Exercise Sheet Deep Learning

Part 3: Attacks and Defenses
Summer 23

This sheet includes a theoretical part and a practical assignment of the third part of the lecture Deep Learning (Attacks and Defenses). It should be handed in as pdf in groups of three via ekvv-Moodle until 1.6.23. at 10.a.m (sharp). Please include a link to the code (e.g. Colab)

name1: Prince Prince

name2: Bhautik Lukhi

name3:

PARTI – THEORY: For the following, you might answer only YES/NO (or abstain), or you can add short arguments (at most two lines per question). If you are not sure, it is better to abstain.

1. For the following attacks, it holds:

yes [✓ **no**] Backdoor attacks try to change the function such that the network can no longer be used.

[✓ **yes**] **no** Data poisoning attacks need access to the training data.

[✓ **yes**] **no** Model inversion attacks can be countered using robust training.

yes [✓ **no**] Universal attacks provide universal backdoors for arbitrary models.

2. For the following adversarial attacks, it holds:

[✓ **yes**] **no** Fast gradient sign and basic iterative method are based on gradient schemes.

yes [✓ **no**] Deepfool uses an adaptive step size.

[✓ **yes**] **no** Universal adversarial perturbations incorporate a regularization against translations of visual objects in scenes.

[✓ **yes**] **no** Adversarial attacks need to have access to the model gradient or estimate it implicitly

3. Defenses against attacks ...

yes [✓ **no**] Data poisoning can be detected by clustering the training set.

[✓ **yes**] **no** Homomorphic privacy enables the public release of models trained on personal data.

yes [✓ **no**] Adversarial training solves the adversarial training loss exactly via computing worst case adversarials in the inner loop.

[✓ **yes**] **no** Certified defenses can rely on Lipschitz bounds.

4. Attacks in reality:

[✓ **yes**] **no** Some training data might deteriorate model accuracy rather than helping it

yes [✓ **no**] Attacks need to address the whole input space

[✓ **yes**] **no** Attacks can be designed such that they hold for different object view points or different scenes in a stream

yes [✓ **no**] Data compression can provably avoid attacks.

5. The following holds:

yes [✓ **no**] Robustified networks have the same accuracy as original ones

yes [✓ **no**] Adversarial risk and minimum perturbation risk are identical for the mean squared error.

[✓ **yes**] **no** Adversarial training complexity strongly depends on the design of adversarial attacks in the inner loop

[✓ **yes**] **no** Adversarial examples do not exist for linear models.

PARTII – PRACTICE: You can use code and models which are publicly available, please clearly reference all sources and tools. Please give a link to your code (e.g. colab). The length of the answer is limited to one page in total for the description of both parts including images. Please provide: short description what you did, how it is done, what is the result. Please be prepared to present the solution in the exercises.

1. Use a deep network for the MNIST data set. Perform at least three different types of targeted attacks on 5 different numbers, including one attack which puts particular effort on the fact that the attacked pattern is indistinguishable from the original one.. Evaluate the performance of the attacks visually (which attack does not change the visual impression) and quantitatively (distance of attack to original sample, success rate of the approach).

2. Use the FashionMNIST data set and a deep model. Create a universal attack, which attacks more than one input at once. Describe how you approach this, and evaluate the performance (success rate). Evaluate whether the universal attack also transfers to other deep learning architectures.