THE UNIVERSITY OF CHICAGO

SaTC: CORE: Medium: Hidden Rules in Neural Networks as Attacks and Adversarial Defenses

PI: Ben Y. Zhao co-PIs: Pedro Lopes, Heather Zheng

Technical Challenges:

- Backdoor attacks train hidden rules into neural network models
- A significant threat against neural network models
- Hard to detect due to inherent "diversity" of backdoor triggers in the physical world

Technical Contributions:

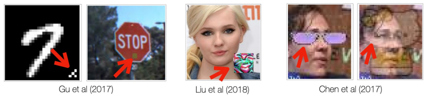
Study *physical backdoors* (physical object as triggers) [1]

1. Highly effective attacks
2. Break assumptions made by existing defenses
3. Clear need for stronger defenses

Using hidden behavior as defenses

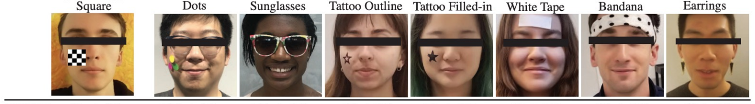
1. As honeypot to detect inference time attacks [2]
2. Invisible perturbations to prevent photos from being used in facial recognition models [3]

Physical World Attacks: Physical Triggers



Prior work: digital triggers

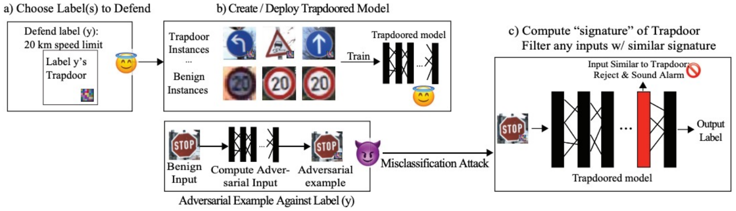
key question:
Can ***everyday physical objects*** serve as triggers in backdoor attacks?



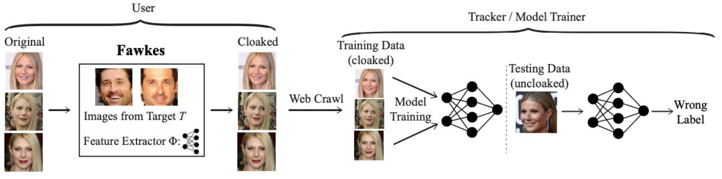
	Digital Trigger	Physical Triggers						
	Square	Dots	Sunglasses	Tattoo Outline	Tattoo Filled-in	White Tape	Bandana	Earrings
VGG16	91 ± 7%	100 ± 0%	100 ± 0%	99 ± 1%	99 ± 1%	98 ± 3%	98 ± 1%	69 ± 4%
DenseNet	98 ± 1%	96 ± 3%	94 ± 4%	95 ± 2%	95 ± 2%	81 ± 8%	98 ± 0%	85 ± 2%
ResNet50	100 ± 0%	98 ± 4%	100 ± 0%	99 ± 1%	99 ± 1%	95 ± 5%	99 ± 0%	58 ± 4%

Attack success rates of physical triggers in facial recognition models trained on various architectures.

Hidden Rules as Defenses



[2] Deploying backdoors as honeypots for adversarial examples



[3] "cloaking" photos to protect against invasive facial recognition

Scientific Impact:

- Demonstrates *physical backdoors* as a real-world threat, especially since they break key assumptions made in SOTA backdoor defenses (designed/tested on digital triggers)
- "Backdoor as a defense" is a new concept for protecting DNN models and user privacy
- Demonstrating similar concepts / mechanisms in voice

Broader Impact:

- Fawkes featured in NYTimes, global media, binary downloads > 841K
- Support student training that combines ML and security, integrated research projects in multiple courses
- Recruited 7 female UG researchers and 1 female HS student

Selected Publications

[1] *Backdoor Attacks Against Deep Learning Systems in the Physical World, CVPR 2021*

[2] *Gotta Catch 'Em All: Using Honeypots to Catch Adversarial Attacks on Neural Networks, CCS 2020*

[3] *Fawkes: Protecting Personal Privacy against Unauthorized Deep Learning Models, USENIX Security 2020*

[4] *Blacklight: Scalable Defense for Neural Networks against Query-Based Black-Box Attacks, USENIX Security 2022*

[5] *"Hello, It's Me": Deep Learning-based Speech Synthesis Attacks in the Real World, CCS 2021*