

Robust Physical-World Attacks on Deep Learning Visual Classification

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

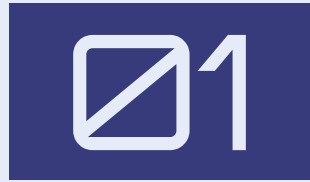
Publication Type : Proceeding

Publication : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

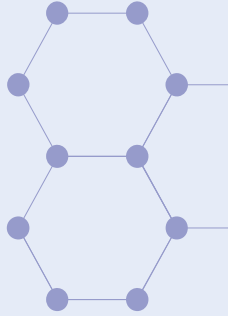
Year : 2018, pp. 1625-1634

Cited by 1582





Introduction



Introduction



- Deep Neural Networks (DNNs) เป็น state-of-the-art ในหลายงาน และบางครั้งก็สามารถแข่งขันกับมนุษย์ได้ในงานด้าน computer vision จำนวนมาก.
- จากความสำเร็จเหล่านี้ มีการนำ Computer Vision ไปใช้มากขึ้น โดยเป็นส่วนหนึ่งของ pipelines ในส่วน physical systems เช่น รถ, อากาศยานไร้คนขับ และและหุ่นยนต์



Introduction

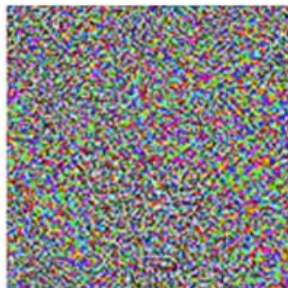
Deep Neural Networks are Useful, But Vulnerable



"panda"

57.7% confidence

+ ϵ



=



"gibbon"

99.3% confidence

Image Credit:
OpenAI

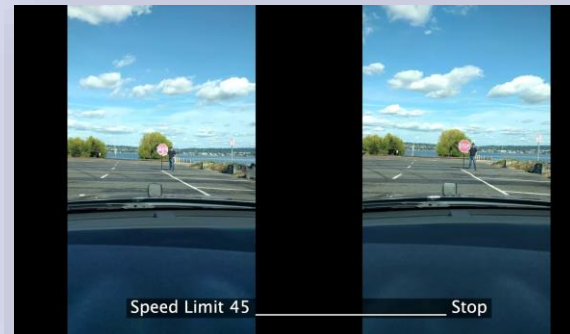
Introduction



- งานนี้มีส่วนช่วยในการทำความเข้าใจ DNN กับตัวอย่างชุดข้อมูลที่มีการเพิ่มสิ่งรบกวนทางกายภาพให้กับวัตถุ โดยเลือกการจำแนกประเภทป้ายจราจรเป็นโดเมนเป้าหมาย เหตุผล คือ
 1. ป้ายจราจรเป็นวัตถุที่เรียบ สามารถมองเห็นชัดเจน ทำให้ยากต่อการซ่อนสิ่งรบกวน
 2. ป้ายจราจรมีอยู่ในสภาพแวดล้อมที่ไม่มีข้อจำกัดและเงื่อนไขทางกายภาพที่เปลี่ยนแปลงอย่างเช่นระยะทางและมุมของกล้อง
 3. ป้ายจราจรมีบทบาทสำคัญในความปลอดภัยในการขนส่ง
 4. Threat model for transportation ที่สมเหตุสมผล คือ Attacker อาจไม่สามารถควบคุมระบบของยานพาหนะได้แต่สามารถปรับเปลี่ยนหรือรบกวนวัตถุทางกายภาพที่ยานพาหนะอาจต้องพึ่งพาเพื่อทำการตัดสินใจด้านความปลอดภัย

