Advbox: a toolbox to generate adversarial examples that fool neural networks

Dou Goodman 1, Hao Xin 1, Wang Yang 1, Xiong Junfeng 1, Zhang Huan 1, Wu Yuesheng 1

1Baidu X-Lab
{liu.yan, haoxin01,wangyang62}@baidu.com

Abstract

In recent years, neural networks have been extensively deployed for computer vision tasks, particularly visual classification problems, where new algorithms reported to achieve or even surpass the human performance. Recent studies have shown that they are all vulnerable to the attack of adversarial examples. Small and often imperceptible perturbations to the input images are sufficient to fool the most powerful neural networks. Advbox is a toolbox to generate adversarial examples that fool neural networks in PaddlePaddle, PyTorch, Caffe2, MxNet, Keras, TensorFlow and it can benchmark the robustness of machine learning models. Compared to previous work, our platform supports black box attacks on Machine-Learning-as-a-service, as well as more attack scenarios, such as Face Recognition Attack, Stealth T-shirt, and Deepfake Face Detect. The code is licensed under the Apache 2.0 license and is openly available at https://github.com/advboxes/AdvBox.

1 Introduction

Deep learning (DL) has made significant progress in a wide domain of machine learning (ML): image classification [Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; He et al., 2016], object detection [Redmon et al., 2016; Redmon and Farhadi, 2017], speech recognition [Graves et al., 2013; Amodi et al., 2016], language translation [Sutskever et al., 2014; Bahdanau et al., 2014], voice synthesis [Oord et al., 2016; Shen et al., 2018].

Szegedy et al. first generated small perturbations on the images for the image classification problem and fooled state-of-the-art deep neural networks with high probability [Szegedy et al., 2013]. These misclassified samples were named as Adversarial Examples. A large number of attack algorithms have been proposed, such as FGSM [Goodfellow et al., 2014], BIM [Kurakin et al., 2016], DeepFool [Moosavi-Dezfooli et al., 2016], JSMA [Papernot et al., 2016b], CW [Carlini and Wagner, 2017], PGD [Madry et al., 2017a].

The scope of researchers’ attacks has also gradually extended from the field of computer vision [Fischer et al., 2017; Xie et al., 2017; Wang et al., 2019a; Jia et al., 2020] to the field of natural language processing [Ebrahimi et al., 2017; Li et al., 2018; Gao et al., 2018] and speech [Carlini and Wagner, 2018; Qin et al., 2019; Yakura and Sakuma, 2019].

Cloud-based services offered by Amazon 1, Google 2, Microsoft 3, Clarifai 4 and other public cloud companies have developed ML-as-a-service tools. Thus, users and companies can readily benefit from ML applications without having to train or host their own models [Hosseini et al., 2017b]. Unlike common attacks against web applications, such as SQL injection and XSS, there are very special attack methods for machine learning applications, e.g., Adversarial Attack. Obviously, neither public cloud companies nor traditional security companies pay much attention to these new attacks and defenses [Goodman and Hao, 2020; Goodman and Wei, 2019; Li et al., 2019; Goodman et al., 2019b; Goodman et al., 2019a; Goodman and Hao, 2019; Goodman et al., 2020; Goodman et al., 2018].

In this paper, we will focus on adversarial example attack, defense and detection methods based on our AdvBox. Our key items covered:

- The basic principles and implementation ideas.
- Adversarial example attack, defense and detection methods.
- Black box attacks on Machine-Learning-as-a-service.
- More attack scenarios, such as Face Recognition Attack, Stealth T-shirt, and Deepfake Face Detect.

2 Related Work

Currently, several attack/defense platforms have been proposed, like Cleverhans [Papernot et al., 2016a], FoolBox [Rauber et al., 2017], ART [Nicolae et al., 2018], DEEPSEC [Ling et al., 2019], etc. For a detailed comparison, see the Table 1.

