THE UNIVERSITY OF CHICAGO

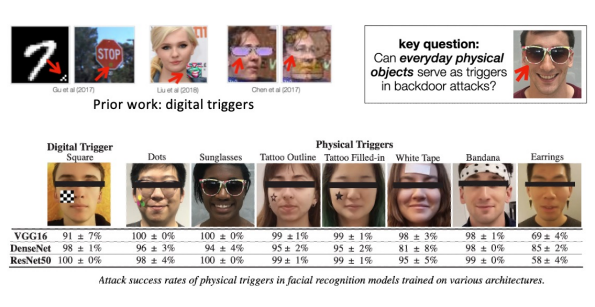
SaTC: CORE: Medium: Hidden Rules in Neural Networks as Attacks and Adversarial Defenses

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

https://sandlab.cs.uchicago.edu/nsf-backdoors

Backdoor attacks train hidden rules into neural network models. It is a significant threat against neural network models, but is also hard to detect due to inherent "diversity" of backdoor triggers in the physical world. In this SaTC project, we study new, novel methods to defend against practical backdoor attacks, and use backdoors to build strong defenses.

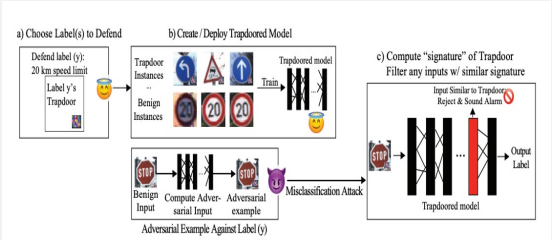
Focus 1: Physical world backdoor attacks (physical object as triggers)



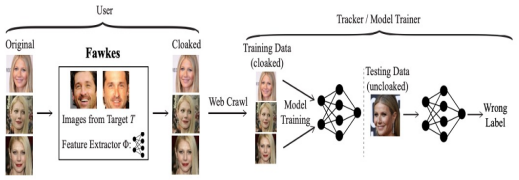
Physical backdoors [1]

- Highly effective attacks in the real world
- Break key assumptions made by SOTA backdoor defenses, which are designed and tested on *digital* triggers
- Clear need for stronger defenses

Focus 2: Hidden rules (backdoors) as defenses



[2] Deploying backdoors as honeypots for adversarial examples

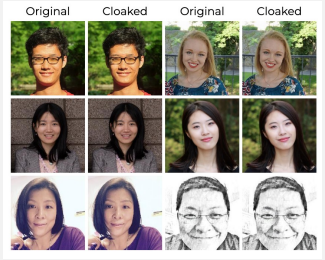


[3] "cloaking" photos to protect against invasive facial recognition

- "Backdoor as a defense" is a new concept for protecting DNN models and user privacy
- Demonstrating similar concepts and mechanisms in voice

Broader Impacts

- Fawkes featured in NYTimes and global media (50+ sources), binary software downloads > 841K https://sandlab.cs.uchicago.edu/fawkes



Education and Outreach

- Supported student training that combines ML and security
- Integrated research projects in multiple courses in security, mobile computing & HCI
- 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

The 5th NSF Secure and Trustworthy Cyberspace Principal Investigator Meeting (2022 SaTC PI Meeting)
June 1-2, 2022 | Arlington, Virginia

Award ID#: 1949650
University of Chicago