Trust Region Based Adversarial Attack on Neural Networks

Abraham Jose

1 Summary

The author tries to provide a better method for adversarial attacks which requires less iterations and which does not involves time-consuming hyper-parameter tuning based on trust region based adversarial attacks which is efficient. The idea of TR based perturbation is to iteratively select the trusted radius to find the adversarial perturbation within this region such that the probability of an incorrect class becomes maximum. Trust region perturbation is found to be effective and effient compared to CW (Carlini-Wagner) attack and DeepFool. The methos is tested in Cifar-10 and ImageNet dataset on classification task using neural netowrk models AlexNet, ResNet-50, VGG-16, and DenseNet-121.

2 Good points

The attack based on trust regions makes it possible to do less perturbation for attack, which reduces the perturbation required by a factor of 4, when compared to DeepFool and CW requires slightly more perturbations. They uses relative perturbation as the metric to validate and study their method which gives us absolute comparison with the existing methods in the field. Trust Region method can adaptively choose the magnitude of perturbation which helps with hyper-parameter tuning. Also, they have tested their model using AlexNet, ResNet-50, VGG-16, and DenseNet-121 which covers all types of neural network models having shallow, deep and multi-layered neural networks with and without skip connections.

3 Weak points

The model is compared w.r.t. the relative perturbation alone. It is important that when an image is perturbed, we should look at the confidence gap between the perturbed class and actual class confidence which will be another parameter recording how effective the method is. This is not compared across the adversarial attacking methods. It might be possible that the results from the DeepFool or CW might give a bigger gap between perturbed image class and the actual class. This is important if the adversarial method is used to increase generalization of the network and for making models robust to attacks.

4 Questions

5 Ideas

As described in the Weak section, we can use the confidence gap between the perturbed class and the actual class or the second best class to calculate the effectiveness of an adversarial attack.