Exercise 01

Adversarial Examples

Reliable and Trustworthy Artificial Intelligence
ETH Zurich

Problem 1 (Coding)**.** In the ZIP file provided on the course webpage, you can find a python skeleton `task1.py` along with a pre-trained MNIST classifier `model`.

Following the documentation in `task1.py`, implement both a targeted (`fgsm_targeted`) and untargeted (`fgsm_untargeted`) FGSM attack for MNIST. Your implementation should clamp the resulting image back to the image domain (i.e., $[0, 1]^{28\times28}$).

Note: The skeleton is based on the PyTorch[1] framework. We strongly recommend that you familiarize yourself with PyTorch now, because the course project will rely heavily on PyTorch. This exercise allows you to gain some initial experience with PyTorch.

Solution 1. See `solution1.py` in the provided ZIP file.

[1] `pytorch.org`