Enhancing Adversarial Robustness through a Purification Network

Mihir Agarwal
Indian Institute of Technology
Gandhinagar, India
agarwalmihir@iitgn.ac.in

Rachit Verma
Indian Institute of Technology
Gandhinagar, India
vermarachit@iitgn.ac.in

Zaqi Momin
Indian Institute of Technology
Gandhinagar, India
zaqimomin@iitgn.ac.in

ABSTRACT

Recent advancements in deep learning have led to significant progress in fields like computer vision and natural language processing, positioning trained classifiers as pivotal components in security-sensitive applications. However, the susceptibility of these classifiers to adversarial attacks, where inputs are manipulated to elicit incorrect model responses, poses a severe challenge, particularly in the domain of computer vision. Traditional adversarial training methods, while effective, often fail to generalize across different types of attacks and are computationally expensive. This report introduces a novel architecture that combines Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to enhance model robustness against a broad spectrum of adversarial inputs. Our architecture leverages the discriminator gradient to guide input images back towards the latent distribution manifold, aiming for a purifying effect that generalizes well across various attacks. We detail the development and implementation of our VAE-GAN-based model, provide a comparative analysis with existing approaches like Defense GAN, and demonstrate its efficacy in maintaining high classification accuracy under adversarial conditions. The report culminates in an experimental section that not only validates the model's performance but also outlines potential future enhancements to further improve classifier resilience.

KEYWORDS

Adversarial Robustness, Purification Network, VAE-GAN

ACM Reference Format:
Mihir Agarwal, Rachit Verma, and Zaqi Momin. 2024. Enhancing Adversarial Robustness through a Purification Network. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent advancements in computer vision and natural language processing place trained classifiers at the forefront of security-sensitive systems. Examples include vision applications in autonomous vehicles, facial recognition, and malware detection. These advancements illustrate the increasing importance of security considerations in machine learning. Specifically, the ability to withstand adversarial inputs is becoming a critical design objective. While trained models are highly effective at classifying benign inputs, recent research demonstrates that adversaries can often manipulate inputs to produce incorrect outputs from the model. Computer vision, in particular, presents a notable challenge: even slight modifications that are imperceptible to the human eye to the images can deceive state-of-the-art classifiers with high confidence.

An important approach to solving this problem is using adversarial training to make the model more resistant to these types of attacks. People in the past have developed a variety of attacks like PGD, FGSM, and other black box attacks to train the classifier to better deal with these kinds of perturbations. During training, the model is also trained with these adversarially altered examples so that when the model encounters them during test time, its weights are theoretically more robust to these perturbations, and hence, the model can classify them correctly.

A major disadvantage of this kind of approach is that the models which are trained to resist against one particular attack and not necessarily resistant to other types of attacks. Moreover, adversarial training is expensive, and every time we change a model in a situation, the new model also needs to be trained adversarially to serve its purpose. As a result, the overall computational costs are high while not being necessarily robust to any attacks by an adversary.

To solve this issue, new kinds of models are being introduced that 'purify' the samples before inputting these images to the target classifier. The advantages of this approach is two fold. Firstly, the purifying model is classifier agnostic; it can be used when classifier which is trained on a similar distribution. Secondly, these have shown to be resistant to a wider range of attacks than during adversarial training.

In this report, we propose our own architecture, which aims to use the discriminator gradient to drive the input image towards the latent distribution manifold. We also present the motivation for this architecture and conclude the report with a section on experiments.

