**Lab 6 Adversarial Examples**

In this Lab, we are going to generate adversarial examples and visualize their effects on the ImageNet dataset using the Foolbox python image. We will use the Anaconda environment from the last lab.

conda activate pytorch\_cpu

conda install -c conda-forge eagerpy

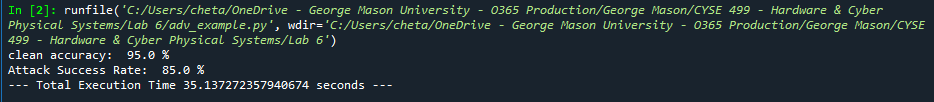
conda install -c conda-forge foolbox

After the installation, activate pytorch-cpu in Anaconda and open Spyder, use Spyder to open adv\_example.py.

**Question 1 Attack Comparisons [25 pts].** Uncomment the parts below #Question 1 in the code and execute the adv\_example.py. It will execute the **FGSM attack**. Record the attack success rate. Change “attack = FGSM()” into “attack = LinfPGD() and “attack = L2CarliniWagnerAttack()”. Compare the Attack Success Rate and Execution Time of the three attack methods. Attach screenshots of the results and add a few sentences of summary. [Note: If the CW Attack exceed 10 mins, simply abort it and explain.]

Text

Description automatically generated



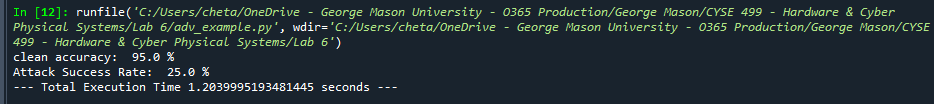
Text

Description automatically generatedGraphical user interface, text, application

Description automatically generated

The FGSM attack and LinfPGD had an accuracy of 95%. FGSM success rate is 65% and an execution time of approximately 4.0669 seconds. LinfPGD had a higher success rate of 85% and a higher execution time of 35.1373 seconds. The L2CarliniWagnerAttack took longer than 10 minutes, so I aborted the execution.

**Question 2 Perturbation Strength [25 pts].** Epsilon represents the perturbation strength we added onto the images. In principle, the larger the strength, the higher the attack success rate and the easier to be detected. Change the epsilons between from 0.0005, 0.001, 0.0015, 0.002, 0.0025, 0.003, 0.0035, 0.004. Use the “attack = FGSM()” and draw a curve in Excel to validate the relation between epsilon and attack success rate. The x axis is the epsilon strength, and the y-axis is the attack success rate.

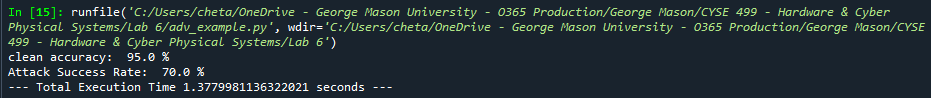


Text

Description automatically generated

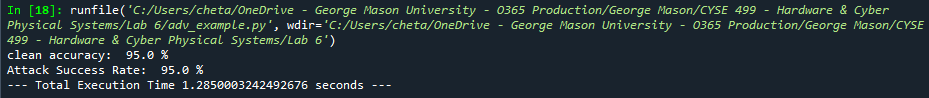
Text

Description automatically generated



Text

Description automatically generated



Text

Description automatically generated

Text

Description automatically generated

**Question 3 Visualize Results [50 pts].** Finally, let us visualize the perturbation and the adversarial images. Uncomment the parts under #Question 3 and execute the program using the FGSM Attack and epsilons = 0.0015. Three different images will be saved. The perturbation is taken as the difference between the two images. If you are not able to see the perturbation (black image), we need to increase it in order to see it by adjusting the value of magnify on some images. (e.g., magnify = 3).

Shape

Description automatically generated with low confidence

**Part a**: Randomly select some images in the mini batch by changing the “image\_num = 0” to a different number, e.g., “image\_num = 5” (has to be under 19 since the batch size is 20). Demonstrate three groups of results. Each group should contain the original image, the perturbation and the adversarial image (a total of 9 images). Answer the question: can you tell the difference by observing on the adversarial image?

A picture containing grass, outdoor, tree, field

Description automatically generatedA picture containing grass, outdoor, tree, house

Description automatically generatedA picture containing flower, colorful, decorated, plant

Description automatically generated

A picture containing text, bottle cap

Description automatically generatedA picture containing text, bottle cap

Description automatically generatedA picture containing grass, colorful

Description automatically generated

A group of stained glass windows

Description automatically generated with medium confidenceA group of stained glass windows

Description automatically generated with medium confidenceA map of a city

Description automatically generated with low confidence

Image numbers I used: 5 (top), 9 (middle), 12 (bottom). The adversary images are on the left, original is in the middle, and the perturbation is on the right. I see no difference between the adversarial and original images, but I do see a difference in the perturbation image.

**Part b**: Now increase the epsilons = 0.0015 to epsilons = 0.01. Select some images by doing **Part a** again. Answer the question: can you tell the difference with a larger epsilon? You can highlight the portion that you find suspicious artifacts.

|  |  |  |
| --- | --- | --- |
| **Adversarial** | **Original** | **Perturbation** |
|  | A picture containing grass, outdoor, tree, house  Description automatically generated |  |
| A picture containing text, bottle cap  Description automatically generated | A picture containing text, bottle cap  Description automatically generated | A picture containing colorful  Description automatically generated |
| A group of stained glass windows  Description automatically generated with medium confidence | A group of stained glass windows  Description automatically generated with medium confidence | A map of a city  Description automatically generated with low confidence |

For the top row, the adversarial and original images are different. One of them is reversed. All the perturbation images are different from the others. The other images look the same.

**Question 4 [Open Discussion +10 Bonus].** Write 1-2 paragraphs about what you have learned through this lab about the existing AI algorithms.

What I have learned about the existing AI algorithms is that it can change the integrity of the files such as an image. One image can be normal while another image could be blurry. It can also calculate the accuracy of a certain cyber-attack. It can also determine the success rate of an attack.