COMP90073 Security Analytics Assignment 2: Red Team Report

Muhammad MD NASREIN

21 September 2024

1 Executive Summary

This report presents the findings of a comprehensive red team assessment conducted to evaluate the robustness and security of our pre-trained PyTorch ResNet-8 model against adversarial attacks such as perturbations, specifically employing L2-bounded Projected Gradient Descent (PGD) attacks. Both targeted and untargeted attack methodologies were scrutinized across varying epsilon values and step size to assess their effectiveness in deceiving the model while maintaining minimal perturbation to input data. The dataset used for testing is primarily derived from CIFAR-10 dataset and comprising 10,000 grayscale images of size 32x32 pixels. The model achieved a baseline accuracy of 93.32% on clean, unperturbed data. The primary objective was to identify potential vulnerabilities that could be exploited to degrade the model's performance, thereby informing strategies to enhance its resilience.

Key findings reveal a direct correlation between increasing epsilon values and the success rate of adversarial attacks, with higher perturbations significantly enhancing the likelihood of misclassification. However, this improvement in attack efficacy comes at the cost of perceptual image quality, as evidenced by reduced Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) scores. To mitigate these vulnerabilities, the project explores swift and resource-efficient defense strategies, notably Image Compression and Feature Squeezing. Implementation of these defenses demonstrated a marked decrease in attack success rates without substantial degradation of input image quality, thereby bolstering the model's resilience against adversarial manipulations.

2 Introduction

Deep neural networks have achieved significant success in applications such as image and speech recognition, autonomous driving, and medical diagnostics. However, they remain vulnerable to adversarial attacks—subtle input perturbations that deceive models into making incorrect predictions, posing serious security risks in critical systems.

Projected Gradient Descent (PGD) is a particularly effective attack strategy, enabling both targeted and untargeted misclassifications by exploiting model decision boundaries. Despite advancements in model architectures and training methods, defending against these attacks remains a pressing challenge, as existing defenses often require substantial computational resources and time. This project addresses the need for efficient and effective defense strategies by implementing and evaluating rapid input preprocessing techniques that do not require retraining the model. By analyzing the impact of these defenses on various attack metrics, the project aims to provide actionable insights for enhancing model robustness in resource-constrained environments.

3 Methodology

3.1 Workflow

The project commenced with meticulous data preparation, where a pre-trained PyTorch model was selected and a labeled image dataset was organized into training and testing subsets. This setup facilitated controlled experimentation for both adversarial attacks and defense mechanisms. The core workflow involved executing Projected Gradient Descent (PGD) adversarial attacks on the test images to assess the model's vulnerability. Subsequently, rapid input preprocessing defense strategies, specifically Image Compression and Feature Squeezing, were implemented to mitigate the impact of these attacks. The entire process was accelerated using GPU resources to ensure efficiency. Comprehensive metrics were collected throughout the experiments, enabling a comparative analysis of the model's performance with and without the applied defenses. This structured approach allowed for the systematic evaluation of defense effectiveness within a resource-constrained environment.

3.2 Adversarial Attack Implementation

Adversarial attacks were systematically crafted using the Projected Gradient Descent (PGD) method, renowned for its efficacy in generating both targeted and untargeted adversarial samples. The implementation began by configuring key parameters, including varying epsilon (ε) values of 0.01, 0.05, 0.1, and 0.2 to control the magnitude of perturbations. For each test image, PGD attacks were executed with a maximum of 100 iterations and a step size proportional to epsilon ($\varepsilon/100$), incorporating a momentum factor of 0.9 to stabilize gradient updates. The attack process involved iteratively adjusting the input images to deceive the model, with early stopping employed to halt the attack upon successful misclassification, thereby conserving computational resources. For every adversarial sample generated, critical metrics such as L-infinity Norm, L2 Norm, Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Success Rate were meticulously recorded. These metrics provided a quantitative basis for evaluating the severity and effectiveness of the adversarial perturbations against the model.

3.3 Defense Strategy Implementation

To swiftly enhance the model's robustness against adversarial attacks without the need for retraining, the project employed two primary input preprocessing defense mechanisms: Image Compression and Feature Squeezing. Image Compression was implemented using JPEG compression techniques, which effectively eliminate subtle adversarial noise while preserving the essential features of the input images. This method was integrated into the attack pipeline by applying compression to each adversarial sample before it was fed into the model for evaluation. Additionally, Feature Squeezing was optionally incorporated to further reduce the complexity of the input data. By limiting the color depth of images, Feature Squeezing constrained the adversary's ability to introduce effective perturbations, thereby enhancing the model's resilience. The defense mechanisms were encapsulated within an $apply\_defense$ function, ensuring seamless integration and application during the adversarial attack process. Following the implementation of these defenses, comparative analyses were conducted by executing attacks both with and without the defenses in place. The resulting metrics were then analyzed to assess the effectiveness of the defense strategies in reducing attack success rates while maintaining high perceptual quality of the input images.

