Experimental Study on the Robustness of Convolutional Neural Networks under FGSM Adversarial Perturbations

Vishakha Shishodia
B. Tech, Computer Science and Engineering (AI & ML)
Dr. A.P.J. Abdul Kalam Technical University
February 2026

Abstract

Neural networks exhibit remarkable performance on standard image classification benchmarks; however, their susceptibility to adversarial perturbations raises concerns regarding reliability in safety-critical applications. This study presents an empirical evaluation of the robustness of a Convolutional Neural Network (CNN) under adversarial perturbations generated using the Fast Gradient Sign Method (FGSM). A baseline CNN trained on the MNIST dataset achieves 98.72% accuracy on clean test samples but degrades to 40.79% under FGSM attack ($\varepsilon = 0.2$). To address this vulnerability, adversarial training is implemented. The robust model achieves 98.37% clean accuracy and 88.89% adversarial accuracy. The results demonstrate that adversarial training substantially enhances robustness while incurring negligible degradation in standard performance. The study highlights the importance of robustness evaluation as a standard component of model assessment.

1. Problem Statement

Deep learning models are typically evaluated under the assumption that training and test distributions are identical. However, in adversarial settings, this assumption fails. Small, structured perturbations—often imperceptible to human observers—can cause substantial degradation in prediction accuracy.

The objective of this work is to:

1. Quantitatively evaluate the vulnerability of a CNN to gradient-based adversarial perturbations.
2. Measure the performance degradation under FGSM attack.
3. Assess the effectiveness of adversarial training in improving robustness.
4. Analyze the trade-off between clean accuracy and adversarial robustness.

This study is experimental in nature and aims to provide empirical evidence supporting the need for robustness-aware training.

2. Theoretical Background

The Fast Gradient Sign Method (FGSM), introduced by Ian Goodfellow, generates adversarial examples by linearizing the loss function with respect to the input. The:
perturbation is computed as:
$X\ adv=x+\epsilon\cdot\mathrm{sign}(\nabla xJ(\theta,x,y))$
Where:
1.x denotes the input image,
2. θ represents model parameters,
3. J is the loss function,
4. ϵ controls perturbation magnitude.

Adversarial training attempts to solve a robust optimization problem:

$\theta\min E(x,y)[\|\delta\|\leq\epsilon\max J(\theta,x+\delta,y)]$

This encourages learning parameters that minimize worst-case loss within an ε-ball around each input.

3. Experimental Methodology

3.1 Dataset

Experiments were conducted on the MNIST dataset, consisting of 60,000 training and 10,000 testing grayscale digit images (28×28 pixels). MNIST provides a controlled environment for evaluating adversarial robustness.

3.2 Model Architecture

A standard Convolutional Neural Network was implemented with:

Two convolutional layers (ReLU activation)
Max pooling layers
Fully connected layer
Softmax output layer

The model was optimized using cross-entropy loss and Adam optimizer.

3.3 Training Protocol

Two training strategies were employed:

Baseline Training: Model trained exclusively on clean data.
Adversarial Training: Model trained using FGSM-generated adversarial examples during optimization.

Both models were evaluated on clean and adversarially perturbed test sets.

4. Experimental Setup

Optimizer: Adam
Learning rate: 0.001
Batch size: 64
Epochs: 5
Attack strength: ε = 0.2
Evaluation metric: Classification accuracy

All experiments were conducted using GPU acceleration on Google Colab.

5. Results

5.1 Quantitative Evaluation

Model	**Clean Accuracy**	**Adversarial Accuracy (ε = 0.2)**
Baseline CNN	98.72%	40.79%
Adversarially Trained CNN	98.37%	88.89%

5.2 Observations

The baseline model exhibits strong generalization on clean data but suffers severe degradation under adversarial perturbation. Accuracy decreases by approximately 58 percentage points under FGSM attack.

Adversarial training significantly improves robustness, increasing adversarial accuracy from 40.79% to 88.89%, while reducing clean accuracy by only 0.35 percentage points.

This indicates that the learned decision boundary becomes substantially more stable under perturbations without sacrificing standard predictive performance.

6. Discussion

The experimental findings confirm that high clean accuracy does not imply robustness. FGSM effectively exploits gradient information to induce misclassification.

The dramatic improvement in adversarial accuracy after adversarial training suggests that incorporating worst-case perturbations during optimization enhances the smoothness and stability of learned representations.

Notably, the trade-off between clean accuracy and robustness is minimal in this experiment. This suggests that, at least for MNIST and moderate ε, adversarial training can provide robustness benefits without significant compromise.

However, robustness against single-step attacks does not guarantee robustness against stronger multi-step attacks. Further evaluation with Projected Gradient Descent (PGD) would provide a more comprehensive assessment.

7. Limitations

Only FGSM attack evaluated
Single perturbation magnitude tested
Evaluation limited to MNIST dataset
No formal robustness certification

These limitations restrict generalizability but provide a focused empirical analysis.

8. Conclusion

This study demonstrates that CNN models are highly vulnerable to gradient-based adversarial perturbations. Adversarial training substantially improves robustness while preserving standard accuracy.

The findings reinforce the necessity of robustness-aware evaluation protocols in deep learning research and deployment.