**Adversarial Learning Course – Ex. 2 Report**

Tzvika Deutsch, Elad Eatah

In this exercise, one first creates and trains a classifier for the MNIST dataset.

Then, a single, an FGSM attack is conducted on 1,000 random samples from MNIST’s test-set, which generates some adversarial examples for the current classifier. Let bet this set of adversarial-examples. The train-set is then augmented with the mislabeled adversarial-examples. Afterwards, another FGSM attack is performed, resulting in new adversarial-examples, denoted by , and once again the (original) train-set is augmented with the mislabeled adversarial-examples.

This process is repeated a total of 5 times, generating 5 sets of adversarial-examples , , , and .

**The Examined Classifier – A Convolutional Neural-Network**

This classifier is the neural-network which was used during lecture no. 2 of the course. This classifier consists of the following layers:

1. Two 2D convolutional-layers of 8 filters of size 3X3, where each layer is followed by a ReLU activation function.
2. A 2D max-pooling layer with kernel-size 2X2.
3. A dropout layer with dropping-rate of 0.25.
4. A flatten-layer.
5. A dense-layer of 128 units.
6. A dropout layer with dropping-rate of 0.5.
7. A dense-layer of 10 units (for the 10 logits), followed by a softmax activation-function.

The 60,000 samples of the original train-set of MNIST are split into batches of size 128. The learning-rate is initialized to 0.1. At the -th epoch, the learning-rate is modified to where denoted the integer-part.

**Commencing Attacks on The Model**

All (untargeted) attacks were performed using .

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Iteration** | **Clean Data Accuracy** | **Attack Success-Rate for** | **Attack Success-Rate for** | **Mean L2 perturbation-distance for** |
| **1** | 0.56 | 0.66 | 0.61 | 0.12 |
| **2** | 0.71 | 0.64 | 0.52 | 0.17 |
| **3** | 0.75 | 0.66 | 0.58 | 0.2 |
| **4** | 0.73 | 0.68 | 0.58 | 0.18 |
| **5** | 0.73 | 0.72 | 0.58 | 0.16 |

**Samples of Adversarial-Examples**

Below are some examples of adversarial-examples, generated through this process. The images in the first row were generated during the first iteration, the images in the second row were generated during the second iteration, and so on.

