Exploring Adversarial Robustness: Strategies for Improved Performance Across Diverse Datasets

Yuwei Liu, 1004735824
yuw.liu@mail.utoronto.ca
Liang Chen, 1005735126
liangliang.chen@mail.utoronto.ca
Wilson Gao, 1005053987
wilson.gao@mail.utoronto.ca

Abstract

Improving the adversarial robustness of deep neural networks has been an active research topic in recent years. Adversarial training and generative models have shown promise in improving the robustness of deep neural networks against adversarial attacks. However, there is still much work to be done in this area, especially when dealing with novel datasets, such as art. In this paper, we investigate the effectiveness of different methods for improving adversarial robustness, image transformation and new network architectures, on different on ResNet and different datasets We evaluate our methods using the Fast Gradient Sign Method (FGSM), which is commonly used in evaluating adversarial robustness. We aim to explore the performance of these methods and analyze the our outputs.

1 Introduction

Deep neural networks have shown remarkable performance in various computer vision tasks,including image classification. However, recent research has shown that these models are vulnerable to adversarial attacks, which can lead to misclassification with severe consequences. As machine learning develops and may be widely used in industry in the future, the safety and robustness of the model will become more important. Therefore, improving the adversarial robustness of these models has become an active research topic.

Since the adversarial examples are generated based on the representation of the image space, the main idea is to make the model more robustness to those noisy perturbation. Once method could be adding a part to denoise the input images for the model, another can be having some transformation to the images after the attack to bring more randomness of the input space. Hence in this paper, we tried to implement these ideas on CIFAR-10. We also exlpored how is the performance of the model under adversarial attack on large art dataset WikiArtt.

2 Related Work

In recent years, there has been significant interest in improving the robustness of image classification models against adversarial attacks. A number of methods have been proposed to mitigate the impact of such attacks on deep neural networks.

One common approach to improve adversarial robustness is adversarial training. This technique involves augmenting the training data with adversarial examples generated during the training process. Goodfellow et al. (2014) proposed the fast gradient sign method (FGSM), which is a simple and effective method for generating adversarial examples. Subsequent work has extended this method,

such as the projected gradient descent (PGD) method proposed by Madry et al. (2018), which is currently considered one of the strongest attacks for evaluating the robustness of a model.

Recent work has also explored the use of generative models to improve adversarial robustness. Samangouei et al. (2018) proposed the use of generative adversarial networks (GANs) to learn a representation of the input space that is robust to adversarial attacks. Other work has focused on using variational autoencoders (VAEs) to generate "clean" reconstructions of adversarial examples (e.g., Song et al., 2018).

While these approaches have shown promise in improving the robustness of deep neural networks, there is still much work to be done in this area. In our project, we mainly consider two papers "Countering Adversarial Image using Input Transformation" by Guo et al. (2018) and "Feature Denoising for Improving Adversarial Robustness" by Xie et al. (2019), which both apply some process for the data before training the model, and applying both ImageNet dataset latter. We want to explore the performance of two methods together in a same size ResNet model and the dataset. We evaluated the performance of these methods using commonly used attack method Fast Gradient Sign Method (FGSM).

3 Comprehensive Method

3.1 Image transformation

We investigated 5 different transformation methods:

Image cropping and rescaling: This transformation has the effect of altering the spatial positioning of the adversarial perturbation, which may help to decrease the influence of attack noise.

Bit-depth reduction: This type of quantization is effective in mitigating small adversarial perturbations in pixel values, effectively removing them from the image.

JPEG compression: This will help remove small perturbations from the adversarial images.

Total variance denoising: This principle is reducing the variance of the signal subject to it being a close match to the original signal; in other word, $\arg_u \min ||u|| + \frac{\lambda}{2} \int (f(x) - u(x))^2 dx$ where λ is a positive parameter and f is the noisy image, we want to estimate the denoised image u. And the split-Bregman optimization method has been implemented to solve the objective function.

Figure 1: Image transformations.

3.2 Feature denoising

A new network architectures can increase adversarial robustness by adding denoising blocks at intermediate layers of the convolutional network improves the robustness of image classification models against adversarial attacks. The denoising blocks are trained jointly with all layers of the network in an end-to-end manner using adversarial training. Three different instantiations of the denoising operation in our denoising blocks are Gaussian filter, mean filter, and median filter.

Gaussian filter: Remove high-frequency noise from the input image, input image becomes more locally homogeneous and robust to small perturbations .

Mean filter: Average pooling with a stride of 1, reduce noise but also smooth structures.

Median filter: Calculate the median over a local region $\sigma(i)$, which is good at removing salt-and-pepper noise and outliers.

