Robust Physical-World Attacks on Deep Learning Models

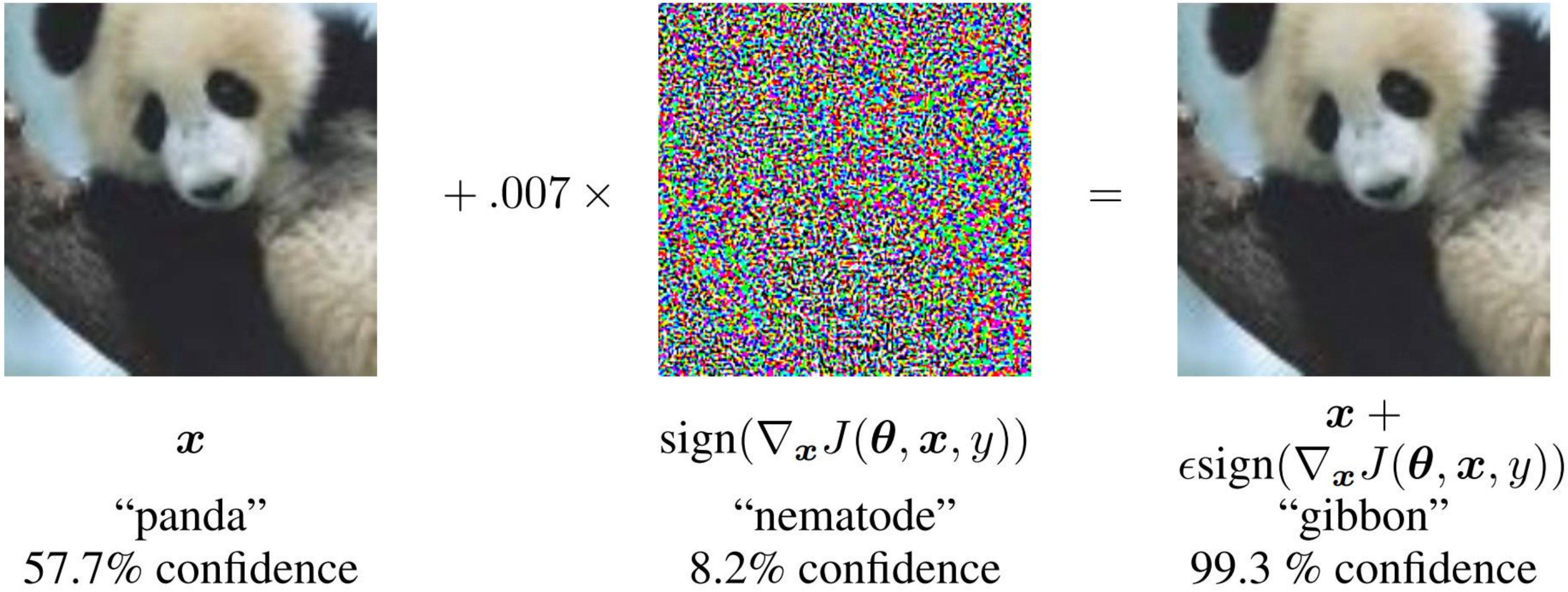
Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati,
Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song

CVPR 2018

UNIVERSITY OF WATERLOO

Introduction

- DNNs are vulnerable to human-imperceptible adversarial perturbations.



UNIVERSITY OF WATERLOO

[1] Explaining and Harnessing Adversarial Examples, Goodfellow et al. 2015

Introduction

- *Objective*: construct robust adversarial perturbations in the physical world.

UNIVERSITY OF WATERLOO

Introduction

- *Objective*: construct robust adversarial perturbations in the physical world.
- *Experimental setting*: Construct printable stickers that can be cut out and placed on physical road signs to cause a DNN classifier to misclassify the road sign.

Introduction

- *Objective*: construct robust adversarial perturbations in the physical world.
- *Experimental setting*: Construct printable stickers that can be cut out and placed on physical road signs to cause a DNN classifier to misclassify the road sign.
- *Adversarial setting*: The proposed method is a targeted white-box attack.

How can we make adversarial examples “work” in the physical world?

Robust physical-world adversarial examples must satisfy the following properties:

How can we make adversarial examples “work” in the physical world?

