COMP90073 – Security Analytics

Week 11 Workshop

The purpose of this tutorial is to help you further understand adversarial training as well as its limitation.

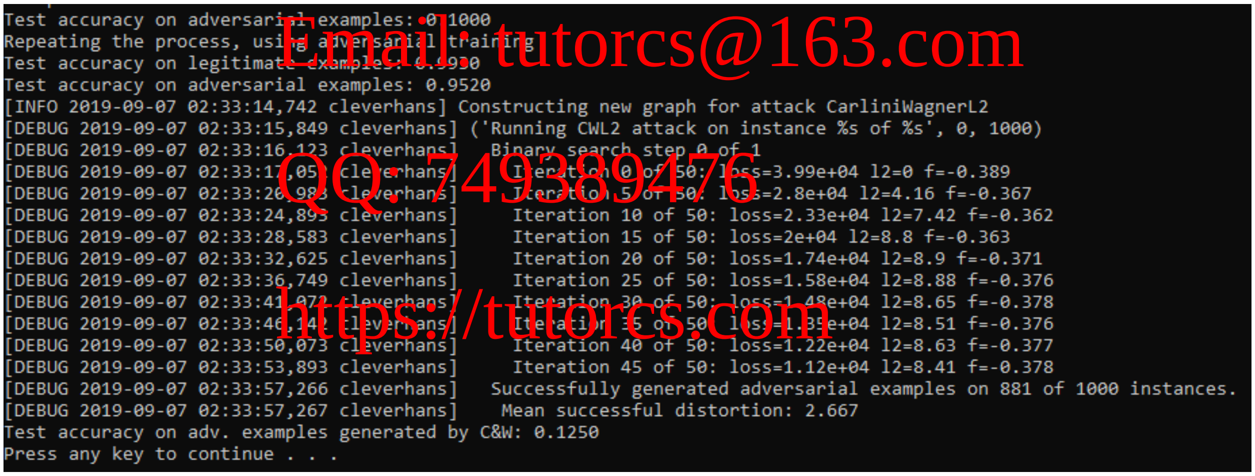
Instructions:

1. Run “mnist_tutorial_[illegible] “cleverhans_tutorials”, and test whether the adversarially trained mode (model2) is als[illegible]al samples generated by the indiscriminate C&W L2 attack. **Hint:** check how ind[illegible]ck is implemented in “mnist_tutorial_cw.py”.

Expected result:

1. The model trained on clean examples (model1) is not robust against adversarial samples generated by the Fast Gradient Sign Method (FGSM) – the accuracy on adversarial samples is around 10%;
2. The model trained on adversarial samples (model2) generated by FGSM is much more robust – the accuracy on adversarial samples increases to over 95%;
3. However, since “model2” is trained on adversarial samples generated by FGSM (a relatively weak form of adversarial attack), it is not robust against adversarial samples generated by the indiscriminate C&W L2 attack – the accuracy goes back to around 10%.

(The percentages may differ on your machine)



```
Test accuracy on adversarial examples: 0.1000
Repeating the process, using adversarial training
Test accuracy on legitimate examples: 0.9930
Test accuracy on adversarial examples: 0.9520
[INFO 2019-09-07 02:33:14,742 cleverhans] Constructing new graph for attack CarliniWagnerL2
[DEBUG 2019-09-07 02:33:15,849 cleverhans] ('Running CWL2 attack on instance %s of %s', 0, 1000)
[DEBUG 2019-09-07 02:33:16,123 cleverhans]   Binary search step 0 of 1
[DEBUG 2019-09-07 02:33:17,059 cleverhans]     Iteration 0 of 50: loss=3.99e+04 l2=0 f=-0.389
[DEBUG 2019-09-07 02:33:20,983 cleverhans]     Iteration 5 of 50: loss=2.8e+04 l2=4.16 f=-0.367
[DEBUG 2019-09-07 02:33:24,895 cleverhans]     Iteration 10 of 50: loss=2.33e+04 l2=7.42 f=-0.362
[DEBUG 2019-09-07 02:33:28,583 cleverhans]     Iteration 15 of 50: loss=2e+04 l2=8.8 f=-0.363
[DEBUG 2019-09-07 02:33:32,625 cleverhans]     Iteration 20 of 50: loss=1.74e+04 l2=8.9 f=-0.371
[DEBUG 2019-09-07 02:33:36,749 cleverhans]     Iteration 25 of 50: loss=1.58e+04 l2=8.88 f=-0.376
[DEBUG 2019-09-07 02:33:41,072 cleverhans]     Iteration 30 of 50: loss=1.48e+04 l2=8.65 f=-0.378
[DEBUG 2019-09-07 02:33:46,142 cleverhans]     Iteration 35 of 50: loss=1.35e+04 l2=8.51 f=-0.376
[DEBUG 2019-09-07 02:33:50,073 cleverhans]     Iteration 40 of 50: loss=1.22e+04 l2=8.63 f=-0.377
[DEBUG 2019-09-07 02:33:53,893 cleverhans]     Iteration 45 of 50: loss=1.12e+04 l2=8.41 f=-0.378
[DEBUG 2019-09-07 02:33:57,266 cleverhans]   Successfully generated adversarial examples on 881 of 1000 instances.
[DEBUG 2019-09-07 02:33:57,267 cleverhans]    Mean successful distortion: 2.667
Test accuracy on adv. examples generated by C&W: 0.1250
Press any key to continue . . .
```