3.3 Model training

The model is trained by the above two different methods and then combined method. When training with the image transformation method, we trained the model with transformed data, and when implemented the attack in the testing process, the adversarial images were transformed again before feeding the data into the trained model. When training with the feature denoising, we just added the model with a denosing block and tested the model after training.

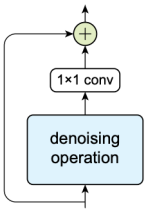

Figure 2: Feature denoising block.

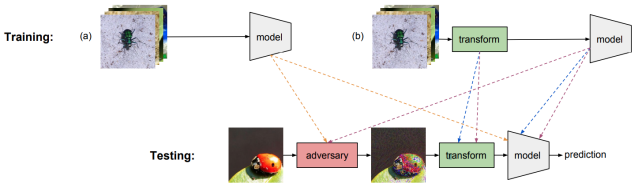

Figure 3: Image transformation training.

3.4 Adversarial Attack

We used the Fast Gradient Sign Method (FGSM) method to implement adversarial attack:

$$x_{adv} = x + \epsilon * sign(\bigtriangledown_x J(\theta, x, y))$$

where x_{adv} = adversarial image, x = original input image, y = original input label, ϵ = multiplier to ensure the perturbations are small, θ = model parameters, J = loss function. The adversarial attack strength can be adjusted by ϵ value, for example, $\epsilon = 0$ implies 0 adversarial noise added on image.

4 CIFAR-10 Experiment

4.1 Experiment output

We first conducted the comparisons between different type of image transformations and filters respectively. Related graphs has been plotted below. The x-axis represents the ϵ value, the larger ϵ gives large attack strength; the y-axis represents the related accuracies of the test set.

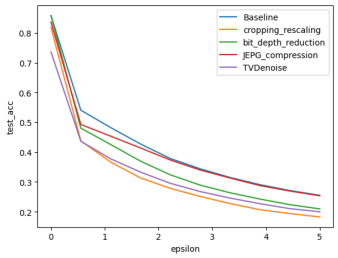

Figure 4: Compare transforms.

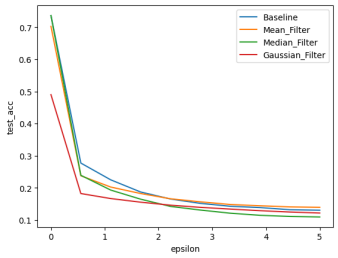

Figure 5: Compare filters.

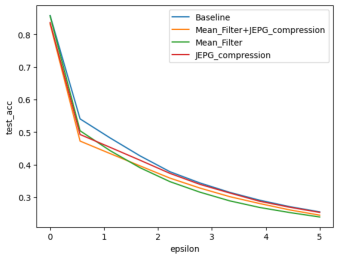

Figure 6: Combination.

Then we picked the best transformation method and the best filter, compared the combination of two methods, as well as the only best filter, only best transformation and the baseline (no method has been add) performances.

During experiments, different models were trained for 20 epochs, the filters were set with kernel size 3, the datasets were preprocessed with random HorizontalFlip, and the specific image transformation methods were added separately to the related model.

4.2 Discussion on the results

The results of this experiment are not very ideal. It can be seen from the figure that the combination of the two methods/the two methods does not improve the robustness of the model.

In Figure 4, JPEG compression has the best performance over several image transformation methods. However, all methods has less accuracies than the baseline model. The reason for the phenomeneon may be that during the training process, due to the loss of information in the transformation part of the images, the generalization ability of the models themselves are lower than the baseline model.

In Figure 5, the mean filter has the best performance over other filters. But the models with filters are also less accurate than the baseline model. The reason for the result may also be that the generalization ability of the model with a denoising block is lower than that of the baseline model.

Although the accuracy rates of the models mentioned in 1 and 2 are lower than the baseline model, it can be seen that as the attack strength increases, their accuracy rates decrease slower than the baseline model. Therefore, the next step of research can consider comparing the performance of different methods for models after they are trained with the same level of generalization ability, or comparing the derivative values of the related curves.

In Figure 6, we can see that the JPEG compression in the combination method improves the performance of the mean filter to a certain extent.

5 WikiArt Experiment

The WikiArt dataset is a large-scale, publicly available dataset of artworks of various art movements and genres. It contains more than 80000 images from 27 different art genres, including Impressionism, Pop Art, Realism, and Romanticism. Each image in the dataset is accompanied with a label, corresponding to its genre. Our goal in using a novel art dataset is to compare our model's performance on a different dataset, besides the well-known and commonly used CIFAR-10 dataset.

