TrustWorthy Machine Learning
Assignment N.2

May 31, 2025

Bar, Alon
205476013

1 Free adversarial training

1. How much time did it take to train each of the models?

   Ans:

   Time (in seconds) to complete standard training: 208.6928

   Time (in seconds) to complete free adversarial training: 180.1277

2. What was the effect of adversarial training on benign accuracy and robustness?

   Ans: It's shown that the benign accuracy decreased a little and the robustness increased significantly.

   Model accuracy:

   - standard : 0.9210

   - adv_trained : 0.8012

   Success rate of untargeted white-box PGD:

   - standard : 0.8955

   - adv_trained: 0.3840

3. Increase the m parameter of free adversarial training, controlling the number of times a mini-batch is repeated, from 4 to 7, and re-train the model. What is the impact of the change on training time, benign accuracy, and robustness?

Ans: It's shown that it didn't changed alot - less accuracy and more robustness.

Model accuracy:

- adv_trained : 0.7610

Success rate of untargeted white-box PGD:

- adv_trained: 0.3950

2 Randomized smoothing

1. Add the resulting plot (randomized-smoothing-acc-vs-radius.pdf) to the write-up and analyze it. What is the outcome of increasing σ? Explain.

Ans: For $\sigma = 0.05$, the smoothed classifier can certify only a relatively small ℓ_2 radius around each test point, whereas for $\sigma = 0.20$, it can certify significantly larger radii.

This can be explained by the fact that increasing σ results in a smoother classifier. Intuitively, adding more Gaussian noise encourages the classifier to become less sensitive to small input perturbations, effectively pushing decision boundaries farther from the data points.

As a result, a larger perturbation is required to change the predicted label, meaning the certified radius is larger. Consequently, the certified accuracy—defined as the fraction of test examples whose certified radius exceeds a given threshold—decreases more slowly as the threshold increases.

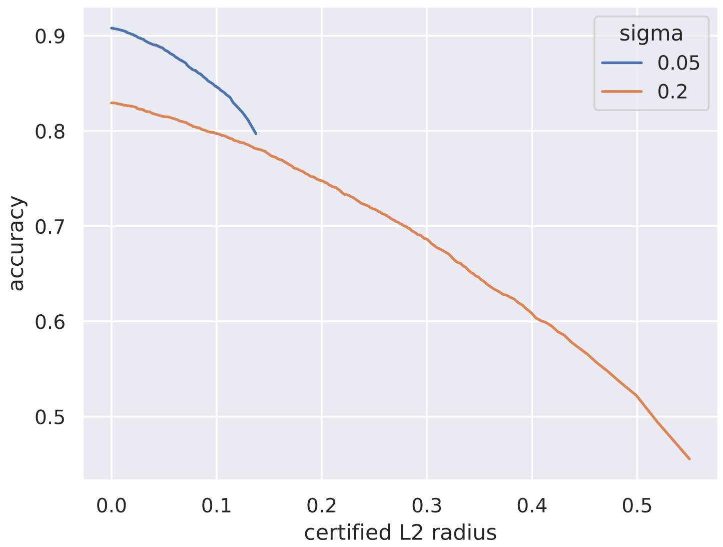

Figure 1: randomized-smoothing-acc-vs-radius

3 Neural Cleanse

1. Which model is backdoored? Which class is the backdoor targeting? Add your responses to the write-up and stdin.

Ans: Accuracy of model 0: 0.9170

Accuracy of model 1: 0.9107

Norm of trigger targeting class 0 in model 0: 171.1302

Norm of trigger targeting class 1 in model 0: 143.9371

Norm of trigger targeting class 2 in model 0: 198.5315

Norm of trigger targeting class 3 in model 0: 185.7896

Norm of trigger targeting class 0 in model 1: 50.8903

Norm of trigger targeting class 1 in model 1: 186.2406

Norm of trigger targeting class 2 in model 1: 188.4279

Norm of trigger targeting class 3 in model 1: 188.9228

Backdoor is at model 1 class 0

Backdoor success rate: 1.0000

2. How does the backdoor look like? Add the images of the mask and trigger to the

write-up.

Ans:

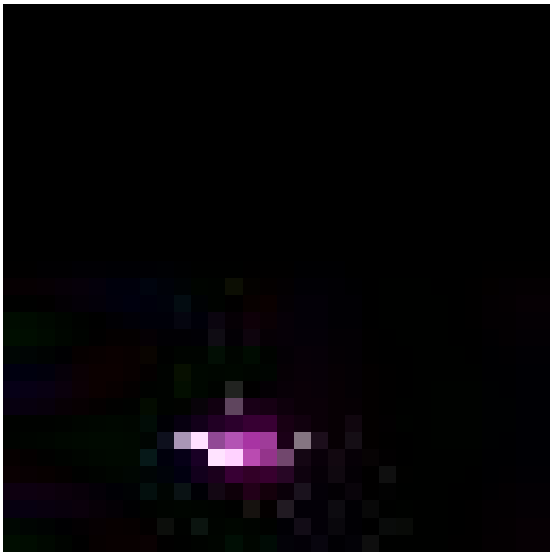

Figure 2: Trigger for the backdoor

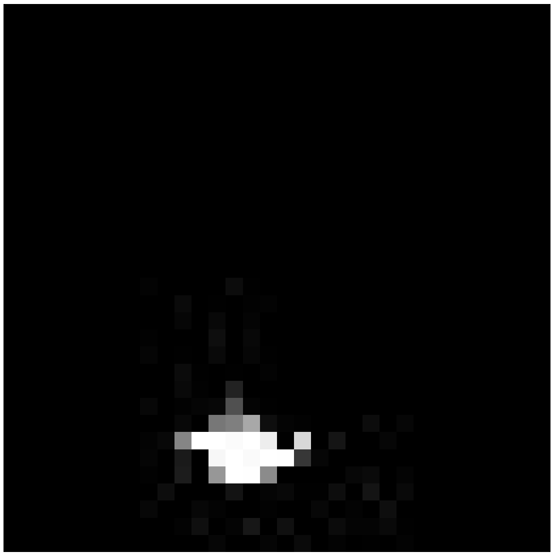

Figure 3: Mask for the backdoor

3. Does the backdoor manage to maintain benign accuracy?

Ans: Yes. The backdoor model achieves almost the exact bengin results as the non-backdoor one (91.7% and 91.07%)

4. How successful is the backdoor at causing misclassification as the target class?

Ans: The backdoor achieves 100% in misclassification as the target class.