TML HW2

Eitan Levinzon
209044841
levinzon@mail.tau.ac.il

Q1 - Free adversarial training

3

Output:
Time (in seconds) to complete standard training: 1279.1875
Time (in seconds) to complete free adversarial training: 377.1758

4

Output:
Model accuracy:
- standard : 0.9175
- adv_trained : 0.8500

Success rate of untargeted white-box PGD:
- standard : 0.8995
- adv_trained: 0.4538

We can see that the effect of adversarial training on the bening accuracy is a negligible drop, whereas the effect on robustness is much more meaningful, as the success rate of PGD is much lower.

5

Output:
Time (in seconds) to complete free adversarial training: 307.9146
Model accuracy:
- standard : 0.9175
- adv_trained : 0.7790

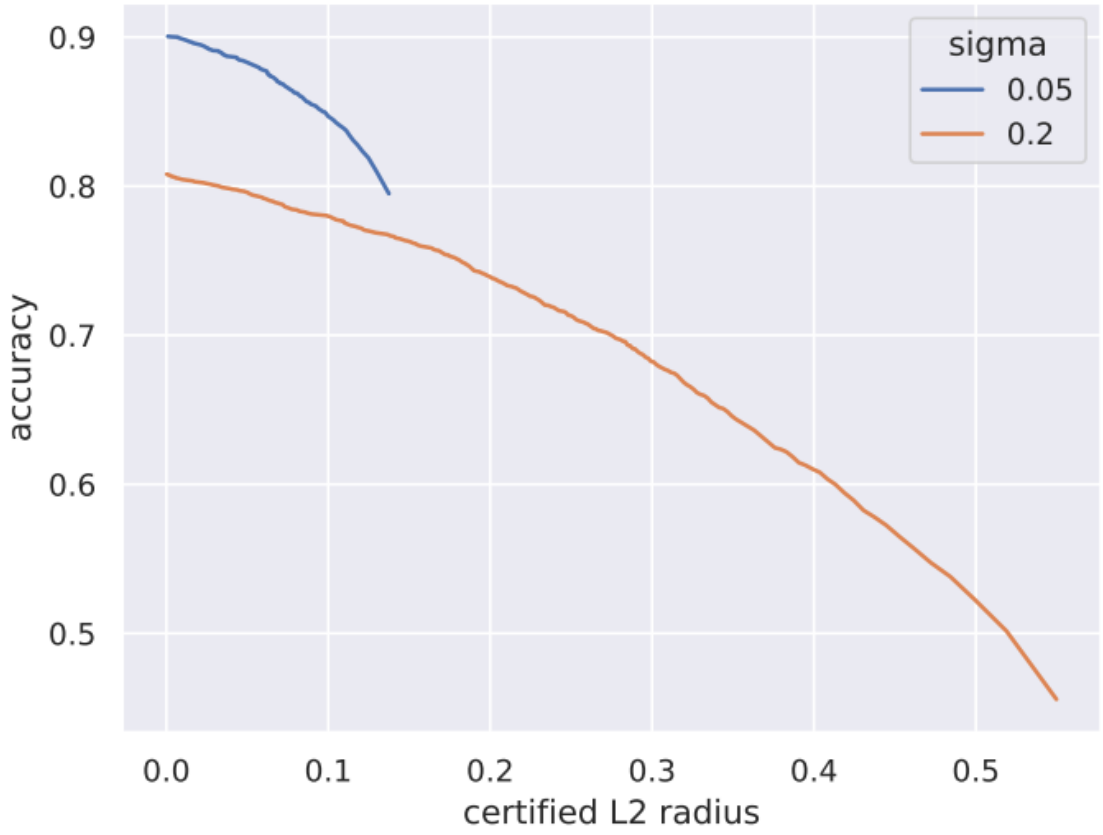
Success rate of untargeted white-box PGD:
- standard : 0.8995
- adv_trained: 0.4475

We can see that increasing m from 4 to 7 decreased (improved) training time. However, this came at the cost of lower bening accuracy, for a small gain in robustness, as the success rate of PGD decreased by just a little.

