Foggy Eye:Adversarial Defense Through Image Preprocessing / Kunlin Team in CAAD 2018

Kunlin Liu, University of Science and Technology in China
Xiaoyu Ye, Sensetime

September 2018

Abstract

Deep neural networks have performed high accuracy on image recognition task in recent years. However, they are extremely vulnerable to adversarial attacks. Adversarial examples which are obtained by adding delicately crafted distortions onto original legal inputs can mislead Deep neural networks easily. In this paper, we proposed a method to defense against adversarial attacks. By combining the proposed preprocessing method with an adversarially trained model, it ranked No.5 in the CAAD2018 defense sub-competition.The code is public available at this https URL.(`https://github.com/0three/CAAD-2018-Kunlin`)

1 Introduction

Convolutional neural networks have demonstrated high accuracy on various tasks in recent years. However, they are extremely vulnerable to adversarial examples. For example, imperceptible perturbations added to clean images can cause convolutional neural networks to fail. In this paper, we propose to utilize gaussian blur and weiner filtering at inference time to mitigate adversarial effects. Our method provides the following advantages: 1) easy to reappearence and extend, 2) compatiable to popular image classification networks. By combining the proposed preprocessing method with an adversarially trained model, it achieves a score of 0.0064 (ranked No.5) in the CAAD2018 adversarial defense sub-competition, which is far better than using adversarial training alone with a score of 0.0002 (ranked No.24). The code is public available at this https URL(`https://github.com/0three/CAAD-2018-Kunlin`)

2 Background

Most existing machine learning classifiers are highly vulnerable to adversarial examples. An adversarial example is a sample of input data which has been modified very slightly in a way that is intended to cause a machine learning classifier to misclassify it. In many cases, these modifications can be so subtle that a human observer does not even notice the modification at all, yet the classifier still makes a mistake.

Adversarial examples pose security concerns because they could be used to perform an attack on machine learning systems, even if the adversary has no access to the underlying model.

To accelerate research on adversarial examples, GeekPwn is partnering with Alexey Kurakin from Google Brain and Dawn Song from UC Berkeley EECS to organize Competition on Adversarial Attacks and Defenses 2018 (CAAD2018).

One of sub-competitions in the CAAD is defense against adversarial attack. The goal of defense is to build machine learning classifier which is robust to adversarial example, i.e. can classify adversarial images correctly.

3 Approach

3.1 Adversarial training

Madry et al.(2017)[3] suggests that PGD(Projected Gradient Descent) is a universal first order adversary – in other words, developing robustness against PGD attacks also implies resistance against many other first order attacks. We use adversarial training with PGD as the underlying basis for our methods:

$$\arg\min_{\theta} E_{(x,y)\in\hat{p}_{data}}\left(\max_{\delta\in S}\mathrm{L}(\theta,x+\delta,y)\right) \tag{1}$$

where $\hat{p}_{data}$ is the underlying training data distribution, $\mathrm{L}(\theta, x+\delta, y)$ is a loss function at data point x which has true classy for a model with parameters θ, and the maximization with respect to δ is approximated using noisy BIM.

As recommended by Goodfellow et al. (2014)[1];Kurakin et al. (2017a)[2], when we train on a mixture of clean and adversarial examples, we can achieve better performance:

$$\arg\min_{\theta}\left[E_{(x,y)\in\hat{p}_{data}}\left(\max_{\delta\in S}\mathrm{L}(\theta,x+\delta,y)\right)+E_{(x,y)\in\hat{p}_{data}}\left(\mathrm{L}(\theta,x,y)\right)\right] \tag{2}$$

3.2 Preprocessing

Cihang Xie utilize randomization to mitigate adversarial effects and achieve No.2 in NIPS2017.[4] A good preprocessing method can effectively reduce the effects of adversarial examples.The main ideal of the defense is to utilize preprocessing to defend adversarial examples.

The pretreatment can be divided into two part, adding gaussian blur and weiner filtering.

In CAAD2018, the input image is 299x299x3 and perturbation limit is 32. Obviously, general image processing method can not work effectively with such a huge perturbation limit. So we take some extreme measures. Pipeline is below:

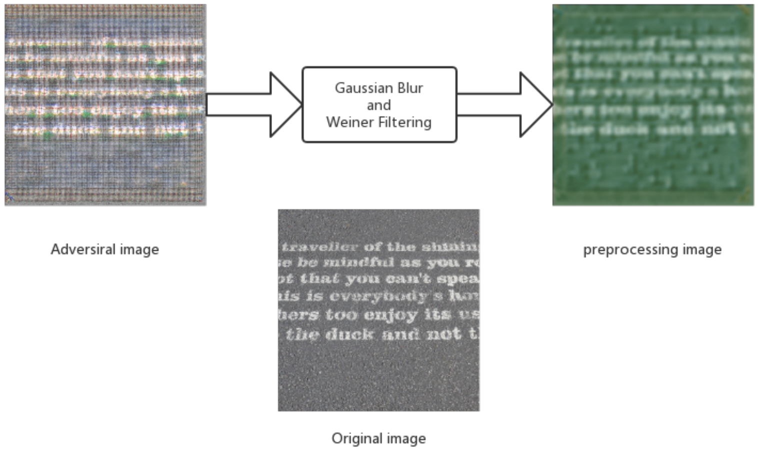

Figure 1: Pipeline

3.2.1 Adding gaussian blur

In image processing, a Gaussian blur (also known as Gaussian smoothing) is the result of blurring an image by a Gaussian function (named after mathematician and scientist Carl Friedrich Gauss). It is a widely used effect in graphics software, typically to reduce image noise and reduce detail. Gaussian blur can interfere the structure of crafted adversarial examples.

3.2.2 Weiner filter

In signal processing, the Wiener filter is a filter used to produce an estimate of a desired or target random process by linear time-invariant (LTI) filtering of an observed noisy process, assuming known stationary signal and noise spectra, and additive noise. The Wiener filter minimizes the mean square error between the estimated random process and the desired process.

4 Conclusion

In this paper, we propose a preprocessing method to mitigate adversarial effects. The proposed method is compatible to popular network structures and can serve as a basic module for defense against adversarial examples. By preprocessing input image, it achieves a score of 0:0064(ranked No.5) in the CAAD2018 defense against adversarial attacks defense challenge,which is far better than using adversarial training alone with a normalized score of 0.0002(ranked No.24).The code is public available at this https URL.(`https://github.com/0three/CAAD-2018-Kunlin`)

References

[1] Shlens Jonathon Goodfellow, Ian J and Chris-tian. Szegedy. Explaining and harnessing adversarial examples. *arXiv*, 2014.

[2] Goodfellow Ian Kurakin, Alexey and Samy. Bengio. Ad-versarial machine learning at scale. *ICLR*, 2017.

[3] Makelov A. Schmidt L. Tsipras D. andVladu A. Madry, A. Towards deep learning models resistant to adversarial attacks. *Technical report,arXiv*, 2017.

[4] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.