EL-GY-9133 Machine Learning for Cyber-Security

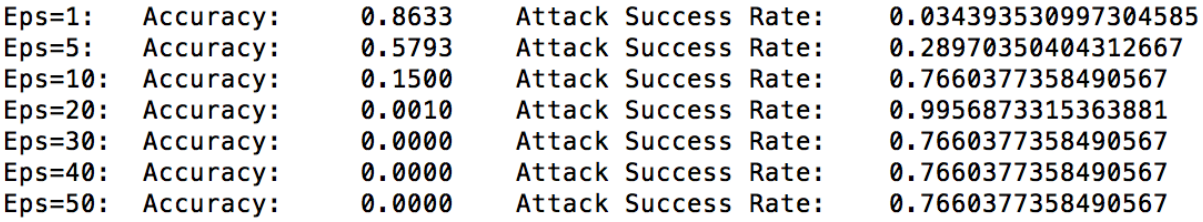
Lab 2: Adversarial Attacks on Deep Neural Networks

Kaihua Cai
kc3172

PART 1: FGSM untargeted attack

Results:

The following shows that for FGSM untargeted, the accuracy rate keep getting lower while attack success rate is higher for higher eps value.

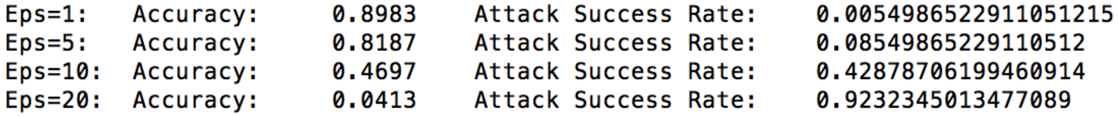


```
Eps=1:   Accuracy:    0.8633   Attack Success Rate:   0.034393530997304585
Eps=5:   Accuracy:    0.5793   Attack Success Rate:   0.28970350404312667
Eps=10:  Accuracy:    0.1500   Attack Success Rate:   0.7660377358490567
Eps=20:  Accuracy:    0.0010   Attack Success Rate:   0.9956873315363881
Eps=30:  Accuracy:    0.0000   Attack Success Rate:   0.7660377358490567
Eps=40:  Accuracy:    0.0000   Attack Success Rate:   0.7660377358490567
Eps=50:  Accuracy:    0.0000   Attack Success Rate:   0.7660377358490567
```

PART 2: Targeted FGSM

Results:

The following graph shows basically the same conclusion as part1, but the attack success rate rises quicker.

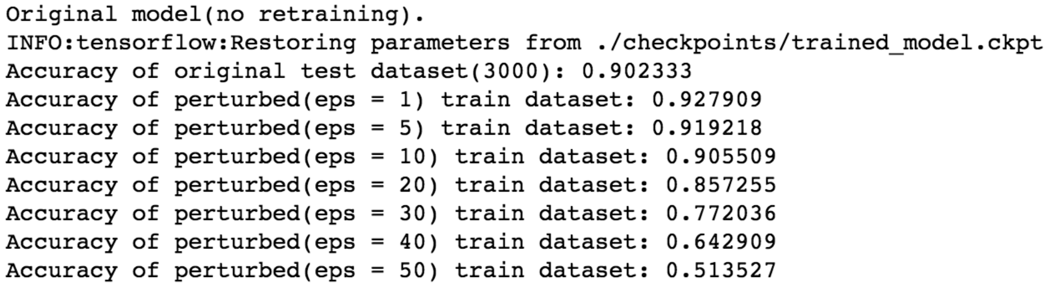


```
Eps=1:   Accuracy:    0.8983   Attack Success Rate:   0.0054986522911051215
Eps=5:   Accuracy:    0.8187   Attack Success Rate:   0.08549865229110512
Eps=10:  Accuracy:    0.4697   Attack Success Rate:   0.42878706199460914
Eps=20:  Accuracy:    0.0413   Attack Success Rate:   0.9232345013477089
```

