Adversarial Signboard against Object Detector

Yi Huang
S160042@ntu.edu.sg
Adams Wai Kin Kong
adamskong@ntu.edu.sg
Kwok-Yan Lam
kwokyan.lam@ntu.edu.sg

School of Computer Science and Engineering
Nanyang Technological University
Singapore

Abstract

Object detector is an indispensable component in many computer vision and artificial intelligence systems, such as autonomous robot and image analyzer for profiling social media users. Analyzing its vulnerabilities is essential for detecting and preventing attacks and minimizing potential loss. Researchers have proposed a number of adversarial examples to evaluate the robustness of object detectors. All these adversarial examples change pixels inside target objects to carry out attacks but only some of them are suitable for physical attacks. According to the best knowledge of the authors, no published work successfully attacks object detector without changing pixels inside the target object. In an unpublished work, the authors designed an adversarial border which tightly surrounds target object and successfully misleads Faster R-CNN and YOLOv3 digitally and physically. Adversarial border does not change pixels inside target object but makes it look weird. In this paper, a new adversarial example named adversarial signboard, which looks like a signboard, is proposed. By putting it below a target object, it can mislead the state-of-the-art object detectors. Using stop sign as a target object, adversarial signboard is evaluated on 48 videos with totally 5416 frames. The experimental results show that adversarial signboard derived from Faster R-CNN with ResNet-101 as a backbone network can mislead Faster R-CNN with a different backbone network, Mask R-CNN, YOLOv3 and R-FCN digitally and physically.

1 Introduction

With the rapid development of deep learning and computational hardware, as well as the availability of big image and video datasets, performance of computer vision methods has very significant improvement in the recent years. Many of them have been deployed to commercial and military applications. Nevertheless, as with other cyber-physical systems, computer vision systems are also subject to attacks. Szegedy et al. [24] found that even though deep neural networks (DNN) based image classifiers achieve great performance, comparing with the traditional hand-crafted feature methods, they are vulnerable to adversarial examples, which are almost no difference from their original images to the naked eyes. Their study draws great concerns from both computer science researchers and engineers, in particular those working in the privacy and security-critical industries. Papernot et al. [20] further

discovered that adversarial examples are transferable, meaning that an adversarial example derived from one network can fool other networks trained on different datasets and even with different architectures. Their findings indicated that attackers can use adversarial examples to perform black-box attacks, even when they do not have knowledge about target network. To identify vulnerabilities in DNN and finally develop more secure DNN-based systems, researchers investigate adversarial examples along four directions: (1) developing new attack schemes for different application scenarios [6, 12, 14, 19, 21, 22, 24], (2) designing countermeasures against adversarial examples [4, 8, 27], (3) studying the existence of adversarial examples and their transferability [10, 16, 25] and (4) deriving bounds of DNNs against zero-day attackss [3, 11, 13]. Different research directions have different emphasises. From the application point of view, developing new attack schemes is vital because once new attacks are discovered, it is easy to design corresponding countermeasures and it also reduces the risk of zero-day attacks. After Szegedy et al. and Papernot et al.'s studies on DNN-based image classifiers, some researchers study vulnerabilities of object detector, which is an essential component in many computer vision systems, such as self-driving cars and unmanned aerial vehicles. However, image classifier and object detector are very different. The former only outputs one label for each image but the latter needs to locate and classify multiple target objects in each image. The state-of-the-art object detectors produce multiple bounding boxes internally for each object and utilize non-maximum suppression or other methods to select a final bounding box. Thus, if an adversarial example is derived to fool object detector, it has to degrade all the internal bounding boxes significantly [18]. To fool object detector, Xie et al. [26] changed every pixel in entire image to craft their adversarial examples but their attack cannot be carried out in the physical world. Some researchers mislead object detector by significantly changing a large amount of pixels inside target objects. Their adversarial examples are very different from the original objects. Fig. 1 shows adversarial examples designed by Lu et al. [17], Chen et al. [7] and Song et al. [23]. According to the best knowledge of the authors, no published adversarial example can attack object detector digitally and physically but without changing target objects. In an unpublished work [7], the authors designed adversarial border and discovered that by surrounding target object with adversarial boarder. It is possible to fool object detector digitally and physically. However, it makes the object look weird (Fig. 2). In this paper, an algorithm is designed to craft a new adversarial example named adversarial signboard, which looks like a signboard. By putting it below a target object, it can mislead object detector digitally and physically. Fig. 3 shows 2 normal signboards from the Internet and Fig. 5 shows 3 adversarial signboards. In this paper, stop sign is used as a target object because it was commonly used in the previous adversarial example studies and is an important object for autonomous vehicles.