Introduction

- ความท้าทายหลักในการสร้างสิ่งรบกวนทางกายภาพ คือความแปรปรวนของสิ่งแวดล้อม
- ความท้าทายในการใช้งานจริงอื่นๆ:
 1. การรบกวนในโลกดิจิทัลอาจมีขนาดเล็กมากจนเป็นไปได้ว่ากล้องจะไม่สามารถรับรู้สิ่งเหล่านั้นได้เนื่องจากความไม่สมบูรณ์ของเซ็นเซอร์
 2. เป็นการยากมากที่จะสร้างสิ่งรบกวนให้กับวัตถุด้วยการปรับเปลี่ยนพื้นหลัง เนื่องจากวัตถุจริงสามารถมีพื้นหลังที่แตกต่างกันได้ขึ้นอยู่กับมุมมอง
 3. กระบวนการผลิต (เช่น การพิมพ์/สร้าง สิ่งรบกวน) ไม่สมบูรณ์



Contributions

01

นำเสนอ Robust Physical Perturbations (RP2) เพื่อสร้างการรบกวนทางกายภาพสำหรับวัตถุทางกายภาพที่สามารถทำให้เกิดการจำแนกประเภทที่ไม่ถูกต้องใน DNN-based classifier ภายใต้สภาวะทางกายภาพแบบไดนามิก รวมถึงมุมมองและระยะทางที่แตกต่างกัน

03

ประเมินการโจมตี/การสร้างสิ่งรบกวนให้กับ Input ให้กับตัวแบบจำแนกประเภทที่สร้างขึ้น: LISA-CNN ที่มีความแม่นยำ 91% ในชุดทดสอบ LISA และ GTSRB-CNN ที่มีความแม่นยำ 95.7% ในชุดทดสอบ GTSRB

02

เนื่องจากขาดวิธีการที่เป็นมาตรฐานในการประเมินการรบกวนทางกายภาพ จึงเสนอวิธีการประเมินเพื่อศึกษาประสิทธิภาพของการรบกวนทางกายภาพในสถานการณ์จริง

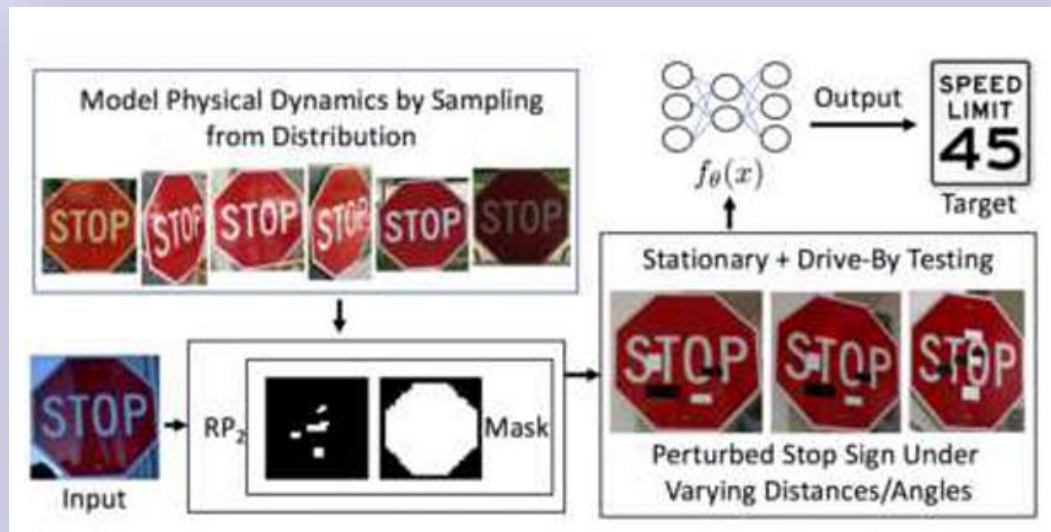
04

เพื่อแสดงถึงแนวทางในลักษณะทั่วไปในงานนี้ ได้สร้างสิ่งรบกวนกับวัตถุทางกายภาพทั่วไป เช่น ไมโครเวฟ ที่แสดงให้เห็นว่า pre-trained Inception-v3 classifier จำแนกประเภทไมโครเวฟเป็น "โทรศัพท์" ผิดโดยเพิ่มสติกเกอร์เดียว



Adversarial Examples for Physical Objects

Robust Physical Perturbation (RP2)