3 Adversarial Attack

3.1 Problem Formulation

The function of a pre-trained classification model $F$, e.g. an image classification or image detection model, is mapping

1https://aws.amazon.com/cn/rekognition/
2https://cloud.google.com/vision/
3https://azure.microsoft.com
4https://clarifai.com
from input set to the label set. For a clean image example $O$, it is correctly classified by $F$ to ground truth label $y \in Y$, where $Y$ including $\{1, 2, \ldots, k\}$ is a label set of $k$ classes. An attacker aims at adding small perturbations in $O$ to generate adversarial example $ADV$, so that $F(ADV) \neq F(O)$, where $D(ADV, O) \leq \epsilon$. $D$ captures the semantic similarity between $ADV$ and $O$, $\epsilon$ is a threshold to limit the size of perturbations. For computer vision, $D$ usually stands for Perturbation Measurement.

### 3.2 Perturbation Measurement

\[ l_p \] measures the magnitude of perturbation by $p$-norm distance:

\[ \|x\|_p = \left( \sum_{i=1}^{n} \|x_i\|^p \right)^{\frac{1}{p}} \]  

\(l_0, l_2, l_\infty\) are three commonly used $l_p$ metrics. $l_0$ counts the number of pixels changed in the adversarial examples; $l_2$ measures the Euclidean distance between the adversarial example and the original sample; $l_\infty$ denotes the maximum change for all pixels in adversarial examples.

### 4 AdvBox

#### 4.1 Overview

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

#### 4.2 Structure

Advbox is based on Python\(^6\) and uses object-oriented programming.

#### Attack Class

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. The code is implemented as `advbox.attack`.

---

\(^6\)https://www.python.org/

---

### Table 1: Comparison of different adversarial attack/defense platforms. "√" means "support".

|                     | Cleverhans | FoolBox | ART | DEEPSEC | Our |
|---------------------|------------|---------|-----|---------|-----|
| Tensorflow[Abadi et al., 2016] | √          |         |     |         | √   |
| PyTorch[Paszke et al., 2019]    | √          |         |     |         | √   |
| MxNet[Chen et al., 2015]       | √          |         |     |         | √   |
| PaddlePaddle\(^5\)            | √          |         |     |         |     |
| Adversarial Attack           | √          |         |     |         |     |
| Adversarial Defense          | √          |         |     |         |     |
| Robustness Evaluation        | √          |         |     |         |     |
| Adversarial Detection        | √          |         |     |         |     |
| Attack on ML-as-a-service    | √          |         |     |         |     |
| Actual attack scenario       | √          |         |     |         |     |

---

**Model Class**

Advbox implements interfaces to Tensorflow[Abadi et al., 2016], PyTorch[Paszke et al., 2019], MxNet[Chen et al., 2015], and PaddlePaddle\(^7\). Additionally, other deep learning frameworks 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.

AdvBox also supports GraphPipe\(^8\), which shields the underlying deep learning platform. Users can conduct black box attack on model files generated by Caffe2\(^9\), CNTK\(^10\), MAT-LAB\(^11\) and Chainer\(^12\) platforms. The code is implemented as `advbox.model`.

**Adversary Class**

Adversary contains the original object, the target and the adversarial examples. It provides the misclassification as the criterion to accept a adversarial example. The code is implemented as `advbox.adversary`.

### 4.3 Adversarial Attack

Advbox supports 6 attack algorithms:

- FGSM[Goodfellow et al., 2014]
- BIM[Kurakin et al., 2016]
- DeepFool[Moosavi-Dezfooli et al., 2016]
- JSMA[Papernot et al., 2016b]
- CW[Carlini and Wagner, 2017]
- PGD[Madry et al., 2017a]

The code is implemented as `advbox.attack`. JSMA are often used as a baseline $l_0$ attack algorithm. CW are often used as a baseline $l_2$ attack algorithm. FGSM and PGD are often used as a baseline $l_\infty$ attack algorithm.

