AI in Cyber Security

Le magicien quantique[1]

2023

[1]Aka Sckathach, Caml Master, Fan2Thermo

A bit of Geopolitics

- Japan: no law on copyright, researchers have full access to data to train their models

A bit of Geopolitics

- Japan: no law on copyright, researchers have full access to data to train their models
- United States: laws, regulations and recommandations

Safe and Effective Systems

Algorithmic Discrimination Protections

Data Privacy

Notice and Explanation

Human Alternatives, Consideration, and Fallback

AI & Defense

AI for Defense (ML4SEC)

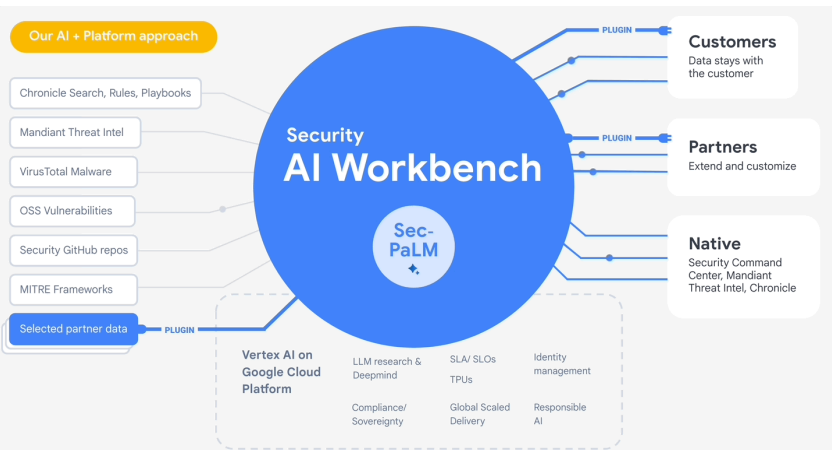
- Spams detection ($\sim$ 1960)
- Network intrusion detection (2000)

AI for Defense (ML4SEC)

- Spams detection ($\sim$ 1960)
- Network intrusion detection (2000)
- Code review

AI for Defense (ML4SEC)

- Spams detection ($\sim$ 1960)
- Network intrusion detection (2000)
- Code review
- Reverse and malware detection $\rightarrow$ Google Sec-PaLM



AI & Attacks

AI for Attack

- Adversarial attacks on AI powered defenses
- Phishing, spear phishing
- Deepfakes
- Targetted missinformation
- Web fuzzing
- Botnet control
- Hiding malwares insides images
- ...

