**Lab 3 Report: Inference Attacks on Deep Neural Networks**

Sui Huang [sh4507@nyu.edu](mailto:sh4507@nyu.edu)

1. **Methodology used to implement the AM algorithm:**

I used gradient decent mythology. The code realized in function named ‘image\_optimizer’. Function image\_optimizer(image\_old, i, lr, print\_=True) takes 4 arguments:

**image\_old** is the initial image I generated using either random or mean value, and this image is generated using function named ‘generate\_x’.

**i:** is the target label we want the initial picture to be updated. For example, i=5, then the target label will be [0,0,0,0,0,1,0,0,0,0]

**lr:** Learing rate

**print\_:** whether I want to print the result( disabled in Q4 because it’s too much)

**Code for optimizing gradient and update the image:**

if (prob\_new[:,i]-prob\_old[:,i]>= 1e-8): (stops when the probability does not increase any more)

gradient = tf.gradients(xs = x, ys = cost) (compute gradient)

image\_new = tf.clip\_by\_value(x - tf.fill([784], step\_size)/gradient,0,1) ***(image = image – lr/gradient, clip from 0 to 1 because I don’t want the value exceed the image range)***

image\_new, gradient = sess.run([image\_new , gradient], feed\_dict={x: image, y: generate\_y(j).reshape(1,10)}) (session run)

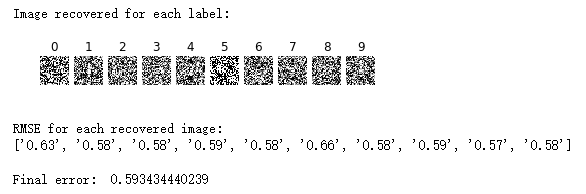
image\_new = image\_new[0,:,:] (decrease one dimension)

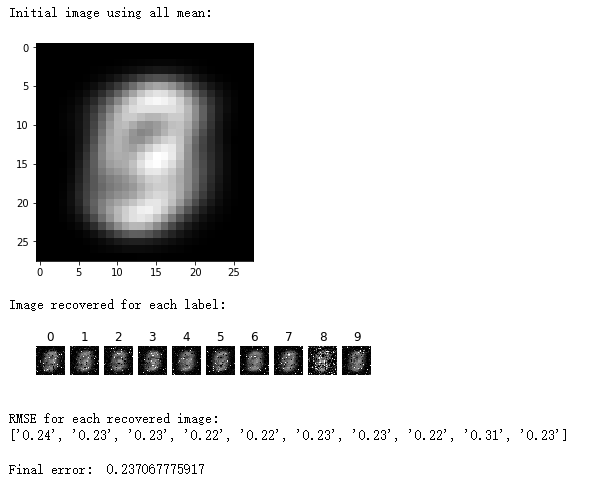
image = image\_new (update image to calculate gradients)

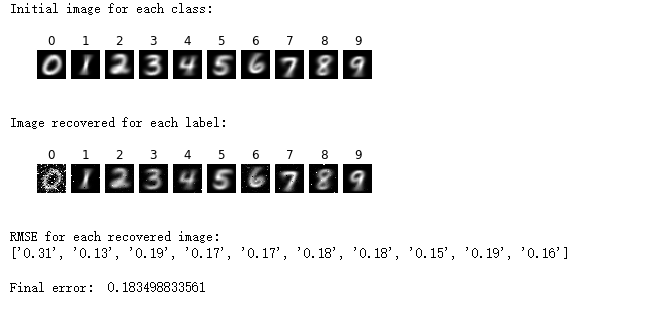
I did not use adaptive learning rate. But used different learning rate (as step size) for different initial images.

PART 1: 1e-1，PART 2：1e-2，PART 3&Q4：1e-5

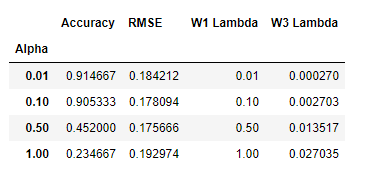
1. **Recovered images for each class (10 per problem) and the computed RMSE for that image. Also indicate the final error.**

**PART 1:** 

**PART 2:** 

**PART 3:** 

1. **Tabulate the prediction accuracy versus the final error for each value of λ. Also indicate the value of λ used.**



1. **Python code along with any instructions required to execute the code.**

All codes and results can be reproduced using the note book named:’ Inference\_ Attacks\_on\_Deep\_Neural Networks’

1. **Some thoughts**

I think I did not do very well in part4. And the result is that adding noise can not increase RMSE in a very large level. Maybe I did the DP wrong. It’s sad because I don’t have more time.