4 Result and Discussion

The baseline model accuracy is 93.32%. The confusion matrix in Figure 1 shows the model's performance on clean data, with most predictions concentrated along the diagonal, indicating strong overall accuracy. Some misclassifications are evident, such as class 5 being confused with class 3 (58 misclassifications), but these are relatively few. This matrix serves as a baseline evaluation for comparing how adversarial attacks impact the model's predictions.

The graphs in Figure 2 illustrate the performance of the L2-bounded PGD attack across various epsilon values for both untargeted and targeted attacks. The epsilon values tested include 0.01, 0.05, 0.1, 0.2, 0.25, and 0.3.

The results of the L2-bounded PGD attack across various epsilon values show that both untargeted and targeted attacks behave predictably as epsilon increases. The L2 norm grows consistently, with larger epsilon values leading to larger perturbations. The untargeted attack maintains a stable success rate (10%) across all epsilon values, while the targeted attack's success rate improves significantly as epsilon increases, reaching approximately 22% at epsilon = 0.3. This demonstrates that larger epsilon values make targeted attacks more successful, though at the cost of introducing more noticeable perturbations.

In terms of image quality, the SSIM and PSNR metrics indicate that while perturbations become more visible at higher epsilon values, the structural similarity of the images remains relatively high. At epsilon = 0.01, the SSIM is nearly perfect ($\approx$1.0), and PSNR remains very high ($\approx$85 dB), suggesting that the perturbations are almost imperceptible. However, by epsilon = 0.3, the SSIM decreases slightly ($\approx$0.9993), and PSNR drops to around 60 dB, indicating some degradation in image quality, though

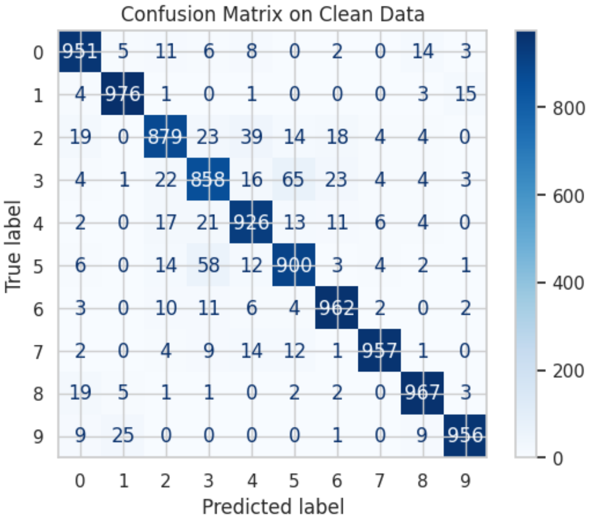


Figure 1: Baseline Confusion Matrix

still within a reasonable range. This suggests that higher epsilon values offer a trade-off between attack success and perceptual distortion.

Given the limited computational resources, we have decided to continue our experiments using epsilon = 0.3, as it provides the highest success rate (although miniscule) for targeted attacks while maintaining acceptable image quality. Although better results may be achieved by exploring larger epsilon values or more advanced attack techniques, our current findings offer valuable insights into the trade-offs between attack effectiveness and image quality. With additional resources, we hypothesize that success rates could be further improved, though this would likely come at the expense of greater image distortions, as indicated by the declining SSIM and PSNR metrics with increasing epsilon.