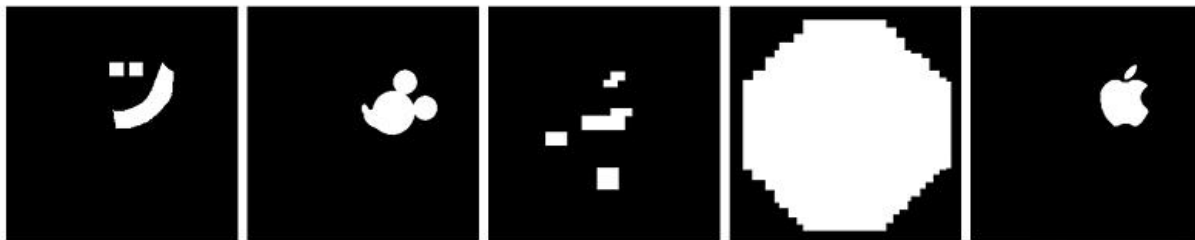
ค้นหาการรบกวน δ ที่จะเพิ่มลงในอินพุต x
โดยที่อินสแตนซ์ที่รบกวน $\hat{x} = x + \delta$ ถูกจำแนกประเภทโดย Target Classifier $f_{\theta}(\cdot)$

Optimizing Spatial Constraints

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

Perturbation/Noise Matrix M_x
 Lp norm (L0, L1, L2, ...) $\|\cdot\|_p$
 Loss Function J
 The target class y^*

Example
Masks



Subtle Poster
Camouflage Sticker

Approximate vandalism



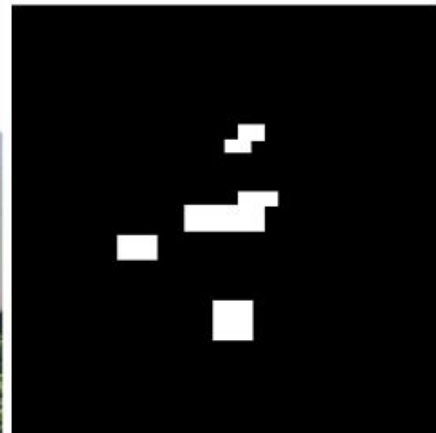
How To Choose A Mask?



Possibility: Mask surface area should be large or should be focused on “sensitive” regions

Use L-1

$$\operatorname{argmin}_{\delta} \lambda ||M_x \cdot \delta||_p$$

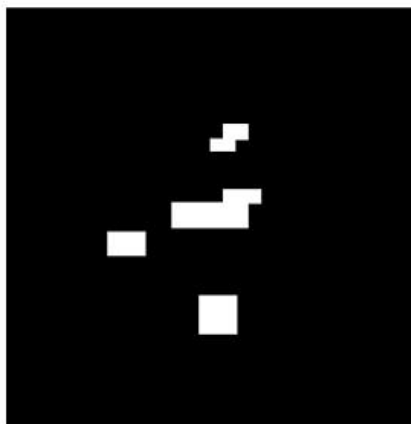


$$+ \mathbb{E}_{x_i \sim X^v} J(f_{\theta}(x_i + T_i(M_x \cdot \delta)), y^*)$$

Process of Creating a Useful Sticker Attack



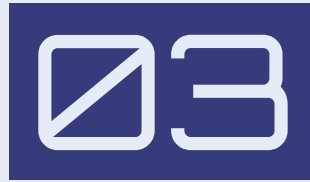
L-1 Perturbation



Result Mask



Sticker Attack!