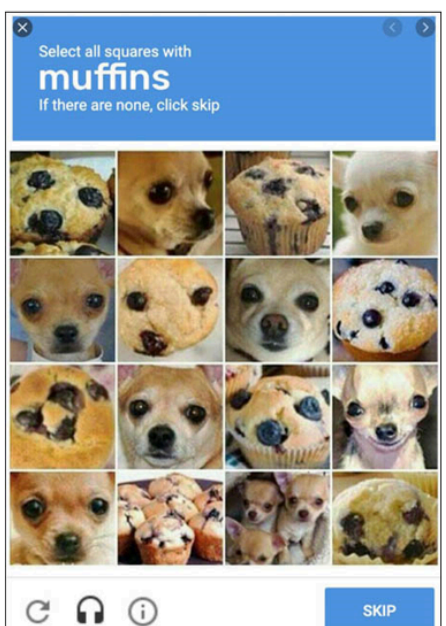
Attacking AI models

Attacking AI models

Data poisoning

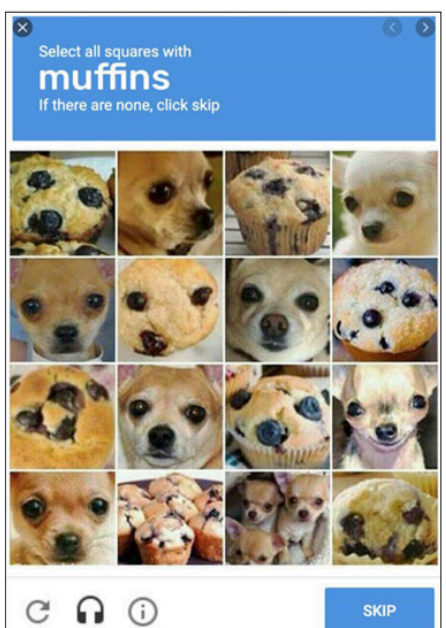
Attacking AI models

Data poisoning



Attacking AI models

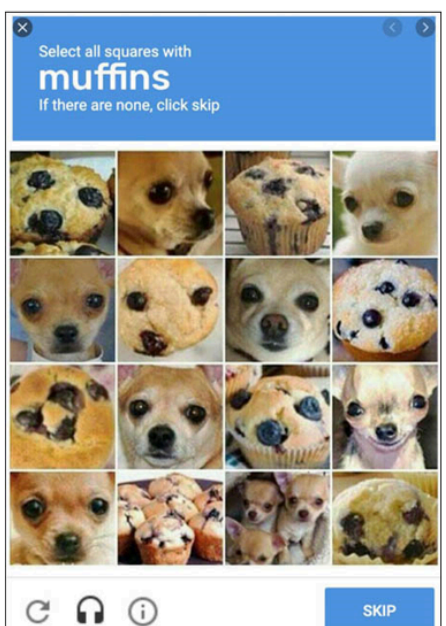
Data poisoning



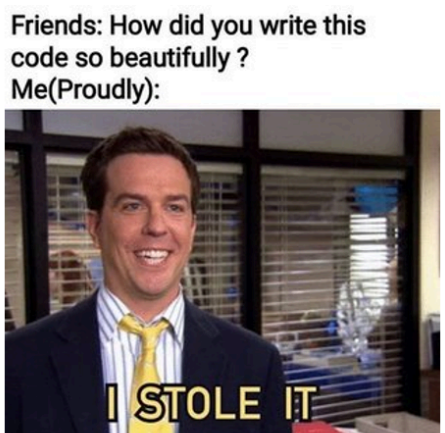