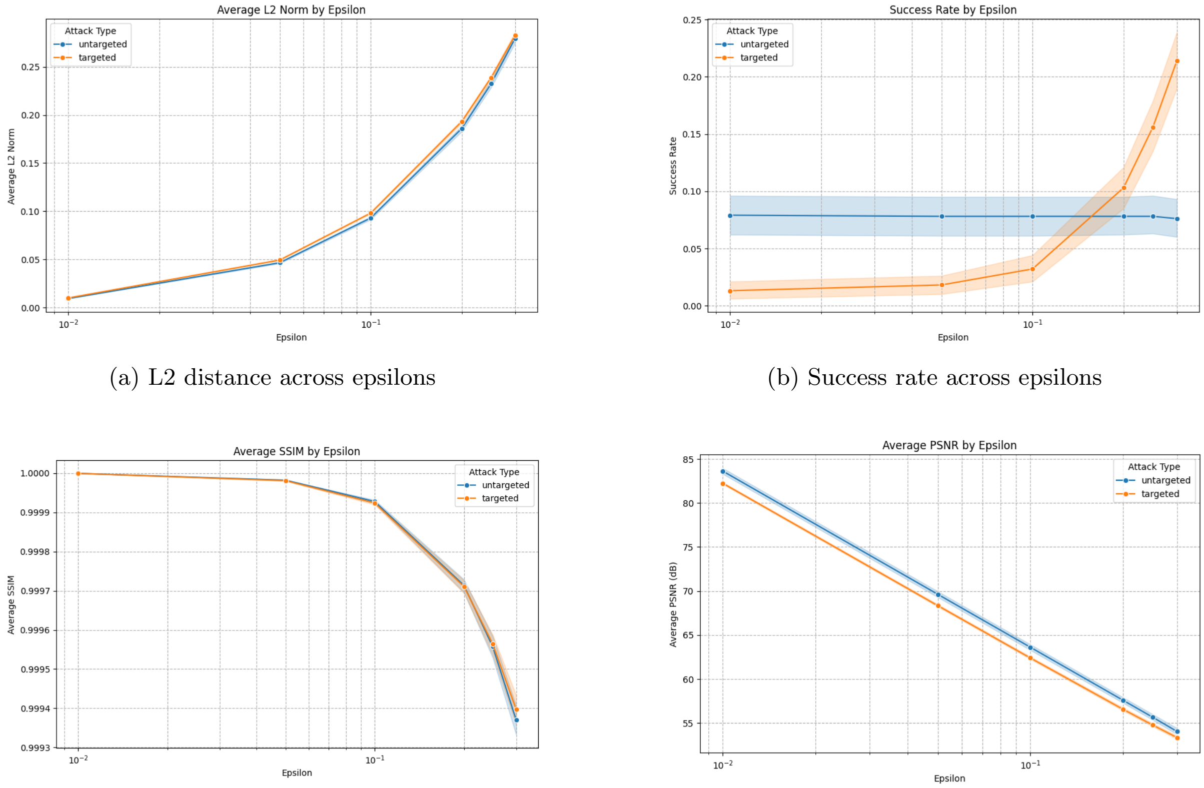
The defense method used combines JPEG compression and feature squeezing to mitigate adversarial perturbations. In this pipeline, the input images are first compressed using JPEG with a quality level of 50, which helps reduce subtle adversarial noise by simplifying the image content. After compression, the images undergo feature squeezing, which reduces the color depth to 4 bits per channel. This process further limits the adversary's ability to introduce effective perturbations by removing fine-grained color variations.

The analysis of the L2-bounded PGD performance before and after applying the defense mechanism as shown in Figure 3, which includes JPEG compression and feature squeezing, reveals significant differences in the adversarial attack behavior. In terms of the L2 norm, the defended model exhibits substantially larger perturbations, with values around 20 for both targeted and untargeted attacks, compared to the negligible L2 norms in the undefended model. This indicates that the defense forces the adversary to create much larger perturbations, making the attacks more detectable. In contrast, the undefended model is easily fooled by minimal, imperceptible perturbations.

When it comes to the number of iterations, there is only a slight difference between the defended and undefended models. Both require approximately 80 to 100 iterations to generate adversarial examples, regardless of the defense being in place. This suggests that while the defense increases the size of the required perturbations, it does not drastically affect the time or number of steps required for an attack to succeed. Thus, the defense mainly increases the perceptibility of adversarial examples rather than delaying their generation.

The defense also has a notable impact on SSIM and PSNR, two metrics related to the perceptual quality of the adversarial images. The SSIM of the defended model drops significantly to 0.5, indicating that the adversarial images are much less structurally similar to the original images. Meanwhile, the undefended model maintains an SSIM close to 1.0, suggesting that adversarial perturbations are almost imperceptible. Similarly, the PSNR for the defended model falls to around 10-15 dB, compared to the undefended model's PSNR of around 50 dB, further confirming that the defense forces the adversary to introduce much more noticeable noise.

