TML A3 17 Robustness Report
Team 17

Maitri Vignesh Shah - 7075780

Yashashri Ajay Balwaik - 7075733

July 6, 2025

1 Task Overview

The objective of this assignment is to develop a robust image classification model that maintains high accuracy on both natural (clean) and adversarially perturbed data. The Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are two examples of adversarial attacks that cause tiny, imperceptible changes to input images that can fool conventional classifiers.
Training a model that is robust to such adversarial examples and performs well on clean test images is the difficult part. A submission server evaluates the finished model by comparing its performance on both clean and FGSM and PGD-attacked images. Because the evaluation takes into account both standard and adversarial accuracy, the model's robustness is crucial.
To meet these requirements, the assignment provides a labeled training set and specifies that only certain torchvision model architectures (ResNet18, ResNet34, ResNet50) are permitted. The model must be trained, adversarially fine-tuned, and submitted in a format compatible with the evaluation system.

2 Data accessible to the Adversary

- Labelled training images from the dataset that was supplied (the greyscale images have integer labels and have the same structure as earlier assignments).
- Validation split of the training data (90% train, 10% validation).
- Test/Submission server that provides clean/FGSM/PGD accuracies and uses private test data to assess robustness.

3 Approach used

3.1 Data Preparation and Augmentation

To guarantee the proper input shape (3×32×32), every image from the supplied training set was preprocessed and suitably normalised. During training, a variety of data augmentation techniques were used to enhance generalisation and reduce the model's susceptibility to overfitting. These included rotation, colour jitter, random erasing, random cropping with padding, and horizontal flipping. Only the minimal preprocessing that the model required was applied to the validation data.

3.2 Model Architecture and Training

We used the ResNet18 architecture from torchvision.models, modifying only the final fully connected layer to produce logits for 10 classes, as required. The training was conducted in two phases. First, the model was trained on clean data for 40 epochs using the Adam optimizer with a relatively high learning rate. This phase was designed to maximize clean accuracy and establish strong baseline feature representations.

3.3 Adversarial Training

The model was subjected to 15 more epochs of adversarial fine-tuning following initial convergence. During this stage, the Projected Gradient Descent (PGD) attack with = 4/255, = 1/255, and 7 iterations was used to create adversarial examples for every batch in real time. Each batch's training loss was calculated as the mean of the adversarial and clean losses (also known as the "mix-clean" strategy), which greatly increases robustness while preserving clean accuracy.

3.4 Validation and Model Selection

Using the same PGD parameters as in training, we assessed the model on a held-out validation split during training, documenting both clean accuracy and adversarial accuracy. For the final submission, the model checkpoints with the highest validation scores were retained. To maximise both clean and robust performance, hyperparameters like attack strength, weight decay, and learning rate were adjusted between the two training phases.

4 Reasoning Behind Design Choices

- Model Architecture: ResNet18 is an approved model for the task and provides a good balance between computational efficiency and capacity.
- Data Augmentation: On a small dataset, aggressive augmentations prevent overfitting and improve generalisation.
- Adversarial Training: Accuracy on clean samples is maintained by combining clean and adversarial losses, which has been shown to be more resilient than utilising only adversarial images.
- Hyperparameters: Adversarial fine-tuning strengthens the model's resistance to disturbances, while initial clean training creates a strong feature extractor.
- Optimizer Tuning: Overfitting is avoided and stability is increased by learning rate reduction and weight decay during fine-tuning.

5 Results

Validation Results

- Clean Accuracy = 57.5%
- FGSM Accuracy = 29.6%
- PGD Accuracy = 10.4%

The model passes the clean accuracy threshold and maintains some robustness to FGSM, but PGD adversarial robustness remains low. This suggests further adversarial training (longer, stronger, or more diverse attacks) could further improve results.

6 Future Improvements and Ideas

- Longer/stronger adversarial training: More epochs, increased PGD steps, or higher epsilon.
- Grid search over PGD/FGSM parameters and data augmentation settings.
- Exploring ResNet34 or ResNet50 for potentially higher capacity (if compute allows).

7 Files and their Descriptions

assignment3_solution.ipynb – Complete code for data loading, model training, adversarial example generation, validation, and submission.

final_submission_model01.pt – Model state dictionary as required for submission.

README.md – Summary of the method, key components in the notebook, and instructions to reproduce the results.