The State of Physical Attacks on Deep Learning Systems

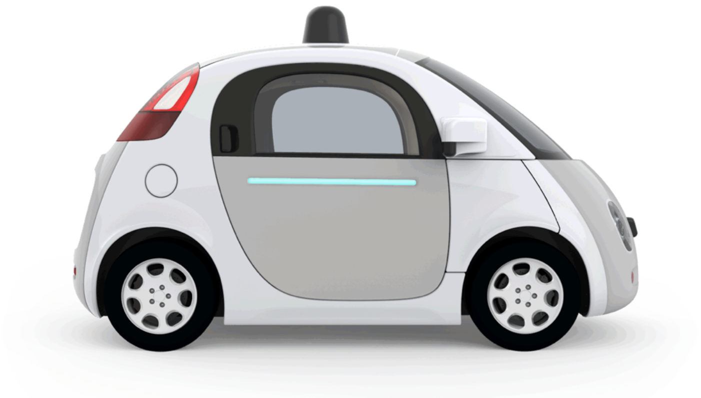
Earlence Fernandes

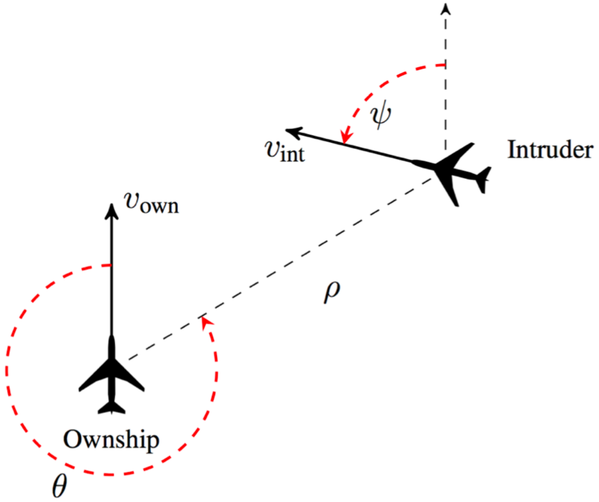
Collaborators:
Ivan Evtimov, Kevin Eykholt, Chaowei Xiao, Amir Rahmati, Florian Tramer, Bo Li, Atul Prakash, Tadayoshi Kohno, Dawn Song

UNIVERSITY of WASHINGTON

Image recognition
Object detection
Scene segmentation
DNA variant calling
Game playing
Speech recognition
Re-enacting politicians
Colorizing photos
Pose estimation
Describing photos
Generating photos
Translation
Music compositions
Creating art
Creating DNNs
Predicting earthquakes
Particle physics
Quantum chemistry
Recommendations
Creating fake news
Fighting fake news
NLP
Automated Surveillance
…

Deep Learning + Cyber-Physical Systems



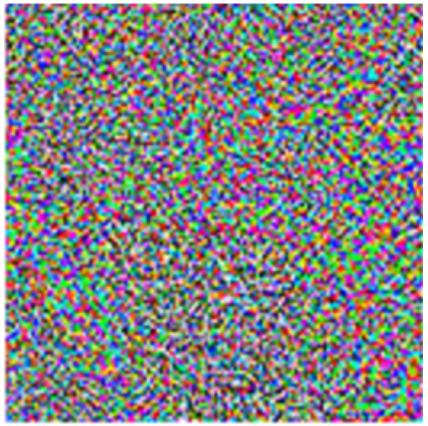


Airborne Collision Avoidance System X unmanned (ACAS Xu)

Apollo (Baidu) Self-Driving Car

The Gibbon-Impersonating Panda
aka, Adversarial Examples

$+ \ \varepsilon$



$=$

Image Credit: OpenAI

"panda"
57.7% confidence

"gibbon"
99.3% confidence

But, an attacker requires pixel-level digital access to the model's input

Explaining and Harnessing Adversarial Examples, Goodfellow et al., arXiv 1412.6572, 2015

How can attackers create physical attacks?

A Compendium of Physical Attacks

Printing out a digitally created adversarial example works, but is less robust to environmental conditions

Fast Gradient Sign Method (FGSM) approach

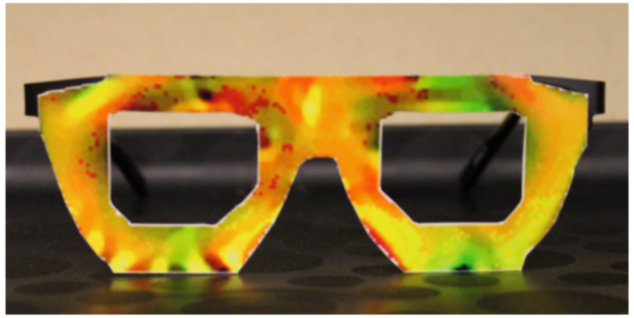
clean

adversarial

Kurakin et al., Adversarial Examples in the Physical World, arXiv 1607.02533, 2016

