Adversarial Noise Extraction

Jaisidh Singh
singh.118@iitj.ac.in

Nakul Sharma
sharma.86@iitj.ac.in

Ishaan Shrivastava
shrivastava.9@iitj.ac.in

Abstract

The problem of denoising adversarially perturbed images by predicting the injected noise is an important task in machine learning. In this project, orthogonal to the prior work in the literature, we propose an approach to predict the adversarial noise added to images and use it to denoise the attacked samples. Our method leverages a fully convolutional neural network (CNN) architecture that is trained on a dataset of clean and adversarially perturbed images to learn the underlying distribution of the attack noises. We evaluate our approach on the CIFAR-10 benchmark dataset and show that our proposed approach achieves promising results on the denoising task. Using adjustments on popular loss functions, our results suggest that predicting the adversarial noise can be an effective strategy for denoising attacked images.

1 Introduction

In recent years, the proliferation of deep learning models has led to significant advances in various computer vision tasks, such as object detection, recognition, and segmentation. However, these models are vulnerable to adversarial attacks, where an attacker perturbs the input data with imperceptible changes to fool the model into making incorrect predictions. Adversarial attacks can pose a significant threat to security and privacy applications, as they can be used to bypass security systems and compromise sensitive information. One of the approaches to counter adversarial attacks is to denoise the attacked samples by removing the adversarial noise added to them. Several methods have been proposed for denoising attacked images, such as filtering-based approaches, gradient-based approaches, and neural network-based approaches. However, these methods often require access to the model architecture or its gradients, which may not be feasible in real-world scenarios.

In this project, we propose a deep learning-based approach to denoise attacked images by predicting the adversarial noise that was added to them. Our approach leverages a convolutional neural network (CNN) architecture that is trained on a dataset of clean and adversarially attacked images to learn the underlying

distribution of the noise added to them. We then use the learned model to predict the noise added to new attacked images and use it to denoise them. Our approach does not require access to the model architecture or gradients, making it more practical for real-world scenarios. The primary motivation for predicting the noise is the ability of CNNs to model noise. This can be seen in recent works as well, like diffusion. As compared to diffusion-based noise removal, adversarial denoising is different, as we are required to predict noise which follows a specific function. Thus, CNNs can be leveraged here for adversarial noise predictiong and denoising.

2 Methodology

2.1 Data Generation

To perform adversarial denoising , we generate attacked samples on two datasets, namely CIFAR10 and CIFAR100. We use 3 popular attacks, namely, Projected Gradient Descent (PGD) [3], Fast Gradient Sign Method (FGSM) [1], and Jitter [4]. Particularly, we use a ResNet-18 [2] CNN model and train it on the outlined datasets. Next, using this model, we construct adversarial samples on the test set of each dataset, as the data to be used in the denoising, described below.

2.2 Adversarial Denoising

Using the inspiration from diffusion models, mentioned above, we construct a UNet. This progressively downsamples the input adversarial image, and then upsamples the learnt features to predict the noise. The learning of this UNet-based denoising autoencoder is heavily contingent on our loss functions which utilize the difference of the predicted noise with the actual adversarial noise. We impose a greater penalty on these differences using higher order mean exponential error as one of our loss functions, which are described as follows.

$$\mathcal{L} = \mathcal{L}_{\text{classification}} + \mathcal{L}_{\text{denoising}} \tag{1}$$

$$\mathcal{L}_{\text{denoising}} = \frac{|(x_{\text{adv}} - x_{\text{clean}}) - f(x_{\text{adv}}, \theta)|^n}{n} \tag{2}$$

which is another way to represent the standard minimization of $(x_{\text{clean}} - x_{\text{denoised}})^2$. We present the loss in this format to show our intuition of penalizing errors in estimating the adversarial noise using n. This is because higher values of n penalize errors greater than lower values, which we leverage. The $\mathcal{L}_{\text{classification}}$ loss is a simply *Cross Entropy Loss* between the prediction of the ResNet-18 model on the denoised image and the image label. Note that the weights ResNet-18 classifier is kept frozen, which only serves to guides the denoiser.

Type of Attack	Clean Sample Accuracy	Attacked Sample Accuracy
PGD	92.4	16.3
FGSM	92.4	32.9
Jitter	92.4	2.9

Table 1: Results of data generation, where we report clean sample accuracies and the attacked sample accuracies for the CIFAR10 dataset.

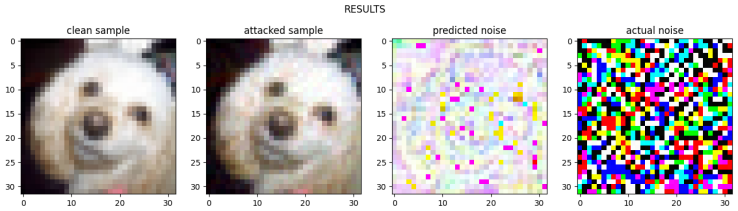

Figure 1: Results of noise prediction on a PGD attacked sample (an image of a fox) from the CIFAR10 test set.

