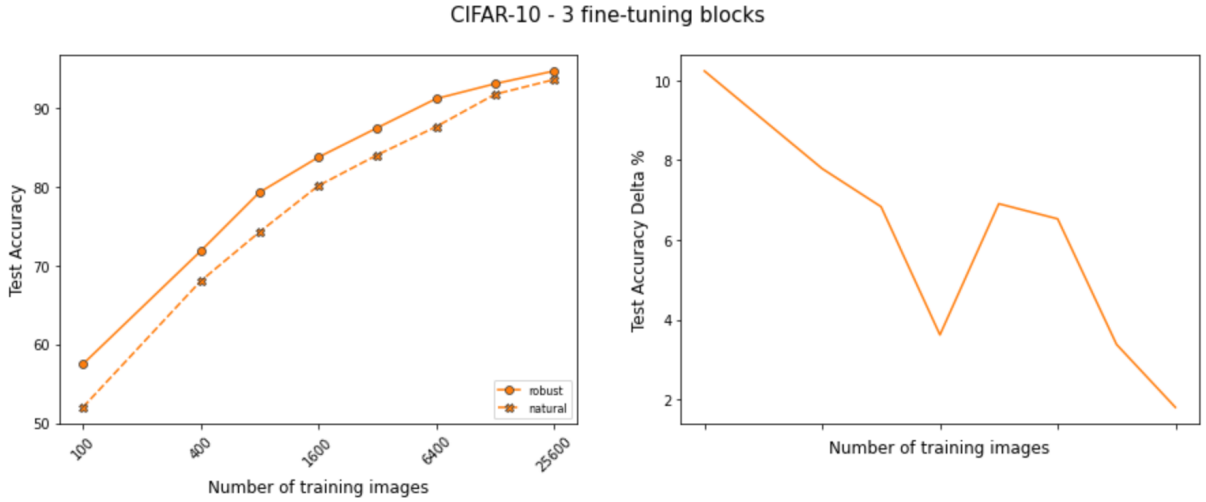


Figure A2: Test accuracy vs number of training images on CIFAR-10 using $\| \delta \|_2 \leq 3$ constrained model and natural model with 3 fine-tuning blocks

Experiment 2:

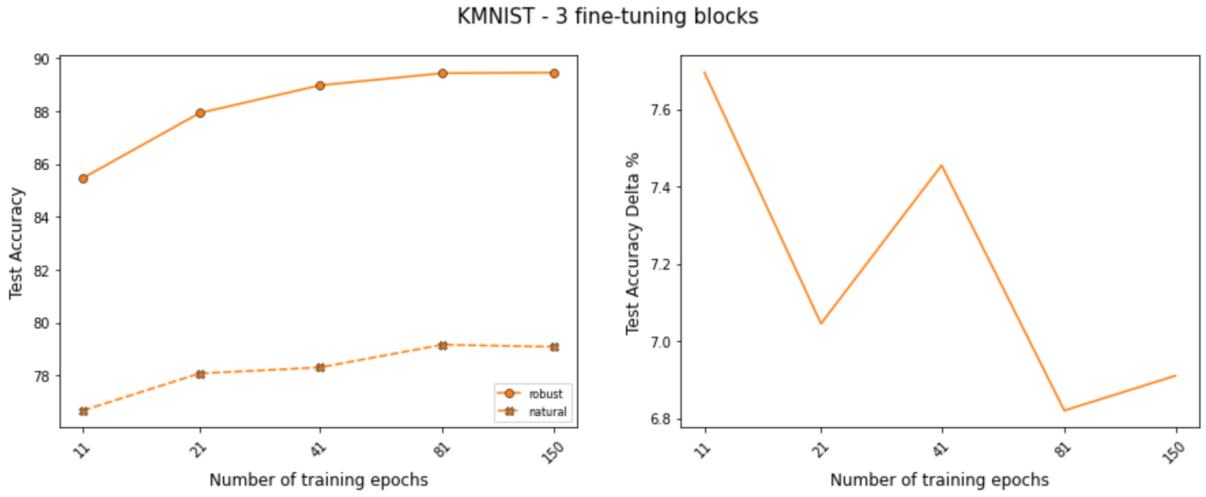


Figure A3: Test accuracy and test accuracy delta vs number of training epochs on KMNIST using $\| \delta \|_2 \leq 3$ constrained model and natural model with 3 fine-tuning blocks and a random training subset of 3200 images