Q2 - Randomized smoothing

4

The results:



As expected, we can see that with larger values of sigma, we can get larger radii guarantees. In addition, the larger the radius in question, the less samples we were able to guarantee it for. Increasing sigma also gave more varied results. Also, small sigma produces radii with more accuracy than larger ones.

Q3

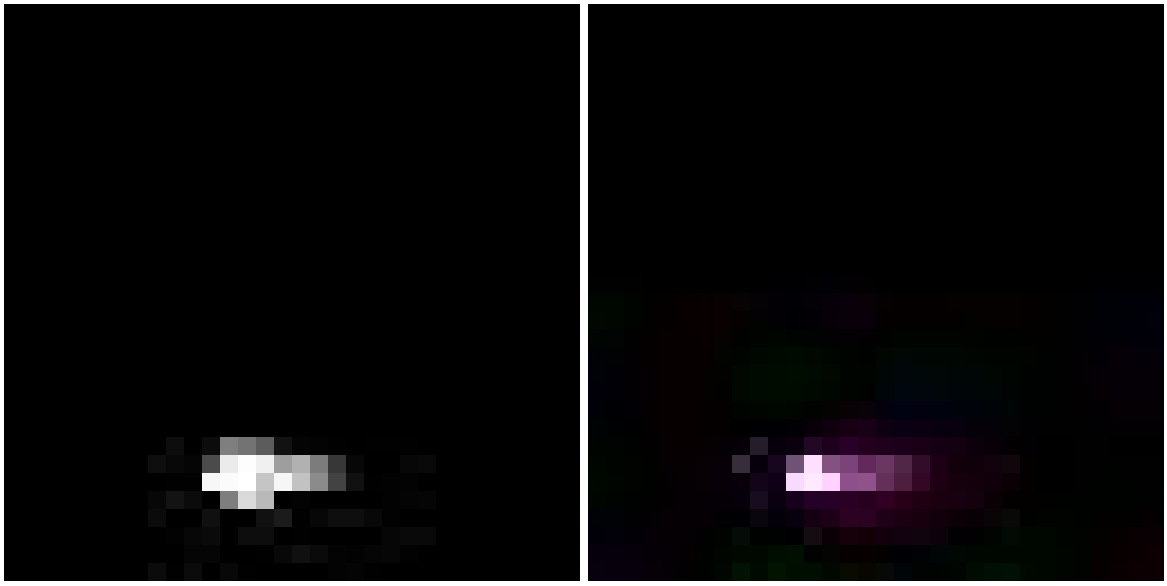
2

Output:
Accuracy of model 0: 0.9168
Accuracy of model 1: 0.9107
Norm of trigger targeting class 0 in model 0: 192.5365
Norm of trigger targeting class 1 in model 0: 132.3472
Norm of trigger targeting class 2 in model 0: 215.8028
Norm of trigger targeting class 3 in model 0: 205.8148
Norm of trigger targeting class 0 in model 1: 54.0424
Norm of trigger targeting class 1 in model 1: 163.6819
Norm of trigger targeting class 2 in model 1: 182.5083
Norm of trigger targeting class 3 in model 1: 211.1251
Which model is backdoored (0/1)? 1
Which class is the backdoor targeting (0/1/2/3)? 0
Backdoor success rate: 0.9982

3

(a)

Mask (left), backdoor (right), scaled up x9.



We can see that the backdoor is indeed quite small, a tiny purple rectangle in the middle bottom part of the image.

(b)

Indeed, the bening accuracy of the backdoored model (1) is almost identical to the non-backdoored model (2). The difference in accuracy between the models can be attributed to randomness as much as to the effect of the backdoor.

(c)

The backdoor is extremely successful at causing misclassification, with an almost perfect success rate of more than 99%.