Data extraction

Attacking AI models

Data poisoning

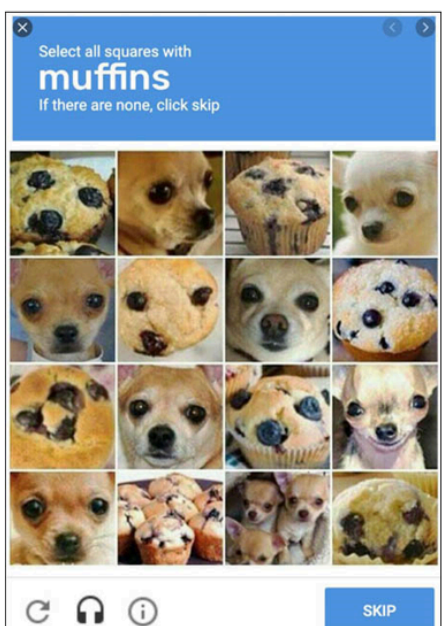


Data extraction

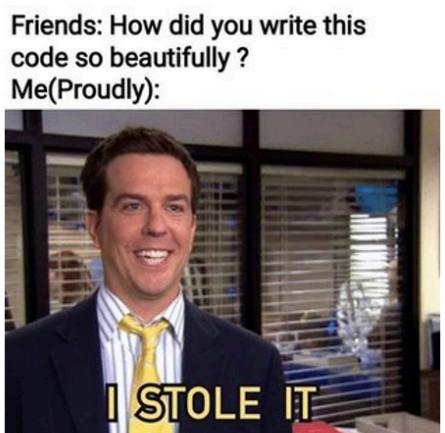


Attacking AI models

Data poisoning



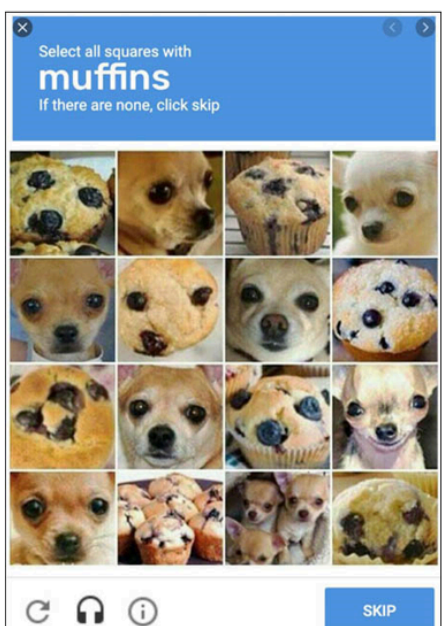
Data extraction



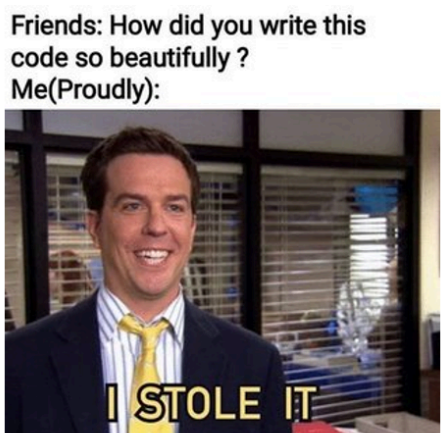
