Adversarial Attacks on Convolutional Neural Networks Using the Fast Gradient Sign Method (FGSM)

Your Name March 4, 2025

# Abstract

Deep learning models, particularly Convolutional Neural Networks (CNNs), are vulnerable to adversarial attacks—small perturbations in in- put images that lead to incorrect classifications. In this paper, we explore the Fast Gradient Sign Method (FGSM) attack on a trained CNN model using the CIFAR-10 dataset. We implement and analyze the impact of FGSM-generated adversarial examples on model performance. Our results demonstrate a significant drop in classification accuracy when adversarial examples are introduced, highlighting the security risks posed by such attacks.

# Introduction

Deep learning models, especially CNNs, have achieved state-of-the-art accuracy in computer vision tasks. However, they remain susceptible to adversarial attack, where carefully crafted perturbations can cause misclassification. The

Fast Gradient Sign Method (FGSM) is a simple yet effective attack that leverages model gradients to generate adversarial examples. This study investigates the vulnerability of a trained CNN to FGSM attacks and visualizes their impact.

# Methodology

## Dataset

We use the CIFAR-10 dataset, consisting of 60,000 images across ten classes. The dataset is split into 50,000 training and 10,000 test images. Images are normalized and converted into PyTorch tensors for processing.

## Model Architecture

The CNN model (SimpleCNN) consists of:

* + - A convolutional layer with 32 filters (ReLU activation).
    - A fully connected (FC) layer for classification into 10 classes.
    - Cross-entropy loss function and the Adam optimizer for training.

## FGSM Attack Implementation

We implement FGSM by computing the gradient of the loss w.r.t. the input image and applying a small perturbation (*ϵ*) in the direction of the gradient:

*x*adv = *x* + *ϵ* · sign(∇*xJ*(*θ, x, y*)) (1)

Where:

* + - *x*adv is the adversarial image.
    - *ϵ* controls perturbation strength.
    - ∇*xJ*(*θ, x, y*) is the gradient of the loss w.r.t. the input.

The perturbation remains imperceptible to humans but significantly affects model predictions.

# Experimental Setup

## Model Training

We trained SimpleCNN on CIFAR-10 using five epochs, achieving an accuracy of 85% on clean images.

## FGSM Attack Execution

We applied FGSM to test images with *ϵ* = 0*.*1, generating adversarial examples.

## Visualization of Adversarial Examples

Original and adversarial images were displayed using the show images() function, revealing slight perturbations that mislead the model.

# Results & Analysis

* The adversarial accuracy dropped significantly from 85.6% to 12.4%.
* Visualizations showed subtle but effective modifications in the images.

|  |  |  |
| --- | --- | --- |
| **Model** | **Clean Accuracy** | **Adversarial Accuracy (***ϵ* = 0*.*1**)** |
| Trained CNN | 85.6% | 12.4% |

Table 1: Effect of FGSM Attack on Model Accuracy

**Below attached image is the result of our model**

A collage of images of vehicles

AI-generated content may be incorrect.

# Conclusion & Future Work

This study confirms the vulnerability of CNNs to FGSM attacks. Future work includes:

1. Exploring stronger attacks (PGD, Carlini & Wagner).
2. Implementing defenses such as adversarial training.
3. Testing on larger datasets (ImageNet, MNIST).

# References

1. Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy. *Explaining and Harnessing Adversarial Examples*, 2015.
2. Alexey Kurakin, Ian Goodfellow, Samy Bengio. *Adversarial Examples in the Physical World*, 2016.