**Hands-on Assignment 11**

**Due Date: See web**

Several pre-trained models are mentioned in HA4:

* AlexNet
* VGG16
* ResNet50
* InceptionV3
* DenseNet121
* MobileNetV2

For this assignment, you need to do two parts about adversarial attack. In the first part, you will generate adversarial examples under different bounds and see if the adversarial examples can be visually distinguishing from original images. In the second part, you will test the transferability of adversarial examples. Details for each part are as follows:

1. For the first part, choose one of the pre-trained models mentioned above. Then choose one adversarial attack. You can use the FGSM attack in the tutorial, or any other adversarial attack you are interested in. Use different parameters (norm bound , types of norms if there are different types of norms) of the chosen adversarial attack to generate adversarial examples. You should use at least three images in the images/ folder and three different parameters to generate at least nine adversarial examples.

Or, you can also simply use CIFAR10 datasets instead of the provided images in images/ folder. In this case, you can use the pre-trained model of ResNet20 in the tutorial for the first part. But you will need to train other models on CIFAR10 on your own for the second part.

1. In the second part, you need to test the transferability of adversarial examples. Assume you have three different trained models. Choose one model and generate a successful adversarial example. Successful adversarial example means that all the three model can correctly classify the original image but the chosen model misclassifies the corresponding adversarial example. Then you need to test if the other two different models can correctly classify the adversarial example. In this part, you can just set the parameters in the first part fixed reasonable values. Test at least three images on at least three different models.

Write a simple report to summary the results and submit the report via Canvas. The summary should contain the followings:

1. Which pre-trained model and adversarial attack were used in the first part?
2. Visualize the original images and corresponding adversarial examples with corresponding parameters.
3. Compare original images with adversarial examples to see if you can visually distinguish adversarial example from original one. Write your conclusion.
4. Test results on the transferability of adversarial examples. For each test image, write the transferability results on different models.
5. Any interesting conclusions and findings.

The report should be organized and neat, and the format should follow either docx or pdf.

Similarity scores will be computed for this assignment.

*Copyright Statement: The materials provided by the instructor in this course are for the use of the students enrolled in the course. Copyrighted course materials may not be further disseminated.*