Writeup hw2

Submitted by: Yahav Lazar

ID – 213762180

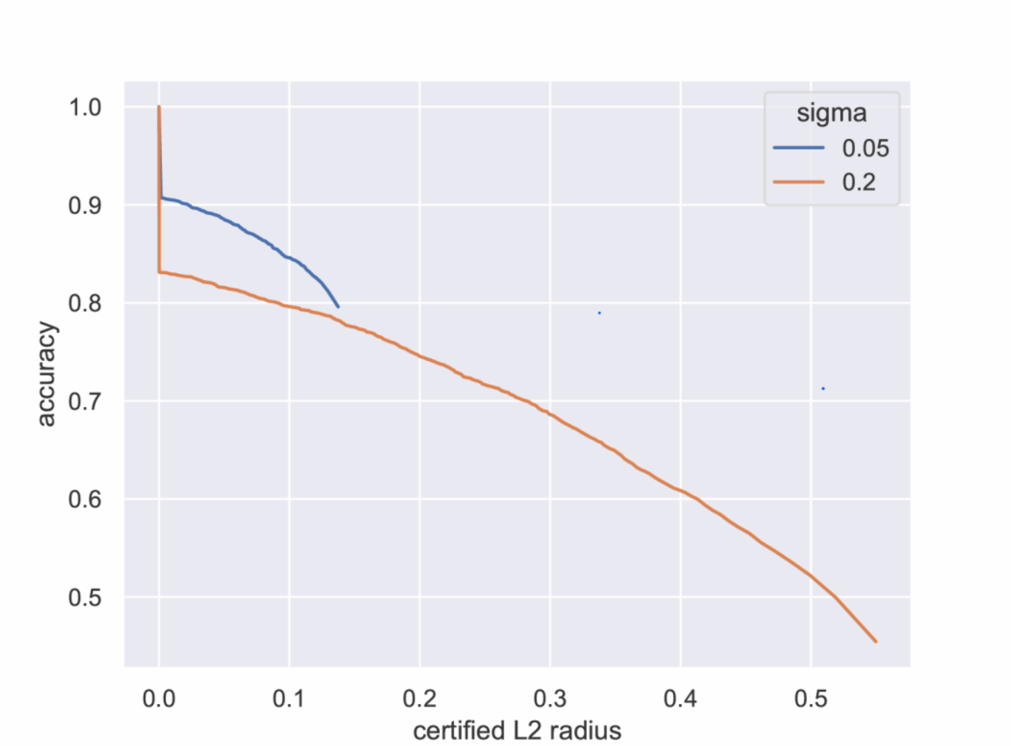
Question 1

Results:

Training	Time(s)	Benign accuracy	PGD success rate
Standard	3099.6511	0.9185	0.8995
Adverserial training m=4	3518.0357	0.8965	0.2838
Adverserial training m=7	3987.4953	0.8953	0.2723

It makes sense that the more we train adversarial training the worst the benign accuracy get and the more robust we are. Although there isn't much of a difference between the m=4 and the m=7 in performance.

Question 2



Looking at the plot, a bigger σ gives you a larger certified radius, but the accuracy on clean inputs gets worse. The orange line for σ=0.2 goes much farther to the right, but it starts at a lower accuracy than the blue line for σ=0.05. This is because a larger σ means adding more noise during the process. More noise makes the model more robust, but it also hurts Benign accuracy.

Question 3

At the start the results were all around 1530 so I could not identify which one has backdoor so I tried to play with the parameters (specifically lambda_c and step_size) and it worked only when I made them bigger.

The backdoor model was model 1 -on class 0. It had 94% success rate. I discovered it as the anomaly lowest norm of trigger.

Both the mask and the trigger and in the zip as well but also trigger and mask screenshots below.

The trigger looks like specific area in the bottom of the picture being blue. Maybe it looks a like a key abit.

The backdoor manages to not affect a lot on the accuracy as we can see:

Accuracy of model 0: 0.9170

Accuracy of model 1: 0.9107

It had 94% success rate so its pretty good backdoor.

The raw main_c.py results:

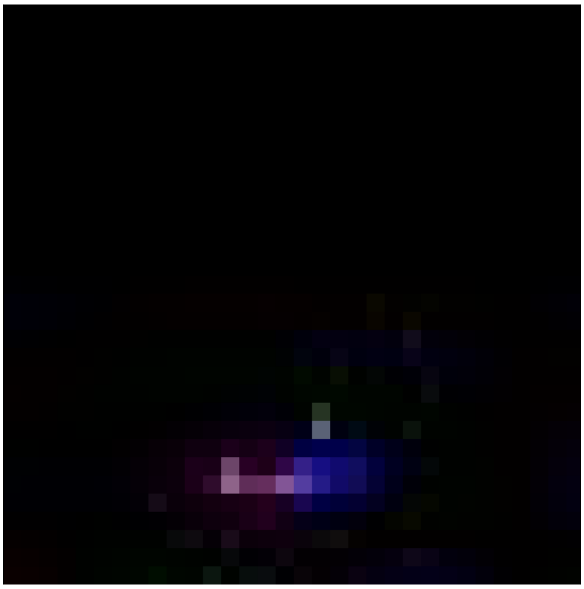
```
Accuracy of model 0: 0.9170

Accuracy of model 1: 0.9107

Norm of trigger targeting class 0 in model 0: 67.9959

Norm of trigger targeting class 1 in model 0: 68.4447

Norm of trigger targeting class 2 in model 0: 94.4063

Norm of trigger targeting class 3 in model 0: 78.5598

Norm of trigger targeting class 0 in model 1: 28.1885

Norm of trigger targeting class 1 in model 1: 64.8257

Norm of trigger targeting class 2 in model 1: 92.2449

Norm of trigger targeting class 3 in model 1: 76.6833
```

Which model is backdoored (0/1)? 1

Which class is the backdoor targeting (0/1/2/3)? 0

Backdoor success rate: 0.9480

The trigger and mask:



Mask-