ECE 661: Homework #5

Adversarial Attacks and Defenses

NetID: wh162

1. True/False Question
   1. Problem 1.1:

false, an evasion attack is to manipulate the input to get a wrong output.

* 1. Problem 1.2:

false, modern defenses aren’t able to realize robustness and accuracy in the same time.

* 1. Problem 1.3:

True, refer lect16

* 1. Problem 1.4:

true, refer to lec16

* 1. Problem 1.5:

true.

* 1. Problem 1.6:

false, there may occur gradient masking

* 1. Problem 1.7:

false, should be gradient-based approaches.

* 1. Problem 1.8:

false, should be input

* 1. Problem 1.9:

true, refer lec18

* 1. Problem 1.10:

true, refer lec 16

1. Lab1: Environment setup and attack implementation
   1. Problem 2.1:

For netA\_standard: the accuracy is 92.48%

Graphical user interface, text

Description automatically generated

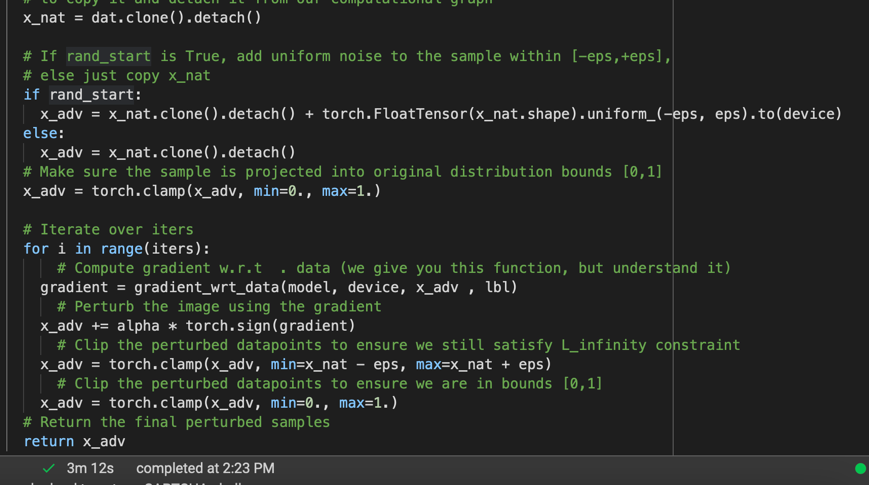
For netB\_standard: the accuracy is 92.37%

Graphical user interface, text

Description automatically generated

No, the two models doesn’t have the same architecture, model A has one 28x28 and one 14x14 conv2d layer, whereas model B has two 28x28 and two 14x14 conv2d layer.

* 1. Problem 2.2:



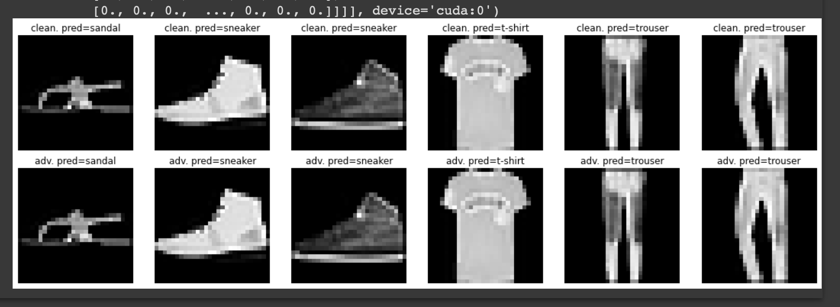
When the value of EPS reaches 0.08, I can start to notice that the difference between the two images. For me, it would be hard to predict or to tell the correctness of the image.

EPS = 0.08

Text

Description automatically generated with medium confidence

EPS = 0.0



* 1. Problem 2.3:

Text

Description automatically generated

We can see that under the same EPS = 0.08, the noise was much noticeable, we can easily spot the difference between the images. As a result, under the same EPS, the noise of PGD and FGSM are visually different.

