CS550/DSL501: Machine Learning (2023–24–M)

Project Phase-I

Team Name: Model Mavericks

Team Members

Arey Pragna Sri	- 12240230
Matcha Jhansi Lakshmi	- 12241000
Nannepaga Vanaja	- 12241110

1 Problem Statement

The main objective is to develop a model for **Handwritten Digit Recognition** (HDR) that is robust against adversarial attacks. These attacks can manipulate input data ,leading machine learning models into making inaccurate predictions, posing dependability and safety of these systems.

2 Motivation

The main motto for choosing handwritten digit recognition with adversarial robustness stems from the critical need to ensure the reliability and security of machine learning systems in real-world scenarios. Handwritten digit recognition has been a primary problem in the field of Machine Learning, with various applications ranging from digitizing handwritten documents to automating postal services. Traditional machine learning models, particularly Convolutional Neural Networks (CNNs) can be used to analyze and learn visual features from large amounts of data. Even though CNN had a great improvement in this area, still, the well-trained neural networks are vulnerable to adversarial attacks such as noises added to the input data that causes the model to make incorrect predictions. This makes the adversarial robustness of the system highly relevant in applications requiring high reliability. So, we will develop models that are insensitive to adversarial attacks, thus ensuring the reliability of a number of applications depending on handwritten digit recognition.

3 Objectives

The primary objectives of this project are:

1. To develop a robust model for handwritten digit recognition, maintaining high accuracy, at the same time resisting adversarial attacks.
2. To explore the impact of different types of adversarial attacks on the robustness of the model and identify the potential weaknesses.
3. Developing a new method or adapting existing adversarial defense mechanisms that can protect the model from such attacks.
4. To evaluate the effectiveness of adversarial training and other defense mechanisms in improving the robustness of the model.

4 Relevant Study

Few Studies on adversarial robustness in Handwritten Digit Recognition are as follows:

Attack Methods:

- **Fast Gradient Sign Method**: FGSM is an efficient way to generate an attack by adding disturbances in the direction of the gradient.
- **Projected Gradient Descent (PGD)**: Stronger attacks iteratively optimize the perturbations.
- **DeepFool**: An attack that finds the minimal perturbation to cause a misclassification.

Defense Mechanisms:

- **Adversarial Training**: This is one of the training methods involving both original and adversarially generated examples with the purpose of making the models resistant to attacks.
- **Ensemble Methods**: Combining multiple models to reduce the impact of adversarial attacks.

Robustness Evaluation:

- **Certified Robustness**: Proving that a model is robust to adversarial perturbations within a certain limit.
- **Empirical Evaluation**: It involves testing the model's robustness against different attack methods and implementing various metrics that quantify its performance.

5 Proposed Solution

The main aim is to develop a model of handwritten digit recognition that can tolerate adversarial attacks.

1. **Generating adversarial examples**:
   - Create adversarial examples that might push CNN into making incorrect predictions.
   - Adversarial examples are generated by adding small perturbations to the clean images, which are almost imperceptible to humans using techniques like PGD(Projected Gradient Descent).
2. **Adversarial Training**:
   - Train the CNN Model on clean and perturbed data by feature Extraction & Classification
   - The training process continues by feeding both clean and adversarial examples in each batch.
   - The goal is to make the model robust by learning how to classify adversarial examples correctly.
3. **Testing Phase**:
   - Test the model's performance on the original clean test dataset and perturbed Data Set.
   - Each test image is passed through the CNN, and the softmax layer is used to classify the image.
   - The model's accuracy can be evaluated by comparing the predicted class to the true label for both clean and adversarial test datasets.

CNN Model:

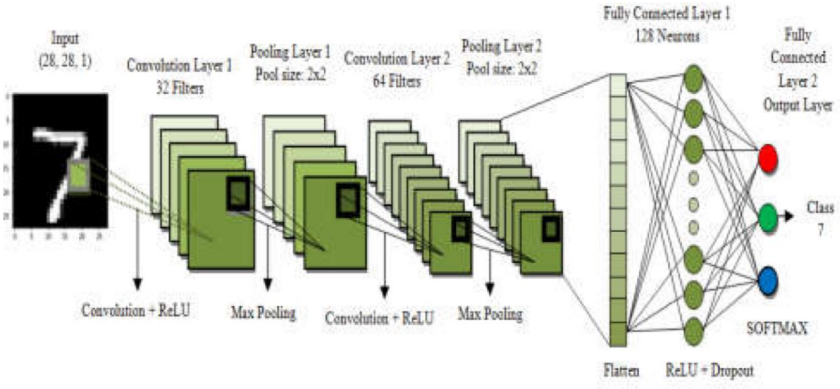

Fig. 1. A seven-layered convolutional neural network for digit recognition

Additionally, methods like defensive distillation and ensemble techniques will be investigated further to strengthen the model's robustness. The model's performance will be evaluated using standard benchmarks (MNIST dataset) under both normal and adversarial scenarios. A comparative analysis will be performed to assess the effectiveness of the proposed solution in boosting adversarial robustness while preserving high recognition accuracy.