Assignment 7
Adversarial Attack on LeNet5

1. Introduction

Given a pre-trained LeNet5 model, your job is to provide jpeg images to fool the network. There are 2 attack scenarios in this assignment, correspond to 2 questions:

Scenario 1: (50%) Make the network think a random image belong to a class with high confidence. Submit a random image, as random as possible (characterized by having high Shannon entropy) that makes the network believe that it belongs to one class (any class of your choice) with high confidence (characterized by having a high predicted max probability).

Scenario 2: (50%) Make the network think an image belongs to one class when it clearly belongs to another class.
For this scenario, submit a modified version of the first image in the test set of the MNIST dataset (see the ADVERSARIAL EXAMPLE GENERATION tutorial) that makes the network think that it belongs to the 0 (zero) class. The modified image must have a high PSNR score (correspond to imperceptible change with respect to the original version). You can get some more help from this tutorial also (How to Intentionally Trick Neural Networks)

2. Submission:

You need to submit **2 jpeg** files: **image1.jpg** and **image2.jpg** corresponding to the two scenarios above to e-class as well as the **result.txt** file created from the given notebook file.
You need to generate the 2 images yourself by any means necessary. The resource folder in this assignment only contains a notebook to do the evaluation, the pretrained model and 2 random jpeg images as an example. There's no skeleton code to generate the images for the questions but you can use the sample code given in the above tutorial as a place to start.
To do the evaluation by yourself, just replace the 2 random jpeg images with your own images by uploading them to google colab.

3. Marking [Worth 5.7% of the total weight]:

Scenario 1 (50%):

- Score depends on the highest probability predicted by the LeNet for your image. Scale from 0.7 to 0.8.
- Your image must satisfy the "randomness" requirement by having a Shannon entropy of at least 6.0

Scenario 2 (50%):

- Score depends on the PSNR score of your image. Scale from 25 to 28 PSNR.
- Your image must fool the network correctly by making it thinks that the image belong to the 0 class.

You can see your score by running the provided notebook using your images. The score you got from there will carry **75%** of the final score. The remaining **25%** depends on the performance of your viva.

DUE DATE: The due date is Friday, November 1 by 11:55 pm. The assignment is to be submitted online on eclass. For late submissions' rules please, check the course information on eclass