Graphical user interface

Description automatically generated

* 1. Problem 2.4:

Text, chat or text message

Description automatically generated

After several experiment on the value of EPS, we can see that the noise of the image grew dramatically, no sooner after the value of 0.2 can’t we see the origin image. Unlike the slowing increasing noise in FGSM and PGD, which we can still see the origin image even after larger value of EPS.

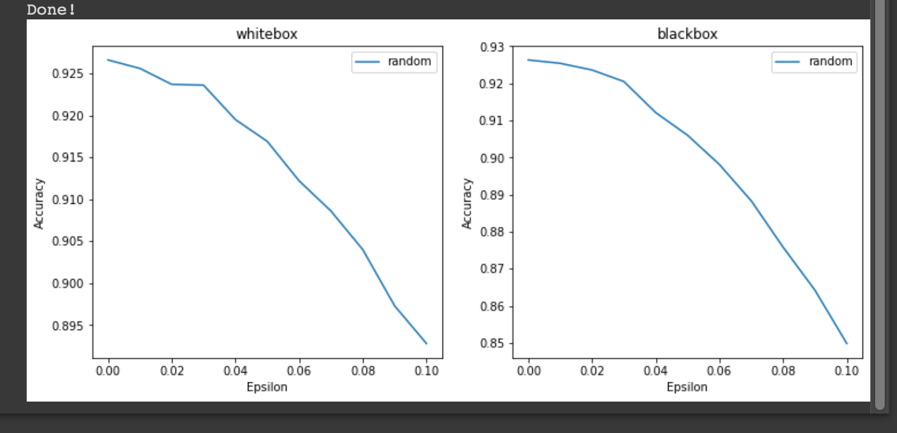
1. Lab2: Measuring attack success rate
   1. Problem 3.1:

A whitebox attack is when the hacker has the access to the model’s weight, whereas a blackbox attack is when the hacker only has access to the input and the output of the model.

Vulnerability

* 1. Problem 3.2:

We can see from the plot that as the eps grows, the accuracy decreases.



* 1. Problem 3.3:

From the plot, it is clearly that PGD has the deepest accuracy drop, whereas random attack maintained the quality of the accuracy. Also, from the plot, FGSM and rFGSM merely different from each other. Though the trigger of the drop of random attack is not obvious, I would guess that around eps=0.05, the decrease of the random attack accuracy starts to get noticeable.

A picture containing chart

Description automatically generated

* 1. Problem 3.4:

From the testing of blackbox, the accuracy starts dropping when eps=0.03, which is different than whitebox testing. The curve remain the same, where random attack has the highest accuracy and PDG drops the fastest.

A picture containing chart

Description automatically generated

* 1. Problem 3.5:

The whitebox attack is more powerful. We can direct access the weight of the model and thus makes it easier and powerful to break down a model. It makes, sense because the main point of a model is the weight, once we can control the weights, we can control the model.

1. Lab3: Adversarial training
   1. Problem 4.1:

The final accuracy of the clean test data is 35.3%. Comparing to the beginning, the accuracy is actually a bit higher than the origin model. We can see from the output below, the whitebox accuracy of FGSM is higher than rFGSM.

FGSM: Text

Description automatically generated with medium confidence

rFGSM: Graphical user interface, text, application

Description automatically generated

* 1. Problem 4.2:

The final accuracy of the clean data is 29.74%, the accuracy is less than the origin model. Also from the results of FGSM and PGD, we can see that FGSM has a higher accuracy than PGD

Graphical user interface, application

Description automatically generated

* 1. Problem 4.3:

We can see that the model is not robust for all type of attacks.

Chart, line chart

Description automatically generated Chart, line chart

Description automatically generated

* 1. Problem 4.4:

From the three plots, it is hard to tell which one is better

Chart, line chart

Description automatically generated

* 1. Problem 4.5:
  2. Problem 4.6: