Robust Physical-World Attacks on Deep Learning Visual Classification

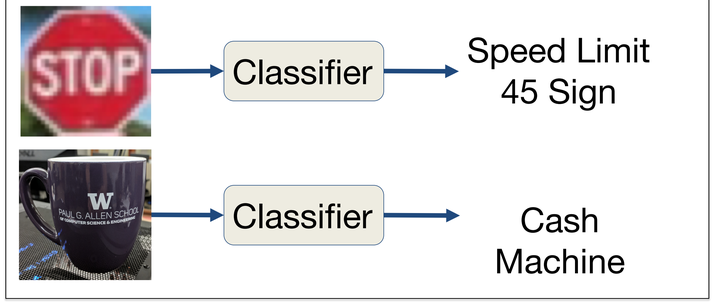
Kevin Eykholt[1], Ivan Etimov[2], Earlence Fernandes[2], Bo Li[3], Amir Rahmati[4], Chaowei Xiao[1], Atul Prakash[1], Tadayoshi Kohno[2], Dawn Song[3]

[1]University of Michigan [2]University of Washington [3] University of California, Berkeley [4]Samsung Research America and Stony Brook University

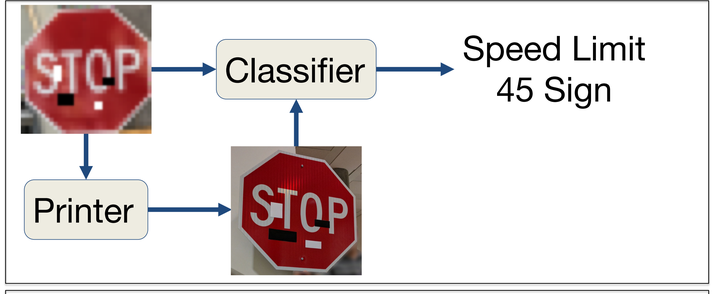
CVPR 2018 SALT LAKE CITY • JUNE 18-22

The Problem

Machine learning is vulnerable to **adversarial examples**



Can we create physical adversarial examples?



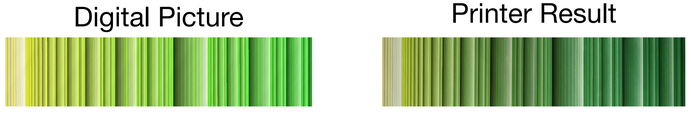
The Challenges of the Physical World

1) Viewer and Environmental Conditions

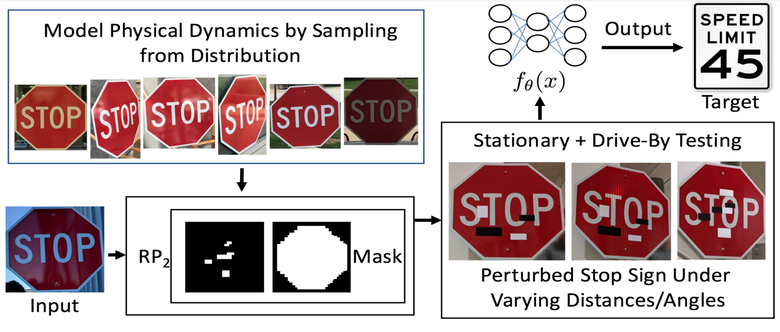
2) Spatial Constraints

3) Limits of Imperceptibility

4) Fabrication Error



Robust Physical Perturbation – RP_2



Objective Function: $\underset{\delta}{\operatorname{argmin}} \lambda \parallel M_x \cdot \delta \parallel + NPS$

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

Experiments

1) Vary Viewing Distances and Angles

2) Vary Mask Shapes

3) Vary Classification Model and Victim Object

Attack Type	Success Rate	Attack Type	Success Rate
Camo Art	100%	Right Turn	73.33%
Camo Art v2	80%	Microwave	90%
Graffiti	66.67%	Cup	71.4%
Subtle	100%		

Visit https://iotsecurity.eecs.umich.edu/\#roadsigns for samples images, videos, and other resources

This research was supported by NSF grants 1422211, 1616575, 1646392, 1740897, and 1565252