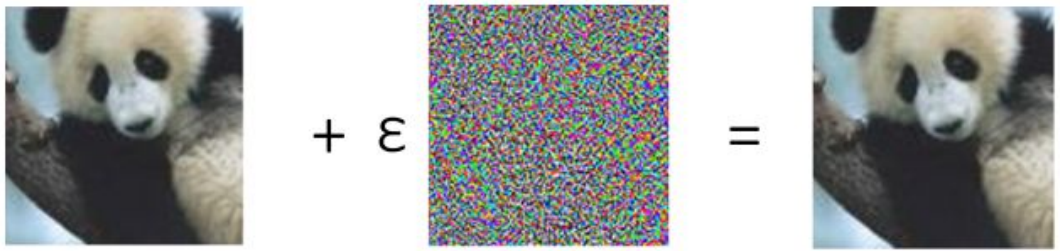
Robust Physical-World Attacks on Deep Learning Visual Classification

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

Introduction

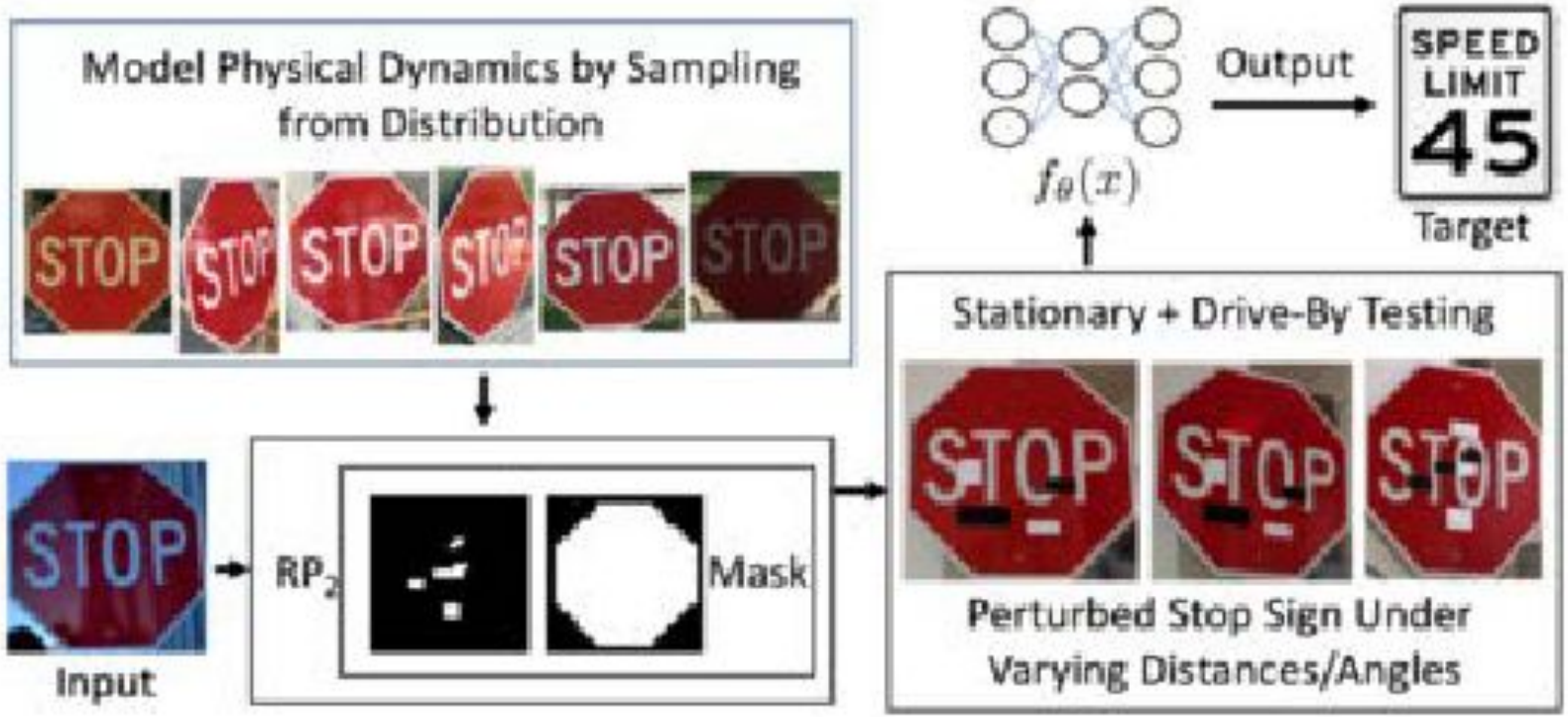
- Computer vision uses Deep Neural Networks (DNN)
  - DNNs are weak to adversarial perturbations
- Most previous adversarial examples work in the digital space
  - What about when adversarial perturbations are added to the physical objects itself?
- Viewpoint of object creates challenging difficulty
  - Adversarial attack must be robust
- Goal: make stickers and posters that lead to misclassification



