DEPENDABLE AND SECURE AI-ML (AI60006)

Session 2022-2023

Assignment 1

You must be alert to (usually minor) changes that may be made to the assignment statement or the guidelines after the assignment is first put up. Refresh this frame and re-read the assignment carefully before you make your final submission.

Aim

* Constructing adversarial examples
* Adversarial training

Problem Description: Attacks on ML systems for image classification and digit classification

Dataset: CIFAR and MNIST

Models: 2-layer fully connected MLP, 6-layer ConvNet, ResNet-50

Adversarial machine learning is a machine learning method that aims to trick machine learning models by providing deceptive input. Hence, it includes both the generation and detection of adversarial examples, which are inputs specially created to deceive classifiers. An Adversarial Attack is a technique to find a perturbation that changes the prediction of a machine learning model. The perturbation can be very small and imperceptible to human eyes. This assignment is intended to familiarise you with adversarial training and example construction.

Exercises

Download the notebooks from the following link:

<https://drive.google.com/file/d/1kPuNF1UnyolHMTw6xVaXRO9zlk2nzAjG/view?usp=sharing>

* Use pretrained Resnet-50 to classify the image(use introduction notebook).

1) Take an image and find its class

2) Create an adversarial example for a targeted attack by adding random noise to the image

3) Feed it to the Resenet for prediction and find the predicted class of the image

(add the image and prediction for both before and after adding the noise in the image to your document)

* Constructing adversarial examples using (use adversarial\_examples notebook )

1)Fast Gradient Sign Method (FGSM) by varying epsilon between [0.1, 0.001](randomly take five values)

2) Projected gradient descent(PGD)

3) PGD with infinity norm

4) PGD infinity norm with a targeted attack

5) PGD l2 norm by varying epsilon, alpha between [0.1, 0.001]

(Take five values at random and do this for each method). Create a table with epsilon, alpha, and iterations as columns and optimization methods as rows.

Compare the misclassification rate relative to all methods. Add screenshots of all outputs in the document.

Anaconda installation Instructions:

<https://docs.anaconda.com/anaconda/install/windows/>

or use google Colab

Marking guidelines:

Submit notebooks and the answers to the questions above in a separate report (word document). Name the report “ROLLNO\_1.doc. Name the archive “ ROLLNO\_1.zip”.

Assignment marking is to be done only after the deadline expires, as submission gets blocked after the assignment is marked.

| Constructing adversarial examples | 5 |
| --- | --- |
| Adversarial training | 15 |
| *Total Marks* | *20* |

You should keep submitting your incomplete assignment from time to time after making some progress, as you can submit any number of times before the deadline expires.

Warning: Cases of copying from internet sources or fellow students will be dealt with seriously and severely, with the recommendation to the Dean to de-register the student from the course.