Adversarial Machine Learning - Home Assignment 1

Overview

Now that we have successfully implemented the FGSM attack, we can try and re-implement the first defense method ever suggested known as Adversarial Re-training.

Generally speaking, adversarial re-training is implemented as follows (although many variants have been suggested) –

I. Train a classifier
II. Use FGSM (or any other attack method) for generating adversarial examples against that classifier
III. Add the newly created adversarial examples into the classifier's training set, while using the correct class labels
IV. Train an updated classifier using the augmented dataset
V. Repeat steps I-IV above for a number of times.

Assignment Steps

1. Implement a base classifier marked C_0 for classifying the MNIST dataset. You can use the network architecture used during the hands on lab session.
2. Use the FGSM attack in order to produce adversarial examples against 1000 randomly chosen images from the testing set. We denote this set of adversarial examples as ${X_0}'$
3. Measure the attack success rate and mean perturbation radius against ${X_0}'$
4. Augment the training dataset with the adversarial examples and continue to train the model for an additional epoch. We denote this fine-tuned classifier as C_1
5. Repeat steps 2-4 above for 5 times altogether. At each iteration report the success rate and mean perturbation radius of C_i for two different sets of adversarial examples - ${X_0}'$, ${X_i}'$
   That is, the resilience against the original set of adversarial examples, as well as to adversarial examples that are crafted against the most updated model.
6. Report all your results in a table as follows –

Iteration	Clean data accuracy	Attack success rate for ${X_0}'$	Attack success rate for ${X_i}'$	Mean L2 perturbation distance for ${X_i}'$
1				
2				
3				
4				
5				

Submission Instructions

- Submission is to be done in pairs
- Submit a MS-Word report accompanied by source code as either a Jupyter notebook or a plain python file.
- Make sure to include in your report at least the following details –
    - Classifier implementation details
    - Hyper parameters for the training and attack process
    - Full details listed in the table above
    - A printout of the adversarial examples where appropriate
- Make sure to include the names of both partners

Have fun

Tzvika & Ziv