The Efficiency of Singular Value Decomposition on Detecting Adversarial Samples

Jeffery Yang, Lavita Zuo
University of California, San Diego

Introduction

Deep neural networks have proved to be very successful in performing daily tasks like autonomous driving(Tesla Autopilot)[1], facial recognition(Apple FaceID), and natural language processing(Siri). Convolutional neural networks(CNNs), one of the most popular neural network architectures, are particularly efficient for image processing.

However, there has been a concern widely about the security and robustness of neural networks given adversarial samples as inputs.[2] Adversarial samples (figure 1) are images or signals that contain noises imperceptible to human eyes, and when fed into neural networks they can cause misclassification of the actual inputs. Such mistakes can be fatal when it comes to critical systems controlling people's lives. Thus, we're inspired to examine the behaviors of adversarial attacks by performing the principal component analysis(PCA), particularly singular value decompolisiotin(SVD) on several existing CNN architectures.

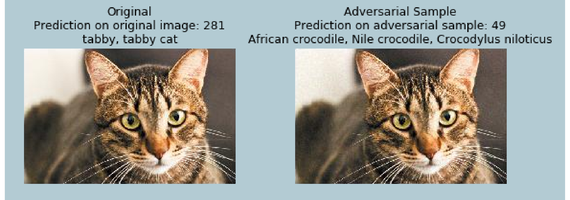

figure1

Goal

Understand how adversarial attacks affect the robustness of neural networks and analyze the adversarial samples in relation to the outputs by performing singular value decomposition.

Experimental Design

In this experiment, we're going to use the CIFAR10 and MNIST dataset and perform singular value decomposition on input images to find out what components are causing the misclassification.

SVD factorizes an input matrix M into three different matrices one of which contains singular values on its diagonal entries. Each singular value represents an independent component of the input matrix.

$$M = U\Sigma V^*$$

We then reconstruct the image using the first k singular values in Python and observe how the neural network reacts. Once we plot the accuracy out, we compare the results from adversarial inputs to normal inputs and see if there's a pattern differentiating between adversarial samples and benign samples.

Results

Using SVD analysis, we get the following results for reconstructing the image using the first k components:

No adversarial samples:.

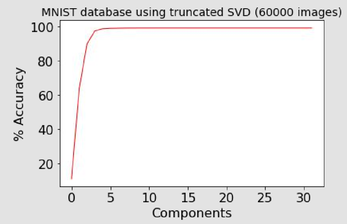

fig2

48 ImageNet validation images with ResNet-50 using PGD with 0.01 perturbation in Foolbox and PyTorch:

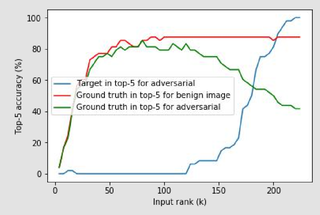

fig3

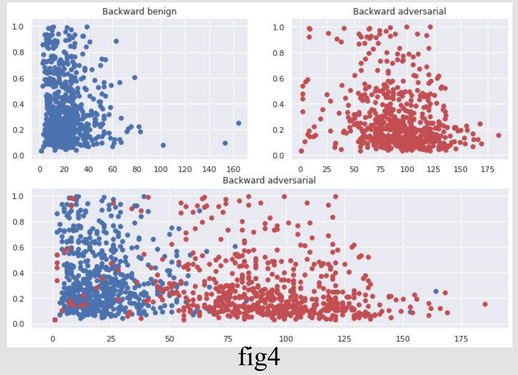

fig4

We can see in figure4 that the k-p point of adversarial samples occur at higher values than those of benign samples, but there are several false-positives.

Conclusion

In general, higher k-p points indicate a larger likelihood of detecting adversarial samples, but they aren't entirely effective since we noticed a few flase positives.

In future work, we plan to look at the transition of the predictions for each component for the adversarial samples and look at these "intermediate classes." Perhaps there is a pattern that occurs leading up to the prediction of the adversarial target.

References

[1]Mariusz Bojarski, Davide Del Testa, et al. End to end learning for self-driving cars. arXiv:1604.07316, 2016.

[2]Youcheng Sun, Min Wu, Wenjie Ruan, et al. Concolic testing for deep neural networks. arXiv:1805.00089, 2018.

[2]Andrew Ilyas, Shibani Santurkar, et al. Adversarial Examples are not Bugs, they are Features. arXiv:1905.02175, 2019.

Acknowledgements

We would like to express our deepest gratitude to Malhar Jere, Shezin Hossain, and Professor Farinaz Koushanfar for guiding us into the research field and providing us with invaluable resources. We would also like to thank Lisa Trahan, Alejandra Arguelles, and many others from the IDEA center for supporting us along the way.