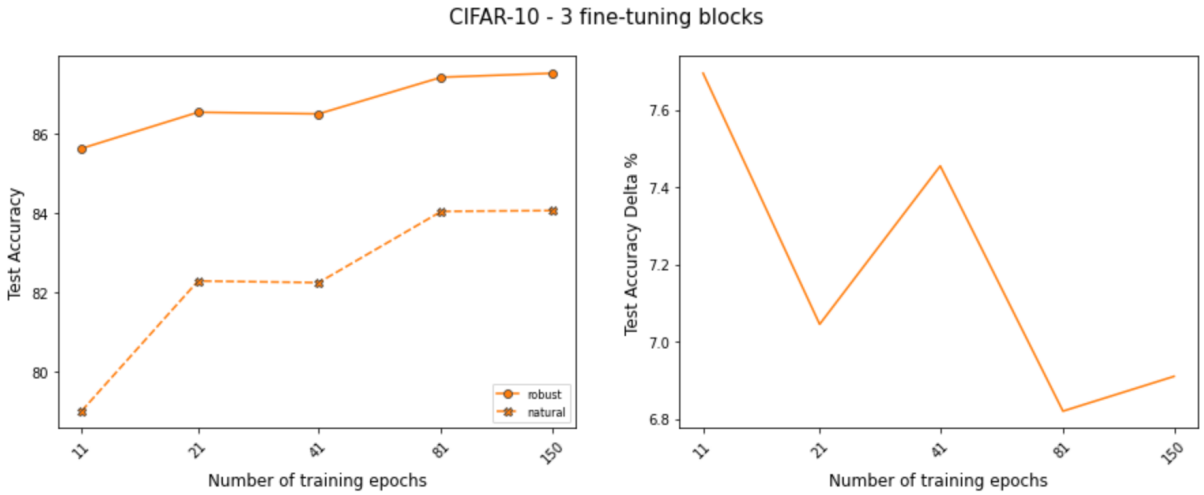


Figure A4: Test accuracy and test accuracy delta vs number of training images on CIFAR-10 using $\| \delta \|_2 \leq 3$ constrained model and natural model with 3 fine-tuning blocks and a random training subset of 3200 images

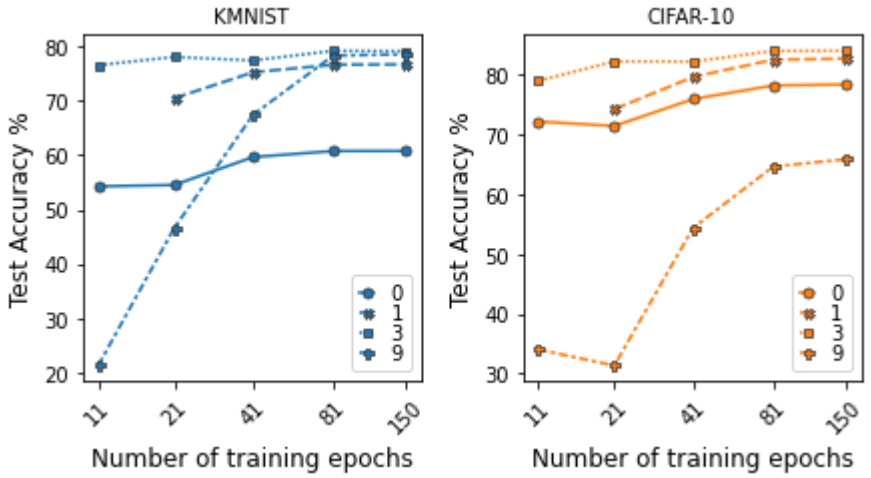


Figure A5: Test accuracy and test accuracy delta vs number of training images on KMNIST and CIFAR-10 using natural model with 1,3 and 9 fine-tuning blocks and a random training subset of 3200 images