Adversarial attacks

Attacking AI models

Data poisoning



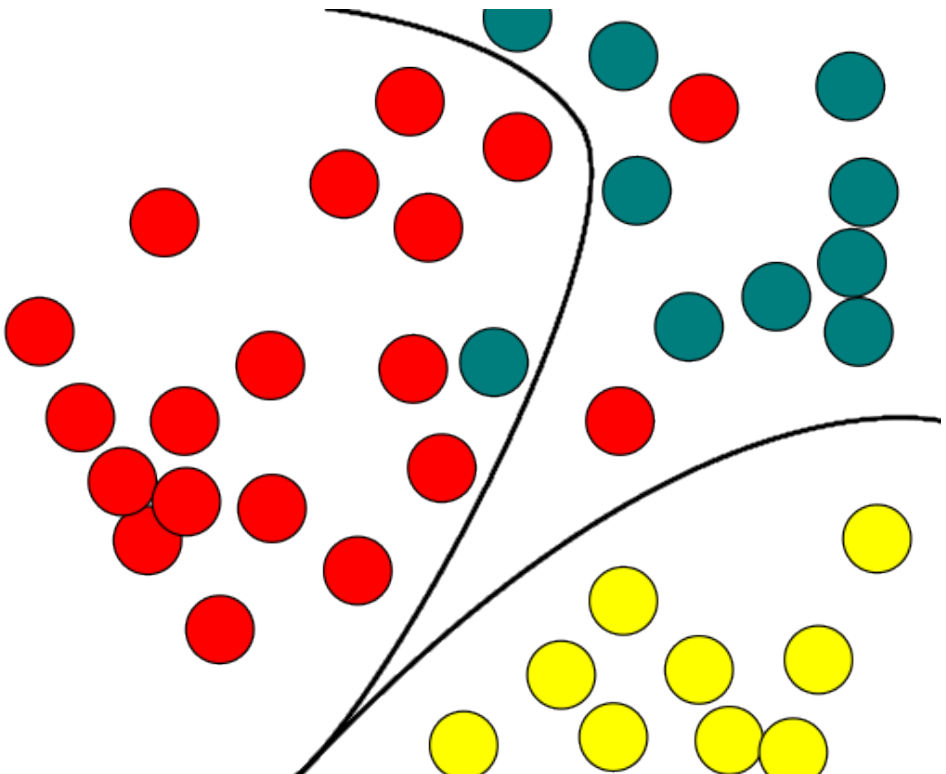
Data extraction



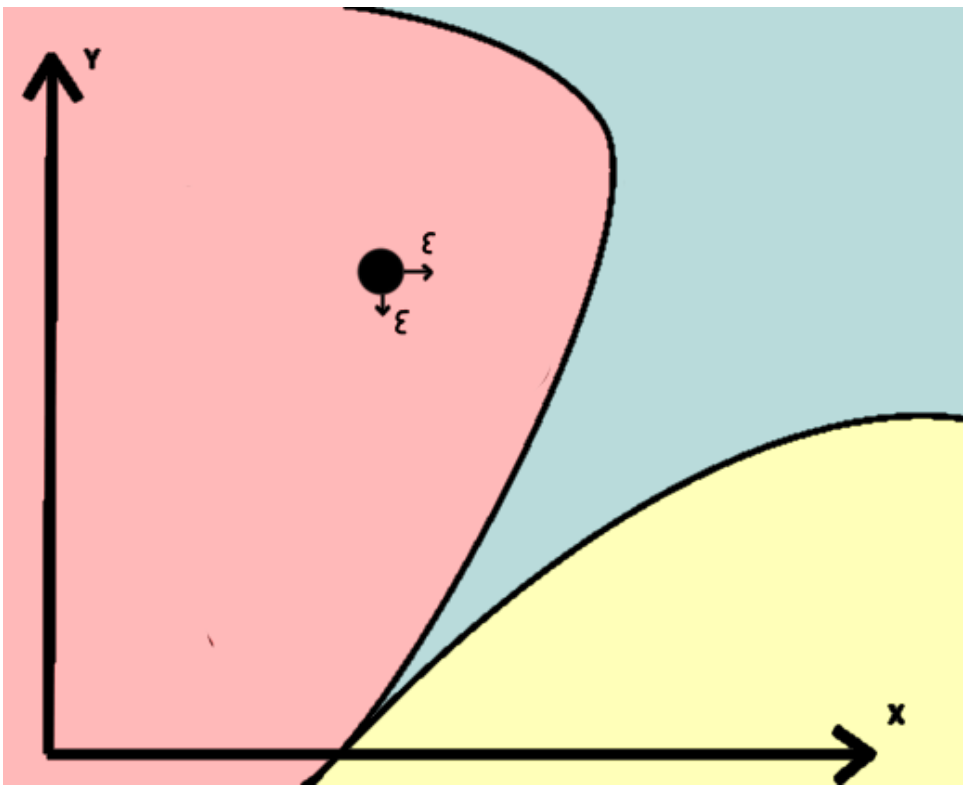
Adversarial attacks

Attacking Classifiers

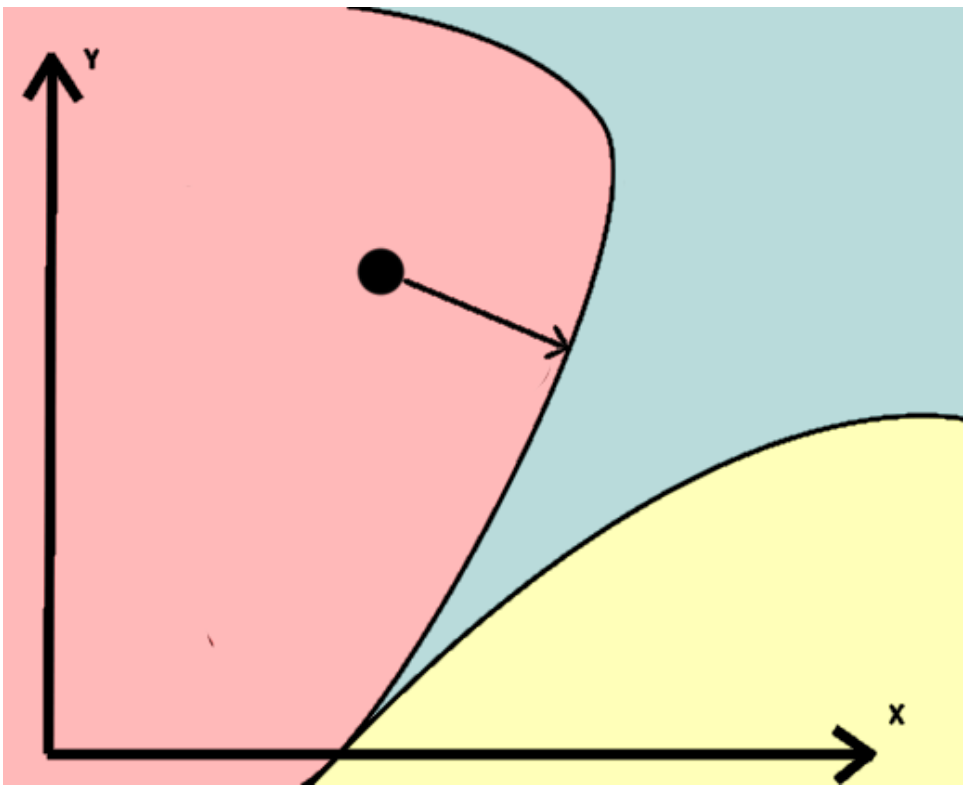
What is a classifier?