The visualized results in Figure 4 compare original images with their adversarial counterparts and the perturbations applied, amplified for better visibility. The first row shows an adversarial image generated with epsilon = 0.1, where the perturbations are more noticeable, affecting a large portion of the image



(a) L2 distance across epsilons

(b) Success rate across epsilons

(c) SSIM score across epsilons

(d) PSNR score across epsilons

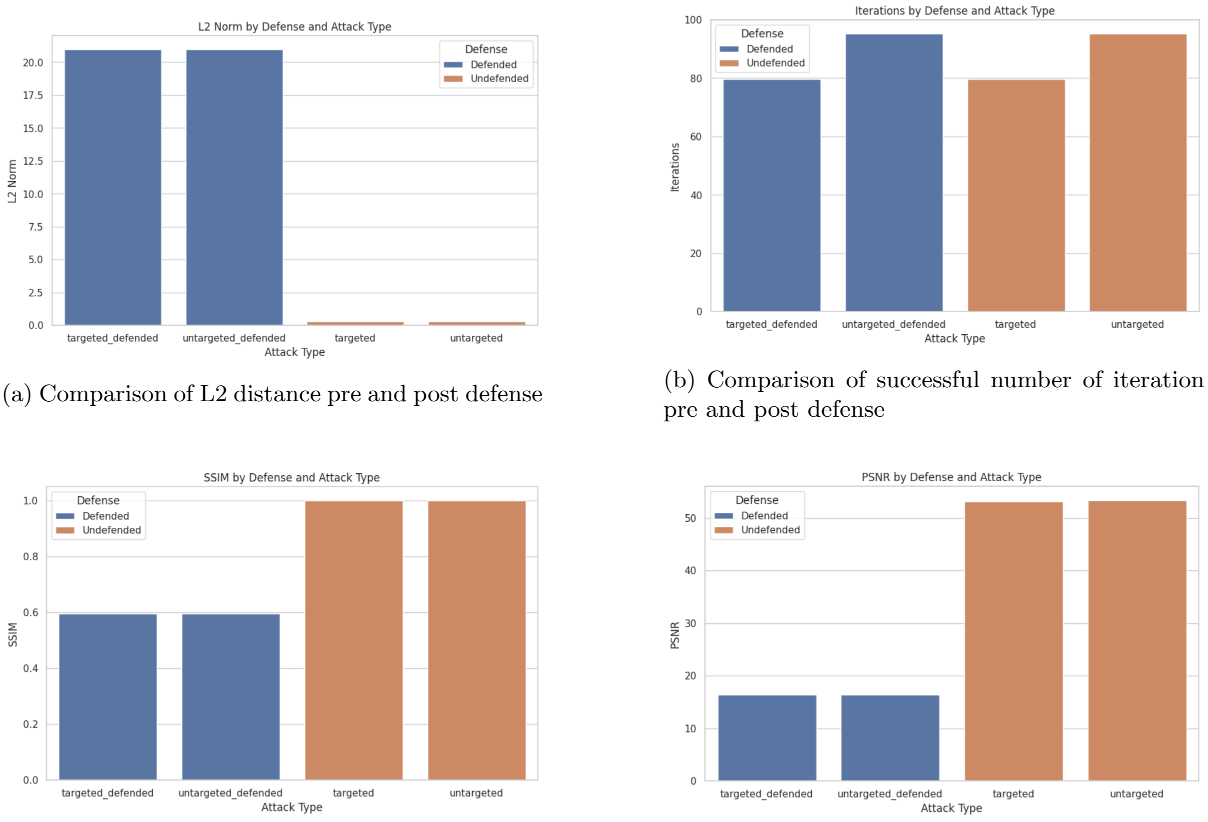
Figure 2: L2-bounded PGD performance across epsilon = 0.01, 0.05, 0.1, 0.2, 0.25, 0.3

and introducing visible noise, though the overall structure of the image is retained. In contrast, the second row with epsilon = 0.01 reveals much subtler perturbations, which are less perceptible even after amplification, resulting in an adversarial image that looks almost identical to the original. This highlights how higher epsilon values lead to more aggressive and visible perturbations, while lower epsilon values create more subtle but still effective changes.

5 Conclusion and Future Work

he defense mechanism combining JPEG compression and feature squeezing demonstrates its effectiveness in increasing the robustness of the model against adversarial attacks. While the number of iterations required to generate adversarial examples remains similar between defended and undefended models, the defense significantly increases the L2 norm of the perturbations, indicating that the adversarial attacks must introduce much larger, more detectable modifications to succeed. This is further supported by the drop in SSIM and PSNR for defended models, showing that the adversarial examples become more perceptible after defense. Additionally, our visual analysis of the adversarial images with different epsilon values reinforces that higher epsilon values result in more aggressive, visible perturbations, while lower epsilon values create subtler changes that are harder to detect. Overall, the defense enhances the model's resilience by making adversarial perturbations more pronounced and less stealthy.

