A study and implementation of ABS (Analysis By Synthesis) model on Cifar-10

Abstract

Deep neural networks (DNNs) are sensitive to minimal adversarial perturbations that are almost imperceptible to humans but can switch the class prediction of DNNs to basically any desired target class. One key problem in finding successful defenses is the difficulty of reliably evaluating model robustness. In this project, our objective is to analyse the performance of the classification model called ABS(Analysis by Synthesis) on CIFAR-10 and perform an evaluation of our defense ABS model which is adversarially robust on many attacks specifically being Decision-based attacks, Gradient-based attacks and Score-based attacks for MNIST dataset [1].

A. Introduction

Recent work has demonstrated that that the existence of adversarial attacks may be an intrinsic weakness of deep learning models. Adversarial examples are the inputs that are almost indistinguishable from natural data and yet classified incorrectly by the network. In paper [2], authors have proposed the solution to overcome the problem of vulnerability against the adversarial attacks of the deep neural networks.They addressed this problem by studying the adversarial robustness of the neural networks through the lens of robust optimization. They have trained networks on MNIST [3] and CIFAR10 [4] that are robust to a wide range of adversarial attacks. Their best MNIST model achieves an accuracy of more than 89% against the strongest adversaries on test data. This was one of the great achievement but the authors in paper [1], raised a question "This is a great success, but does the model really learn more causal features to classify MNIST? So, authors showed that MNIST is unsolved from the point of adversarial robustness and have proposed a new robust classification model, "Analysis By Synthesis" (ABS). ABS model learn generative distribution i.e., considering the casual features of input and trying to generate and learn the given input using these features and then classify new inputs using Bayes Theorem. They have demonstrated this model robustness against different white-box and black-box adversarial attacks on MNIST dataset.
In this work, ABS model [5] is used which aims to analyse the robustness of the classification model on CIFAR-10 dataset in a white-box setting. The ABS model is trained using a variational autoencoder (VAE) [6] with Adam optimizer and KL-Divergence loss. A variational autoencoder is an autoencoder whose training is regularized to avoid overfitting and ensure that the latent space has good properties that enable generative process. In other words, VAE is trained to minimize the reconstruction error between the encoded-decoded data and the initial data. We have evaluated the trained ABS model on attacks like FGSM, I-FGSM, and PGD.
Our contribution in this work are as follows:

- Implement and train the ABS model on Cifar-10 dataset.
- Attack the trained model using L2 perturbation.
- Compare the results with a CNN model and original ABS model trained on MNIST dataset.

B. Background

Not all defenses suggested in the literature increase robustness over undefended neural networks except a few ones which are based on data augmentation with adversarials found by iterative projected gradient descent with random starting points(However overfits on $L\infty$). Also there are many defenses using the generative model to project the input or the hidden activations onto the (learned) manifold of "natural" inputs. This is particularly implemented in [7], [8] and [9], all of which project an image onto the manifold defined by a generator network G. The generated image is then classified by a discriminator in the usual way. A similar idea is used by [10] which uses an autoregressive probabilistic method(the output variable depends linearly on its own previous values and on an imperfectly predictable term) to learn the data manifold. Other ideas in similar directions include the use of denoising autoencoders which project or reject inputs depending on their distance to the data manifold. All of these proposed defenses have not been found effective. Reason being, many adversaries still look like normal data points to humans and mainly, the classifier on top of the projected image is as vulnerable to adversarial examples as before. Hence, for any data set with a considerable amount of variation there will almost always be a certain perturbation against which the classifier is vulnerable and which can be fooled by inducing the right inputs.

This is where the idea of increasing the adversarial robustness by studying the classifier arises. We study the adversarial robustness of the first adversarially robust neural network model on MNIST that is ABS which uses learned class-conditional data distributions. Here the input distribution is modeled within each class (instead of modeling a single distribution for the complete data), and by classifying a new sample according to the class under which it has the highest likelihood(Bayesian classifier) unlike any conventional CNN's.
$p(y|x) = p(x|y)p(y)p(x) \propto p(x|y)p(y)$

It combines several elements(Class-conditional distributions,Optimization-based inference and Classification and confidence) to simultaneously achieve high accuracy and robustness against adversarial perturbations.