2 ADVERSARIAL ROBUSTNESS

Adversarial robustness can be defined as the capability of a machine learning model to maintain its performance and accuracy even when subjected to intentionally crafted perturbations or attacks designed to deceive the model. In the context of image classification, adversarial robustness refers to a model's ability to correctly classify images even when they have been modified with imperceptible alterations that are specifically crafted to mislead the model's predictions.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

To illustrate this concept, we can utilize the pre-trained ResNet50 model available in the PyTorch framework to classify an image of a pig. Given that ImageNet, the dataset on which ResNet50 is trained, has a class corresponding to both "hog" and "pig", the correct label for the image of the pig would be formally as Hog shown in the figure1.

```
import json
with open("imagenet_class_index.json") as f:
    imagenet_classes = {int(i):x[1] for i,x in json.load(f).items()}
print(imagenet_classes[pred.max(dim=1)[1].item()])
```

```
hog
```

Figure 1: Classification of image on Resnet50 as "Hog"

2.1 Importance of Adversarial Robustness

Adversarially robust networks play a important role in the domain of machine learning due to several critical reasons, primarily aimed at bolstering the security and dependability of models when operating under adversarial conditions. These networks are intricately engineered to withstand or rectify the detrimental impacts of adversarial attacks, which are manipulations of input data designed to mislead models. For instance, in domains such as finance, healthcare, and autonomous vehicles, ensuring the integrity and reliability of predictions made by machine learning models is of much importance. In the context of medical diagnosis, for example, a model erroneously classifying malignant tumors as benign owing to slight perturbations in input data could precipitate fatal consequences.

The significance of adversarially robust networks thus lies in their ability to mitigate the risks associated with such adversarial attacks, thereby This makes machine learning systems more dependable and trustworthy across different areas like finance, healthcare, and self-driving cars.

2.2 FMNIST Dataset

The Fashion-MNIST (FMNIST) dataset comprises 70,000 grayscale images categorized into 10 distinct fashion classes. It serves as a more intricate alternative to the conventional MNIST dataset, utilized for handwritten digit recognition tasks. Each image within the FMNIST dataset is sized at 28x28 pixels, employed for evaluating the performance of machine learning algorithms in image classification tasks. The images from FMNIST classes are shown in the figure3.

For the purpose of constructing an adversarial dataset, a subset comprising 10,000 images from the training dataset and 2,000 images from the testing dataset is selected. This subset serves as the training and testing sets respectively. Subsequently, these datasets undergo a defined process, as elucidated in subsequent presentation slides, to generate the adversarial dataset. This process involves specific manipulations or perturbations aimed at evaluating the robustness of machine learning models against adversarial attacks in the context of image classification.

2.3 Adversarial Data Generation

In the context of machine learning, particularly in the realm of adversarial attacks, we have introduce some formal notations. The

Figure 2: FMNIST Classes

model, or hypothesis function, denoted as $h_\theta : X \to \mathbb{R}^k$, which maps the input space X (e.g., a three-dimensional tensor representing images) to the output space, a k-dimensional vector. Here, k represents the number of classes being predicted. The parameter vector θ encompasses all the parameters defining this model, including convolutional filters, fully-connected layer weight matrices, biases, etc. These parameters (θ) are typically optimized during the training phase of a neural network.

Furthermore, we define a loss function, denoted as $\ell : \mathbb{R}^k \times \mathbb{Z}^+ \to \mathbb{R}^+$, which maps the model predictions and true labels to a non-negative number. Specifically, for an input $x \in X$ and the true class $y \in \mathbb{Z}$, the notation $\ell(h_\theta(x), y)$ signifies the loss incurred by the classifier in its predictions for x, assuming the true class is y.

A commonly employed loss function in deep learning is the cross-entropy loss (or softmax loss), defined as:

$$\ell(h_\theta(x), y) = \log\left(\sum_{j=1}^{k} \exp(h_\theta(x)_j)\right) - h_\theta(x)_y$$

where $h_\theta(x)_j$ denotes the jth element of the vector $h_\theta(x)$.

The goal of training a classifier typically involves minimizing the average loss over a training set $\{(x_i, y_i)\}_{i=1}^m$, which can be formulated as the optimization problem:

$$\text{minimize} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$$

This optimization problem is commonly addressed using gradient descent or its variants. The key component here is the gradient $\nabla_\theta \ell(h_\theta(x_i), y_i)$, which computes how adjustments to the parameters θ affect the loss function.