The rest of this paper is organized as follows. Section 2 describes existing adversarial examples designed to fool image classifiers and object detectors. Section 3 presents the proposed algorithm for crafting adversarial signboard. Section 4 reports the experimental results on 48 videos with totally 5416 frames for digitally and physically attacking Faster R-CNN with VGG-16 and ResNet-101 as backbone networks, YOLOv3, Mask R-CNN and F-RCNN. Section 5 gives some conclusive remarks.

2 Related Work

Adversarial examples have drawn great attention from the scientific community and industries. Diverse research has been carried out to understand, analyse and protect DNN from

adversarial examples. Adversarial signboard is designed to fool object detector and therefore only adversarial examples designed to fool image classifier and object detector are described in this section.

2.1 Adversarial examples against image classifier

In 2014, Szgedy et al. [24] proposed a method named L-BFGS to craft adversarial examples and found that their adversarial examples and the original images are almost the same to the naked eye but DNN image classifiers classify them differently. Their work pinpointed that although DNNs offer excellent performance in many applications and outperform traditional approaches, they are vulnerable to attacks based on adversarial examples. To speed up L-BFGS, in 2015, Goodfellow et al. [12] proposed another method called Fast Gradient Sign Method (FGSM) to generate adversarial examples against DNN image classifier. In 2016, Rozsa et al. [22] replaced the sign of the gradient in FGSM with the original gradient in order to obtain more accurate optimization direction and finally derive stronger adversarial example. Their method was named Fast Gradient method (FGM). These methods change every pixels in images to produce their adversarial examples. In 2016, Papernot et al. [21] demonstrated that by changing a small amount of pixels, it is possible to mislead DNN image classifier, but their attack has higher computation cost, comparing with the previous methods. In addition to these attacks, researchers designed other attacks, such as Deepfool [19], Zeroth Order Optimization (ZOO) attack [6] and C&W attack [5] to study vulnerability of DNN image classifier. All these attacks are based on the assumption that attackers can input their adversarial examples to targeted DNN image classifier directly. They have no effect on the systems which take input images directly from cameras. In other words, these attacks are only suitable for digital attacks but not physical attacks. To achieve better robustness of the adversarial examples, Kurakin et al. [15] used a finer optimizer to improve FGSM and Athalye et al. [1] took distance between camera and object, variations and other noise in the physical world into consideration and used a function to simulate these variations in the training process. As with other DNN training, training adversarial examples with diverse images can improve their robustness. Eykholt et al .[9] took this approach to strengthen their adversarial examples for physical attacks. More clearly, they used images taken from different viewpoints, distances and illumination environments to train their adversarial examples. They also employed a function to synthesize the distortion in the physical world for further enhancing the robustness of their adversarial examples.

Figure 1: Adversarial stop signs generated by [7, 17, 23]

Figure 2: Adversarial border

Figure 3: Normal signboards collected from the Internet.

