# Machine Learning for Cyber-Security

Lab 3: Adversarial Attacks on Deep Neural Networks
Release Date: 10/18/2023; Due Date: Midnight, 11/18/2022

## Overview

In this lab, you will investigate adversarial perturbation attacks on Deep Neural Networks using the MNIST digits dataset as a benchmark which is a commonly used “toy” benchmarks for machine learning. It contains 28X28 grayscale images with a label from 10 classes, along with the associated labels. The dataset is available as part of the Tensorflow or PyTorch package. You will then evaluate adversarial retraining as a defense against adversarial perturbations.

## What You Have to Do

- You will implement a simple deep learning model with TensorFlow or PyTorch for MNIST digit classification. The DNN has a 784 (28x28) dimensional input, a 10-dimensional output (prediction probabilities for each of the 10 classes). You will implement your attacks and defenses on this baseline DNN. Note that the baseline DNN first normalizes pixel values to lie between [0,1] by dividing each pixel by 255.
- FGSM based untargeted attacks: Your first goal is to implement FGSM based untargeted attacks using images from the test set on the baseline DNN. That is, your goal is to adversarially perturb each image in the test set using with different values of epsilon (budget of the attack). Report the success rate of your attack, i.e., the fraction of test images correctly classified by the baseline DNN that are misclassified after adversarial perturbation for each ε.
- FGSM based targeted attacks: Next, you will repeat the step above, except this time perform targeted attacks where digit i is classified as (i+1)%10 on the baseline DNN. (Here, I refers to the true ground-truth label of the test images, and you can assume that the attacker has access to these labels). Report the attack’s success rate as a function of parameter ε, where the success rate is defined as the fraction of test images correctly classified by the baseline DNN that are misclassified after adversaria lperturbations with label (i+1)%10.
    - Adversarial Retraining against Untargeted FGSM Attacks: For this step, you can assume ε = 125/255 throughout. To defend against adversarial perturbations, the defender adversarially perturbs each image in her training set using the attacker’s strategy in Step 1. She then appends the adversarially perturbed images to her training set, but using their correct labels. Then, the defender retrains the baseline DNN with a new training dataset containing both images from the original training dataset and the new adversarially perturbed images. We call the new DNN the adversarially retrained DNN. • Report the classification accuracy of the adversarially retrained DNN on the original test dataset that contains only clean inputs.
    - Is the adversarially trained DNN robust against adversarial perturbations? Implement FGSM based untargeted attacks using images from the clean test set on the adversarially retrained DNN. Report the success rate of your attack.

## What to Submit

Colab Notebook and a pdf file containing the results of your experiments.