Paper: <https://arxiv.org/pdf/1808.09413.pdf> Code: <https://github.com/turned2670/DLFuzz>

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 8890:8890 --name dlfuzz -v /home/daryln/adversarial\_attacks:/root --gpus=1 daryln/dlfuzz:latest**

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

1. DLFuzz attack directory is under /root as DLFuzz
2. Install required libraries: (pip install …)
   1. opencv-python==3.4.0.12
   2. keras==2.1.3
   3. h5py==2.7.1
   4. pillow==5.0.0
   5. tensorflow-gpu==1.2.1
3. **MNIST** and **ImageNet** directoriescontains the mnist and imagenet dataset with the DLFuzz attacks in them
4. **cd MNIST** (to run dlfuzz attack on models trained with mnist)
5. Run attack using

**python gen\_diff.py [2] 0.5 5 0602 5 model1**

[2] -> the list of neuron selection strategies

0.5 -> the activation threshold of a neuron

5 -> the number of neurons selected to cover

0602 -> the folder holding the adversarial examples generated

5 -> the number of times for mutation on each seed

model1 -> the DL model under test

1. The adversarial images will be generated as .png files under the directory as specified above (in this case, it would be “0602”) under “generated\_inputs” directory.

To view the imageset

To view them, it is recommended to open a Jupyter notebook for this. There is already one generated in the directory (**Compare adversarial output.ipynb)**.

To use Jupyter

Command to run: **jupyter lab --no-browser --ip=0.0.0.0 --port=8890 --allow-root**

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

