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]

[1]Baidu 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; Amodei *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].

Success of ML algorithms has led to an explosion in demand. To further broaden and simplify the use of ML algorithms, cloud-based services offered by Amazon, Google, Microsoft, Clarifai 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]. For example, Google introduced the Cloud Vision API for image analysis. A demonstration website has been also launched, where for any selected image, the API outputs the image labels, identifies and reads the texts contained in the image and detects the faces within the image. It also determines how likely is that the image contains inappropriate contents, including adult, spoof, medical, or violence contents. 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, 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.

Table 1: Comparison of different adversarial attack/defense platforms. "$\surd$" means "support".

	Cleverhans	FoolBox	ART	DEEPSEC	Our
Tensorflow[Abadi *et al.*, 2016]	$\surd$	$\surd$	$\surd$		$\surd$
PyTorch[Paszke *et al.*, 2019]	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
MxNet[Chen *et al.*, 2015]		$\surd$	$\surd$		$\surd$
PaddlePaddle[1]					$\surd$
Adversarial Attack	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Adversarial Defense	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Robustness Evaluation		$\surd$	$\surd$	$\surd$	$\surd$
Adversarial Detection		$\surd$	$\surd$	$\surd$	$\surd$
Attack on ML-as-a-service					$\surd$
Actual attack scenario					$\surd$

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 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 $D(ADV, O) < \epsilon$, D captures the semantic similarity between ADV and O, ϵ 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:

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

l_0, l_2, l_∞ 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_∞ 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[2] 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 $advbox.attack$.

Model Class

Advbox implements interfaces to Tensorflow[Abadi *et al.*, 2016], PyTorch[Paszke *et al.*, 2019], MxNet[Chen *et al.*, 2015], and PaddlePaddle[3]. 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.

AdvBox also supports GraphPipe[4], which shields the underlying deep learning platform. Users can conduct black box attack on model files generated by Caffe2[5], CNTK[6], MATLAB[7] and Chainer[8] 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]**

[2]https://www.python.org/

[3]https://github.com/paddlepaddle/paddle

[4]https://github.com/oracle/graphpipe

[5]https://caffe2.ai/

[6]https://docs.microsoft.com/en-us/cognitive-toolkit/

[7]https://www.mathworks.com/products/deep-learning.html

[8]https://chainer.org/

(a) Labeled as "Bill Gates" (b) Labeled as "Michael Jordan" (c) Labeled as "Michael Jordan"

Figure 1: Robustness of our Stealth T-shirt.

- **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_∞ attack algorithm.

4.4 Adversarial Attack Mitigation

Advbox supports 6 defense algorithms:

- **Feature Squeezing[Xu *et al.*, 2017]**
- **Spatial Smoothing[Xu *et al.*, 2017]**
- **Label Smoothing[Xu *et al.*, 2017]**
- **Gaussian Augmentation[Zantedeschi *et al.*, 2017]**
- **Adversarial Training[Madry *et al.*, 2017b]**
- **Thermometer Encoding[Buckman *et al.*, 2018]**

The code is implemented as $advbox.defense$. Adversarial Training is often used as a baseline defense algorithm.

4.5 Robustness Evaluation Test

We independently developed a sub-project *Perceptron*[9] 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[Carlini and Wagner, 2017]**
- **Gaussian Noise[Hosseini *et al.*, 2017a]**
- **Uniform Noise[Hosseini *et al.*, 2017a]**
- **Pepper Noise[Hosseini *et al.*, 2017a]**
- **Gaussian Blurs[Goodman, 2020; Yuan *et al.*, 2019]**
- **Brightness[Goodman *et al.*, 2019b; Yuan *et al.*, 2019]**
- **Rotations[Engstrom *et al.*, 2017]**
- **Bad Weather[Narasimhan and Nayar, 2000]**

[9]https://github.com/advboxes/perceptron-benchmark

5 Attack scenario

Compared to previous work[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.

5.1 Scenario 1: Face Recognition Attack

We chose a pre-trained FaceNet[Schroff *et al.*, 2015] model that is a state-of-the-art and widely used face recognition system as our white box attacked model. We used gradient-based attacks methods and modify its loss function using FaceNet embedding distance.

As shown in Fig. 1, Fig. (a) and Fig. (b) can be correctly identified, but Fig. (c) is incorrectly identified.

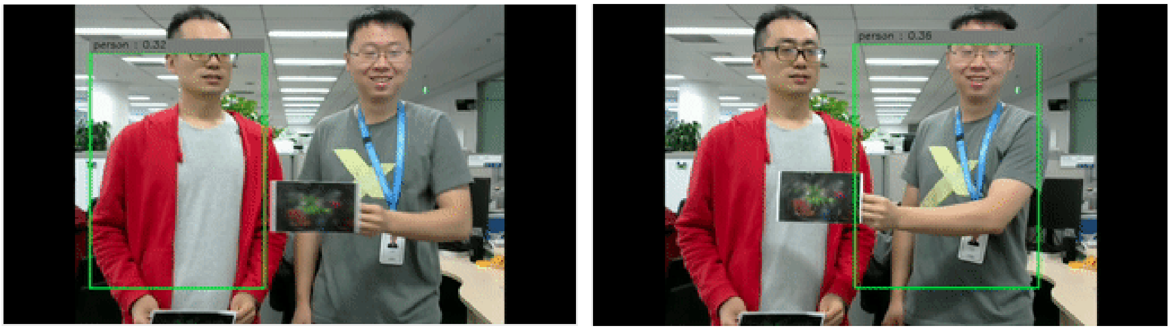
5.2 Scenario 2: Stealth T-shirt

On Defcon China[Goodman *et al.*, 2019a], we demonstrated T-shirts that can disappear under smart cameras. We open source the programs and deployment methods of smart cameras for demonstration.