PART 3 & PART 4: Retrained Untargeted FGSM

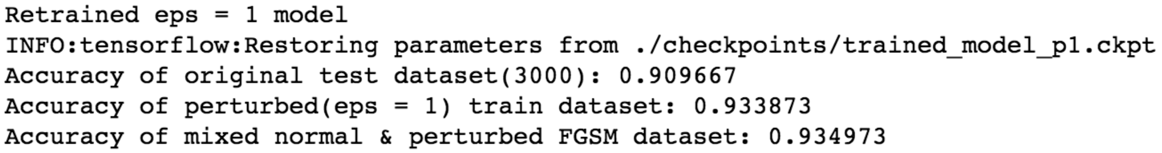
Results:

The following graph is the accuracy result of the original model (without being retrained), it is used as a comparison graph. It includes the different values of eps. This shows that the original model is very much affected by the changes in eps.

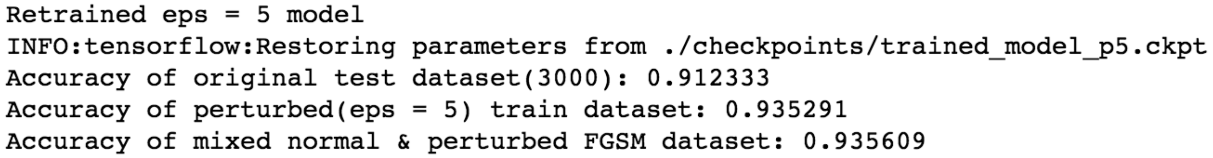


```
Original model(no retraining).
INFO:tensorflow:Restoring parameters from ./checkpoints/trained_model.ckpt
Accuracy of original test dataset(3000): 0.902333
Accuracy of perturbed(eps = 1) train dataset: 0.927909
Accuracy of perturbed(eps = 5) train dataset: 0.919218
Accuracy of perturbed(eps = 10) train dataset: 0.905509
Accuracy of perturbed(eps = 20) train dataset: 0.857255
Accuracy of perturbed(eps = 30) train dataset: 0.772036
Accuracy of perturbed(eps = 40) train dataset: 0.642909
Accuracy of perturbed(eps = 50) train dataset: 0.513527
```

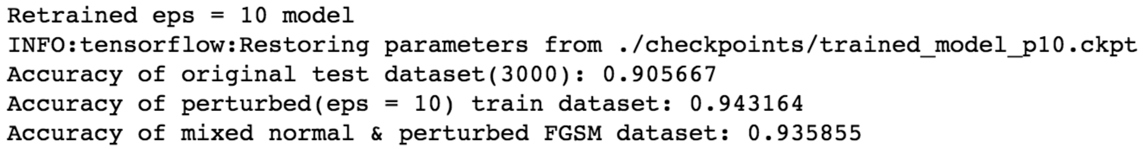
Now the following graphs are the resulting accuracy for every eps, the third accuracy is simply applying the above 2 datasets to the newly trained model. As it can be seen from the below 6 graphs, no matter how the eps value changes, the perturbed dataset can always remain at a high value. It can be even do better than the accuracy of the original dataset. This is, I believe, caused by the perturbations actually make it easier for the new retrained model to classify the 10 digits.



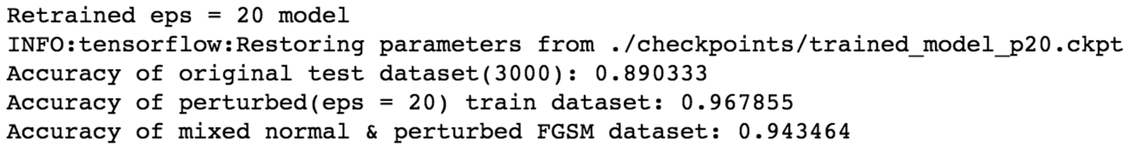
```
Retrained eps = 1 model
INFO:tensorflow:Restoring parameters from ./checkpoints/trained_model_p1.ckpt
Accuracy of original test dataset(3000): 0.909667
Accuracy of perturbed(eps = 1) train dataset: 0.933873
Accuracy of mixed normal & perturbed FGSM dataset: 0.934973
```