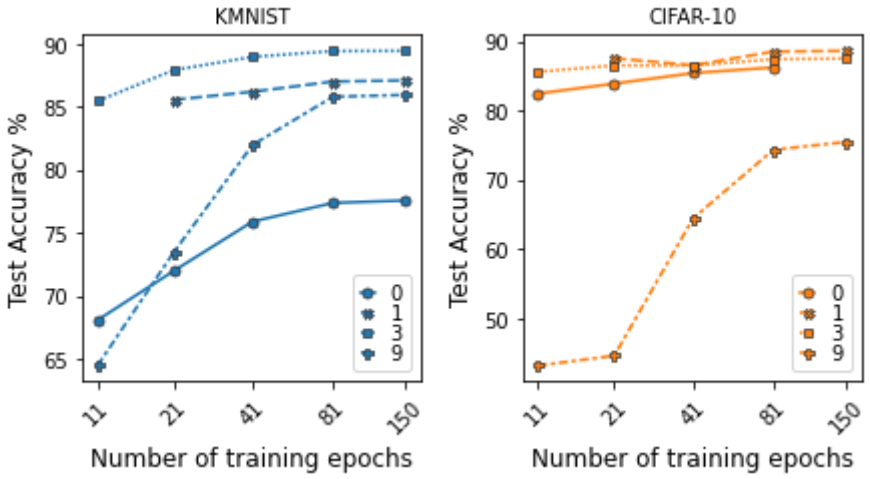


Figure A6: Test accuracy and test accuracy delta vs number of training images on KMNIST and CIFAR-10 using the $\| \delta \|_2 \leq 3$ robust model with 1,3 and 9 fine-tuning blocks and a random training subset of 3200 images

Experiment 3

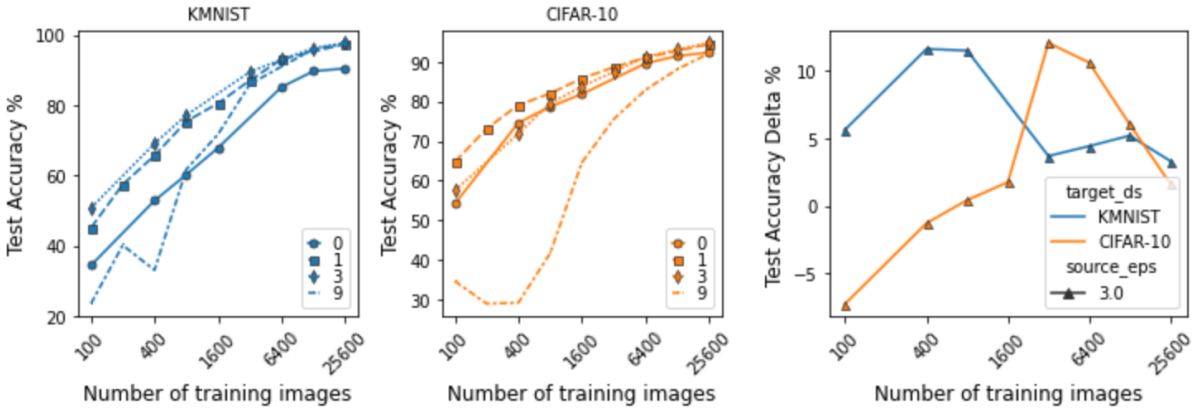


Figure A7: Test accuracy and test accuracy delta vs number of training images on KMNIST and CIFAR-10 using $\| \delta \|_2 \leq 3$ constrained model with 0,1,3,9 fine-tuning blocks

