Labs
Machine Learning Course
Fall 2019

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Rüdiger Urbanke
www.epfl.ch/labs/mlo/machine-learning-cs-433

Problem Set 13, Dec 13, 2019 (Adversarial Robustness)

Security and robustness of machine learning models have traditionally been often overlooked, for the sake of greater performance and accuracies. It is usually quite easy for an adversary to create inputs (e.g. images) that fool an ML model into thinking they are something else, however at the same time look much like the originals to a human.

In this lab you will:

1. Learn how to make small modifications in handwritten digit images that result in dramatic errors by ML models. However, humans can still recognize these adversarial examples.
2. Implement a simple defense against this attack.

Setup It is the easiest to run this notebook in Google Colab. You can make use of a free GPU there to train the models faster. If you want to run the notebook locally, you can also use `template/ex13.ipynb`. However, expect to have much longer running-time if you don't have GPU's.

1. Open the colab link for the lab 13:
   `https://colab.research.google.com/drive/1U3YJCVVy3aV5myzMRWplDas7MS1s0zRm`
2. To save your progress, click on *"File > Save a Copy in Drive"* to get your own copy of the Notebook.
3. Click 'connect' on top right to make the notebook executable (or 'open in playground')
4. Start solving the missing parts.