Now, to generate an adversarial example, we aim to manipulate the input x such that the classifier misclassifies it. This involves solving the optimization problem:

$$\text{maximize} \ell(h_\theta(x + \delta), y)$$

where δ represents a perturbation to the input x, and Δ defines an allowable set of perturbations. In practice, we typically restrict δ within an ℓ_∞ ball, defined as:

$$\Delta = \{\delta : \|\delta\|_\infty \leq \epsilon\}$$

This ensures that the perturbation remains close to the original input x. Techniques such as projected gradient descent (PGD) are often employed to iteratively adjust δ while ensuring it stays within the ℓ_∞ ball.

By maximizing the loss function over δ, we can generate an adversarial example that appears similar to the original input but is misclassified by the classifier. This phenomenon highlights the vulnerability of machine learning models to adversarial attacks, despite the imperceptibility of the introduced perturbations.

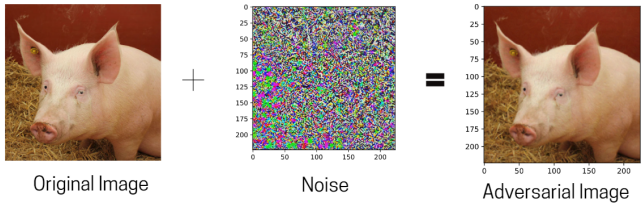

Figure 3: Showing the Adversarial attack on Pig Image

2.4 Targeted Attacks

In the context of adversarial attacks in deep learning, the technique of crafting adversarial examples to mislead a classifier extends beyond altering an image's classification to resemble a different class. This form of attack, known as a "targeted attack," involves maximizing the loss of the correct class while simultaneously minimizing the loss of a target class. The optimization problem associated with targeted attacks can be expressed as:

$$\text{maximize } \delta \in \Delta \left(\ell(h_\theta(x+\delta), y) - \ell(h_\theta(x+\delta), y_{\text{target}})\right)$$

where Δ represents the allowable set of perturbations, y_{target} denotes the target class, and ℓ represents the loss function. This optimization problem seeks to maximize the difference between the loss of the correct class and the loss of the target class when applied to the perturbed input. Simplifying this expression yields:

$$\text{maximize } \delta \in \Delta \left(h_\theta(x+\delta)_{y_{\text{target}}} - h_\theta(x+\delta)_y\right)$$

This process demonstrates the ability to manipulate the classifier's decision to classify the image of a pig as an airliner. Despite the perturbations applied to the image, it still resembles a typical pig visually.

The ease with which such attacks can be executed prompts considerations about the development of deep learning classifiers that are resilient to adversarial attacks. While progress has been made in this regard, achieving robustness against adversarial attacks remains an ongoing challenge in the field of deep learning.

2.5 Adversarial FMNIST

In a similar vein, the Adversarial Fashion-MNIST (FMNIST) dataset was generated by maximizing the loss function with respect to our model. This process entails crafting adversarial examples from the original FMNIST dataset with the aim of perturbing the inputs in such a way that the model's loss is maximized.

Formally, given an input image x from the FMNIST dataset and its corresponding true label y, the objective is to find a perturbation δ that maximizes the loss function $\ell(h_\theta(x+\delta), y)$, where h_θ represents the hypothesis function mapping the input space to the output space.

The optimization problem associated with generating adversarial examples from the FMNIST dataset can be expressed as:

$$\text{maximize } \delta \in \Delta \left(\ell(h_\theta(x+\delta), y)\right)$$

where Δ represents the allowable set of perturbations. This process aims to identify perturbations that lead to misclassification or higher loss values, thereby creating adversarial examples.

This formal representation encapsulates the methodology employed to construct the Adversarial FMNIST dataset, which serves as a resource for evaluating the robustness of machine learning models against adversarial attacks in the context of image classification.

