Assignment 03-Report  
CS 487

**Part One**

1. **Classification Accuracy:** 91%

**Training Accuracy:** 99%

**Validation Accuracy:** 91%

**A graph of loss and accuracy

AI-generated content may be incorrect.**

**Task Two:**

The Boundary Attack is a decision-based attack that starts with an adversarial example and systematically deforms it to resemble the target image while ensuring that it remains misclassified. It does not utilize any gradients and only relies on the classification results. In the Code a TensorFlowV2Classifier is applied, and the attack is performed on a test image during five iterations, each one consisting of 200 steps. The estimated L2 norm of the perturbation and the predicted class for each perturbation is recorded. The attack gradually reduces the distortion while misclassifying the image until the classifier is deceived. The last image of the adversarial example is shown while the time taken is logged.

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**A red and white stop sign

AI-generated content may be incorrect.**

**TASK THREE:**

The Boundary attack is an adversarial attack that modifies an image to minimize the required perturbation while still being classified into a predetermined class. In contrast to the untargeted variant, this version initiates with a **target class image** and attempts to progressively distort it back to the original image whilst ensuring all classifications during the process remain within the target class.

A target class image, along with an attack-image, is selected in the provided code. The attack is started on the target image and is refined over the course of **five iterations of 200 refinement steps each**. The perturbation process decrements the achieved **L2 norm** of the perturbation while restricting the output of the classifier to the desired class so that the perturbed image can be shown during every step and the time taken can also be captured.

A red and white stop sign

AI-generated content may be incorrect.

Before targeted attack

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**A red and white stop sign

AI-generated content may be incorrect.**

After Targeted attack

***PART TWO***

***TASK ONE***

a.

Test dataset accuracy: **71.930%**

Training Accuracy: **100.00%**

Validation Accuracy: **69.436%**

b.

A graph of training and training loss

AI-generated content may be incorrect.

1. The training results indicate significant overfitting, as the model achieves **100% training accuracy** but only **69.436% validation accuracy**. This suggests that the model is memorizing the training data rather than learning generalizable patterns. The test accuracy of **71.930%** further confirms that the model struggles to perform well on unseen data, highlighting the generalization issue.

***Task 2,3,4***

a.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy on test  dataset | Accuracy on FGSM  attacked subset of test  images | Accuracy on PGD  attacked subset of 200  images |
| Standard classifier | 71.930% | 2.506% | 0% |
| Adversarially  trained classifier | 64.787% | 50.877% | 22% |

b. The tradeoff regarding accuracy and robustness in my solution is dictated by the magnitude of adversarial perturbations and the type of attack used during training. In my case, I achieved **64.787% clean test accuracy**, **50.877% FGSM accuracy**, and **22.000% robust PGD accuracy** while training the VGG-16 model with FGSM adversarial examples with (epsilon = 10/255). These numbers suggest that the model is capable of resisting FGSM attacks, but not PGD, which is much more powerful as it employs multiple steps. The model would achieve greater clean accuracy while still being susceptible to adversarial perturbations if **smaller perturbations ((\epsilon = 2/255))** were used. On the other hand, **training with PGD generated adversarial examples** would lead to higher PGD robustness than 40-50% but would incur a penalty in the form of reduced clean accuracy and increased training duration.