Adversarial Perturbations for Physical Objects

- Attack must fit on sign
  - Cannot modify background
- Attack must be printable
  - Printers have limitations and tolerances
- Attack should ‘blend in’
  - Perturbation should look subtle or like graffiti
- Regularize the optimization using Lagrangian-relaxed form (L_1)
  - Makes the optimization sparse, meaning focus on hitting the model where it is weakest
- Sample many different signs in many different conditions
  - Different distances, backgrounds, and angles
  - Samples are randomly cropped, brightness is changed, and spatially transformed
  - Helps make the attack more robust

Pipeline


Model Physical Dynamics by Sampling from Distribution
STOP
STOP
STOP
STOP
STOP
STOP
$f_\theta(x)$
Output
SPEED LIMIT 45
Target
Stationary + Drive-By Testing
STOP
STOP
STOP
Perturbed Stop Sign Under Varying Distances/Angles
STOP
Input
RP_2
Mask

Optimization

1) Untargeted : $\arg min_{\delta} \lambda \|\delta\|_p - J(f_\theta(x+\delta), y)$
2) Targeted : $\arg min_{\delta} \lambda \|\delta\|_p + J(f_\theta(x+\delta), y^*)$

δ : perturbation, λ: regularization coefficient, x : input,

y : authentic class, y* : target class, $\|.\|_p$: 2D p-norm($\sum_{i,j}(|\delta_{(i,j)}|^p)^{1/p}$), J : cross entropy, θ : hyper parameter

Credit: SYSSEC ChoManGi

Optimization

Consider various distances, angles, brightness for loss calculations.

1) Untargeted : $\arg min_{\delta} \lambda \|\delta\|_p - J(f_\theta(x+\delta), y)$
2) Targeted : $\arg min_{\delta} \lambda \|\delta\|_p + J(f_\theta(x+\delta), y^*)$

1) Untargeted : $\arg min_{\delta} \lambda \|\delta\|_p - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+\delta), y)$
2) Targeted : $\arg min_{\delta} \lambda \|\delta\|_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+\delta), y^*)$

Average for distance, angle, brightness!

Credit: SYSSEC ChoManGi

Optimization

Use mask matrix to modify specific areas only.

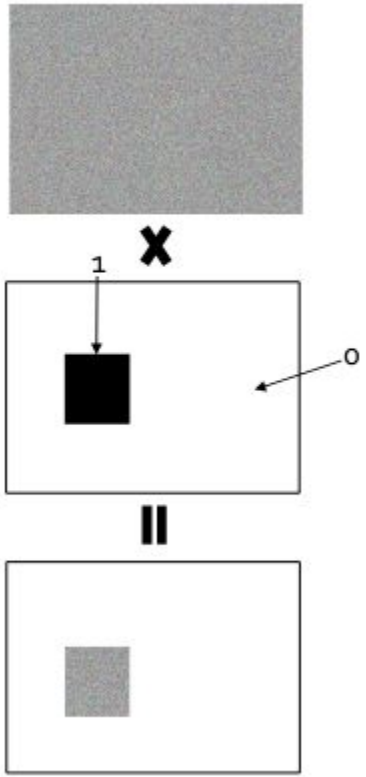
1) Untargeted : $\arg min_{\delta} \lambda \|\delta\|_p - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+\delta), y)$