To raise people's awareness of techniques that can deceive deep learning models, we designed this "Stealth T-shirt" with the adversarial pattern to fool an object detector. The T-shirt is capable of hiding a person who wears it from an open-source object detector. By wearing it and showing the adversarial pattern in front of a camera and its object detector behind it, the person who wears it disappears, whereas the person who doesn't wear the T-shirt is still under object detector detection.

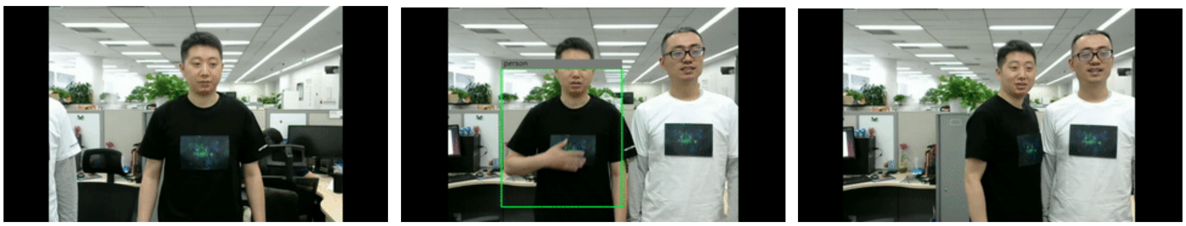
When the smart camera recognizes a human body in the video, it uses a green box to mark the range of the human body. Assume that the black piece of paper in the Fig. 2 is part of the T-shirt. As shown in Fig. (a), the black piece of paper covered the gray man, but did not cover the man in red. The man in red was identified, and the man in gray was not identified. As shown in Fig. (b), the black piece of paper blocked the man in red, and the man in gray was not covered. The man in gray was identified, and the man in red was not identified.

As shown in the Fig. 3, unlike the previous work[Thys *et al.*, 2019; Xu *et al.*, 2019], the picture we need to print in the T-shirt is smaller, facing the distortion, folding, turning, the attack effect is more robust.



(a) (b)

Figure 2: Screenshots to demonstrate our Stealth T-shirt.



(a) Normal (b) Occlusion (c) Turning

Figure 3: Robustness of our Stealth T-shirt.

5.3 Scenario 3: Deepfake Face Detect

We have opened the Deepfake detection capability for free, and you can remotely call our cloud detection detection api by using the Python script we provide. Details of our related 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.

[Gao *et al.*, 2018] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE, 2018.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2014.

[Goodman and Hao, 2019] Dou Goodman and Xin Hao. Transferability of adversarial examples to attack real world porn images detection service. In *HITB CyberWeek Conference*, 2019.

[Goodman and Hao, 2020] Dou Goodman and Xin Hao. Attacking and defending machine learning applications of public cloud. In *Blackhat Asia Conference*, 2020.

[Goodman and Wei, 2019] Dou Goodman and Tao Wei. Cloud-based image classification service is not robust to simple transformations: A forgotten battlefield, 2019.

[Goodman *et al.*, 2018] Dou Goodman, Xin Hao, Yang Wang, Junfeng Xiong, and Yuesheng Wu. Advbox:a toolbox to generate adversarial examples that fool neural networks, March 2018.

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

[Goodman *et al.*, 2019b] Dou Goodman, Xin Hao, Yang Wang, Jiawei Tang, Yunhan Jia, Pei Wang, and Tao Wei. 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.

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

[Graves *et al.*, 2013] 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.

[He *et al.*, 2016] 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.

[Hosseini *et al.*, 2017a] 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.

[Hosseini *et al.*, 2017b] 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.

[Jia *et al.*, 2020] 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.

[Krizhevsky *et al.*, 2012] 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.

[Kurakin *et al.*, 2016] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world, 2016.

[Li *et al.*, 2018] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*, 2018.

[Li *et al.*, 2019] Xurong Li, Shouling Ji, Meng Han, Juntao Ji, Zhenyu Ren, Yushan Liu, and Chunming Wu. Adversarial examples versus cloud-based detectors: A black-box empirical study. *IEEE Transactions on Dependable and Secure Computing*, page 1–1, 2019.

[Ling *et al.*, 2019] Xiang Ling, Shouling Ji, Jiaxu Zou, Jiannan Wang, Chunming Wu, Bo Li, and Ting Wang. Deepsec: A uniform platform for security analysis of deep learning model. In *IEEE S&P*, 2019.

[Madry *et al.*, 2017a] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2017.

[Madry *et al.*, 2017b] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[Moosavi-Dezfooli *et al.*, 2016] 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.

[Narasimhan and Nayar, 2000] 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.

[Nicolae *et al.*, 2018] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, 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.

[Oord *et al.*, 2016] 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.

[Papernot *et al.*, 2016a] Nicolas Papernot, Ian Goodfellow, Ryan Sheatsley, Reuben Feinman, and Patrick McDaniel. 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, Chaoran 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.