Project 2

Adversarial Crafting I

Due date: February 13 (two weeks)

In this project you will use the Numpy, the Keras/Tensorflow backend and the precooked foolbox package which allows you to easily craft adversarial images using various adversarial crafting techniques some of which we covered in class, e.g. FGSM, SaliencyMap etc. The goal is the build rogue images that with high confidence fool ML classifiers. Specifically, we will target the ResNet50 DNN model using random images from the ImageNet image library. Both can be easily access from Keras.

1. Make sure you have python and keras installed with a backend of Tensorflow (or Theano) on your programming platform. The Keras website is here
   https://keras.io/
   It is recommended that you use python3

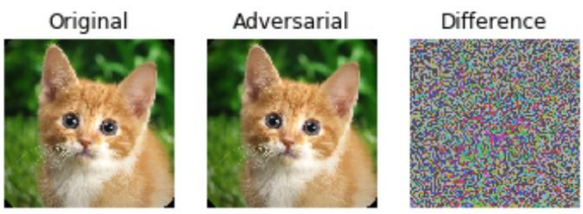
2. Install foolbox and familiarize yourself with the software interface
   http://foolbox.readthedocs.io/en/latest/

3. Install Randomstate, h5py, Pillow and matplotlib packages to python via pip or conda.

4. Run the minimal working example which can be found here
   https://github.com/bethgelab/foolbox#example
5. Study the adversarial crafting attacks supported by foolbox:
   http://foolbox.readthedocs.io/en/latest/modules/attacks.html
6. For 10 images that you pick from ImageNet, use the foolbox to craft adversarial samples using the following attacks:

   - BlendedUniformNoiseAttack
   - ContrastReductionAttack
   - FGSM
   - SinglePixelAttack
   - SaliencyMapAttack

7. For each attack print the original image the crafted image and the noise in one row as done in the foolbox minimal example, i.e.



8. Turn in a single document (PDF) which contains
   - the rows of images labeled with the name of attack and suitable parameter values
   - the listing of the code you have written to generate the adversarial samples.