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

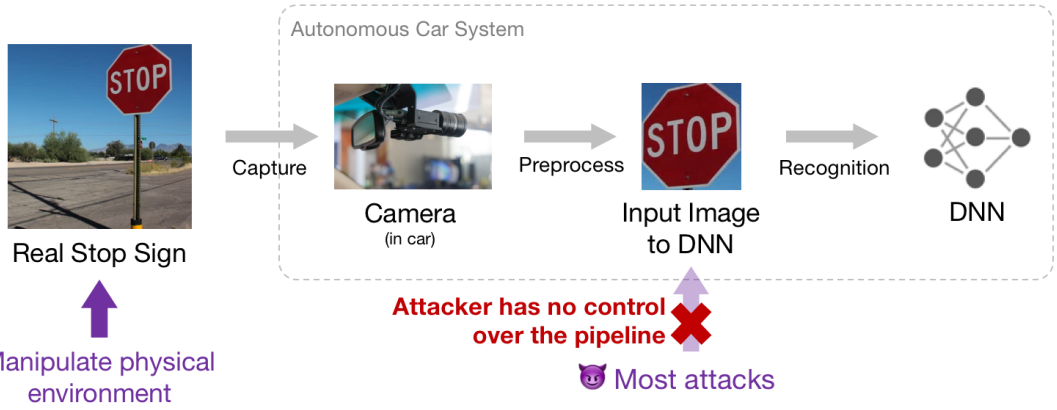

Figure 1: Illustration motivating the need of physical adversarial attack, from attackers' perspectives, as they typically do not have full control over the computer vision system pipeline.

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.