Printed patterns on eyeglass-shaped cut-outs can compromise face recognition

Optimization approach



Lujo Bauer

Mila Jovovich (87%)

Sharif et al., Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition, CCS 2016

A Compendium of Physical Attacks

Stickers on Stop signs can fool object classifiers and detectors in a range of physical conditions

Optimization approach

My work

Eykholt et al., Robust Physical-World Attacks on Deep Learning Visual Classification, CVPR 2018

Eykholt et al., Physical Adversarial Examples for Object Detectors, WOOT 2018

Chen et al., Robust Physical Adversarial Attack on Faster-RCNN Object Detector, arXiv 1804.05810, 2018

Attackers can backdoor DNNs so that special stickers cause specific behavior

Training-time attack

Gu et al., BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, arXiv 1708.06733, 2017

A Compendium of Physical Attacks

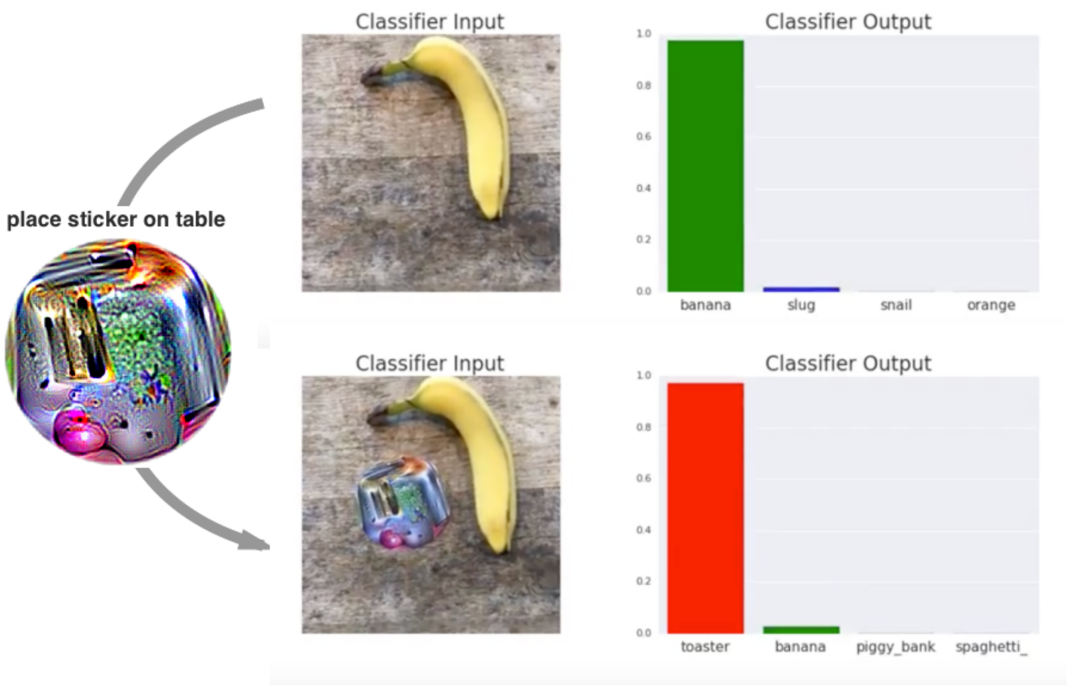
3D printed turtles can be rifles to a state-of-the-art classifier

Expectation-over-Transformations approach (optimization)

Athalye et al., Synthesizing Robust Adversarial Examples, ICML 2018

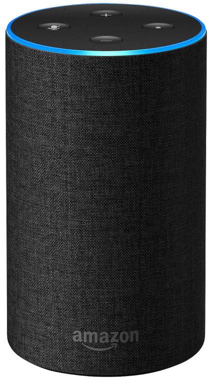
Patches that camouflage any object as a toaster exist

Expectation-over-Transformations approach (optimization)



Brown et al., Adversarial Patch, arXiv 1712.09665, May 2018

Adversarial Examples can hide in music



Carlini et al., Audio Adversarial Examples: Targeted Attacks on Speech-to-Text, DLS Workshop 2018

Yuan et al., CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition, USENIX Security 2018

Open Questions

- Are there other physical domains where we can explore adversarial examples?
- Current attacks only look at a single model. But, a model is only a part of the whole CPS. Do these attacks have system-wide effects?
- Is there anything specific about physical adversarial examples that make them easier or more difficult to defend against?
- Should we only depend on “pure ML” techniques for defense?
- What aspects of CPSs can we leverage to defend (defense in depth)?

Thank you!

- Are there other physical domains where we can explore adversarial examples?
- Current attacks only look at a single model. But, a model is only a part of the whole CPS. Do these attacks have system-wide effects?
- Is there anything specific about physical adversarial examples that make them easier or more difficult to defend against?
- Should we only depend on "pure ML" techniques for defense?
- What aspects of CPSs can we leverage to defend (defense in depth)?

Earlence Fernandes, earlence@cs.washington.edu, earlence.com