Adversarial Examples on Image Recognition with Deep Neural Networks

Ran Liu, Zheng Qu, Yong Wu

Background

After AlexNet was devised in 2012, deep neural networks classifiers have been widely exploited in industry, such as autonomous driving, mobile check deposit. The accuracy of deep neural network is greater than any traditional classifier like SVM, logistic regression. Nowadays the convolutional neural networks architectures become deeper and deeper, such as VGG and ResNet, and the accuracy becomes better and better, even better than human experts in ImageNet recognition. Whereas, recently, adversarial examples[1-4] have been found to be able to cause almost all machine learning classifiers to misclassify, especially those with linearity. The attacker is able to generate adversarial examples which is very close to the original ones, the adversarial examples are indistinguishable to the human eyes, but a variety of classifiers would misclassify them, in some cases, the adversarial examples are unidentifiable to human, but machine learning classifiers could be led to make prediction with high confidence.

Dataset

In this project, we will use ImageNet dataset, the commonly used dataset for image processing and benchmarking. Fortunately, Google has launched three adversarial attack competition[5] three months ago and they are still ongoing. The public dataset in the competitions is also from ImageNet, we plan to use it in our project. All the images are

299x299 pixels (RGB) with 1000 classes.

Our Project

In this project, we will pay most of our attention on the adversarial attack toward image recognition, trying to figure out a general adversarial attack, especially on the state-of-the-art classifier with convolutional neural networks like inception-v3. We will benchmark our performance. In our work, we will show how the adversarial attack is going to be effective, and why it works. And we will propose a new algorithm based on fast gradient method to generate adversarial examples with acceptable confidence.

References

[1] Huang, Sandy, et al. "Adversarial attacks on neural network policies." *arXiv preprint arXiv:1702.02284* (2017).

[2] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).

[3] Tramèr, Florian, et al. "Ensemble Adversarial Training: Attacks and Defenses." *arXiv preprint arXiv:1705.07204* (2017).

[4] Shen, Shiwei, et al. "AE-GAN: adversarial eliminating with GAN." *arXiv preprint arXiv:1707.05474* (2017).

[5] NIPS 2017: Non-targeted Adversarial Attack, https://www.kaggle.com/c/nips-2017-non-targeted-adversarial-attack