2.2 Adversarial examples against object detector

The risk in DNN image classifiers motivated some researchers to investigate adversarial examples against object detector. In 2017, Xie et al. [26] proposed a method named Dense Adversary Generation (DAG), which derives adversarial examples from Faster R-CNN. They first replaced the original label of each object with an adversarial label and then maximized the confidence to the adversarial label and minimized the confidence to the original label. Experimental results showed that DAG can fool Faster R-CNN. In the same year, Lu et al. [18] used a target vector with all elements close to one which represents a background label and a method designed for attacking DNN image classifier to craft adversarial examples against object detector. Lu et al.'s method is effective for attacking YOLO digitally, but not physically. Lu et al. [17] proposed another adversarial example named adversarial stop sign against Faster R-CNN. Because stop sign is a regular octagon, they used a shape matching function to map an adversarial stop sign in a root coordinate system to the stop signs in the training frames. Their objective function is to minimize the mean score of the stop signs detected by Faster R-CNN. As with Eykholt et al.'s training approach [9], they also trained adversarial stop sign on diverse frames from a video to increase its robustness. Experimental results showed that their adversarial stop sign can physically fool Faster R-CNN, but looks very different from normal stop sign (Fig.1 the first column). Because of the shape matching function, their method is not applicable to objects without well-defined shapes, e.g., desks. Chen et al. [7] used expectation over transformation to design another adversarial stop sign, which can successfully mislead Faster R-CNN in physical attacks. Their method changes every pixel inside stop sign, except for those in the word, STOP. Although the word, STOP, is clear, the rest region, which should be pure red, is totally replaced with other patterns (Fig. 1 the middle column). Song et al. [23] further limited the attack region and produced adversarial sticker, which can successfully attack both YOLO and Faster R-CNN digitally and physically (Fig. 1 the last column). According to the best knowledge of the authors, all published adversarial examples which are effective in physical attacks change target object significantly and create noticeable patterns inside it. The adversarial border designed by the authors in an unpublished article [2] can successfully fool Faster R-CNN and YOLOv3 digitally and physically but without changing any pixels inside target object. It pinpointed that it is possible to perform attacks even without changing target object. However, to carry out such an attack, target object has to be placed at the centre of the adversarial border such that it is surrounded by adversarial patterns. It makes the target object look weird (Fig. 2). To

address this problem, in this paper, a new adversarial example named adversarial signboard is proposed. It looks like a signboard and by putting it below target object, it can perform physical and digital attacks.

3 Algorithm

The proposed adversarial signboard is constructed using Faster R-CNN (Fig. 4), which is composed of three main sub-networks - feature extraction network, region proposal network (RPN) and detection network. The feature extraction network is used to compute features consumed by the RPN and the detection network. It is also called a backbone network and VGG-16 and ResNet are commonly employed as the feature extraction network. The RPN takes features from the feature extraction network and determines a set of region proposals which have high probability of containing objects. The detection network takes both outputs from the feature extraction network and the RPN as input and computes final bounding boxes and their corresponding classes using respectively the box regression network and the box classification network (see Fig. 4). More precisely, the box regression network computes parameters to refine the coordinates of the input region proposals and determines the final bounding boxes and the box classification network produces a probability matrix P, each of whose row and column correspond respectively to one input region proposal and one specific category. The element in the i^{th} row and the j^{th} column in P indicates the probability of the i^{th} region proposal belonging to the j^{th} category.

Adversarial signboard is designed to mislead the box classification network into outputting low probabilities for target object, i.e., stop sign in this study. To enhance its robustness to scale, distance and lighting variations in the physical world, n images, each of which contains one target object T, are sampled from k videos $V_{1,\ldots,k}$ as a training set. Given an input image I, let $B(I)$ be the bounding boxes outputted by the box regression network and $P_T(I)$ be the corresponding probabilities of the target object T. More clearly, $P_T(I)$ is the column of P corresponding to the class of the target object. Since not all the bounding boxes in $B(I)$ is generated by the target object, only the bounding box $b_t(I)$ with the highest probability in $P_T(I)$ is selected for training. It is assumed that in each training frame, there is one target object and Faster R-CNN can roughly locate it. To train adversarial signboard on

