Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *arXiv preprint arXiv:2106.10151*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Thomas Tanay and Lewis Griffin. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. On the convergence and robustness of adversarial training. *arXiv preprint arXiv:2112.08304*, 2021.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

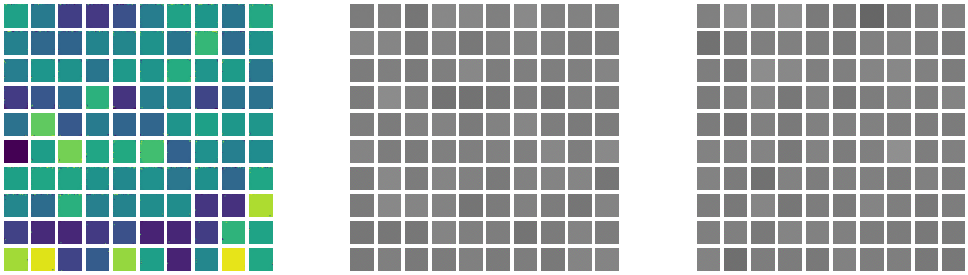
A Appendix

SVD decomposition on MNIST, CIFAR-10 and CIFAR-100. Figure 8 and 9 show the first 100 and last 100eigenvectors obtained after SVD decomposition on MNIST, CIFAR-10 and CIFAR-100, respectively. It can be seen that large-scale features have rich semantic information, mainly including shape, color, etc. The small-scale features contain more high-frequency features that are difficult for humans to recognize semantics.

(a) MNIST (b) CIFAR-10 (c) CIFAR-100

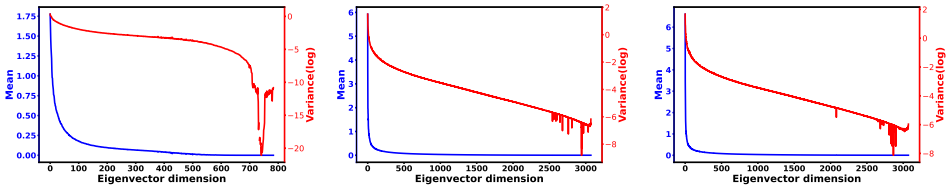
Figure 8: The first 100 eigenvectors obtained after SVD decomposition on the three data of MNIST, CIFAR10 and CIFAR100.

Projecting the samples onto all the eigenvectors, we take the absolute value of the projection on each eigenvector and calculate its mean and variance. The results are shown in Figure 10. The cumulative distributions of MNIST, CIFAR-10 and CIFAR-100 on the eigenvectors, respectively, are shown in Figure 11. We can see that the cumulative distribution of these three datasets on the top 300 feature vectors has exceeded 95%, which shows that the main features of these datasets are concentrated on large-scale features. Figure 12 depicts the large-scale features and small-scale features in the original samples, which are decomposed by utilizing the first 300-dimensional eigenvectors. It can be seen that the small-scale features in the sample are difficult for us to identify, so we do not want the model to rely too much on this part of the features.



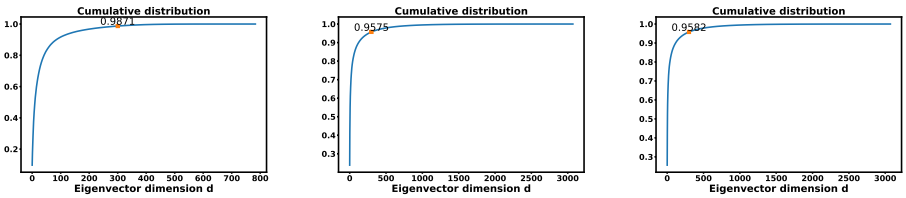
(a) MNIST (b) CIFAR10 (c) CIFAR100

Figure 9: The last 100 eigenvectors obtained after SVD decomposition on the three data of MNIST, CIFAR10 and CIFAR100.