Experiments

Experimental Design

Attack the LISA-CNN and GTSRB-CNN

91%

accuracy on test-set



LISA-CNN

ฝึกอบรมตัวแบบด้วยชุดข้อมูล U.S. traffic sign ประกอบด้วย 17 ป้ายจราจรที่พบบ่อยที่สุด

95.7%

accuracy on test-set




























GTSRB-CNN

ฝึกอบรมตัวแบบด้วยชุดข้อมูล ป้ายจราจรของประเทศเยอรมัน และ ภาพ U.S. Stop sign จากชุดข้อมูลที่ใช้ฝึกอบรมในตัวแบบ LISA-CNN

Results for LISA-CNN

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%

➤ The attack's target class
Speed Limit 45

Results for LISA-CNN

Table 2: Targeted physical perturbation experiment results on LISA-CNN using a poster-printed Stop sign (subtle attacks) and a real Stop sign (camouflage graffiti attacks, camouflage art attacks). For each image, the top two labels and their associated confidence values are shown. The misclassification target was Speed Limit 45. See Table 1 for example images of each attack. Legend: SL45 = Speed Limit 45, STP = Stop, YLD = Yield, ADL = Added Lane, SA = Signal Ahead, LE = Lane Ends.

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)

Results for GTSRB-CNN

Table 3: 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)

- The attack's target class Speed Limit 80

Results for Inceptionv3



Figure 3: Physical adversarial example against the Inception-v3 classifier. The left shows the original cropped image identified as microwave (85.2%) while the right shows the cropped physical adversarial example identified as phone (77.8%).

➤ ผีกรอบมตัวแบบบนข้อมูล

ImageNet

Run the code

evtimovi / **robust_physical_perturbations** Public

Watch 7 Fork 31 Star 91

Code Issues 4 Pull requests 2 Actions Projects Security Insights

master 3 branches 0 tags

Go to file Add file Code

evtimovi adding mapping.txt with indices to class labels 3810af0 on Dec 10, 2019 15 commits

gtsrb-cnn-attack	GTSRB readme	5 years ago
imagenet-attack	README updates	5 years ago
lisa-cnn-attack	adding mapping.txt with indices to class labels	3 years ago
.gitignore	added imagenet attack readme and updated inception download	5 years ago
LICENSE	license time	5 years ago
README.md	README updates	5 years ago

README.md

This repository holds the code (and some results) used in [Robust Physical-World Attacks on Deep Learning Visual Classification](#). The software carries an MIT license.

The folders are as follows:

- `lisa-cnn-attack` holds the code to attack the LISA-CNN that classifies US road signs from the LISA dataset.

About

Public release of code for Robust Physical-World Attacks on Deep Learning Visual Classification (Eykholt et al., CVPR 2018)

iotsecurity.eecs.umich.edu/#roadsigns

Readme MIT license 91 stars 7 watching 31 forks

Releases

No releases published

Packages

No packages published

Official: https://github.com/evtimovi/robust_physical_perturbations.git

Run the code

shangtse / robust-physical-attack Public

Watch 7 Fork 46 Star 145

Code Issues 1 Pull requests Actions Projects Security Insights

master 1 branch 0 tags

Go to file Add file Code

Physical adversarial attack for fooling the Faster R-CNN object detector

computer-vision faster-rcnn object-detection adversarial-examples adversarial-attacks

Readme BSD-3-Clause license 145 stars 7 watching 46 forks

Releases No releases published

Packages

dxoigmn Fix bug in Makefile f8c5510 on Jan 14, 2020 21 commits

File	Description	Time
data	Merged commit includes the following changes:	3 years ago
imgs	Merged commit includes the following changes:	3 years ago
.gitignore	Add 2d ShapeShifter examples	3 years ago
LICENSE	Add 2d/3d ShapeShifter code	3 years ago
Makefile	Fix bug in Makefile	3 years ago
Pipfile	Add 2d/3d ShapeShifter code	3 years ago
Pipfile.lock	Add 2d/3d ShapeShifter code	3 years ago
README.md	Merged commit includes the following changes:	3 years ago
lucid.diff	Merged commit includes the following changes:	3 years ago
object_detection_api.diff	Add 2d/3d ShapeShifter code	3 years ago
robust_physical_attack.ipynb	initial commit	5 years ago
shapeshifter.py	Merged commit includes the following changes:	3 years ago

Official: <https://github.com/shangtse/robust-physical-attack.git>



THANK
YOU