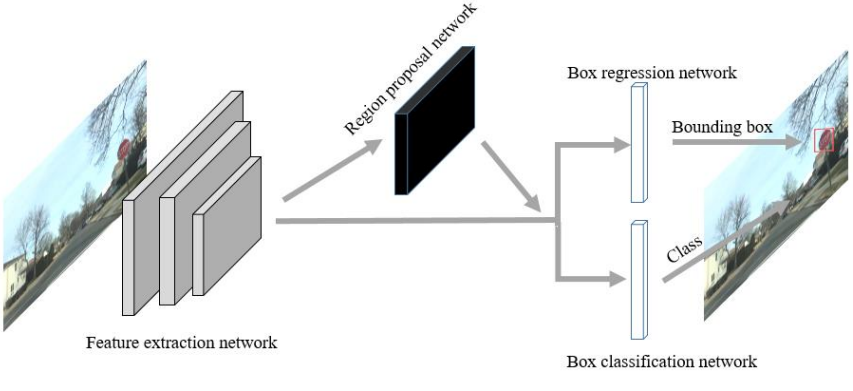

Figure 4: Illustration of Faster R-CNN.

a large database without ground truth bounding boxes, $b_t(I)$ is used as a reference to scale and place an adversarial signboard in I. Let the adversarial signboard be a patch Λ with a fixed size of $h \times w$ and the size of $b_t(I)$ be $h_t \times w_t$. The adversarial signboard is scaled to $\alpha h_t \times \beta w_t$. In the experiments, α and β are set to 1 and 2 respectively. If the bounding box given by Faster R-CNN is not accurate enough and if the adversarial signboard is placed right below $b_t(I)$, it will cover a part of the target object. To address this problem, the adversarial signboard is placed with a distance below $b_t(I)$. In the experiments, the distance D is set to $1/10 \times h_t$. Furthermore, it also makes the position of the adversarial signboard more natural. In physical attacks, it is hard to put the adversarial signboard in an accurate position as digital attacks. Therefore, random shifts in both vertical and horizontal directions are added into the training process to improve its robustness against this inaccuracy. The range of the random shifts in the horizontal direction is from $-1/10 \times w_t$ to $1/10 \times w_t$ and the range of the random shifts in the vertical direction is from 0 to $1/10 \times h_t$. After the random shifts, the distance between the adversarial signboard and the target object varies from $1/10 \times h_t$ to $2/10 \times h_t$. To mimic real signboard, a text mask is applied to the adversarial signboard and only the pixels outside the text mask are involved in the training. Let $m(\Lambda)$ be the process of putting the text mask in the adversarial signboard and $f(I, m(\Lambda), b_t(I))$ be the process of scaling and placing the adversarial signboard with the text mask on I. Note that the output of $f(I, m(\Lambda), b_t(I))$ is an image with the adversarial signboard. Instead of minimizing the mean probability, the proposed objective function minimizes the expected maximum original class probability of all the bounding boxes in B, which can be written as:

$$\min_{I \in V_{1,\ldots,k}} \mathbb{E}(max(P_T(f(I, m(\Lambda), b_t(I))))) \tag{1}$$

Though both adversarial border [2] and adversarial signboard are placed outside target object to perform attacks, they are derived by different algorithms. Adversarial border is designed to fool the box regression network, while adversarial signboard is designed to fool the box classification network. The former minimizes the mean difference between the bounding box parameters produced by the box regression network and predefined target values, while the latter minimizes the expected maximum original class probability. Furthermore, adversarial border has to surround target object tightly to perform attacks and makes it look weird (Fig. 2). Adversarial signboard significantly lightens the positing requirement. By placing adversarial signboard below target object with some distance, it is enough to perform an attack. Besides, it camouflages itself as a normal signboard, which is much more natural than adversarial border. Fig . 3 shows normal signboards and Fig. 5 shows adversarial signboards with different text designs. Furthermore, in the previous study, adversarial border was only evaluated on Faster R-CNN with VGG-16 as a backbone network and YOLOv3, while in this study, adversarial signboard is evaluated on Faster R-CNN with VGG-16 and ResNet-101 as backbone networks, Mask R-CNN, YOLOv3 and R-FCN.

