**Adversarial Machine Learning – Home Assignment 3**

Goal and Objective

The goal of this home assignment is to get a better understanding of defensive distillation and the way it influences the defended network. As part of this assignment you will be re-implementing defensive distillation, training a classifier network using this defense method, and eventually testing the resilience of the resulting classifier against diversified attack scenarios.

Assignment Steps

1. Re-implement the defensive distillation approach we have discussed in class
2. Train a simple convolutional neural network for classifying the MNIST dataset. You can reuse the same network architecture and training code used in our lab session
3. Attack your non defended classifier network using FGSM, TGSM, targeted and non-targeted PGD (you can use the notebook we have shared after the lab session).
4. Re-train your classifier network, this time while using the Defensive Distillation defense you have implemented in step 1. A distillation temperature of 30 will do the job, but you are more than welcome to try lower and higher temperatures as well.
5. Attack the classifier you have trained in step 4 using the same set of attack methods.
6. Randomly choose two out of the ten classes in the MNIST dataset. Filter out the dataset so that it includes only the two selected classes. Train a defensively distilled binary classifier over the filtered dataset, and eventually attack your classifier with the 4 different attack methods we have listed earlier.
7. Visually analyze the effect of defensive distillation by creating 4 two dimensional charts as detailed below (your result should conceptually resemble the charts in slide #15 from today’s presentation) –
   1. The first two charts should refer to a ‘regular’ classifier for solving the two class MNIST problem, while the second pair will refer to a defensively distilled version.
   2. Within each pair, one chart should visualize the values of the logits layer, while the other will visualize the Softmax output
   3. All charts are two-dimensional scatter graphs. One axis will reflect the values associated with one class and the second axis will be associated with the other. Each point in this plot will reflect the logit/softmax values associated with a single input sample
   4. Add the x=y line to your charts
   5. All of the input samples used for plotting your charts should be of the same class
8. Look at the charts you have created, identify the phenomenon reflected in them and give a short explanation to each phenomenon

Submission guidelines

1. Your submission should include the full source code as well as a single Word file with all results, diagrams and conclusions.
2. For each of the relevant steps report the attack parameters you have used (epsilon, number of iterations, etc)
3. For each of the attack steps report the following statistics –
   1. Percentage of cases where the original class label was changed following the attack
   2. Percentage of cases where you have managed to reach the target class (for targeted attacks only)
   3. The mean Euclidean perturbation distance

Good luck!

Tzvika & Ziv