AdvBox:A toolbox to take off the emperor's new clothes of neural networks

Dou Goodman [1] **Hao Xin** [1] **Wang Yang** [1]

Abstract

(Szegedy et al., 2013). discovered an intriguing properties of deep neural networks in the context of image classification for the first time. They showed that despite the state-of-the-art deep networks are surprisingly susceptible to adversarial attacks in the form of small perturbations to images that remain (almost) imperceptible to human vision system. These perturbations are found by optimizing the input to maximize the prediction error and the images modified by these perturbations are called as adversarial examples. The profound implications of these results triggered a wide interest of researchers in adversarial attacks and their defenses for deep learning in general. Advbox is a toolbox to generate adversarial examples that fool neural networks in PaddlePaddle, PyTorch, Caffe2, MxNet, Keras, tensorFlow and Advbox can benchmark the robustness of machine learning models. Advbox give a command line tool to generate adversarial examples with Zero-Coding.Code and examples available at https://github.com/baidu/AdvBox,licensed under the Apache License 2.0.

1. Adversarial Examples Overview

(Szegedy et al., 2013) discovered an intriguing properties of deep neural networks in the context of image classification for the first time. They showed that despite the state-of-the-art deep networks are surprisingly susceptible to adversarial attacks in the form of small perturbations to images that remain (almost) imperceptible to human vision system. These perturbations are found by optimizing the input to maximize the prediction error and the images modified by these perturbations are called as adversarial examples. The profound implications of these results triggered a wide interest of researchers in adversarial attacks and their defenses for deep learning in general.The initially involved computer vision task is image classification. For that, a variety of attacking methods have been proposed, such as FGSM of (Goodfellow et al., 2014), deepfool of (Moosavidezfooli et al., 2016) ,CW of (Carlini & Wagner, 2017) and so on. Transferability of adversarial examples was first examined by (Szegedy et al., 2013), which studied the transferability between different models trained over the same dataset.The study of transferability was followed by (Goodfellow et al., 2014), which attributed the phe nomenon of transferability to the reason that the adversarial perturbation is highly aligned with the weight vector of the model. Again, this hypothesis was tested using MNIST and CIFAR-10 datasets. We show that this is not the case for models trained over ImageNet.(Liu et al., 2016) show that while existing approaches are effective to generate non-targeted transferable adversarial examples , only few targeted adversarial examples generated by existing methods can transfer .They propose novel ensemble-based approaches to generate adversarial examples . Their approaches enable a large portion of targeted adversarial examples to transfer among multiple models for the first time.(Wei et al., 2018) propose the Unified and Efficient Adversary (UEA) for attacking image and video object detection. UEA is the first attacking method that can not only efficiently deal with both image and video data, but also simultaneously fool the proposal based detectors and regression based detectors.

[1]Baidu X-Lab. Correspondence to: Dou Goodman <liu.yan@baidu.com>.

2. AdvBox Overview

Advbox is a toolbox to generate adversarial examples that fool neural networks in PaddlePaddle,PyTorch,Caffe2,MxNet,Keras,TensorFlow and Advbox can benchmark the robustness of machine learning models. Advbox give a command line tool to generate adversarial examples with Zero-Coding.Advbox improves upon the existing Python package cleverhans by (Papernot et al., 2016) in three important aspects:

1. Advbox is a toolbox to generate adversarial examples that fool neural networks in PaddlePaddle,PyTorch,Caffe2,MxNet,Keras,TensorFlow and Advbox can benchmark the robustness of machine learning models.

2. Advbox give a command line tool to generate adver-

sarial examples with Zero-Coding.

3. Advbox also support white-box and black-box attacks and defence algorithms.

2.1. advbox.attack

Advbox implements several popular adversarial attacks which search adversarial examples. Each attack method uses a distance measure(L1, L2, etc.) to quantify the size of adversarial perturbations. Advbox is easy to craft adversarial example as some attack methods could perform internal hyperparameter tuning to find the minimum perturbation.

2.2. advbox.model

Advbox implements interfaces to PaddlePaddle. Additionally, other deep learning framworks such as TensorFlow can also be defined and employed. The module is use to compute predictions and gradients for given inputs in a specific framework.

2.3. advbox.adversary

Adversary contains the original object, the target and the adversarial examples. It provides the misclassification as the criterion to accept a adversarial example.

3. Implemented A White-box attack methods

- L-BFGS
- FGSM
- BIM
- ILCM
- MI-FGSM
- JSMA
- DeepFool
- C/W

4. Implemented A Black-box attack methods

- Single Pixel Attack
- Local Search Attack

5. Implemented A Defense methods

- Feature Fqueezing
- Spatial Smoothing
- Label Smoothing
- Gaussian Augmentation
- Adversarial Training
- Thermometer Encoding

6. Implemented Attack AI application

Attack Face recognition

References

Carlini, Nicholas and Wagner, David. Towards evaluating the robustness of neural networks. In *Security and Privacy*, 2017.

Goodfellow, Ian J, Shlens, Jonathon, and Szegedy, Christian. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Liu, Yanpei, Chen, Xinyun, Chang, Liu, and Song, Dawn. Delving into transferable adversarial examples and black-box attacks. 2016.

Moosavidezfooli, Seyed Mohsen, Fawzi, Alhussein, and Frossard, Pascal. Deepfool: a simple and accurate method to fool deep neural networks. In *Computer Vision and Pattern Recognition*, 2016.

Papernot, Nicolas, Goodfellow, Ian, Sheatsley, Ryan, Feinman, Reuben, and Mcdaniel, Patrick. cleverhans v1.0.0: an adversarial machine learning library. 2016.

Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian, and Fergus, Rob. Intriguing properties of neural networks. *Computer Science*, 2013.

Wei, Xingxing, Liang, Siyuan, Cao, Xiaochun, and Zhu, Jun. Transferable adversarial attacks for image and video object detection. *arXiv preprint arXiv:1811.12641*, 2018.