Physical Adversarial Examples for Image Classifiers and Object Detectors

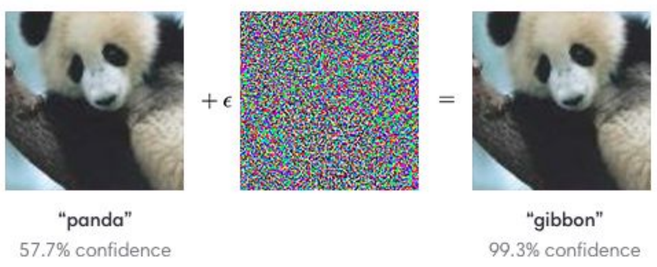
Ivan Evtimov*
ie5@cs.washington.edu
https://ivanevtimov.eu/

PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

*This work was performed in collaboration with Kevin Eykholt, Chaowei Xiao, and Atul Prakash (University of Michigan); Earlence Fernandes and Tadayoshi Kohno (University of Washington); Bo Li and Dawn Song (UC Berkeley); Florian Tramer (Stanford University); and Amir Rahmati (Stony Brook University and Samsung Research America).

Adversarial Examples

Deep neural networks are vulnerable to slight perturbations of their inputs which cause incorrect classifications, a.k.a. ***adversarial examples***.



This canonical example shows that images altered ***digitally*** can mislead a neural network image classifier. Image Courtesy: OpenAI

Does the Physical World Protect Us?

The physical world introduces extra challenges to the adversary.

Challenge: Road signs are perceived at large distances and angles and in varying lighting conditions.

Challenge: With physical adversarial examples, the adversary does not control the exact pixel values.

Digital Picture

Printer Result

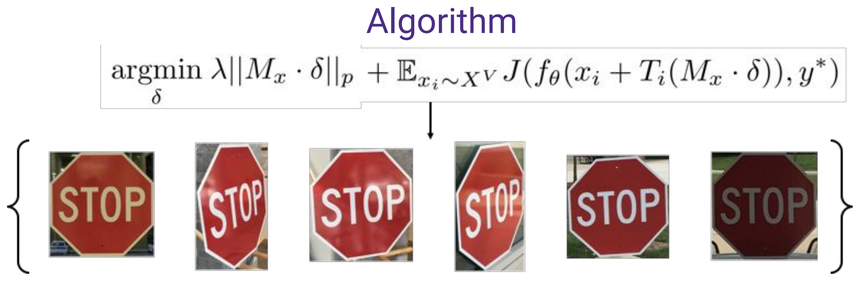
Question: Are these challenges enough to provide a reliable defense against adversarial examples in the physical world?

Robust Physical Perturbations (RP2)

Answer: No. Optimizing over simulated and real physical variations of the image produces adversarial examples effective against the classifier at a distance and from a wide angle.

Algorithm

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||M_x \cdot \delta||_p + \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$$



Results

Speed Limit 45 Sign
(by 95% accurate *CNN on GTSRB*)

Stop Sign
(by *91% accurate CNN on LISA*)

Phone
(by *InceptionV3 on ImageNet*)

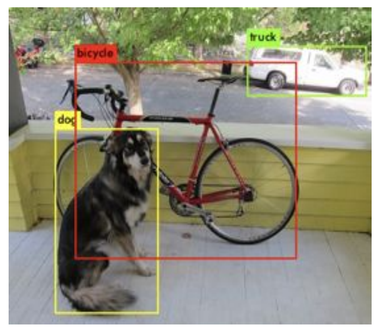
Cash Machine
(by *InceptionV3 on ImageNet*)

See our full paper [Eykholt et al. *Robust Physical World Attacks on Deep Learning Visual Classification*, CVPR 2018] for details and https://youtu.be/1mJMPqi2bSQ for a video.

This work was supported in part by NSF grants 1422211, 1565252, 1616575, 1646392, 1740897, Berkeley Deep Drive, the Center for Long-Term Cybersecurity, FORCES (which receives support from the NSF), the Hewlett Foundation, the MacArthur Foundation, a UM-SJTU grant, and the UW Tech Policy Lab.

Object Detectors and Other Vision Models

Classification: What is the dominant object in this cropped photo?



Object Detection: What are the objects in this scene, and where are they?

Semantic Segmentation: What are the precise shapes and locations of objects?

Are Object Detectors Less Vulnerable?

Detectors are more sophisticated models, so they present extra challenges for the adversary, on top of those arising from the uncertainties of the physical world.

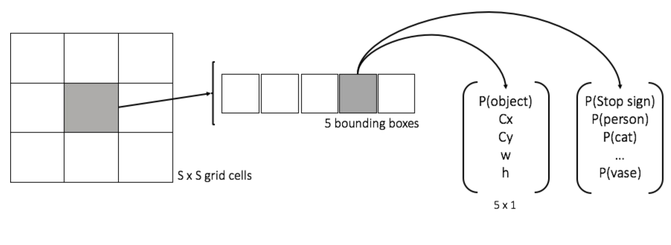
Challenge: Detectors process the entire scene, so they have contextual information at their disposal.

Challenge: The location of the target object within the scene can vary widely.

Extensions of RP2 to Attack Detectors

Detectors are not any more safe. Our algorithm can be extended to produce robust physical perturbations that compromise those models as well.
We demonstrate an attack on the YOLOv2 detector [Redmond and Farhadi *YOLO9000: Better, Faster, Stronger*. 2016. arXiv:1612.08242].

Yolo's Output

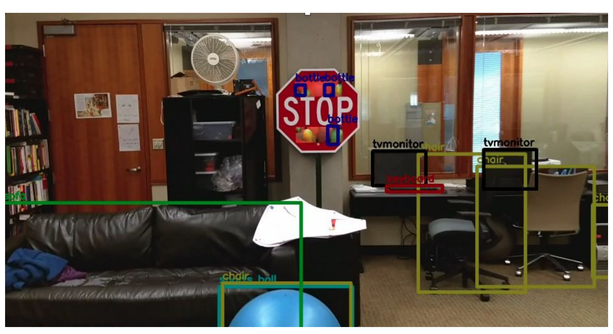


Adapted Attack Optimization Loss

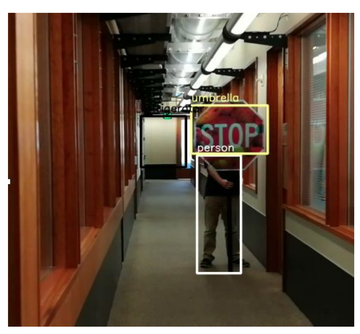
$$J_d(x,y) = \max_{s \in S^2, b \in B} P(s,b,y,f_\theta(x))$$

Prob. of object being class 'y'
Input scene
Output of YOLO, 19 x 19 x 425 tensor

Results from the Extended RP2 Algorithm



YOLO **fails to detect** the stop sign. Video at https://youtu.be/zSFZyzHdTO0



YOLO detects an **umbrella instead of a stop sign**. Video at https://youtu.be/gkKyBmULVvM

See our full paper [Eykholt et al. *Physical Adversarial Examples*, WOOT 2018] for more details.