---

\(^5\)https://github.com/paddlepaddle/paddle

\(^6\)https://github.com/oracle/graphpipe

\(^7\)https://www.python.org/

\(^8\)https://www.mathworks.com/products/deep-learning.html

\(^9\)https://chainer.org/

\(^10\)https://docs.microsoft.com/en-us/cognitive-toolkit

\(^11\)https://caffe2.ai/

\(^12\)https://www.mathworks.com/products/deep-learning.html
4.4 Adversarial Attack Mitigation
Adbox supports 6 defense algorithms:

- Feature Squeezing\cite{Xu et al., 2017}
- Spatial Smoothing\cite{Xu et al., 2017}
- Label Smoothing\cite{Xu et al., 2017}
- Gaussian Augmentation\cite{Zantedeschi et al., 2017}
- Adversarial Training\cite{Madry et al., 2017b}
- Thermometer Encoding\cite{Buckman et al., 2018}

The code is implemented as \texttt{advbox.defense}. Adversarial Training is often used as a baseline defense algorithm.

4.5 Robustness Evaluation Test
We independently developed a sub-project \textit{Perceptron}\footnote{https://github.com/advboxes/perceptron-benchmark} to evaluate the robustness of the model. Perceptron is a robustness benchmark for computer vision DNN models. It supports both image classification and object detection models as AdvBox, as well as cloud APIs. Perceptron is designed to be agnostic to the deep learning frameworks the models are built on.

Perceptron provides different attack and evaluation approaches:

- CW\cite{Carlini and Wagner, 2017}
- Gaussian Noise\cite{Hosseini et al., 2017a}
- Uniform Noise\cite{Hosseini et al., 2017a}
- Pepper Noise\cite{Hosseini et al., 2017a}
- Gaussian Blurs\cite{Goodman, 2020; Yuan et al., 2019}
- Brightness\cite{Goodman et al., 2019b; Yuan et al., 2019}
- Rotations\cite{Engstrom et al., 2017}
- Bad Weather\cite{Narasimhan and Nayar, 2000}

5 Attack scenario
Compared to previous work\cite{Abadi et al., 2016; Rauber et al., 2017; Nicolae et al., 2018; Ling et al., 2019}, our platform supports more attack scenarios, such as Face Recognition Attack, Stealth T-shirt, and Deepfake Face Detect.
work can refer to the conference [Wang et al., 2019c; Wang et al., 2019b].

Acknowledgement

Thanks to every code submitter. Thanks to everyone who uses or cites AdvBox in their papers.

References

[Abadi et al., 2016] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pages 265–283, 2016.

[Amodei et al., 2016] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In {International conference on machine learning}, pages 173–182, 2016.

[Bahdanau et al., 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[Buckman et al., 2018] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. 2018.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP), pages 39–57. IEEE, 2017.

[Carlini and Wagner, 2018] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. 2018 IEEE Security and Privacy Workshops (SPW), May 2018.

[Chen et al., 2015] Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274, 2015.

[Ebrahimi et al., 2017] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. Hotflip: White-box adversarial examples for text classification. arXiv preprint arXiv:1712.06751, 2017.

[Engstrom et al., 2017] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. arXiv preprint arXiv:1712.02779, 2017.

[Fischer et al., 2017] Volker Fischer, Mummadi Chaithanya Kumar, Jan Hendrik Metzen, and Thomas Brox. Adversarial examples for semantic image segmentation, 2017.
Cloud-based image classification service is not robust to adversarial examples. A forgotten battlefield, 2020.

Dou Goodman and Wei, 2019. At-

Dou Goodman and Tao Wei. Google’s cloud vision api is not robust to noise. In Proceedings of the 2019 ACM SIGSAC Conference on Cloud Computing Security Workshop, pages 43–43, 2019.

