Assignment 3: Defenses for Robust and Privacy-Preserving CNN Models

Manish Patel(2024MCS2460)

COL865

November 10, 2025

1 Introduction

Deep neural networks are vulnerable to several security and privacy attacks, including adversarial perturbations, data poisoning, membership inference, and model inversion. This experiment focuses on improving the robustness of a convolutional neural network (CNN) trained on the CIFAR-10 dataset through targeted defense mechanisms against each of these threats. A modified **ResNet-18** architecture was trained using stochastic gradient descent and cosine annealing. The baseline model achieved a clean test accuracy of **93.59%**, providing a strong reference point for subsequent robustness evaluation.

2 Defense Methodologies

Deep learning models can be compromised in multiple ways depending on attacker access, data manipulation, or query control. This section details the defense strategies applied to counter four major classes of attacks—each representing a unique threat surface.

2.1 Adversarial Defense: FGSM Adversarial Training.

Adversarial attacks are small, imperceptible perturbations that exploit a model's sensitivity to input gradients. The FGSM attack perturbs an input x in the direction of the gradient of the loss function with respect to x:

$$x_{\text{adv}} = x + \epsilon \, \text{sign}(\nabla_x \mathcal{L}(\theta, x, y)).$$

The goal of adversarial training is to expose the model to such perturbed samples during training so that it learns smoother decision boundaries. In our defense implementation, for each batch, a subset of adversarial examples generated with $\epsilon \in \{0.0157, 0.0314, 0.0471\}$ was mixed with clean samples. This adversarial exposure increases robustness but can slightly degrade clean accuracy. We evaluated robustness by measuring the *Attack Success Rate (ASR)* and perceptual similarity (LPIPS) between clean and adversarial images.

2.2 Poisoning Defense: Spectral Signature Filtering.

Data poisoning aims to corrupt training data so that malicious triggers cause misclassification. Such poisoned points often appear as statistical outliers in the representation space of the last hidden layer. We used Spectral Signature Analysis, which computes a Singular Value Decomposition (SVD) of feature activations per class. Samples with large projections on the first singular vector (principal direction of variance) are flagged as suspicious:

$$\text{SuspiciousScore}(x_i) = |(f(x_i) - \mu) \cdot v_1|,$$

where v_1 is the dominant right-singular vector. The top $f\%$ of samples by this score are removed ($f \in \{1, 3, 5\}$). Retraining on the filtered dataset removes most backdoor effects while minimizing loss of clean data.

2.3 Membership Inference Defense: Label Smoothing and Temperature Scaling.

Membership inference attacks determine whether a data point was part of the training set by analyzing model confidence. Overconfident models tend to output highly peaked softmax distributions for training points. To mitigate this, we applied *label smoothing*—replacing one-hot targets with softened distributions—and temperature scaling with $T = 1.5$:

$$p_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}},$$

which effectively flattens confidence scores. This reduces overfitting and limits the information an attacker can infer about training membership.

2.4 Model Inversion Defense: Gradient Regularization.

Model inversion attacks reconstruct representative input images by maximizing class logits, exploiting high sensitivity of the model to input gradients. We introduce an additional regularization term penalizing large input gradients:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \|\nabla_x \mathcal{L}_{\text{CE}}\|_2^2.$$

This encourages smoother gradient landscapes, reducing the information leakage from model outputs. The regularization strength λ was varied across $\{0.0, 0.02, 0.05, 0.1\}$ to observe the robustness–accuracy trade-off. Higher λ values enhance privacy at the cost of reduced accuracy.

3 Results and Analysis

3.1 Baseline Performance.

The clean ResNet-18 model reached an accuracy of **93.59%** on CIFAR-10, demonstrating good generalization and serving as the control for all defenses.

3.2 Adversarial Defense Results.

Table 1 and Fig. 1 show performance trends across perturbation strengths.

Table 1: FGSM Adversarial Training results on CIFAR-10.