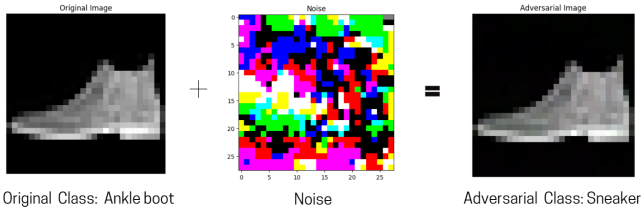

Figure 4: Showing the Adversarial attack on FMNIST Dataset Image

3 VAE GANS

The essence of VAE-GANs lies in their ability to leverage learned representations to measure similarities in data space more effectively than traditional methods. By integrating a VAE as an encoder and a GAN as a generative model, VAE-GANs introduce a novel approach to encoding, generating, and comparing dataset samples simultaneously. This integration allows the VAE to encode input data into a latent representation and reconstruct it using a learned similarity metric from the GAN discriminator rather than relying on simplistic element-wise error measures.

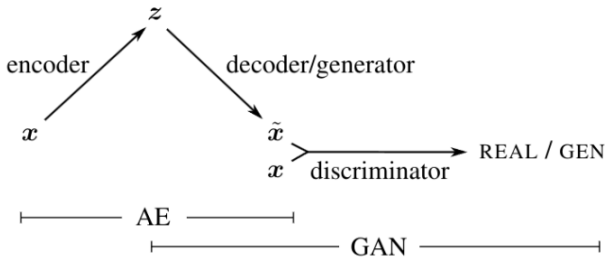

Figure 5: Overall training pipeline for a VAE GAN

This shift enables the VAE to capture data distribution more accurately, offering invariance to transformations like translation and enhancing visual fidelity in image generation tasks. By utilizing a learned similarity measure from the GAN discriminator, VAE-GANs produce sharper and more realistic images compared to traditional VAE models that rely on simplistic error metrics. This forms the most essential component of our architecture. We pass the image through a VAE since VAEs are shown to have inherent adversarial robustness properties. Moreover, using a VAE GAN was important because we also wanted to learn a discriminator that could identify out-of-distribution samples. We will talk about this in more detail in further sections.

4 DEFENSE GAN

Defense GAN was a major inspiration for our mode. They operate by utilizing a generator network to reconstruct clean, unperturbed images from adversarially perturbed ones. The key idea is that the generator network, trained on legitimate data, can produce reconstructions that are closer to the original, clean images than to the adversarially perturbed ones. This closeness is measured using Mean Squared Error (MSE) as a metric, where a lower MSE indicates a higher similarity to the original image.

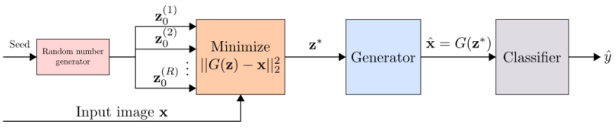

Figure 6: Overview of the Defense GAN architecture

As is visible from the figure above, the input image is run through the generator a number of times(while backpropagating the gradient) to project the image back to the latent space. The idea is that adversarial samples are viewed as lying out of the latent distribution. They are constructed in the first place by taking a sample and making perturbations to push it out of the latent space. The most optimal direction is shown to be perpendicular to the latent space[5]. Hence, by minimizing the the distance between the given sample and the generated sample, the models hopes to learn the projection of the input point into the data manifold.

5 OUR ARCHITECTURE

Defense GAN was a major inspiration for our mode. They operate by utilizing a generator network to reconstruct clean, unperturbed images from adversarially perturbed ones. The key idea is that the generator network, trained on legitimate data, can produce reconstructions that are closer to the original, clean images than to the adversarially perturbed ones. This closeness is measured using Mean Squared Error (MSE) as a metric, where a lower MSE indicates a higher similarity to the original image.

As is visible from the figure above, the input image is run through the generator a number of times(while backpropagating the gradient) to project the image back to the latent space. The idea is that adversarial samples are viewed as lying out of the latent distribution. They are constructed in the first place by taking a sample and making perturbations to push it out of the latent space. The most optimal direction is shown to be perpendicular to the latent space. Hence, by minimizing the the distance between the given sample and the generated sample, the models hopes to learn the projection of the input point into the data manifold.