To be able to work on a large dataset, our training set was consisted of sampling one-fourth of the entire data. In order to unify the shapes of our images, we first cropped the image into a 512 by 512 pixel square, and then resized it to a 32 by 32 pixel image. During the training process, the model still needed large memory space and time cost and took roughly 30 minutes per epoch.

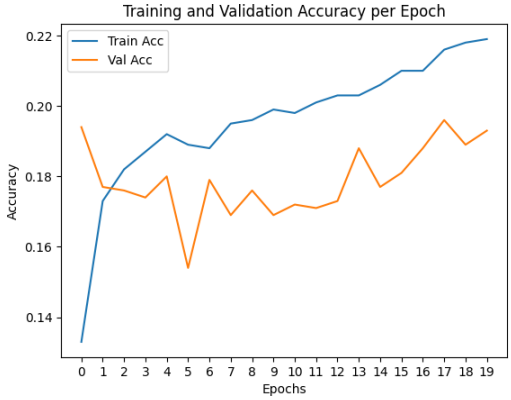

Figure 7: Training on WikiArt Dataset

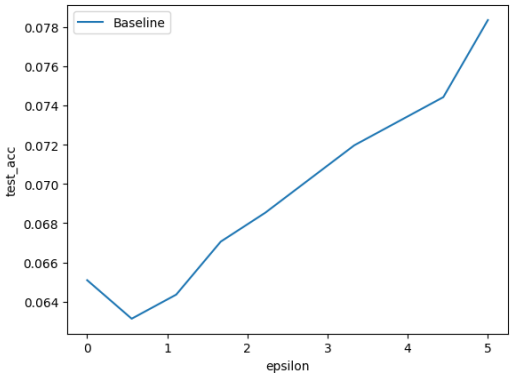

Figure 8: Testing on WikiArt Dataset

Both training and validation accuracy when using the WikiArt dataset was significantly lower than that of the CIFAR-10 dataset. This is to be expected as even in the same art genre, the varying colors used or stylistic choices that each artist makes can all add up to vastly different looking images. Besides, two images can result in almost no resemblance at all, which could hurt our model during taining. And the accuracy of the model on the test set was extremely low, though still better than if the model were randomly guessing, as 1/27 is approximately 0.037. In Figure 8, we can see that the FGSM attack ended up improving the accuracy of the model on the test dataset when the strength

was increased, which is unexpected. It is possible that the FGSM attack removed some confusing aspect about the images, leading to an increase in accuracy. However, further experimentation is needed to rule out the possibility that these observations are a result of random chance, as the values are extremely small.

6 Summary

In this project we explored advsarial attack and different defense strategies. We have explored the spatial characteristics of image data and different noise reduction methods. Our experiments also have a lot of ways for improvement, such as training and comparing models more rigorously, or using devices with more powerful computing power to further explore features from large art datasets, also using other adversarial attack algorithms to compare attack effects. Adversarial robustness is an interesting and broad topic, when we try toc help the model to be more robust, we also got a deeper understanding of the principles behind the model.

References

[1] Danielczuk, M., Matl, M., Gupta, S., Li, A., Lee, A., Mahler, J., & Goldberg, K. (2019). Segmenting Unknown 3D Objects from Real Depth Images using Mask R-CNN Trained on Synthetic Data. In Proc. IEEE Int. Conf. Robotics and Automation (ICRA), pp. 7560-7566.

[2] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations. https://arxiv.org/abs/1412.6572

[3] Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018, January 25). Countering adversarial images using input transformations. arXiv.org. Retrieved February 16, 2023, from https://arxiv.org/abs/1711.00117

[4] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations. https://arxiv.org/abs/1706.06083

[5] Samangouei, P., Kabkab, M., & Chellappa, R. (2018). Defense-GAN: Protecting classifiers against adversarial attacks using generative models. International Conference on Learning Representations. https://arxiv.org/abs/1805.06605

[6] Song, J., Kim, T., Nowozin, S., Ermon, S., & Kushman, N. (2018). Adversarial examples for generative models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7224-7233). https://arxiv.org/abs/1702.06832

[7] Xie, C., Wu, Y., Maaten, L. van der, Yuille, A., & He, K. (2019, March 25). Feature denoising for improving adversarial robustness. Retrieved February 16, 2023, from https://arxiv.org/abs/1812.03411

Appendix

Contributions

Liang Chen: Construct proposal and report, experiment on CIFAR-10 Dataset
Wilson Gao: Construct proposal and report, experiment on WikiArt Dataset
Yuwei Liu: Construct proposal and report, experiment on CIFAR-10 Dataset, construct model structure and adversarial attack function

Source code

https://github.com/oolonglilfox/CSC2516Project