2019-1 Deep Learning Homework #1

Dae-Ki Kang

March 21, 2019

(Deadline : March 28)

You should train a neural network and perform experiments from the questions below. You should do a white-box attack for Question 1. You may reference to the following link for this homework: https://github.com/bethgelab/foolbox or https://github.com/tensorflow/cleverhans

1. Generate 10 adversarial images (1 per class) from each dataset (i.e. MNIST, Fashion-MNIST and CIFAR-10 dataset) by using different attack techniques.
   (a) Fast Gradient Sign Method
   (b) Basic Iterative Method
   (c) Momentum Iterative Method
   (d) Saliency Map Attack
   (e) Carlini & Wagner L2 Attack
2. Generate adversarial noises (in image form) from the difference between the adversarial images (generated from Question 1) and original images. Afer that, analyze the noises in terms of L_2-distance.
3. Analyze the result after applying a defense technique, namely advesarial training.
4. Perform and analyze between a white-box attack and a black-box attack by using Basic Iterative Method attack technique.