Pluribus One
seeing one in many

Pattern Recognition and Applications Lab

University of Cagliari, Italy

Attacks on Machine Learning

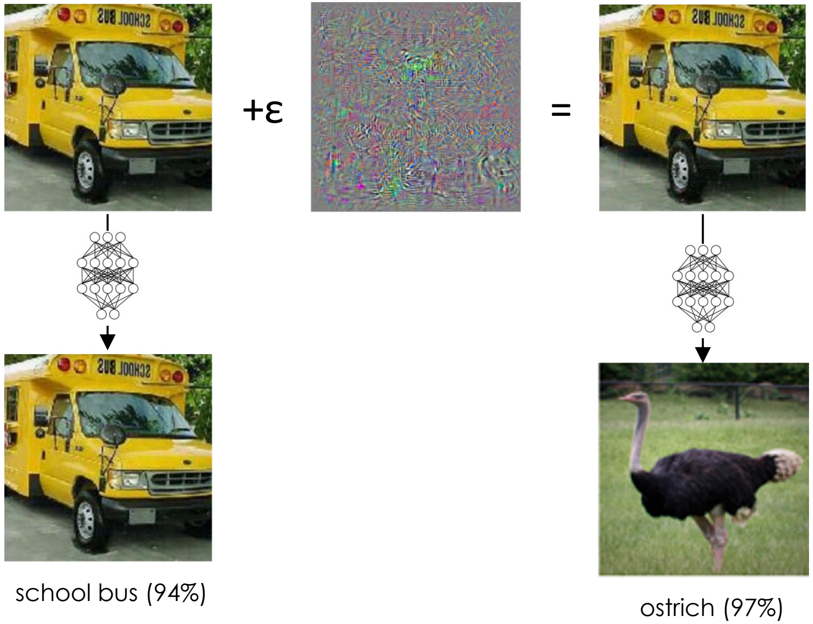
Battista Biggio
battista.biggio@unica.it
@biggiobattista
University of Cagliari, Italy

ENISA-ETSI Joint Workshop on Remote Identity Proofing – May 3, 2022

The Elephant in the Room: *Adversarial Examples*

- AI/ML successful in many applications
  - Computer Vision
  - Speech Recognition
  - Cybersecurity
  - Healthcare

- ... but extremely *fragile* against *adversarial examples*
  - Carefully-perturbed inputs that mislead classification



Biggio et al., Evasion attacks against machine learning at test time, **ECML-PKDD 2013**
Szegedy et al., Intriguing properties of neural networks, **ICLR 2014**

Attacks against AI are Pervasive!

Sharif et al., *Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,* ACM CCS 2016

"without the dataset the article is useless"

"okay google browse to evil dot com"

Carlini and Wagner, *Audio adversarial examples: Targeted attacks on speech-to-text*, DLS 2018 https://nicholas.carlini.com/code/audio_adversarial_examples/

Eykholt et al., *Robust physical-world attacks on deep learning visual classification*, CVPR 2018

- Demetrio, Biggio, Roli et al., *Adversarial EXEmples: ...*, ACM TOPS 2021
- Demetrio, Biggio, Roli et al., *Functionality-preserving black-box optimization of adversarial windows malware*, IEEE TIFS 2021
- Demontis, Biggio, Roli et al., *Yes, Machine Learning Can Be More Secure!...*, IEEE TDSC 2019

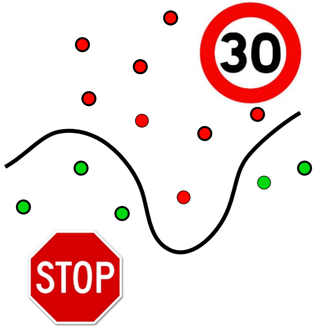
Attacks against Machine Learning

Attacker's Goal

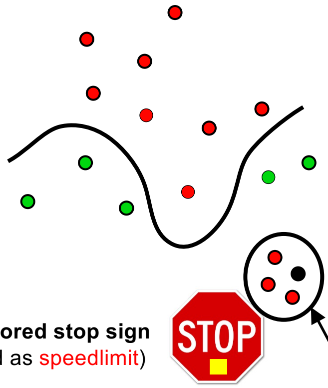
Attacker's Capability	Misclassifications that do not compromise normal system operation **Integrity**	Misclassifications that compromise normal system operation **Availability**	Querying strategies that reveal confidential information on the learning model or its users **Privacy / Confidentiality**
Test data	**Evasion (a.k.a. adversarial examples)**	Sponge attacks	Model extraction / stealing Model inversion (hill climbing) Membership inference
Training data	Backdoor poisoning (to allow subsequent intrusions) – e.g., backdoors or neural trojans	DoS poisoning (to maximize classification error)	-

Backdoor/Poisoning Attacks

Training data (no poisoning)



Training data (poisoned)



Backdoored stop sign
(labeled as speedlimit)

What Is the Magic Behind These Attacks?

- Adversarial attacks work as they generate out-of-distribution samples (i.e., something quite different from the known training samples used to build your model)

- Optimizing the perturbation requires substantial knowledge of the targeted system/training data or, alternatively, querying it multiple times (~ tens of thousands)
  - Trivial mechanisms to detect whether MLaaS is being abused can be easily set up (e.g., detecting similar and repeated input queries coming from the same IP)

- For remote ID proofing, I would be more concerned about *deepfakes* and other impersonating mechanisms (presentation attacks)
  - They can still be detected if generated with known techniques (there are even visible artefacts...)
  - But their combination with adversarial techniques may enable them to stay undetected / become much more realistic