4 Experiments

To evaluate adversarial signboard, stop sign is used as a target object because it was commonly used in the previous adversarial example studies and an important object for autonomous vehicles. In the experiments, 529 images sampled from 4 videos and Faster R-CNN with ResNet-101 as a backbone network are used to train three adversarial signboards with a size of 90 by 180 pixels. The three adversarial signboards with different text masks

(a) AS-W (b) AS-B (c) AS-BW

Figure 5: The first row is the three adversarial signboards for digital attacks and the second row is their printouts for physical attacks.

shown in Fig. 5 (a)-(c) are respectively denoted as AS-W, AS-B and AS-BW. In the experiments, both white-box and black-box attacks are studied. For white-box attacks, Faster R-CNN with ResNet-101 as its backbone network is used to evaluate the adversarial signboards. For black-box attacks, four object detectors YOLOv3, R-FCN, Faster R-CNN using VGG-16 as its backbone network and Mask R-CNN, are used to evaluate the adversarial signboards. The backbone networks of YOLOv3, R-FCN and Mask R-CNN are respectively Darknet-53, ResNet-101 and ResNet-101. For the sake of convenience, the two Faster-CNN detectors are denoted as Faster R-CNN(VGG-16) and Faster R-CNN(ResNet-101). In addition to white-box and black-box attacks, they are also examined in digital and physical attacks.

4.1 Digital attacks

In digital attacks, 36 videos collected from the Internet are employed. Since adjacent frames in the videos are highly similar, each alterative frame is selected and forms a testing set with 4,073 images. The adversarial signboards are scaled and placed in the testing images. Fig. 6 gives sample detection results with and without adversarial signboards. The successful attack rate AR for one video is defined as the number of successful attacks divided by the number of original detected images. Note that AR is computed from each video and the average AR from the 36 videos is used as a final performance index for digital attacks. Table. 1 lists the original detection rates (DR-ORG) and the average successful attack rates for white-box and black-box attacks. For white-box attacks, the adversarial signboards achieve average successful attack rates over 84%. For black-box attacks, they can mislead detectors with similar architectures, i.e., Faster R-CNN(VGG-16) and Mask R-CNN at high successful attack rates. The mean average ARs of these two detectors are 69% and 70%, respectively. For R-FCN, the average AR varies between 70% and 59%. For YOLOv3, the average AR varies between 58% and 50%. The experimental results also show that AS-W performs the best among the three adversarial signboards and its mean average AR across all the detectors is 72%. These experimental results demonstrate that the adversarial signboards can significantly reduce the number of detected stop signs in both white-box and black-box attacks.

	Faster R-CNN (ResNet-101)	Faster R-CNN (VGG-16)	Mask R-CNN	YOLOv3	R-FCN
DR-ORG	89	70	95	95	76
AS-B	87	58	75	50	59
AS-W	88	78	67	57	70
AS-BW	84	71	68	58	64
Mean AR	86	69	70	55	64

Table 1: The average successful attack rates in digital attacks (%)

	Faster R-CNN (ResNet-101)	Faster R-CNN (VGG-16)	Mask R-CNN	YOLOv3	R-FCN
Stop sign	98	80	94	68	96
AS-B	20	47	16	18	47
AS-W	30	43	18	13	38
AS-BW	32	57	19	16	59

Table 2: The average detection rates in physical attacks (%)