2.3 Component for XAI

Since neural networks and adversarial perturbations are virtually black-box, i.e., are difficult to comprehend by humans, we incorporate the popular technique of visualizing what the classifier looks at in an image, for assigning it a particular label. This technique is Grad-CAM [5], and presents a layer of transparency which the user can query for greater understanding of the process. Thus, we present the activation maps from the first layer of the ResNet-18 classifier (as deeper layers make the input vanish, yielding no map no for images of resolution 32x32) for the adversarial image and the denoised image to provide explanability.

3 Results

Type of Attack	Clean Sample Accuracy	Attacked Sample Accuracy
PGD	67.4	10.2
FGSM	67.4	37.6
Jitter	67.4	0.2

Table 2: Results of data generation, where we report clean sample accuracies and the attacked sample accuracies for the CIFAR10 dataset.

We present the results of our experiments, i.e., for the generation of data and the denoising of adversarial images.

Data Generation We train a ResNet-18 classifier on the clean samples of CIFAR10 and CIFAR100, achieving a test accuracy of 92.4 % and 67.4 %.

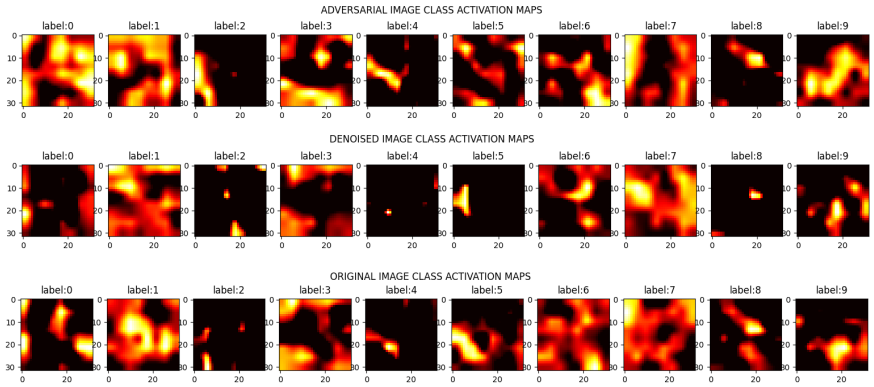

Figure 2: Results of the explainable component of our project, where we present the regions of focus or attention of the ResNet-18 classifier in the adversarial setting for the CIFAR10 dataset.

Upon attacking the these models using the PGD, FSGM and Jitter attacks, we obtain lower classification accuracies as presented in Table 1.

Denoising Using Adversarial Noise Prediction To mitigate the adversarial attacks performed in the previous experiment, we conduct adversarial denoising by predicting the noise in the attacked sample. We can see that the UNet learns to model the adversarial noise well, as can be seen by the results of adversarial denoising for the CIFAR10 dataset in Table 2 and for CIFAR100 in Table 3. We further present our results on a sample image in Figure 1.

4 Ablation Study

Dataset	Accuracy with $n = 1$	Accuracy with $n = 2$	Accuracy with $n = 3$	Accuracy $n = 5$
CIFAR10	90.4	90.5	91.0	91.4
CIFAR100	61.0	61.3	61.9	62.0

Table 3: Results showing the effect of n, which aligns with our the mathematical intuition imposing greater penalties on incorrect predictions as the power of the function, i.e., the value of n goes up.

We explore the effects of our the components of our algorithm in this section, namely the effect of the penalty incurred with the value of n in the denoising loss, on the results of the denoising process.

Effect of n We present the results of the effect of n in the denoising loss

function in Table 3. As can be seen, there is a steady, statistically significant increase in the accuracy of denoised samples, which is comparable to the original accuracy of the ResNet-18 classifier on the clean samples. Thus, the choice of the modification from a naive Mean Squared Error loss to a hyper-parameter governed denoising loss is both intuitively and empirically sound.

5 Discussion

We show the Grad-CAM class activation maps in Figure 2. There are noticeable differences between the class activations of the adversarial images from the other two, however, the original and denoised activation maps have several similarities. This verifies that our approach is able to successfully defend the victim classifier against attacks.

However, it is tough to fully appreciate the efficacy of out approach as we use the first layer activation maps, which do not carry enough high level spatial information in order to show stark differences between the activation maps of the denoised and the adversarial images. Our approach is expected to exhibit significant visual improvements in the class activation maps on datasets of larger images such as ImageNet.

6 Conclusion

We present a framework for denoising adversarially attacked images using an autoencoder which leverages the power of CNNs to model noise governed by a mathematical function. This results in robust denoising, which we achieve via controlled penalties in the loss function of the training process. Further, we incorporate an explainable component into our framework, which shows the activation maps of the adversarial image, and the denoised image, supporting our inferences.

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[3] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[4] Leo Schwinn, René Raab, An Nguyen, Dario Zanca, and Bjoern Eskofier. Exploring misclassifications of robust neural networks to enhance adversarial attacks. *Applied Intelligence*, pages 1–17, 2023.

[5] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from

deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019.