Paper: <https://arxiv.org/pdf/1706.06083.pdf>

Code for MNIST: <https://github.com/MadryLab/mnist_challenge>

Code for CIFAR10: <https://github.com/MadryLab/cifar10_challenge>

1. From terminal/cmd:

**ssh** [**daryln@10.4.0.15**](mailto:daryln@10.4.0.15) **or ssh daryln@<server ip address>**

1. In server
   1. If container is not already running

command to run:

**docker run -it -p 8895:8895 --name pgdattack -v /home/daryln/adversarial\_attacks/pgdattack:/home --gpus '"device=1"' tensorflow/tensorflow:1.15.0-gpu-py3**

* 1. If container is already running (should already be named)
     1. **docker start pgdattack**
     2. **docker attach pgdattack**

1. PGD attack directory is under /home containing attack on models for MNIST and CIFAR-10 in 2 directories
2. **cd mnist\_challenge**
3. Run attack using

**python pgd\_attack.py**

1. An **attack.npy** file will be generated from running either of the above attacks. This **attack.npy** file contains the adversarial images of the respective imageset used (MNIST)

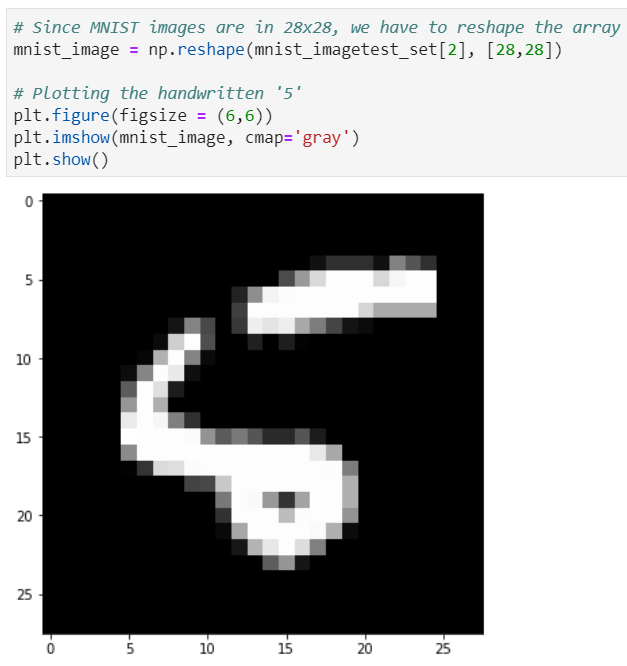
To view the imageset

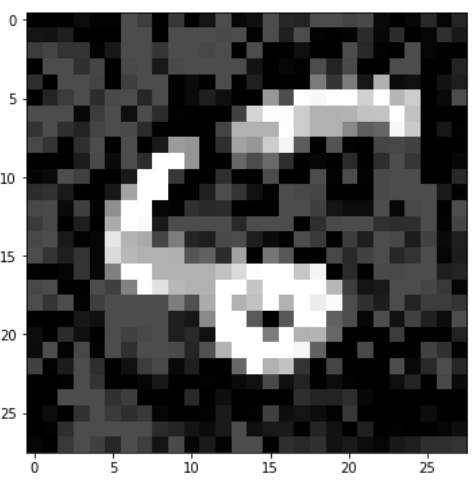
The adversarial images are stored in an imageset which is converted into a numpy array (hence .npy). To view them, it is recommended to open a Jupyter notebook for this. There is already one generated in the directory (**Compare Adversarial Images.ipynb)**.

To use Jupyter

Command to run: **jupyter lab --no-browser --ip=0.0.0.0 --port=8895 --allow-root**

Open the notebook and the code should already be written for you.

1. To evaluate the attack

**python run\_attack.py**

1. To evaluate a neural network

**python run\_attack.py** and a **pred.npy** file will be created

1. This **pred.npy** indicates the model predictions on the adversarial images
2. For CIFAR-10

**cd cifar10\_challenge**

1. Repeat steps 5 to 9
2. Adversarial images compared with original