Robust physical-world adversarial examples must satisfy the following properties:

1. Robust to varying environmental conditions (lighting/distance/angle/weather)

How can we make adversarial examples “work” in the physical world?

Robust physical-world adversarial examples must satisfy the following properties:

1. Robust to varying environmental conditions (lighting/distance/angle/weather)
2. Account for spatial constraints of the adversarial perturbation

How can we make adversarial examples “work” in the physical world?

Robust physical-world adversarial examples must satisfy the following properties:

1. Robust to varying environmental conditions (lighting/distance/angle/weather)
2. Account for spatial constraints of the adversarial perturbation
3. Imperceptible to humans, but perceptible to cameras

How can we make adversarial examples "work" in the physical world?

Robust physical-world adversarial examples must satisfy the following properties:

1. Robust to varying environmental conditions (lighting/distance/angle/weather)
2. Account for spatial constraints of the adversarial perturbation
3. Imperceptible to humans, but perceptible to cameras
4. Account for fabrication errors (e.g., error introduced when printing the perturbation)

General Attack Method

- Constrained Optimization Problem:

$$\min ||\delta||_p \text{ s.t. } f_\theta(x+\delta) = y^*$$

General Attack Method

- Constrained Optimization Problem:

$$\min ||\delta||_p \text{ s.t. } f_\theta(x+\delta) = y^*$$

- Lagrangian-relaxed form:

$$\underset{\delta}{\operatorname{argmin}}\ \lambda \underbrace{||\delta||_p}_{\text{L-}p\text{ norm of }\delta} + \underbrace{J(f_\theta(x+\delta), y^*)}_{\text{Loss function}}$$

UNIVERSITY OF WATERLOO

Robustness to varying environmental conditions (lighting/distance/angle/weather)

Robustness to varying environmental conditions (lighting/distance/angle/weather)

- Collect set of images X^V of object class o (e.g., stop sign) consisting of:
  - *Physical transformations*: real-world images in varying physical conditions, such as lighting, distance, angle and weather
  - *Synthetic transformations:* random crops, varying brightness levels

Robustness to varying environmental conditions (lighting/distance/angle/weather)

- Collect set of images X^V of object class o (e.g., stop sign) consisting of:
  - *Physical transformations*: real-world images in varying physical conditions, such as lighting, distance, angle and weather
  - *Synthetic transformations:* random crops, varying brightness levels

$$\underset{\delta}{\operatorname{argmin}}\ \lambda||\delta||_p + J(f_\theta(x+\delta), y^*)$$

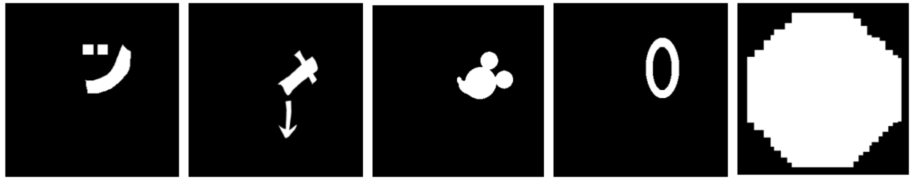
Alignment transformation

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

Accounting for spatial constraints and limits of physical perceptibility

Accounting for spatial constraints and limits of physical perceptibility

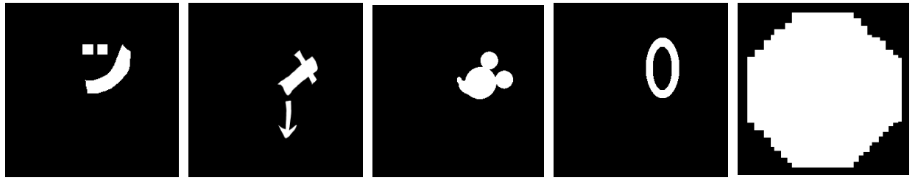
- Utilize a mask $M_x \in \mathbb{R}^d$ to constrain the region of the image where the perturbation can exist.



- Motivated by graffiti on road signs, perturbations hidden “in the human psyche”.

Accounting for spatial constraints and limits of physical perceptibility

- Utilize a mask $M_x \in \mathbb{R}^d$ to constrain the region of the image where the perturbation can exist.