2) Targeted : $\arg min_{\delta} \lambda \|\delta\|_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+\delta), y^*)$

3) Untargeted : $\arg min_{\delta} \lambda \|M_x \delta\|_p - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+M_x\delta), y)$

4) Targeted : $\arg min_{\delta} \lambda \|M_x \delta\|_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x+M_x\delta), y^*)$

Perturbate only matrix area!



Credit: SYSSEC ChoManGi

Optimization

NPS(p) = $\prod_{\hat{p}\in P}|p - \hat{p}|$, P = set of printable colors, p = color of each pixels

1) Untargeted : $\arg min_{\delta} \lambda \|M_x\delta\|_p - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x + M_x\delta), y)$

2) Targeted : $\arg min_{\delta} \lambda \|M_x\delta\|_p + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x + M_x\delta), y^*)$

3) Untargeted : $\arg min_{\delta} \lambda \|M_x\delta\|_p + NPS(M_x\delta) - \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x + M_x\delta), y)$

4) Targeted : $\arg min_{\delta} \lambda \|M_x\delta\|_p + NPS(M_x\delta) + \frac{1}{k}\sum_{i=1}^{k} J(f_\theta(x + M_x\delta), y^*)$

Don't use non-printable color!

Credit: SYSSEC ChoManGi

Experiments

- Two classifiers used
  - LISA-CNN and GTSRB-CNN, with 91% and 95% accuracy respectively
  - Both classifiers use LISA stop sign images
- Two different test types
  - Stationary lab test
  - Moving vehicle test
  - Images taken at various distances and angles
- Lighting not controlled in different settings

Experiments

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%

Experiments

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

LISA-CNN

Experiments

Distance & Angle	Poster-Printing		Sticker			
	Subtle		Camouflage–Graffiti		Camouflage–Art	
5′ 0°	SL45 (0.86)	ADL (0.03)	STP (0.40)	SL45 (0.27)	SL45 (0.64)	LE (0.11)
5′ 15°	SL45 (0.86)	ADL (0.02)	STP (0.40)	YLD (0.26)	SL45 (0.39)	STP (0.30)
5′ 30°	SL45 (0.57)	STP (0.18)	SL45 (0.25)	SA (0.18)	SL45 (0.43)	STP (0.29)
5′ 45°	SL45 (0.80)	STP (0.09)	YLD (0.21)	STP (0.20)	SL45 (0.37)	STP (0.31)
5′ 60°	SL45 (0.61)	STP (0.19)	STP (0.39)	YLD (0.19)	SL45 (0.53)	STP (0.16)
10′ 0°	SL45 (0.86)	ADL (0.02)	SL45 (0.48)	STP (0.23)	SL45 (0.77)	LE (0.04)
10′ 15°	SL45 (0.90)	STP (0.02)	SL45 (0.58)	STP (0.21)	SL45 (0.71)	STP (0.08)
10′ 30°	SL45 (0.93)	STP (0.01)	STP (0.34)	SL45 (0.26)	SL45 (0.47)	STP (0.30)
15′ 0°	SL45 (0.81)	LE (0.05)	SL45 (0.54)	STP (0.22)	SL45 (0.79)	STP (0.05)
15′ 15°	SL45 (0.92)	ADL (0.01)	SL45 (0.67)	STP (0.15)	SL45 (0.79)	STP (0.06)
20′ 0°	SL45 (0.83)	ADL (0.03)	SL45 (0.62)	STP (0.18)	SL45 (0.68)	STP (0.12)
20′ 15°	SL45 (0.88)	STP (0.02)	SL45 (0.70)	STP (0.08)	SL45 (0.67)	STP (0.11)
25′ 0°	SL45 (0.76)	STP (0.04)	SL45 (0.58)	STP (0.17)	SL45 (0.67)	STP (0.08)
30′ 0°	SL45 (0.71)	STP (0.07)	SL45 (0.60)	STP (0.19)	SL45 (0.76)	STP (0.10)
40′ 0°	SL45 (0.78)	LE (0.04)	SL45 (0.54)	STP (0.21)	SL45 (0.68)	STP (0.14)