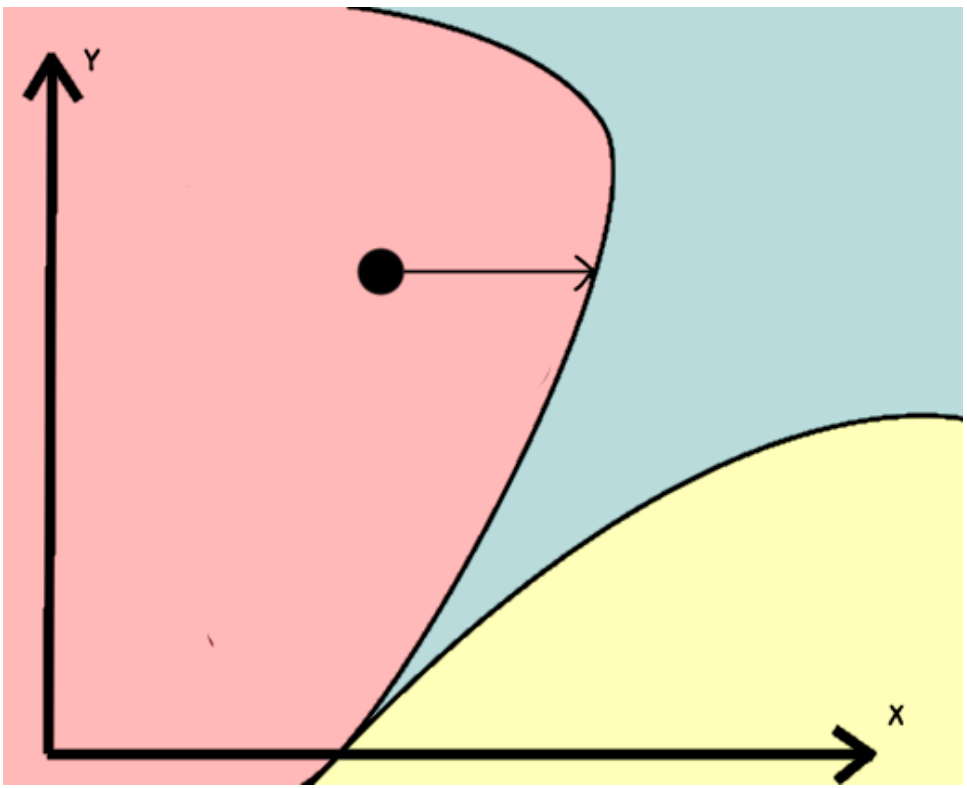
L_∞ norm



L_2 norm



L_0 norm



Whitebox Adversarial Attacks

- Fast Gradient Sign Method (FGSM): L_∞
- Basic Iterative Method (BIM): L_∞

Whitebox Adversarial Attacks

- Fast Gradient Sign Method (FGSM): L_∞
- Basic Iterative Method (BIM): L_∞
- DeepFool (DF): L_2

Whitebox Adversarial Attacks

- Fast Gradient Sign Method (FGSM): L_∞
- Basic Iterative Method (BIM): L_∞
- DeepFool (DF): L_2
- Jacobian-based Saliency Maps Attacks (JSMA): L_0

Whitebox Adversarial Attacks

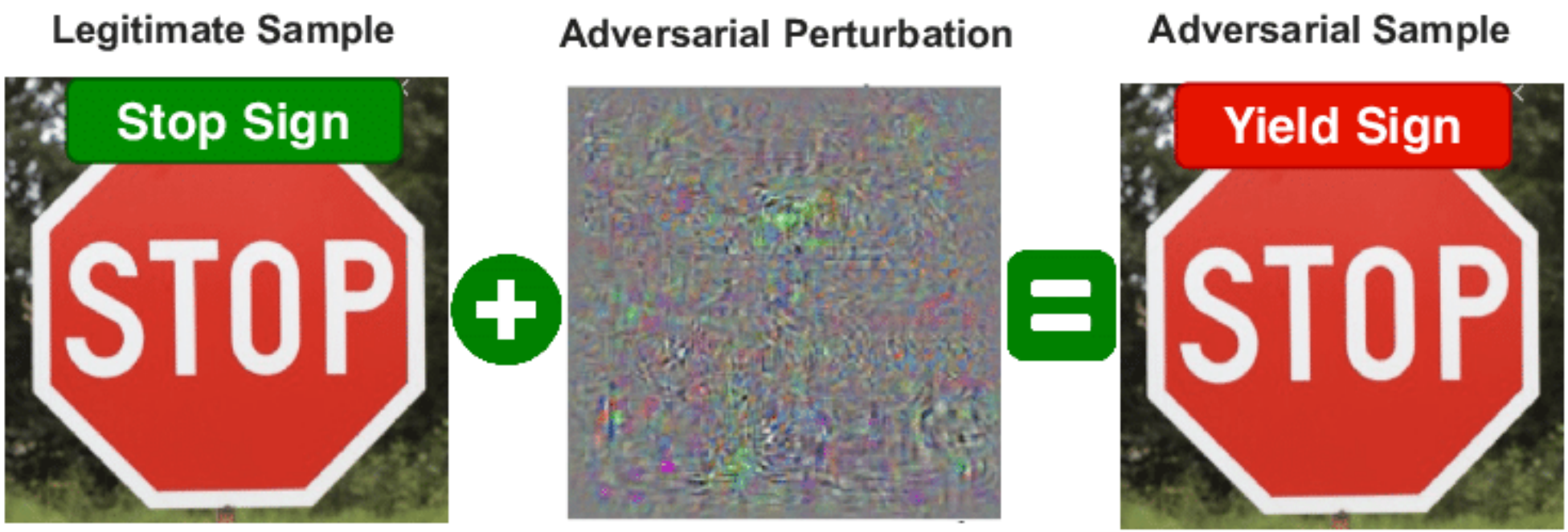
- Fast Gradient Sign Method (FGSM): L_∞
- Basic Iterative Method (BIM): L_∞
- DeepFool (DF): L_2
- Jacobian-based Saliency Maps Attacks (JSMA): L_0
- Carlini and Wagner (CW): L_∞, L_2, L_0