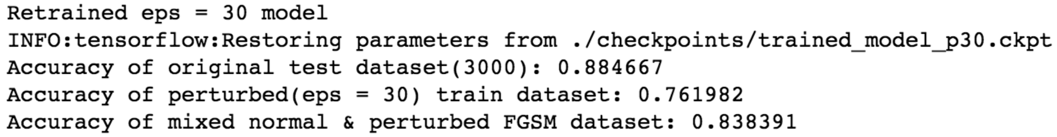
```
Retrained eps = 5 model
INFO:tensorflow:Restoring parameters from ./checkpoints/trained_model_p5.ckpt
Accuracy of original test dataset(3000): 0.912333
Accuracy of perturbed(eps = 5) train dataset: 0.935291
Accuracy of mixed normal & perturbed FGSM dataset: 0.935609
```



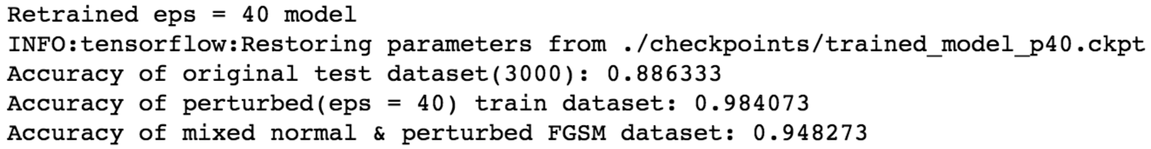
```
Retrained eps = 10 model
INFO:tensorflow:Restoring parameters from ./checkpoints/trained_model_p10.ckpt
Accuracy of original test dataset(3000): 0.905667
Accuracy of perturbed(eps = 10) train dataset: 0.943164
Accuracy of mixed normal & perturbed FGSM dataset: 0.935855
```



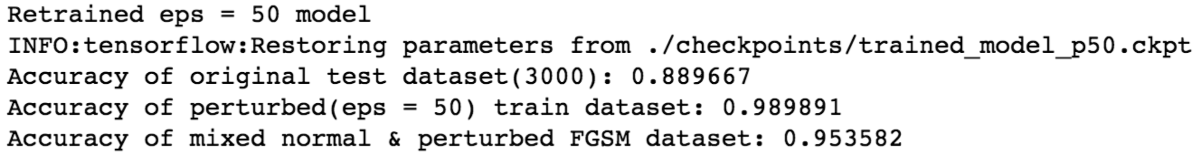
```
Retrained eps = 20 model
INFO:tensorflow:Restoring parameters from ./checkpoints/trained_model_p20.ckpt
Accuracy of original test dataset(3000): 0.890333
Accuracy of perturbed(eps = 20) train dataset: 0.967855
Accuracy of mixed normal & perturbed FGSM dataset: 0.943464
```



```
Retrained eps = 30 model
INFO:tensorflow:Restoring parameters from ./checkpoints/trained_model_p30.ckpt
Accuracy of original test dataset(3000): 0.884667
Accuracy of perturbed(eps = 30) train dataset: 0.761982
Accuracy of mixed normal & perturbed FGSM dataset: 0.838391
```



```
Retrained eps = 40 model
INFO:tensorflow:Restoring parameters from ./checkpoints/trained_model_p40.ckpt
Accuracy of original test dataset(3000): 0.886333
Accuracy of perturbed(eps = 40) train dataset: 0.984073
Accuracy of mixed normal & perturbed FGSM dataset: 0.948273
```



```
Retrained eps = 50 model
INFO:tensorflow:Restoring parameters from ./checkpoints/trained_model_p50.ckpt
Accuracy of original test dataset(3000): 0.889667
Accuracy of perturbed(eps = 50) train dataset: 0.989891
Accuracy of mixed normal & perturbed FGSM dataset: 0.953582
```