LISA-CNN

Experiments

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
5' 0°	Stop (0.39)	Speed Limit 45 (0.10)
5' 15°	Yield (0.20)	Stop (0.18)
5' 30°	Stop (0.13)	Yield (0.13)
5' 45°	Stop (0.25)	Yield (0.18)
5' 60°	Added Lane (0.15)	Stop (0.13)
10' 0°	Stop (0.29)	Added Lane (0.16)
10' 15°	Stop (0.43)	Added Lane (0.09)
10' 30°	Added Lane (0.19)	Speed limit 45 (0.16)
15' 0°	Stop (0.33)	Added Lane (0.19)
15' 15°	Stop (0.52)	Right Turn (0.08)
20' 0°	Stop (0.39)	Added Lane (0.15)
20' 15°	Stop (0.38)	Right Turn (0.11)
25' 0°	Stop (0.23)	Added Lane (0.12)
30' 0°	Stop (0.23)	Added Lane (0.15)
40' 0°	Added Lane (0.18)	Stop (0.16)

LISA-CNN

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)

GTSRB-CNN

Attack Generalized

Distance/Angle	Image	Distance/Angle	Image
8″0°		8″15°	
12″0°		12″15°	
16″0°		16″15°	
20″0°		20″15°	
24″0°		24″15°	
28″0°		28″15°	
32″0°		32″15°	

Distance/Angle	Image	Distance/Angle	Image
2′ 0°		2′ 15°	
5′ 0°		5′ 15°	
7′ 0°		7′ 15°	
10′ 0°		10′ 15°	
15′ 0°		20′ 0°	

Attack Generalized

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
2' 0°	Phone (0.78)	Microwave (0.03)
2' 15°	Phone (0.60)	Microwave (0.11)
5' 0°	Phone (0.71)	Microwave (0.07)
5' 15°	Phone (0.53)	Microwave (0.25)
7' 0°	Phone (0.47)	Microwave (0.26)
7' 15°	Phone (0.59)	Microwave (0.18)
10' 0°	Phone (0.70)	Microwave (0.09)
10' 15°	Phone (0.43)	Microwave (0.28)
15' 0°	Microwave (0.36)	Phone (0.20)
20' 0°	Phone (0.31)	Microwave (0.10)

Inception-v3

Distance & Angle	Top Class (Confid.)	Second Class (Confid.)
8" 0°	Cash Machine (0.53)	Pitcher (0.33)
8" 15°	Cash Machine (0.94)	Vase (0.04)
12" 0°	Cash Machine (0.66)	Pitcher (0.25)
12" 15°	Cash Machine (0.99)	Vase (<0.01)
16" 0°	Cash Machine (0.62)	Pitcher (0.28)
16" 15°	Cash Machine (0.94)	Vase (0.01)
20" 0°	Cash Machine (0.84)	Pitcher (0.09)
20" 15°	Cash Machine (0.42)	Pitcher (0.38)
24" 0°	Cash Machine (0.70)	Pitcher (0.20)
24" 15°	Pitcher (0.38)	Water Jug (0.18)
28" 0°	Pitcher (0.59)	Cash Machine (0.09)
28" 15°	Cash Machine (0.23)	Pitcher (0.20)
32" 0°	Pitcher (0.50)	Cash Machine (0.15)
32" 15°	Pitcher (0.27)	Mug (0.14)

Inception-v3

Thoughts

- Paper focuses on white-box setting
  - Model accessible
  - Model can be extracted from black-box
  - Technique not useful against systems without meaningful access
- Attack focuses on single model. Model is only part of a cyber-physical system
- Hard to prevent detection, Jiajun Lu et al. 2017

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