NIPS 2017:

Defense Against Adversarial Attack

Contents

1 Goal & Background

2 Data

3 Our Work

4 Preliminary Results

Goal & Background

Goal

Our goal is to construct a robust machine learning classifier for antagonistic examples, that is, to correctly classify input samples that have been slightly modified.

Background

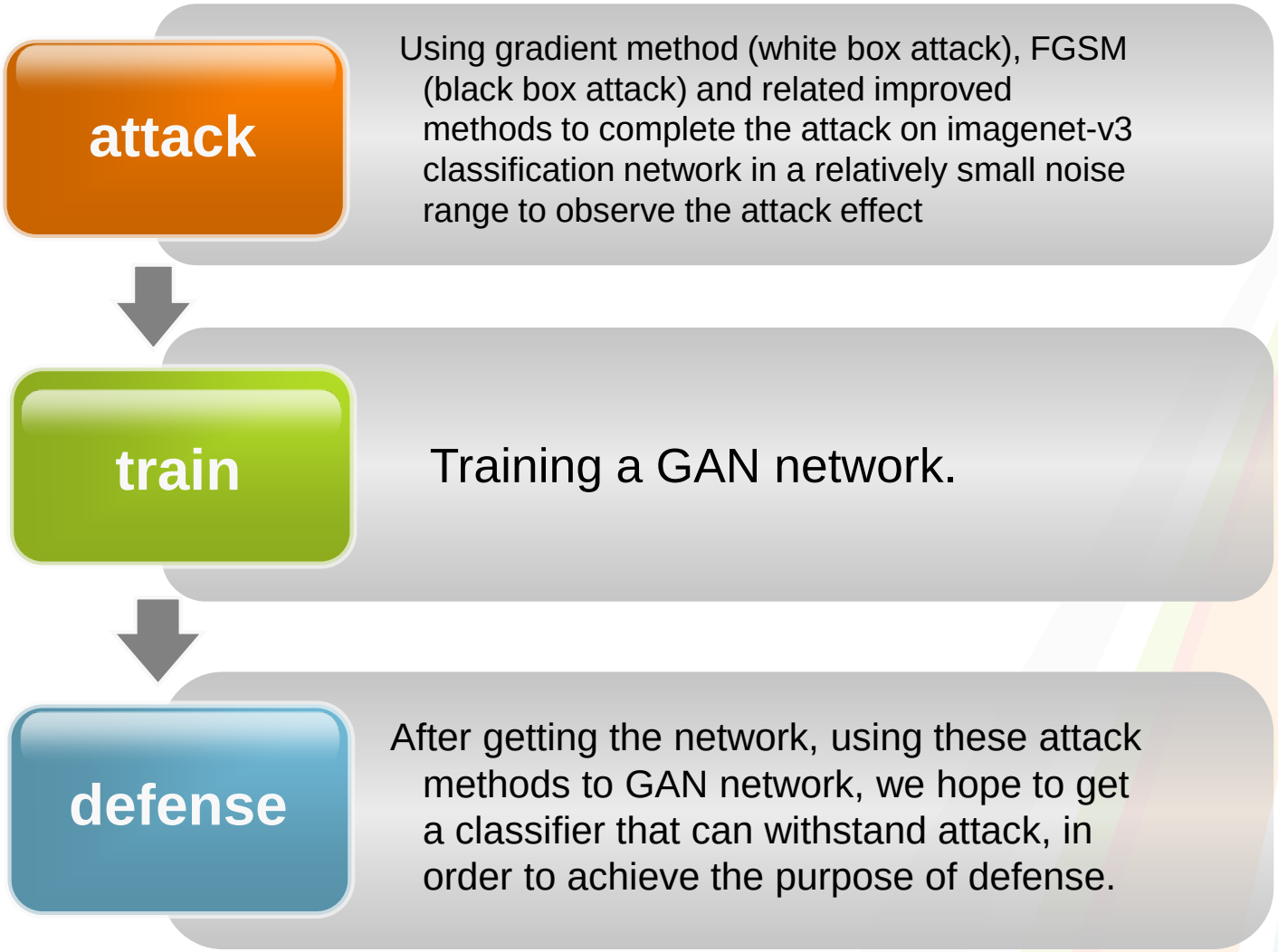
The purpose of counterattack is to induce the machine learning classifier to classify errors. The adversarial examples involved can cause security problems that cannot be ignored, because attackers can attack machine learning systems without accessing the underlying model.

Data

Our Data

We will use 949 images from the Imagenet data set provided by the kaggle competition website, each of which is 299*299 images with 3 channels. In addition, each of these images has its correct tag (category 1001) and target tag to be attacked.

Our Work


attack
Using gradient method (white box attack), FGSM (black box attack) and related improved methods to complete the attack on imagenet-v3 classification network in a relatively small noise range to observe the attack effect
train
Training a GAN network.
defense
After getting the network, using these attack methods to GAN network, we hope to get a classifier that can withstand attack, in order to achieve the purpose of defense.

Preliminary Results

We first completed the attack on the original classification network. The attack effect is quite good, and the success rate of white-box attack is 100%. This actually shows that the existing image classification network has little resistance to countermeasure samples, and it really needs to be defended.

Some pictures of the effects of white box attack are shown next page:

White box attack effect show

figure 1-1 original image

figure 1-2 Attack-generated image

figure 1-3 Image of the target class

figure 2-1 original image

figure 2-2 Attack-generated image

figure 2-3 Image of the target class

Preliminary Results

For the original inception-v3 network, the effect of black-box attack is worse, only 60.379% of the success rate, may need to be improved in the follow-up work, in order to get a better black-box attack to test the strength of GAN network.

Literature & Team

Literature

https://arxiv.org/pdf/1711.00117.pdf

https://arxiv.org/pdf/1608.04644.pdf

https://arxiv.org/pdf/1710.10766.pdf

Team

于淼 1801210068　李兴远 1801210051　赵越 1801210072

北 京 大 学 数 学 科 学 学 院

Thank You!