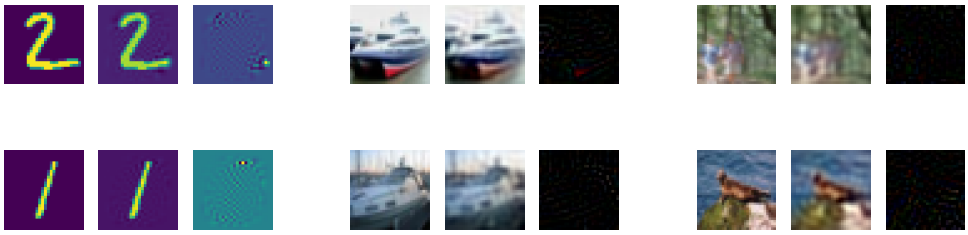
(a) MNIST (b) CIFAR10 (c) CIFAR100

Figure 10: Take the absolute value of the sample projection on each eigenvector and visualize the mean and variance(log) in each dimension.



(a) MNIST (b) CIFAR10 (c) CIFAR100

Figure 11: Cumulative distributions of MNIST, CIFAR-10, and CIFAR-100 over eigenvectors, respectively.



(a) MNIST (b) CIFAR10 (c) CIFAR100

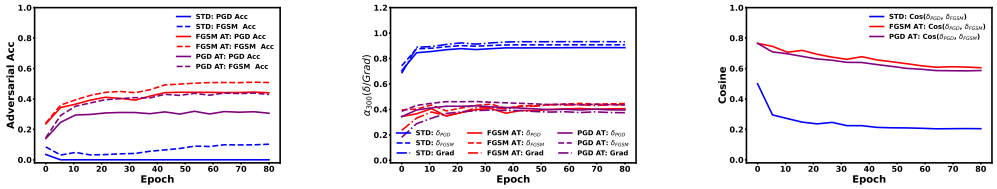
Figure 12: For a given original image (the first column), the large-scale features in the sample (the second column) are obtained by reconstructing the sample with the first 300-dimensional feature vectors, and the remaining features in the sample, that is, the small-scale features (third column).

B APPENDIX

The size of the adversarial perturbation is set to $\varepsilon = 0.3$ on MNIST and $\varepsilon = 8/255$ on CIFAR-10 and CIFAR-100, respectively. For the PGD AT we iterate 10 times with a step size of $\alpha = 2/255$ on CIFAR-10 and CIFAR-100, and 40 iterations on MNIST with a step size of 0.01. All adversarial training methods do not employ random restarts. To evaluate the robustness of the model under PGD attack, we adopt PGD-20-10 with 20 iterations and 10 restarts with a stride $\alpha = \varepsilon/4$. Since adversarial training is prone to adversarial overfitting (Rice et al., 2020), the results we report in this paper are the best robust performance of the model before adversarial overfitting occurs. In addition, we also test the performance of the models trained on the CIFAR-10 and CIFAR-100 datasets under the AutoAttack attack (Croce & Hein, 2020).

C APPENDIX

We conduct experiments on the VGG16 model on CIFAR-10. The changes in the adversarial attack recognition accuracy, the proportion of small-scale features, etc. of the VGG16 model trained by STD, FGSM AT, and PGD AT are summarized in Figure 13. Surprisingly, it can be seen from Figure (a) that the VGG16 model of FGSM AT does not undergo catastrophic overfitting. Moreover, the robust accuracy of the VGG16 model of FGSM AT is similar to that of PGD AT (Table 5). It is worthy of our further research to analyze whether the skip connection will affect the robustness results of FGSM AT training.



(a) (b) (c)

Figure 13: Visualization of different metrics in the training process of VGG16 models for STD, PGD AT, and FGSM AT on CIFAR-10. All statistics are computed on the test set.

Table 5: The recognition accuracy of the VGG16 on CIFAR-10 clean test set and under different attack algorithms. We report mean values of the accuracy in three independent experiments. The adversarial perturbation size is $8/255$. '_' indicates that we did not test the performance of the model.

Training Method	STD	PGD AT	FGSM AT	STD FDA	FGSM AT GradAlign	FGSM AT FDA
Test Acc	92.24%	77.53%	75.02%	92.18%	78.26%	77.65%
FGSM Acc	9.51%	50.64%	50.21%	10.02%	49.64%	49.65%
PGD Acc	0.00%	46.97%	45.70%	0.12%	43.88%	44.21%
AutoAttack	_	43.90%	_	_	38.60%	40.40%