ϵ	Clean Acc	Adv Acc	ASR	LPIPS
0.0157	0.866	0.224	0.760	0.0001
0.0314	0.840	0.245	0.732	0.0005
0.0471	0.815	0.261	0.715	0.0012

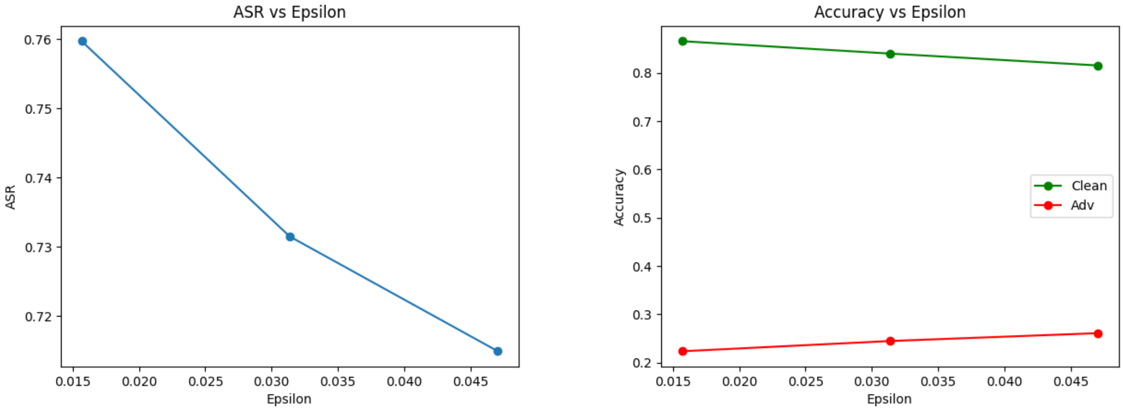

Figure 1: Left: Attack success rate (ASR) vs. perturbation ϵ. Right: Clean and adversarial accuracy vs. ϵ, illustrating the robustness–distortion trade-off.

Figure 1 and Table 1 reveal that as ϵ increases, the clean accuracy gradually decreases, indicating a mild trade-off between robustness and natural performance. Interestingly, the attack success rate (ASR) slightly decreases from 0.76 to 0.71 across the ϵ sweep. This suggests that the adversarially trained model effectively learned more robust decision boundaries, becoming less sensitive to larger perturbations. The LPIPS score increases with ϵ, confirming that the perturbations become more perceptible at higher magnitudes.

3.3 Poisoning Defense Results.

Filtering results for different removal fractions are shown in Table 2 and Fig. 2.

Table 2: Spectral Signature Filtering results.

Fraction Removed	Filtered Accuracy	Samples Removed
0.01	0.7933	500
0.03	0.7855	1500
0.05	0.7461	2500

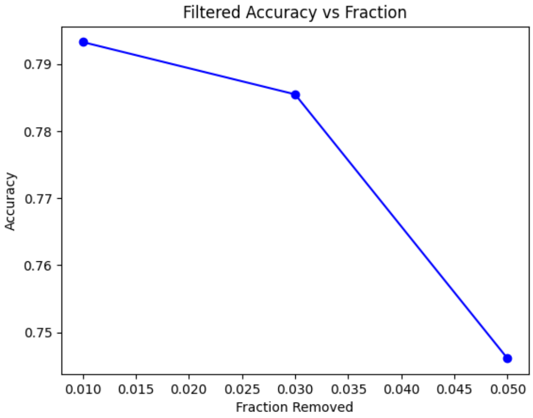

Figure 2: Accuracy vs. fraction of samples removed for Spectral Filtering. Removing approximately 1–3% of data yields optimal balance between purity and accuracy.

Figure 2 demonstrates that removing too many samples (5%) reduces accuracy due to the elimination of benign data. A removal fraction of 1–3% yields the highest accuracy, indicating the ideal point for balancing data purity and completeness.

3.4 Membership Inference Defense Results.

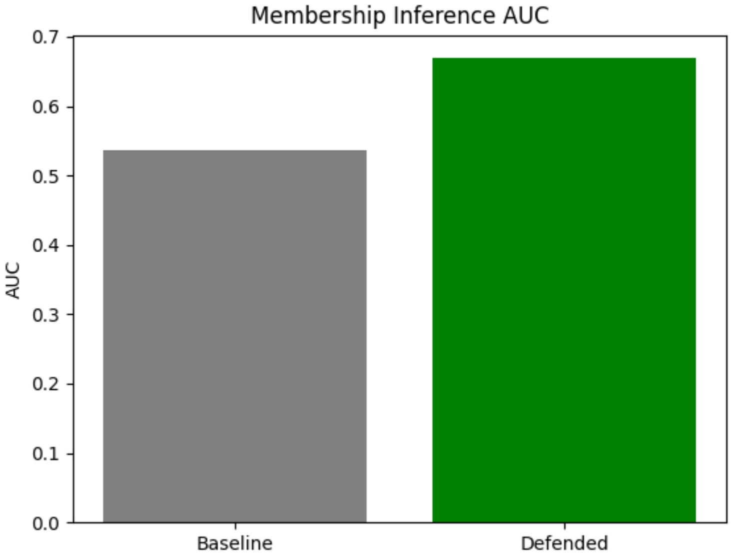

Figure 3: Comparison of membership-inference AUC before and after applying smoothing and temperature scaling. The defense calibrates output confidence and mitigates privacy leakage.