Experiment 4

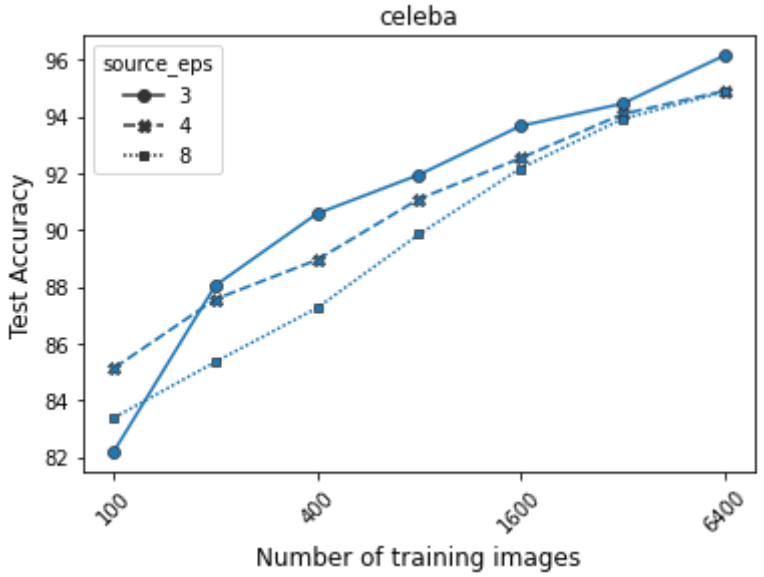


Figure A8: Test accuracy vs number of training images on CelebA using $\| \delta \|_2 \leq 3$, $\| \delta \|_\infty \leq 4/255$ and $\| \delta \|_\infty \leq 8/255$ constrained models with 3 fine-tuning blocks

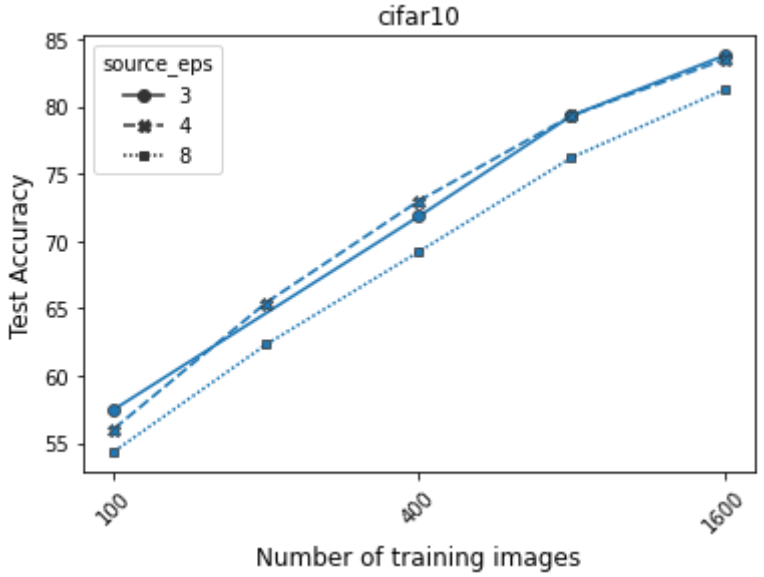


Figure A9: Test accuracy vs number of training images on CIFAR-10 using $\| \delta \|_2 \leq 3$, $\| \delta \|_\infty \leq 4/255$ and $\| \delta \|_\infty \leq 8/255$ constrained models with 3 fine-tuning blocks

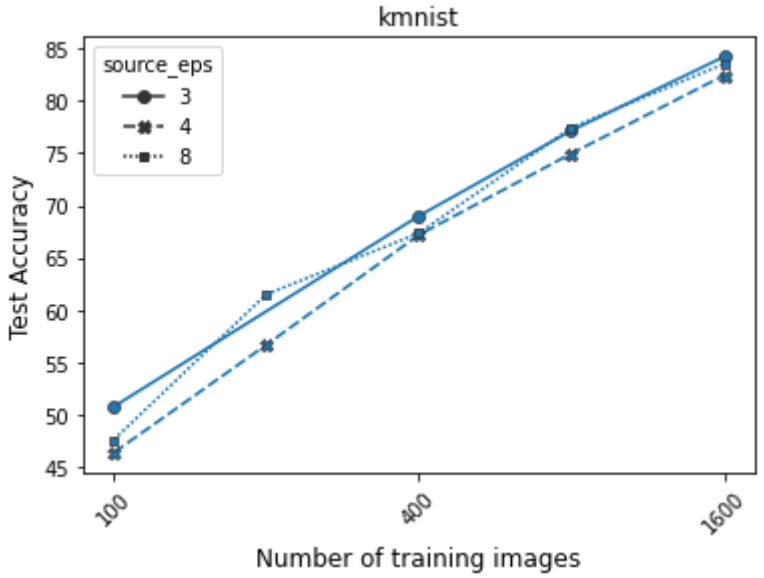


Figure A10: Test accuracy vs number of training images on KMNIST using $\|\delta\|_2 \leq 3$, $\|\delta\|_\infty \leq 4/255$ and $\|\delta\|_\infty \leq 8/255$ constrained models with 3 fine-tuning blocks

