Physical Adversarial Attack on Object Detectors

(Extended Abstract)

Shang-Tse Chen
Georgia Institute of Technology
Atlanta, GA, USA
schen351@gatech.edu

Cory Cornelius
Intel Corporation
Hillsboro, OR, USA
cory.cornelius@intel.com

Jason Martin
Intel Corporation
Hillsboro, OR, USA
jason.martin@intel.com

Duen Horng (Polo) Chau
Georgia Institute of Technology
Atlanta, GA, USA
polo@gatech.edu

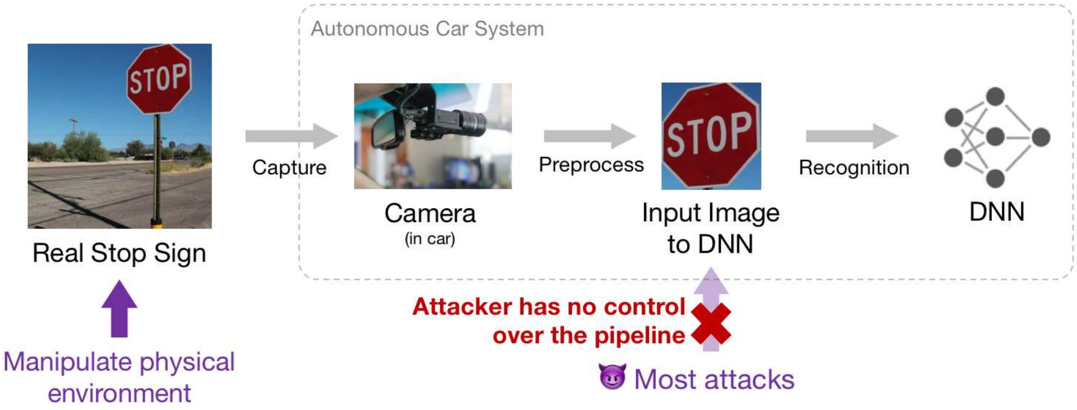


Figure 1: Illustration motivating the need of physical adversarial attack, from attackers' perspectives, as they typically do not have full control over the computer vision system pipeline.

ABSTRACT

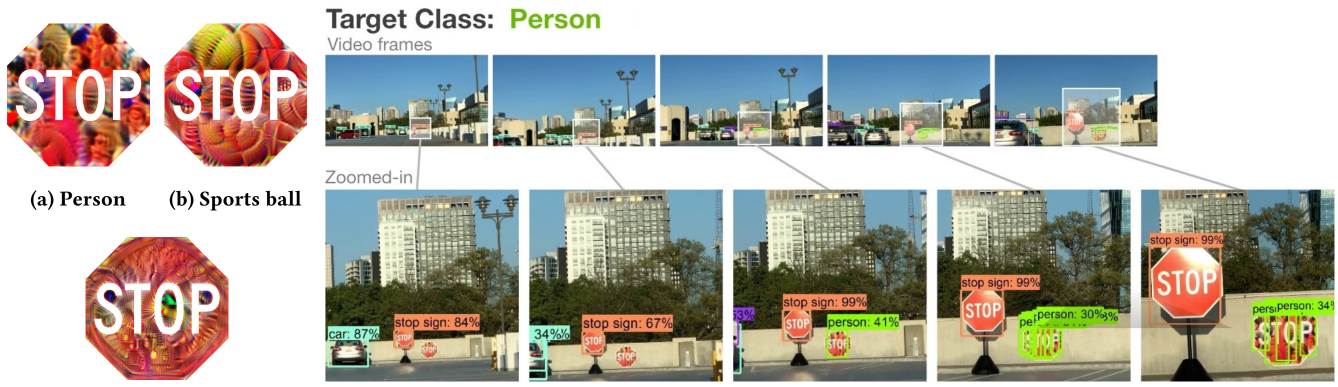
Given the ability to directly manipulate image pixels in the digital input space, an adversary can easily generate imperceptible perturbations to fool a deep neural network image classifier, as demonstrated in prior work. In this work, we tackle the more challenging problem of crafting physical adversarial perturbations to fool image-based object detectors like Faster R-CNN. Attacking an object detector is more difficult than attacking an image classifier, as it needs to mislead the classification results in multiple bounding boxes with different scales. Extending the digital attack to the physical world adds another layer of difficulty, because it requires the perturbation to be robust enough to survive real-world distortions due to different viewing distances and angles, lighting conditions, and camera limitations. In this showcase, we will demonstrate the first robust physical adversarial attack that can fool a state-of-the-art Faster R-CNN object detector. Specifically, we will show various perturbed stop signs that will be consistently mis-detected by an object detector as other target objects. The audience can test in real time the robustness of our adversarially crafted stop signs from different distances and angles. This work is a collaboration between Georgia Tech and Intel Labs and is funded by the Intel Science & Technology Center for Adversary-Resilient Security Analytics at Georgia Tech.

