HW2 - Neural Networks and Deep Learning

due March 3, 2023

This is solely a programming homework. The programming part must be done on colab. The submission for programming part is one python notebook file that you should upload to canvas. Both problems must be done in one notebook. Use text to clearly separate out the two problems. For both the problems, I have provided you **a saved model file and five images**. You must ensure that these files are in a folder "/content/drive/My Drive/Colab Notebooks/Data/" when you read them from within colab. This is so that it matches that location in my colab.

1 Problem 1 (Feature Maps) - 4 marks

In this problem, you will plot the feature maps from the first two convolutions layers. The model that you must use is provided as attachment to the HW - it is the same model that I trained on my laptop. Hence, when you load it you must use the lines of code in "Adversarial Example Pytorch.ipynb", specifically

```
model.load_state_dict(torch.load(pretrained_model,map_location="cuda").state_dict())
```

Use the attached 000015.jpg as input for which you plot the feature maps. You must write code to read this image file and provide as input to the neural network model that you load. You may not want to use the Dataset or Dataloader class as that is an overkill for inputting one image to an already trained neural network.

You need to do the following:

1. Plot **all the feature maps** after first convolution, after the first relu, after second convolution and after second relu. The plots must be in order of the layers mentioned above - note that each layer outputs multiple feature maps, you should plot all of them. You should plot using the imshow function in matplotlib.

2. Your figures **must have 8 plots in one row** and then however many rows you need for the feature maps from each layer.

3. Your python noteboook should have your name at the top and all images must be showing when you save it for submission.

2 Problem 2 (Adversarial Example) - 6 marks

You saw that we use the PFD attack to produce adversarial examples for images in class. In this problem, you implement another attack called FGSM. In fact, I will point you to the explanation and code for this attack `https://broutonlab.com/blog/adversarial-attacks-on-deep-learning-models` (the blog has mnay more things other than FGSM, you can ignore those). You are free to use any material from internet.

You need to attack the same network as in Problem 1 (the saved network is attached to the homework). The five images that you attack are also attached as 000015.jpg, 000016.jpg, 000017.jpg, 000018.jpg, 000019.jpg. Do note that if you just copy the code from the above blog it will not work as is, you need to use the loss function that was used to train the network (which is not crossentropy in torch).

You need to do the following:

1. Plot the five original images in one row with the ground truth label and the prediction of the model (note prediction of model on the original image may not match ground truth label). In the next row, plot the adversarial examples and the prediction of the model for these adversarial examples.

2. Your python noteboook should have your name at the top and all images must be showing when you save it for submission.