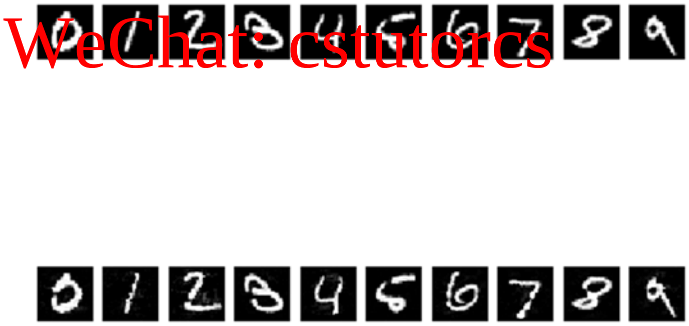
COMP90073 – Security Analytics
Week 10 Workshop

The purpose of this tutorial is to help you gain some hands-on experience of generating adversarial samples. You will be running examples provided by CleverHans (https://github.com/tensorflow/cleverhans/releases/tag/v.3.0.1), and compare adversarial samples generated by the fast gradient sign method (FGSM) and the C&W attack introduced in the lecture.

1. Prerequisite:
   (1) Python3 (https://www.python.org/downloads/);
   (2) Tensorflow (https://www.tensorflow.org/install/).

2. Install CleverHans:
   (1) Download CleverHans from https://github.com/tensorflow/cleverhans/releases/tag/v.3.0.1. **Do not use the latest main branch**.
   (2) Unzip the file and navigate to the folder.
   (3) Run “pip install -e .”.

3. Run tutorials:
   (1) Run “mnist_tutorial_tf.py”, “mnist_tutorial_cw.py” in the subfolder of “cleverhans_tutorials”;
   (2) Add the functionality of saving the trained model in “mnist_tutorial_tf.py”;
   Hint: (1) refer “mnist_tutorial_cw.py” for the similar functionality;
   (2) add two more parameters to “mnist_tutorial()”: i) model_path: path to save or load the model trained on clean examples; ii) model_adv_path: path to save or load the model trained on adversarial samples.
   (3) Compare the adversarial samples generated by FGSM and C&W under the **indiscriminate** setting.
   Hint: (1) Change “TARGETED = True” to “TARGETED = False” in “mnist_tutorial_cw.py”, and re-run the code. You should be able to get the following image:



(2) Replace “adv = cw.generate_np(adv_inputs, **cw_params)” in “mnist_tutorial_cw.py” with how FGSM generates adversarial samples (refer “mnist_tutorial_tf.py”), and re-run the code. You should be able to get the following image: