THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Standard detectors aren’t (currently) fooled by physical adversarial stop signs

Authors: Jiajun Lu, Hussein Sibai, Evan Fabry, David Forsyth
Presenter: Dev Shah

Classifier and Detector

A classifier is a machine learning model that takes in input data and assigns it to a specific class or category. In this paper, the classifiers used are LISA-CNN and a publicly available implementation of a classifier demonstrated to work well at road signs.

On the other hand, a detector is a machine learning model that takes in input data and outputs the location of objects in the data, such as the location of a stop sign in an image. In this paper, the detectors used are YOLO and Faster RCNN.

The main difference between the two is that a classifier assigns a label to input data, while a detector outputs the location of objects in the data.

Outline

- The paper presents a construction of physical adversarial stop signs
- The construction of the physical adversarial stop signs is shown to be effective against classifiers, but not against detectors
- The effectiveness of adversarial examples against detectors is an important question because applications usually require detection, not classification
- The paper suggests that adversarial examples that can fool detectors may not exist because the adversarial pattern would need to be invariant to a wide range of parametric distortions

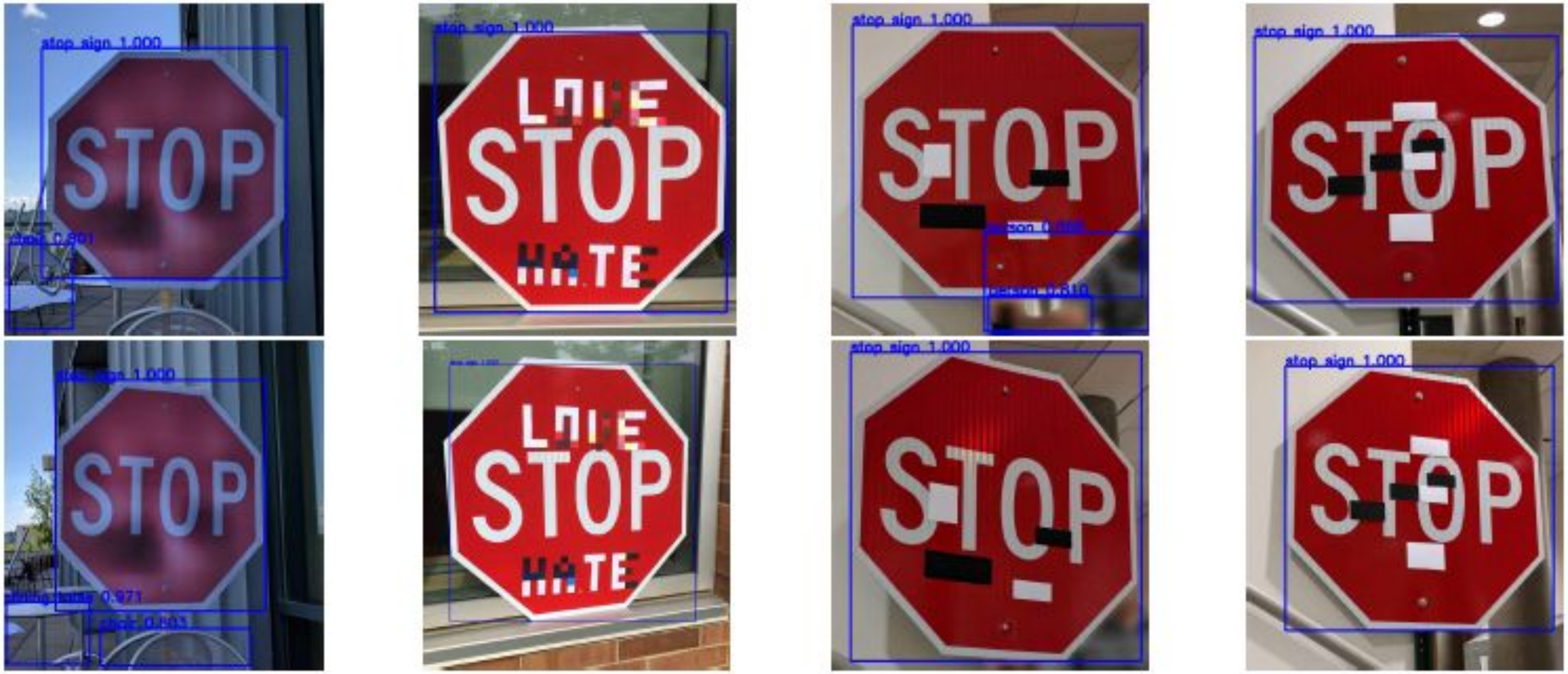
Drawback of previous research

- Their attack is demonstrated on stop-signs that are cropped from images and presented to a classifier.
- By cropping, they have proxied the box-prediction process in a detector
- Their attack is not intended as an attack on a detector.

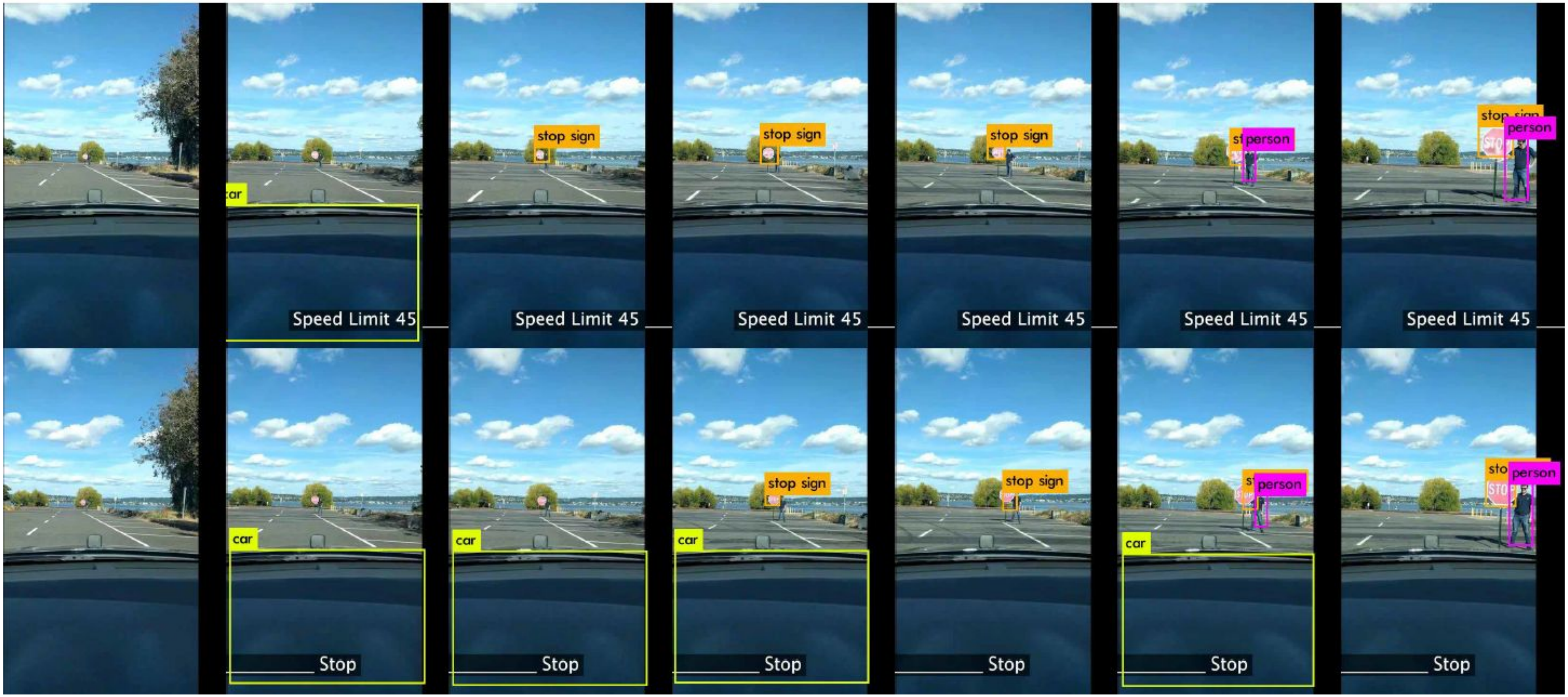
Results

- According to the paper, the authors used two standard detectors, YOLO and Faster RCNN, and applied them to images and videos provided by Evtimov et al.
- They found that both detectors successfully detected adversarial stop signs produced by poster attacks and sticker attacks.
- In addition, they found that Faster RCNN detected stop signs more accurately than YOLO, and that both detectors had difficulty detecting small stop signs.
- These results indicate that adversarial stop signs may not pose a significant threat to modern detection systems.

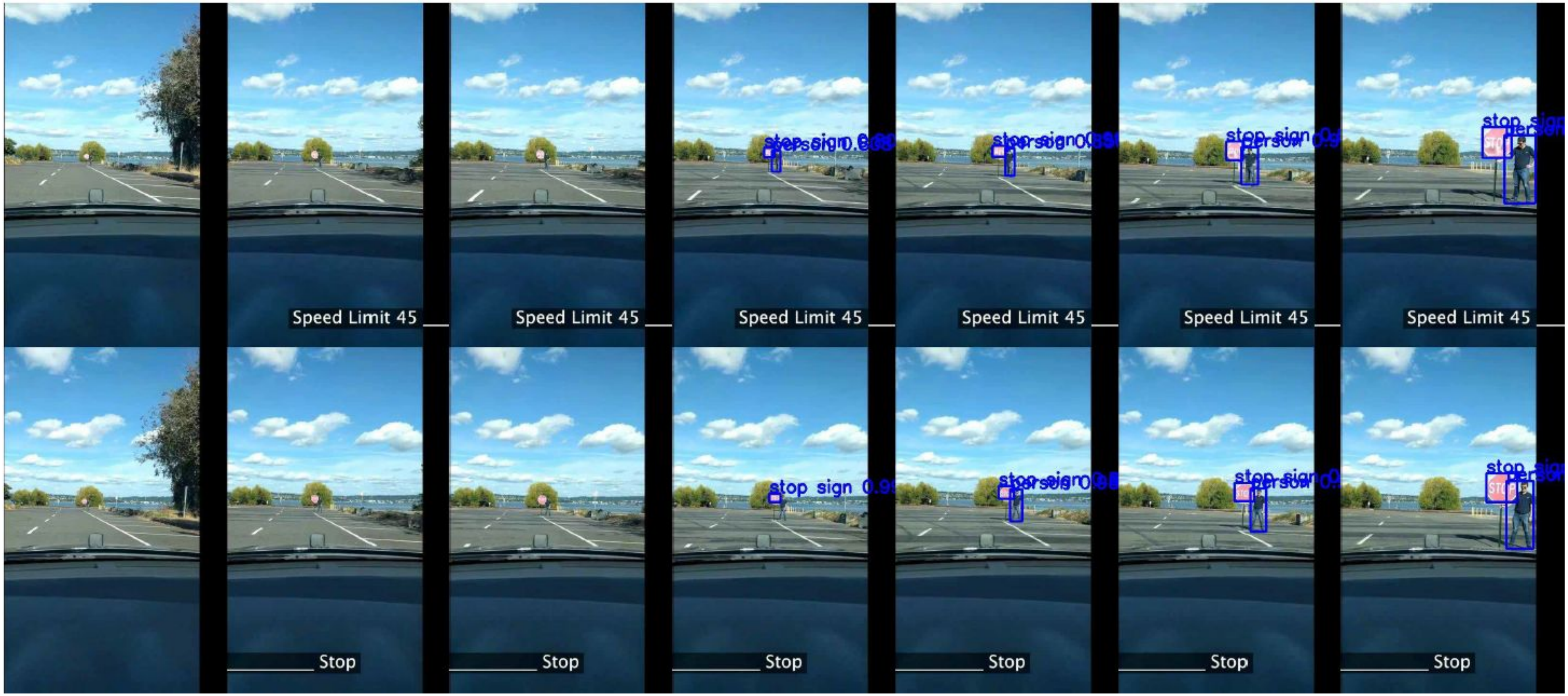
Results


stop sign 1.000
chair 0.901
STOP
stop sign 1.000
dining table 0.971
chair 0.203
stop sign 1.000
LOVE
STOP
HATE
stop sign 1.000
STOP
person 0.888
person 0.810
stop sign 1.000
STOP
stop sign 1.000
STOP
stop sign 1.000
STOP

Results



Results


Speed Limit 45
Speed Limit 45
Speed Limit 45
Speed Limit 45
Speed Limit 45
Speed Limit 45
stop sign
person
stop sign 0.9
person
stop sign
person
Stop
Stop
Stop
Stop
Stop
Stop

Result Analysis

There are several reasons that adversarial examples that can fool classifiers may not exist for detectors.

1. One reason is that the cropping procedure used in a detector often removes the effects of scale and translation, making it difficult for an adversarial pattern to remain effective.
2. Additionally, the statistics of how boxes are placed in a detector are complex and poorly understood, making it difficult to construct an effective adversarial pattern.
3. Finally, the internal structure of the adversarial space, as learned from training examples, may not generalize well across viewing parameters, making it difficult to construct an adversarial pattern that is effective across a wide range of distortions.

Discussion

- The paper presents an analysis of the effectiveness of adversarial examples in fooling object detectors.
- The authors argue that previous research has focused on adversarial examples for classifiers, but not for detectors.
- They present experimental results showing that state-of-the-art object detectors are not fooled by adversarial examples, even when those examples are presented in physical form.
- The authors speculate that this may be because the adversarial patterns are not robust to a wide range of viewing conditions, or because the internal structure of the network allows it to generalize across viewing parameters better than it generalizes across labels.
- They call for further research to explore the existence and potential of adversarial examples for object detectors.

Thank you.