Future work could explore the use of adaptive defenses that dynamically adjust based on the attack type or intensity, potentially further reducing the success rate of adversarial attacks while maintaining image quality. Additionally, testing the defense pipeline on more sophisticated attacks such as adaptive or gradient-free adversarial methods would provide deeper insights into its generalizability. Further investigation into per-class vulnerabilities could help in identifying and strengthening areas where the model remains weak even after defense. Finally, expanding the range of epsilon values and fine-tuning the balance between attack effectiveness and visual detectability could lead to more refined defense strategies that offer better protection against a broader spectrum of adversarial threats.



(a) Comparison of L2 distance pre and post defense

(b) Comparison of successful number of iteration pre and post defense

(c) Comparison of SSIM pre and post defense

(d) Comparison of PSNR pre and post defense

Figure 3: L2-bounded PGD performance before and after image compression and feature squeezing

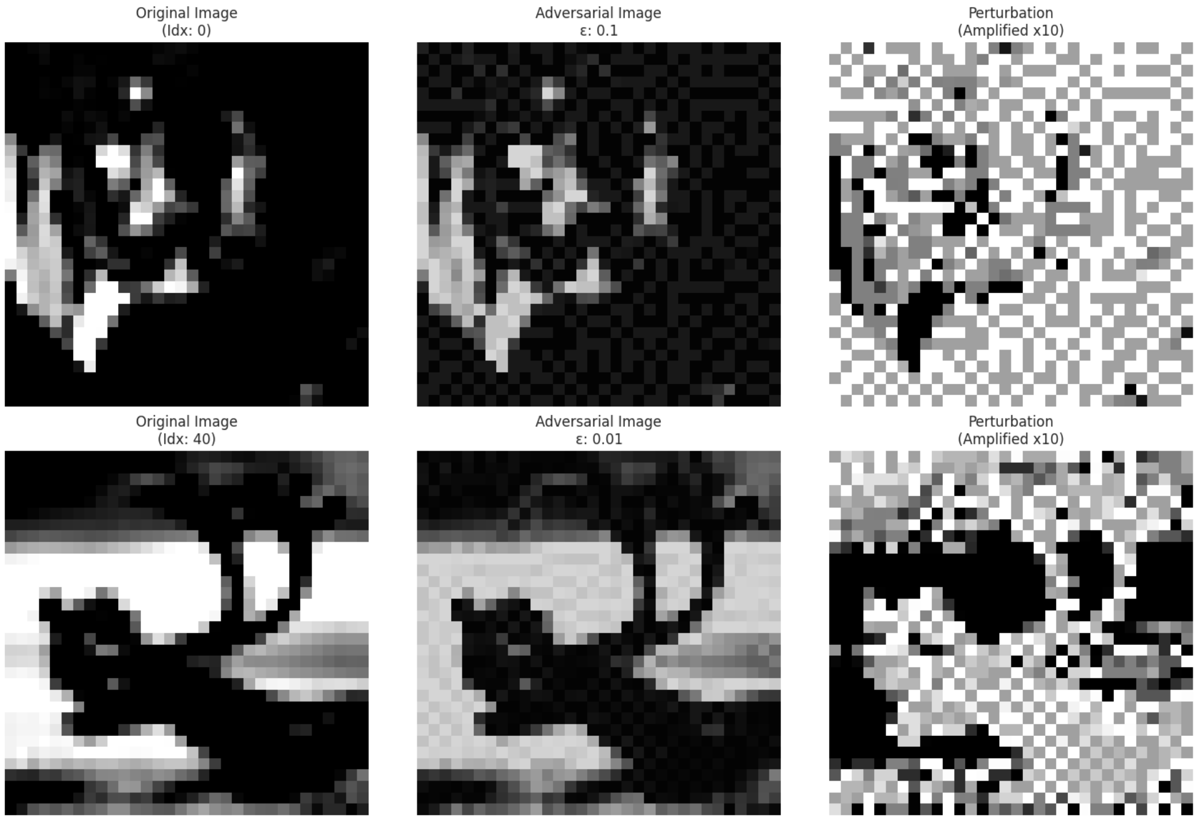


Figure 4: Visual Comparison of Adversarial Images and Perturbations at Different Epsilon Values