- Motivated by graffiti on road signs, perturbations hidden "in the human psyche".

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

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

How do we choose a good mask?

How do we choose a good mask?

1. Train model with octagonal mask with L_1-norm

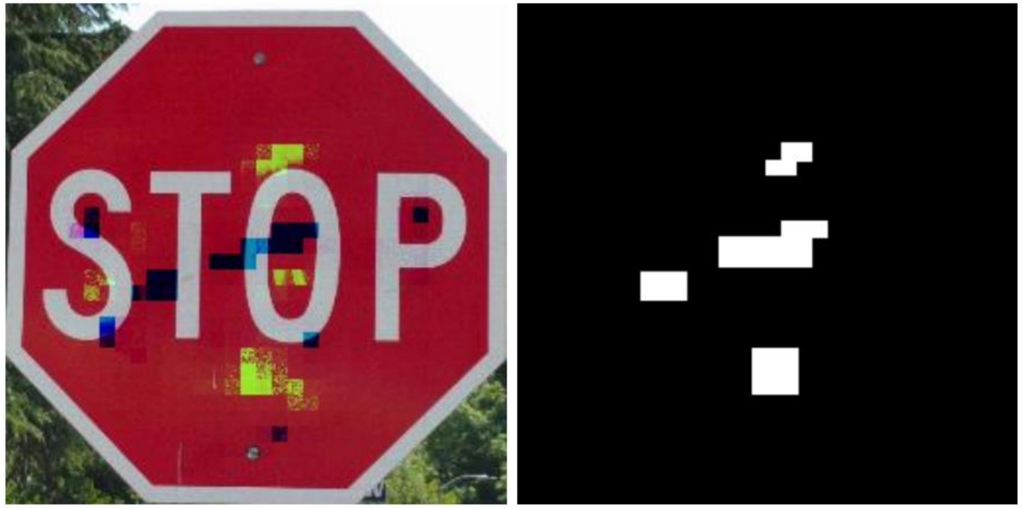
$$\underset{\delta}{\operatorname{argmin}}\, \lambda ||M_x \cdot \delta||_1 + \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$$

How do we choose a good mask?

1. Train model with octagonal mask with L_1-norm

$$\underset{\delta}{\operatorname{argmin}}\, \lambda ||M_x \cdot \delta||_1 + \mathbb{E}_{x_i \sim X^V} J(f_\theta(x_i + T_i(M_x \cdot \delta)), y^*)$$

2. Threshold the highly-activated perturbation regions.



Final Mask

UNIVERSITY OF WATERLOO

Account for fabrication errors

Account for fabrication errors

- Add an additional term to the objective function that encourages the perturbation to be reproducible by the printer.

Account for fabrication errors

- Add an additional term to the objective function that encourages the perturbation to be reproducible by the printer.
- Let $R(\delta)$ be the set of RGB triplets used in perturbation δ and let P be the set of printable RGB triplets:

$$NPS = \sum_{\hat{p} \in R(\delta)} \prod_{p' \in P} |\hat{p} - p'|$$

Account for fabrication errors

- Add an additional term to the objective function that encourages the perturbation to be reproducible by the printer.
- Let $R(\delta)$ be the set of RGB triplets used in perturbation δ and let P be the set of printable RGB triplets:

$$NPS = \sum_{\hat{p}\in R(\delta)} \prod_{p'\in P} |\hat{p} - p'|$$

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

$$\Downarrow$$

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

Account for fabrication errors

- Add an additional term to the objective function that encourages the perturbation to be reproducible by the printer.
- Let $R(\delta)$ be the set of RGB triplets used in perturbation δ and let P be the set of printable RGB triplets:

$$NPS = \sum_{\hat{p} \in R(\delta)} \prod_{p' \in P} |\hat{p} - p'|$$

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

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

Experiments

Experiments

- Attack two trained classifiers:
    - LISA-CNN: Trained on LISA road sign classification dataset. 91% accuracy on test set.
    - GTSRB-CNN: Trained on GT-SRB road sign classification dataset. 95.7% accuracy on test set.