We are implementing this ABS classifier on CIFAR 10 to study the domain adaptation of the model and further to check the adversarial robustness of the model against CIFAR 10 on L 2 norm measures. We also study the model's perception of the CIFAR 10's input manifold and data distribution. We check if the classifier learns the necessary features as claimed or is it too specific to the domain which in this case, concluding that the classifier is still vulnerable under certain circumstances.

C. Experimental Settings

We compare ABS against recent classification model simple DLA using the CIFAR 10 Dataset. Firstly, we train a CNN classifier named simple DLA(Deep Layer Aggregation) [11] on CIFAR 10. This is the current SOTA as it achieves 94.89% accuracy on CIFAR 10 outperforming the scores of ResNeXt29,MobileNetV2, ResNet101 etc. We train this classifier for 25 epochs and obtain an accuracy of 85.46 on test data. Later we train the ABS on the same and achieve an accuracy of 45.38 on 25 epochs and both measured on L2 norms. Later we tested both the models on Gradient-based attacks such as Fast Gradient Sign Method(FGSM), Iterative Fast Gradient Sign Method (IFGSM) and Projected Gradient Descent(PGD) for L2 norms alone. For each model and L2 norm, we show how the accuracy of the models decreases with increasing adversarial perturbation size $\epsilon = [0.1, 0.2, 0.3]$ and report metrics model's accuracy against bounded adversarial perturbations. Clean samples that are already misclassified are counted as adversarial with a perturbation size equal to 0. We report the measures in the table.

Clean Accuracy ABS = 45			
ϵ	FGSM	I-FGSM	PGD
0.10	0.257	0.22	0.20
0.20	0.2	0.19	0.14
0.30	0.166	0.14	0.10

Table 1. Epsilon vs Accuracy for FGSM, I-FGSM, and PGD attacks on ABS Model

Clean Accuracy ABS = 45			
ϵ	FGSM	I-FGSM	PGD
0.10	0.17	0.123	0.23
0.20	0.143	0.09	0.132
0.30	0.135	0.074	0.063

Table 2. Epsilon vs Accuracy for FGSM, I-FGSM, and PGD attacks on DLA Model

D. Results

Our robustness evaluation results of all models are reported in Table. As we see, ABS performs poorly on the CIFAR 10 dataset as a classifier although it is a Bayesian classifier built to have high accuracy and to be robust. In the Paper [1] its accuracy on MNIST was 99% but as we see, when trained on CIFAR 10, it's accuracy drops to 45%. On FGSM attack, having perturbation 0.1, the Accuracy of the DLA classifier drops to 0.177 whereas on ABS, it drops to 0.257. On IFGSM attack, the accuracy of the former, accounts to 0.143 and that of the latter to0.22. On PGD-L2, DLA has accuracy of 0.122 and ABS has accuracy of 0.20. Here, although the classification accuracy of ABS is poor, it definitely performs considerably better against attacks compared to the DLA. The drop in the accuracy is found to be huge in comparison with that of ABS. The corresponding accuracy-distortion graphs are reported in the figure 2 and figure 3.