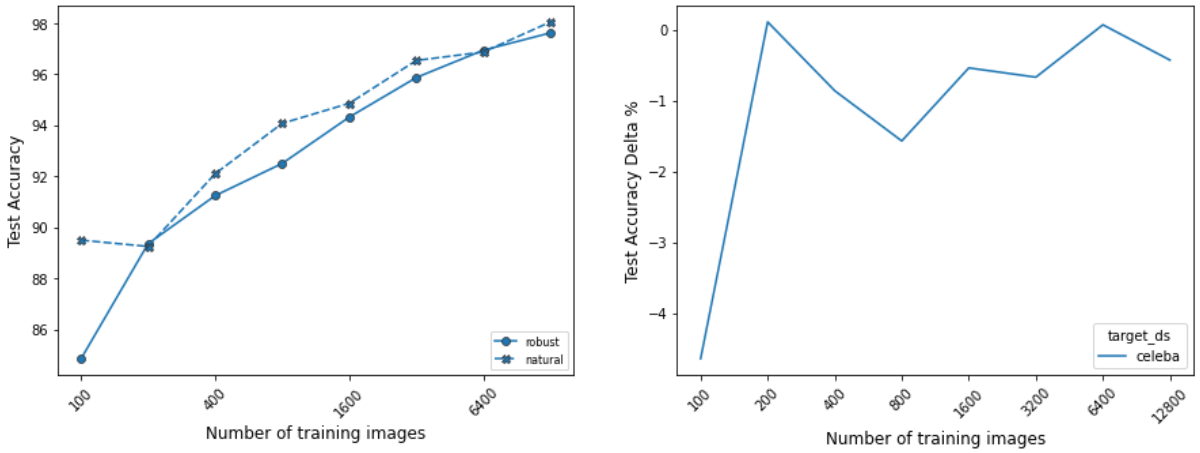


Figure A11: Test accuracy and test accuracy delta vs number of training images on high resolution Celeb A using $\|\delta\|_2 \leq 3$ constrained model with 3 fine-tuning blocks

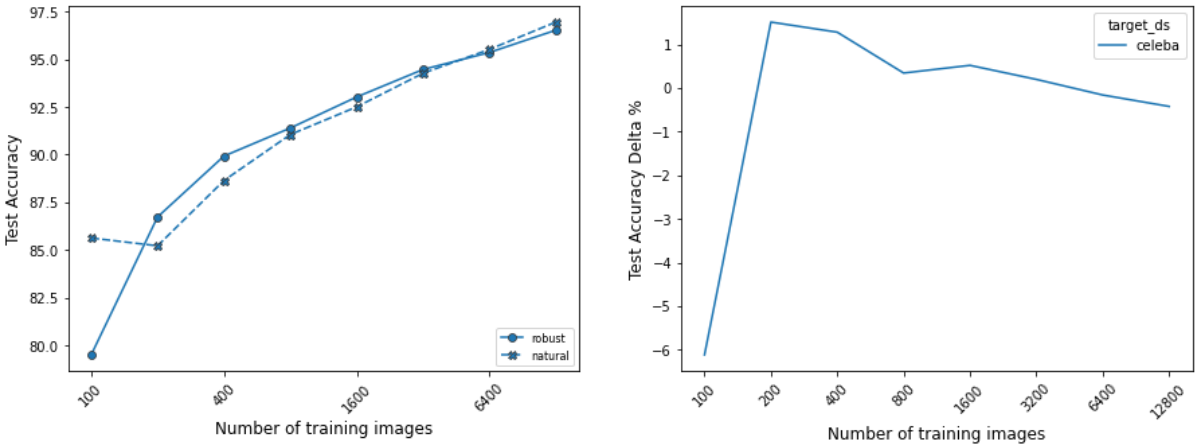


Figure A12: Test accuracy and test accuracy delta vs number of training images on low resolution Celeb A using $\|\delta\|_2 \leq 3$ constrained model with 3 fine-tuning blocks

Figure A13: Image Predictions for the robust and natural model on CIFAR-10