Experiments

- Attack two trained classifiers:
  - LISA-CNN: Trained on LISA road sign classification dataset. 91% accuracy on test set.
  - GTSRB-CNN: Trained on GT-SRB road sign classification dataset. 95.7% accuracy on test set.
- Two types of experiments:
  - Stationary (lab) tests
  - Drive-by (field) tests

Results: Lab Test

Table 1: Sample of physical adversarial examples against LISA-CNN and GTSRB-CNN.

Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Table 5: A camouflage art attack on GTSRB-CNN. See example images in Table 1. The targeted-attack success rate is 80% (true class label: Stop, target: Speed Limit 80).

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
5′ 0°	Speed Limit 80 (0.88)	Speed Limit 70 (0.07)
5′ 15°	Speed Limit 80 (0.94)	Stop (0.03)
5′ 30°	Speed Limit 80 (0.86)	Keep Right (0.03)
5′ 45°	Keep Right (0.82)	Speed Limit 80 (0.12)
5′ 60°	Speed Limit 80 (0.55)	Stop (0.31)
10′ 0°	Speed Limit 80 (0.98)	Speed Limit 100 (0.006)
10′ 15°	Stop (0.75)	Speed Limit 80 (0.20)
10′ 30°	Speed Limit 80 (0.77)	Speed Limit 100 (0.11)
15′ 0°	Speed Limit 80 (0.98)	Speed Limit 100 (0.01)
15′ 15°	Stop (0.90)	Speed Limit 80 (0.06)
20′ 0°	Speed Limit 80 (0.95)	Speed Limit 100 (0.03)
20′ 15°	Speed Limit 80 (0.97)	Speed Limit 100 (0.01)
25′ 0°	Speed Limit 80 (0.99)	Speed Limit 70 (0.0008)
30′ 0°	Speed Limit 80 (0.99)	Speed Limit 100 (0.002)
40′ 0°	Speed Limit 80 (0.99)	Speed Limit 100 (0.002)

Results: Field Test

Perturbation	Attack Success	A Subset of Sampled Frames $k = 10$
Subtle poster	100%	
Camouflage abstract art	84.8%	

In the Press

IEEE Spectrum FOR THE TECHNOLOGY INSIDER

Type to search

NEWS | TRANSPORTATION

Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms › Minor changes to street sign graphics can fool machine learning algorithms into thinking the signs say something completely different

BY EVAN ACKERMAN | PUBLISHED 04 AUG 2017 | 5 MIN READ

Auto Tech

It is surprisingly easy to bamboozle a self-driving car

Researchers confused cameras into misinterpreting signs with a few small tricks and a lot of math.

Andrew Krok

Aug. 7, 2017 1:35 p.m. PT

2 min read

EMILY DREYFUSS SECURITY AUG 5, 2017 7:00 AM

Security News This Week: A Whole New Way to Confuse Self-Driving Cars

Each Saturday we roundup the major security news of the week.

You can confuse self-driving cars by altering street signs

It doesn't take much to send autonomous cars crashing into each other.

Researchers Find a Malicious Way to Meddle with Autonomous Cars

C/D MARK HARRIS AUG 4, 2017

HOME · CARS · NEWS

Stickers on street signs can confuse self-driving cars, researchers show

By Trevor Mogg

August 6, 2017

SHARE

ars TECHNICA

Hacking street signs with stickers could confuse self-driving cars

Conclusions

- Are self-driving cars at risk based solely on *this work*?
- No! This work did not conduct any experiments with an autonomous vehicle. To make this conclusion, a more complete attack must be proposed that targets the full autonomous driving pipeline.

Conclusions

- Are self-driving cars at risk based solely on *this work*?
- No! This work did not conduct any experiments with an autonomous vehicle. To make this conclusion, a more complete attack must be proposed that targets the full autonomous driving pipeline.
- Are self-driving cars potentially at risk based solely on *this work*?
- Absolutely!

Conclusions

- Are self-driving cars at risk based solely on *this work*?
- No! This work did not conduct any experiments with an autonomous vehicle. To make this conclusion, a more complete attack must be proposed that targets the full autonomous driving pipeline.
- Are self-driving cars potentially at risk based solely on *this work*?
- Absolutely!

Any questions? Please send me an email! l6rowe@uwaterloo.ca