4.2 Physical attacks

In this experiment, a stop sign with a size of 19.2 cm by 19.2 cm and the three adversarial signboards with a size of 19.2 cm by 38.4 cm are printed out. Fig. 5 shows the stop sign and the adversarial signboards. Four groups of videos are taken in an open car-park by a smartphone camera and each group has one video for the stop sign without adversarial signboard and three videos for the three adversarial signboards with the stop sign above. In each group, the videos are taken from roughly the same viewpoint, location and camera direction. Thus, in total, 16 videos with a resolution of 1920 by 1080 (or 1080 by 1920) pixels are collected and the videos in the same group are more comparable. On average, each video has 115 frames used in this experiment. In the videos, the stop sign is on the right hand side, which is the same as in the training videos. Fig. 7 shows two sample frames from the videos. The average successful attack rate cannot be employed as a performance index in physical attacks, because the videos with and without adversarial signboards are not collected from the exact same viewpoint and at the same time and have different numbers of frames. Thus, the average detection rates given in Table. 2 are used a performance index. All the detectors except for YOLOv3 perform well on the original stop sign with detection rates of over 80%. The adversarial signboards significantly reduce the detection rates of Faster R-CNN (ResNet-101), YOLOv3, and Mask R-CNN. All their detection rates are below 32%. Seven out of the nine detection rates from these detectors are even below 21%. The adversarial signboards also decrease the detection rates of R-FCN substantially. Its average detection rate is 48%. Comparing with the original detection rate, the adversarial signboards reduce the detection rates by 48%. The adversarial signboards have less impact on Faster R-CNN(VGG-16), but they still can reduce its average detection rate by 31%. As with the digital experiment, AS-W performs the best.

Figure 6: Detection results with and without adversarial signboards from Faster R-CNN(ResNet-101)

Figure 7: Sample frames with and without adversarial signboard from the videos in the same group.

5 Conclusion

Object detector is an essential component in many computer vision and AI systems. The previous adversarial studies pointed out that by significantly changing pixels inside target object, it is possible to fool object detector. In an unpublished work, the authors demonstrated that it is possible to attack object detector digitally and physically through adversarial border, which does not require changing pixels insider target object. However, it surrounds target object tightly and makes it look weird. In this paper, a new adversarial example named adversarial signboard is proposed. It is placed below target object with a distance and looks more natural. Extensive experiments on 48 videos and five object detectors have been carried out and demonstrated that adversarial signboard can perform digital and physical attacks in white-box and black-box settings.

6 Acknowledgment

This work is partially supported by the Ministry of Education, Singapore through Academic Research Fund Tier 2, MOE2016-T2-1-042

References

[1] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, 2018.

[2] Authors. Attacking object detectors without changing the target object, 2019. IJCAI-19 submission ID 2455. Supplied as additional material `IJCAI-19.pdf`.

[3] Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya V. Nori, and Antonio Criminisi. Measuring neural net robustness with constraints. In *NIPS*, 2016.

[4] Arjun Nitin Bhagoji, Daniel Cullina, Chawin Sitawarin, and Prateek Mittal. Enhancing robustness of machine learning systems via data transformations. *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–5, 2018.

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

[6] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *AISec@CCS*, 2017.

[7] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *ECML/PKDD*, 2018.

[8] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E. Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *CoRR*, abs/1705.02900, 2017.

[9] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[10] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Fundamental limits on adversarial robustness. In *ICML 2015*, 2015.

[11] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning*, 107:481–508, 2017.

[12] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1412.6572`.

[13] Xiaowei Huang, Marta Z. Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks. In *CAV*, 2017.

[14] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *CoRR*, abs/1611.01236, 2016.

[15] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017. URL https://arxiv.org/abs/1607.02533.

[16] Yanpei Liu, Xinyun Chen andChang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of 5th International Conference on Learning Representations*, 2017.

[17] Jiajun Lu, Hussein Sibai, and Evan Fabry. Adversarial examples that fool detectors. *CoRR*, abs/1712.02494, 2017.

[18] Jiajun Lu, Hussein Sibai, Evan Fabry, and David A. Forsyth. No need to worry about adversarial examples in object detection in autonomous vehicles. *CoRR*, abs/1707.03501, 2017.

[19] 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)*, pages 2574–2582, 2016.

[20] Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016.

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

[22] Andras Rozsa, Ethan M. Rudd, and Terrance E. Boult. Adversarial diversity and hard positive generation. *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 410–417, 2016.

[23] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018.

[24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

[25] Pedro Tabacof and Eduardo Valle. Exploring the space of adversarial images. *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 426–433, 2016.

[26] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan L. Yuille. Adversarial examples for semantic segmentation and object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1378–1387, 2017.

[27] Stephan Zheng, Yang Song, Thomas Leung, and Ian J. Goodfellow. Improving the robustness of deep neural networks via stability training. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4480–4488, 2016.