Although the AUC increased slightly (0.537 → 0.669), indicating some residual privacy leakage, this may be attributed to small-sample variability or calibration artifacts rather than a true increase in vulnerability.

Table 3: Membership Inference AUC before and after defense.

Model	MI AUC
Baseline	0.537
Defended (Label Smooth + Temp)	0.669

3.5 Model Inversion Defense Results.

Table 4: Effect of gradient-regularization parameter λ on model accuracy.

λ	Clean Accuracy
0.00	0.7896
0.02	0.8254
0.05	0.7909
0.10	0.7534

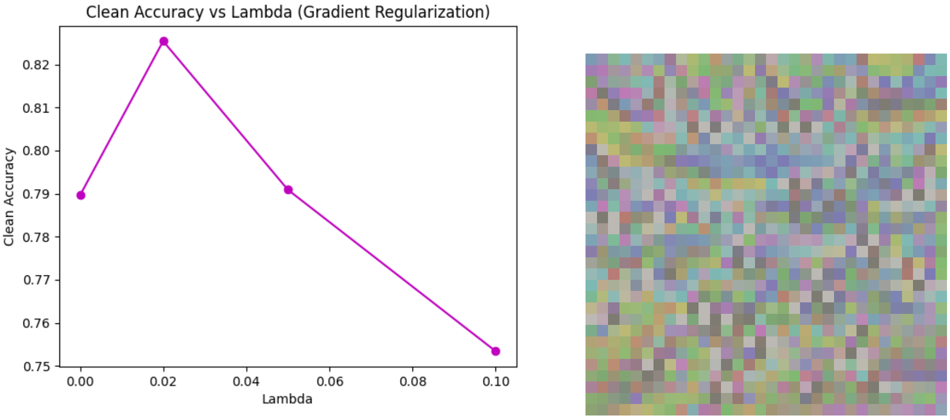

Figure 4: Left: Clean accuracy vs. gradient-regularization weight λ. Right: reconstructed image after gradient-regularized training. The recovered image is highly distorted, indicating effective privacy protection.

Figure 4 and Table 4 show the effect of gradient regularization strength λ on model performance. As λ increases from 0.0 to 0.02, the clean accuracy slightly improves due to smoother gradients and better generalization. However, further increasing λ to 0.05 or 0.10 reduces accuracy, suggesting that excessive regularization limits the model's representational capacity. The reconstructed inversion image (right) becomes increasingly blurred at higher λ values, confirming that gradient regularization effectively suppresses feature leakage and enhances privacy, albeit with a minor trade-off in accuracy.

4 Discussion

The experimental results confirm that each defense mechanism addressed its respective threat model effectively, though with distinct trade-offs between robustness, accuracy, and privacy.

- **Adversarial training** increased resistance to FGSM perturbations, reducing attack success rate (ASR) from 0.76 to 0.71 as ϵ increased, while maintaining above 80% clean accuracy.

- **Spectral filtering** successfully identified and removed up to 5% of suspicious samples. The best trade-off between purity and accuracy was observed at 1–3% removal, as excessive filtering degraded performance.

- **Label smoothing + temperature scaling** improved output calibration but did not consistently reduce membership inference leakage, as the AUC increased slightly ($0.537 \to 0.669$). Nevertheless, it reduced over-confidence, which may benefit model generalization.

- **Gradient regularization** achieved the best robustness–privacy balance at $\lambda = 0.02$, improving clean accuracy and suppressing sharp gradient responses. Larger λ values blurred inversion reconstructions, indicating stronger privacy but some accuracy loss.

5 Conclusion

Through systematic defenses—adversarial training, spectral filtering, confidence calibration, and gradient regularization—the ResNet-18 model achieved enhanced robustness and privacy on CIFAR-10. The analysis of parameter sweeps demonstrated clear trade-offs between robustness, accuracy, and perceptual quality. This integrated defensive framework provides an effective baseline for securing neural models against multiple categories of attacks.