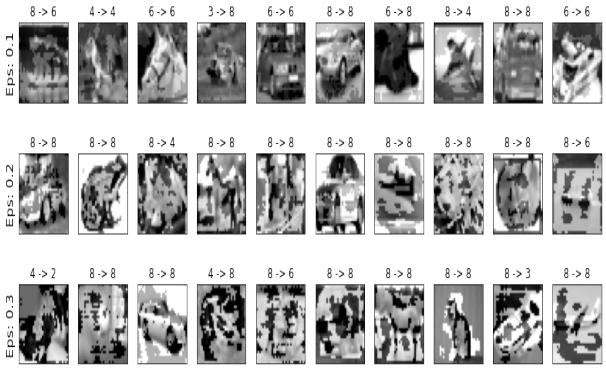

Figure 1. ABS: Predicted Label before and after the FGSM attack

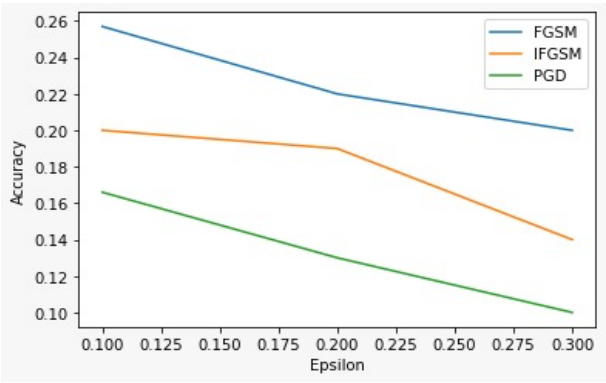

Figure 2. Accuracy distortion graph for ABS Model

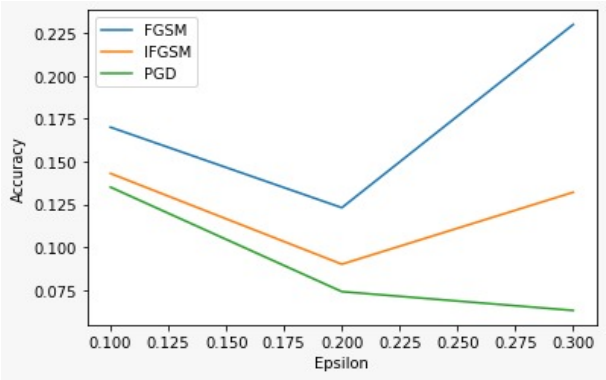

Figure 3. Accuracy distortion graph for DLA Model

E. Conclusion

Firstly, we need to address the issue found in the differences brought by accuracy scores of the ABS as a classifier on MNIST and CIFAR-10. There are many possible explanations to substantiate this inference. We should note that MNIST is a much simpler problem set than CIFAR-10, and can get 98% from a fully-connected (non-convolutional) network with very little effort. A very simple CNN with just one or two convolutional layers can likewise get to the same level of accuracy. But to perform with the same measure on a dataset that requires visual recognition of rich representations that span levels from low to high, scales from small to large, and resolutions from fine to coarse, it becomes difficult. Even with the depth of features in a convolutional network, a layer in isolation is not enough to learn the classification of the object under scrutiny. Also, there are many dimensions that contribute to the learning of a classifier (eg. Color RGB channelization that helps in better feature maps representation). Hence, ABS although claims to learn the casual features of input generatively, it does not perform according to standards when the domain changes. So, this could be a discussion open to research and development as there certainly is a room for improvement when we speak of its architecture. What is more interesting is that, although it is a poor classifier on CIFAR-10, compared to MNIST, it exhibits a better performance when attacked. Compared to DLA classifier or any clean CNN architecture, ABS certainly performs better and is robust to perturbations caused even if it is not that applaudable. Hence, we conclude, that maybe if the ABS could learn the domain adaptation, it can certainly be used as a robust classifier that could make significant changes in the Machine Learning paradigms pertaining to security.

References

[1] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist, 2018. 1, 2

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019. 1

[3] Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005. 1

[4] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012. 1

[5] Bethgelab. Bethgelab/analysisbysynthesis: Adversarially robust neural network on mnist. 1

[6] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. 1

[7] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models, 2018. 1

[8] Shiwei Shen, Guoqing Jin, Ke Gao, and Yongdong Zhang. Ape-gan: Adversarial perturbation elimination with gan, 2017. 1

[9] Ajil Jalal, Andrew Ilyas, Constantinos Daskalakis, and Alexandros G. Dimakis. The robust manifold defense: Adversarial training using generative models, 2019. 1

[10] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples, 2018. 1

[11] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation, 2019. 2