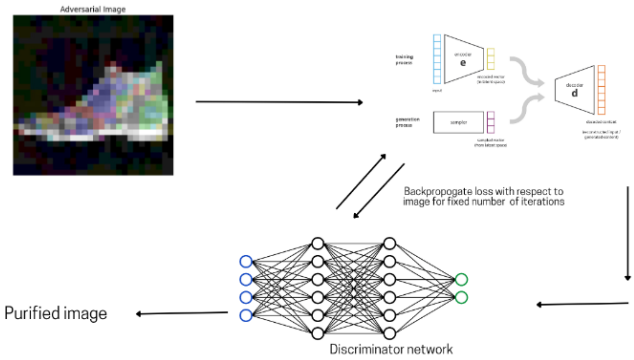

Figure 7: Overview of the proposed architecture

6 RESULTS

The following table illustrates the performance of the VGG16 model on an adversarial dataset, both before and after a purification process was applied to enhance its robustness against adversarial attacks. The results highlight the significant improvement in accuracy post-purification.

Table 1: Accuracy of VGG16 Model on Adversarial Dataset

Models	Accuracy on Adversarial Dataset (%)
VGG 16	19.64
VGG 16 Model after Purification	76.24

7 CONCLUSION

This report presented a novel architecture inspired by Defense GAN aimed at enhancing the robustness of classifiers against adversarial attacks. Our approach utilized a generator network to reconstruct clean images from adversarially perturbed ones, effectively projecting adversarial inputs back to the latent space of legitimate data distributions. The effectiveness of this architecture was measured using Mean Squared Error (MSE), with lower MSE values indicating greater similarity to the original, unperturbed images.

7.1 Analysis of Results

The results clearly demonstrate the potential of our proposed architecture in mitigating the effects of adversarial attacks. The VGG16 model, initially vulnerable to such attacks with an accuracy of only 19.64% on the adversarial dataset, showed significant improvement after the purification process, achieving an accuracy of 76.24%. This substantial increase underscores the ability of the purification mechanism to filter and correct perturbations introduced by adversarial inputs.

7.2 Implications

The marked improvement in accuracy post-purification not only validates the efficacy of incorporating a generative adversarial network in the training loop but also highlights the importance of continual learning and adaptation in neural networks. By leveraging the generative capabilities of GANs, our model adapts to the evolving nature of adversarial attacks, offering a dynamic defense mechanism that is not static but improves through interaction with new and complex adversarial samples.

8 FUTURE WORKS

The Future works are as follows

- Generating stronger white box attacks such as PGD.
- Experiment with black box attacks.
- Currently, the VAE GAN is trained for 50 epochs only due to computational constraints. Usually VAE GANs need to be trained longer to obtain better results.
- Exploring latent space attacks and adversarially training the VAE to neutralize these attacks.

9 REFERENCES

(1) Cemgil, Taylan, et al. "Adversarially robust representations with smooth encoders." International Conference on Learning Representations. 2019

(2) Khan, Asif, and Amos Storkey. "Adversarial robustness of VAEs through the lens of local geometry." International Conference on Artificial Intelligence and Statistics. PMLR, 2023.

(3) Pouya, Samangouei. "Defense-GAN: protecting classifiers against adversarial attacks using generative models." Retrieved from https://arXiv: 1805.06605 (2018).

(4) Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy (Dj) Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. 2019. Adversarial robustness through local linearization. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 1240, 13842–13853.

(5) U. Hwang, J. Park, H. Jang, S. Yoon and N. I. Cho, "PuVAE: A Variational Autoencoder to Purify Adversarial Examples," in IEEE Access, vol. 7, pp. 126582-126593, 2019, doi: 10.1109/ACCESS.2019.2939352.

(6) Bai, Tao, et al. "Recent advances in adversarial training for adversarial robustness." arXiv preprint arXiv:2102.01356 (2021).

(7) Camuto, Alexander, et al. "Towards a theoretical understanding of the robustness of variational autoencoders." International Conference on Artificial Intelligence and Statistics. PMLR, 2021.

(8) Ye, Shaokai, et al. "Adversarial robustness vs. model compression, or both?." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

(9) Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).