Blackbox Adversarial Attacks

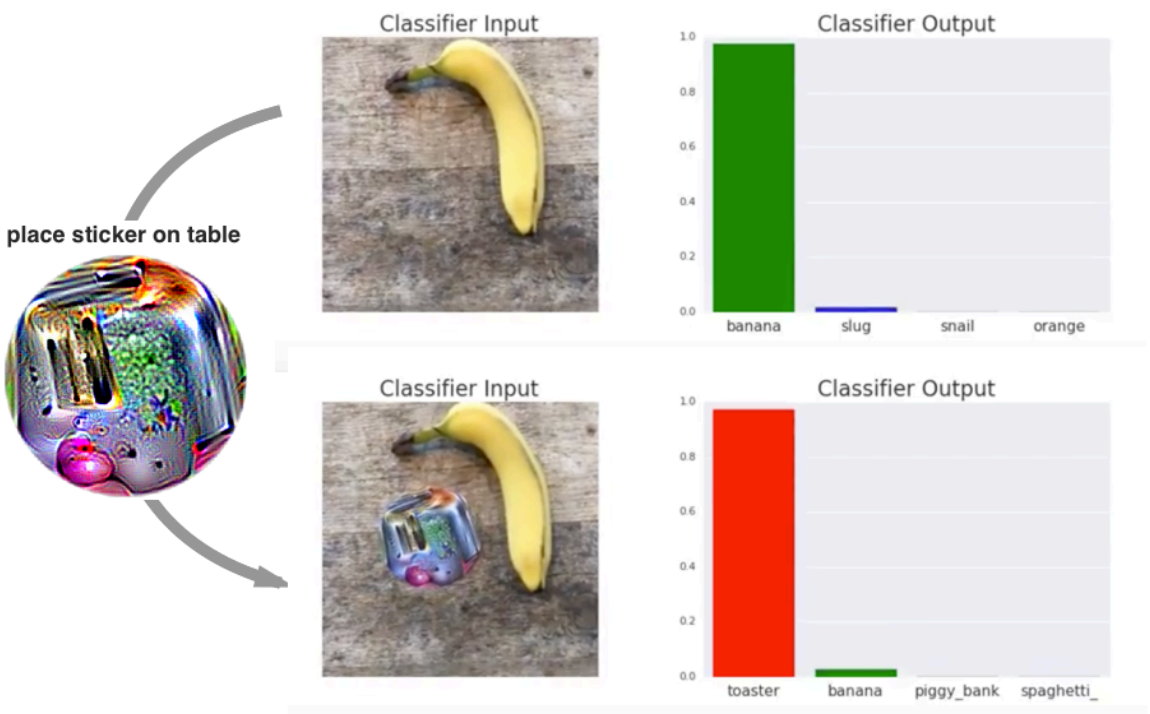
- Zeroth Order Optimisation attack (ZOO)
- Transferable Attentive Attack (TAA)
- Attack Inspired GAN (AI-GAN)

Examples

FGSM noise attack: L_∞



JSMA patch attack: L_0



Practice!