Overview

Adversarial examples are input instances that are intentionally designed to fool a machine learning model into producing a chosen prediction [9]. Although many adversarial attack algorithms have been proposed, attacking a real-world computer vision system is difficult [7], because attackers usually do not have the ability to directly manipulate data inside such systems (Figure 1), and so far the existing attempts to physically attack object detectors remain unsatisfactory [4, 6]. In this showcase, we present ShapeShifter [3] — the first robust targeted attack that can fool a state-of-theart Faster R-CNN object detector [8]. The perturbed stop signs (Figure 2 (a)-(c)) are consistently mis-detected by Faster R-CNN as arbitrary target objects like *person* or *sports ball*, or undetected for the untargeted attack case, in real drive-by tests (Figure 2d). We will demonstrate the robustness of our attack by allowing the audience to hold these stop signs in front of a real-time object detection system with different distances and angles. All our code and demo videos are publicly available at https://github.com/shangtse/robust-physical-attack.

Attack Method

Our attack algorithm is based on the Carlini-Wagner attack [2], which was originally proposed for the task of image classification.



(a) Person (b) Sports ball

(c) Untargeted

(d)

Figure 2: (a)-(c): Perturbed stop signs we created with different target classes. (d) Snapshots of the drive-by test results for the *person* targeted attack. The top images are the original video frames, and the bottom ones are the zoomed-in views that more clearly show the detection results. The real stop sign was correctly detected with high confidence in all of the frames, while the perturbed one was detected as a person with medium confidence in 47% of the frames and only once as a stop sign.

We first describe the method of [2] before showing how to extend it to attack the Faster R-CNN object detector.

Let $F : [-1, 1]^{h\times w\times 3} \to \mathbb{R}^K$ be an image classifier that outputs a probability distribution over K classes for each input image. Denote $L_F(x, y) = L(F(x), y)$ as the loss function that calculates the distance between the model output $F(x)$ and the target label y. Given an original input image x and a target class y', [2] proposes the following optimization-based attack.

$$\underset{x' \in \mathbb{R}^{h\times w\times 3}}{\arg\min} \; L_F(\tanh(x'), y') + c \cdot || \tanh(x') - x||_2^2. \tag{1}$$

The use of *tanh* ensures that each pixel is between $[-1, 1]$. The constant c controls the similarity between the modified object x' and the original x. In practice, c can be determined by binary search.

For object detection, Faster R-CNN first generates several region proposals, and performs classification within each of the region proposals. We focus on attacking the final classifiers and change the loss in (1) to the sum of losses from all the regions. To make the attack more robust, we adopt the *Expectation over Transformation* technique [1] by adding random translation, rotation, and scaling of the perturbed object in each iteration of the optimization.

Evaluation

We evaluate our method by fooling a pre-trained Faster R-CNN model. The model was trained on the MS-COCO dataset [5] and is publicly available[1]. We generate perturbed stop signs by only allowing the perturbation to change the red part of the stop sign, leaving the white text intact. The drive-by tests were done as follows. We put a purchased real stop sign as a control and our printed perturbed stop sign side by side. Starting from about 200 feet away, we slowly drove towards the signs and recorded video from the vehicle's dashboard at 4K resolution and 24 FPS using an iPhone 8 Plus. For each video frame, we obtained the detection results from Faster R-CNN. We only consider detection with confidence value higher than 30%. In all the drive-by tests, the real stop sign was correctly detected in all of the frames.

Crafted with the *person* target class, the perturbed stop sign (Figure 2a) was detected as a person in 190 out of 405 frames and only correctly detected as a stop sign once. For the rest of the 214 frames the object detector failed to detect anything around the perturbed stop sign. See Figure 2d for snapshots of the video. The video we took with the *sports-ball-perturbation* (Figure 2b) had 445 frames. The perturbed stop sign was never detected as a stop sign. As the vehicle (video camera) moved closer to the perturbed stop sign, 160 of the frames were detected as a sports ball. One frame was detected as *apple* and *sports ball* and the remaining 284 frames had no detection around the perturbed stop sign. Finally, the video of the untargeted perturbation (Figure 2c) totaled 367 frames. The perturbed stop sign was detected as *bird* 6 times and never detected for the remaining 361 frames.

[1] http://download.tensorflow.org/models/object_detection/faster_rcnn_inception_v2_coco_2017_11_08.tar.gz

REFERENCES

[1] Anish Athalye and Ilya Sutskever. 2017. Synthesizing robust adversarial examples. *arXiv:1707.07397* (2017).

[2] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*. 39–57.

[3] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng Chau. 2018. ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector. In *ECML-PKDD*.

[4] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Dawn Song, Tadayoshi Kohno, Amir Rahmati, Atul Prakash, and Florian Tramer. 2017. Note on Attacking Object Detectors with Adversarial Stickers. *arXiv:1712.08062* (2017).

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*. 740–755.

[6] Jiajun Lu, Hussein Sibai, and Evan Fabry. 2017. Adversarial Examples that Fool Detectors. *arXiv:1712.02494* (2017).

[7] Jiajun Lu, Hussein Sibai, Evan Fabry, and David Forsyth. 2017. No need to worry about adversarial examples in object detection in autonomous vehicles. *arXiv:1707.03501* (2017).

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*. 91–99.

[9] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR*.