Dou Goodman, 2020. Co-

Dou Goodman and Xin Hao. Transferability of adversarial examples to attack cloud-based image classifier service. In Defcon China Conference, 2019.

Dou Goodman, Xin Hao, and Yang Wang. Transferability of adversarial examples to attack cloud-based image classifier service. In Defcon China Conference, 2019.

Dou Goodman, Xin Hao, and Yang Wang. Cloud-based image classification service is not robust to affine transformation: A forgotten battlefield. In Proceedings of the 2019 ACM SIGSAC Conference on Cloud Computing Security Workshop, pages 43–43, 2019.

Dou Goodman. Cloud-based image classification service is not robust to adversarial examples: A forgotten battlefield, 2020.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In 2013 IEEE international conference on acoustics, speech and signal processing, pages 6645–6649. IEEE, 2013.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2016.

Hossein Hosseini, Baicen Xiao, and Radha Poovendran. Google’s cloud vision api is not robust to noise. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 101–105. IEEE, 2017.

Hossein Hosseini, Baicen Xiao, and Radha Poovendran. Google’s cloud vision api is not robust to noise. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Dec 2017.

Yunhan Jia, Yantao Lu, Junjie Shen, Qi Alfred Chen, and Hao Chen. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In International Conference on Learning Representations, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

Alessandro Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017.

Alessandro Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2016.

Srinivasa G Narasimhan and Shree K Nayar. Chromatic framework for vision in bad weather. In Proceedings IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2000 (Cat. No. PR00662), volume 1, pages 598–605. IEEE, 2000.

Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Bussier, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.1.0. CoRR, 1807.01069, 2018.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499, 2016.
cleverhans v2. 0.0: an adversarial machine learning library. *arXiv preprint arXiv:1610.00768*, 10, 2016.

[Papernot et al., 2016b] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016.

[Paszke et al., 2019] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035, 2019.

[Qin et al., 2019] Yao Qin, Nicholas Carlini, Ian Goodfellow, Garrison Cottrell, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition, 2019.

[Rauber et al., 2017] Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

[Redmon and Farhadi, 2017] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[Redmon et al., 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[Schroff et al., 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[Shen et al., 2018] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Sutskever et al., 2014] I Sutskever, O Vinyals, and QV Le. Sequence to sequence learning with neural networks. *Advances in NIPS*, 2014.

[Szegedy et al., 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[Thys et al., 2019] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection, 2019.

[Wang et al., 2019a] Derui Wang, Chaoan Li, Sheng Wen, Xiaojun Chang, Surya Nepal, and Yang Xiang. Daedalus: Breaking non-maximum suppression in object detection via adversarial examples, 2019.

[Wang et al., 2019b] Yang Wang, Junfeng Xiong, Dou Goodman, and Tao Wei. Face swapping video detection with cnn. In *Defcon China Conference*, 2019.

[Wang et al., 2019c] Yang Wang, Junfeng Xiong, Dou Goodman, and Tao Wei. How to detect fake faces (manipulated images) using cnns. In *HITB GSEC Conference*, 2019.

[Xie et al., 2017] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[Xu et al., 2017] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[Xu et al., 2019] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Evading real-time person detectors by adversarial t-shirt. *arXiv preprint arXiv:1910.11099*, 2019.

[Yakura and Sakuma, 2019] Hiromu Yakura and Jun Sakuma. Robust audio adversarial example for a physical attack. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, Aug 2019.

[Yuan et al., 2019] Kan Yuan, Di Tang, Xiaojing Liao, Xiaofeng Wang, Xuan Feng, Yi Chen, Menghan Sun, Haoran Lu, and Kehuan Zhang. Stealthy porn: Understanding real-world adversarial images for illicit online promotion. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 952–966. IEEE, 2019.

[Zantedeschi et al., 2017] Valentina Zantedeschi, Maria-Irina Nicolae, and Ambrish Rawat. Efficient defenses against adversarial attacks. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 39–49. ACM, 2017.