Khwopa College of Engineering

ADVERSIAL ATTACK

RAM KATWAL

Adversary attack on Tesla Cars

A two inch piece of tape fooled Tesla's cameras.

It made Tesla Cars Autonomously Accelerate Up To 85 In a 35 Zone.

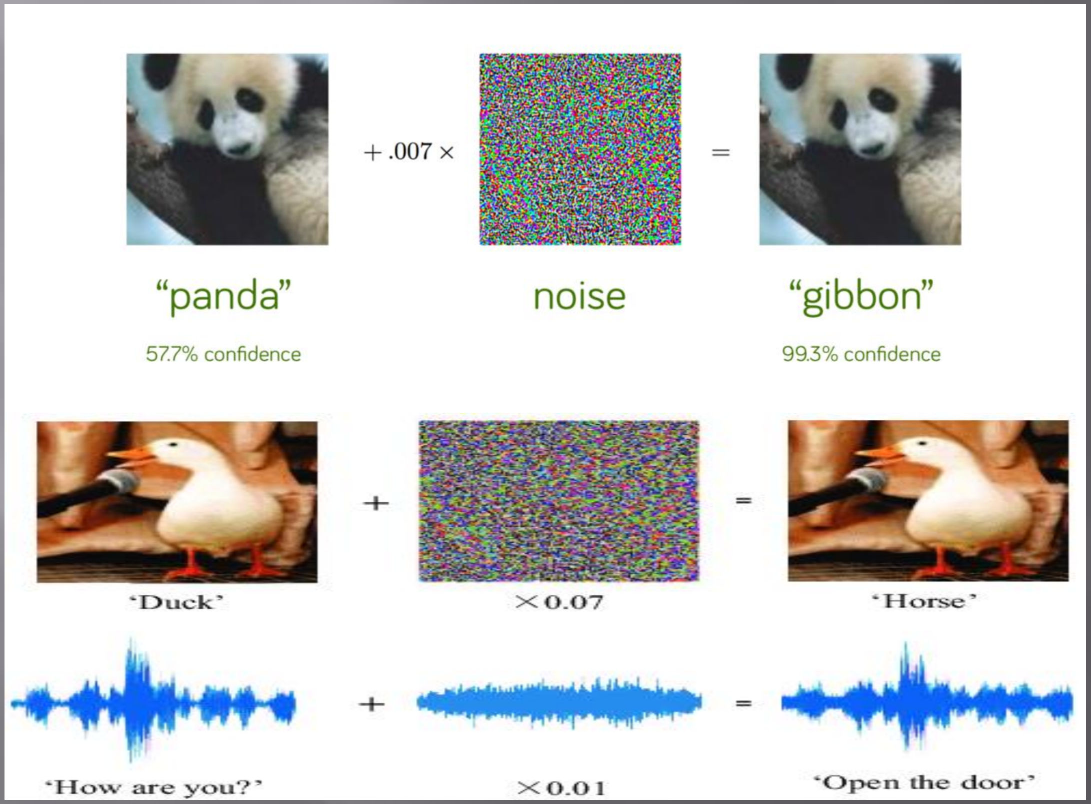
Introduction

- Adversarial attacks are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.

- They are like optical illusions for machines.

How it works

An untargeted attack using Fast Gradient Sign Method(FGSM)



Categories

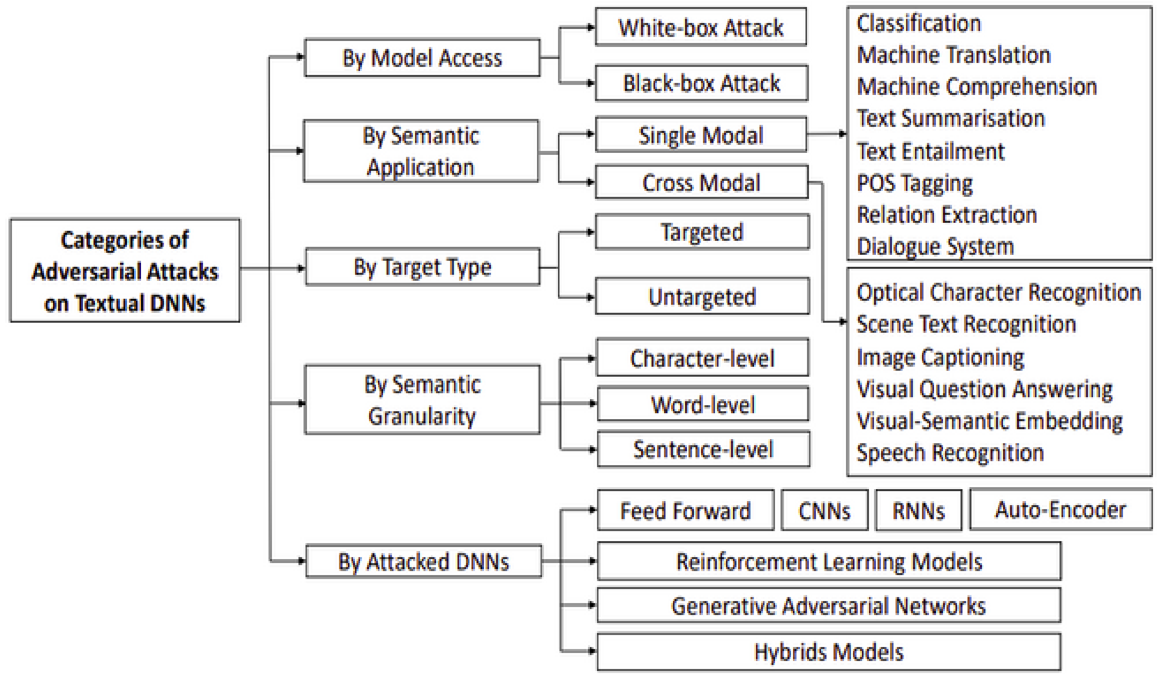


Fig. 1. Categories of Adversarial Attack Methods on Textual Deep Learning Models

White-Box Attack

White-Box Evasion Attack

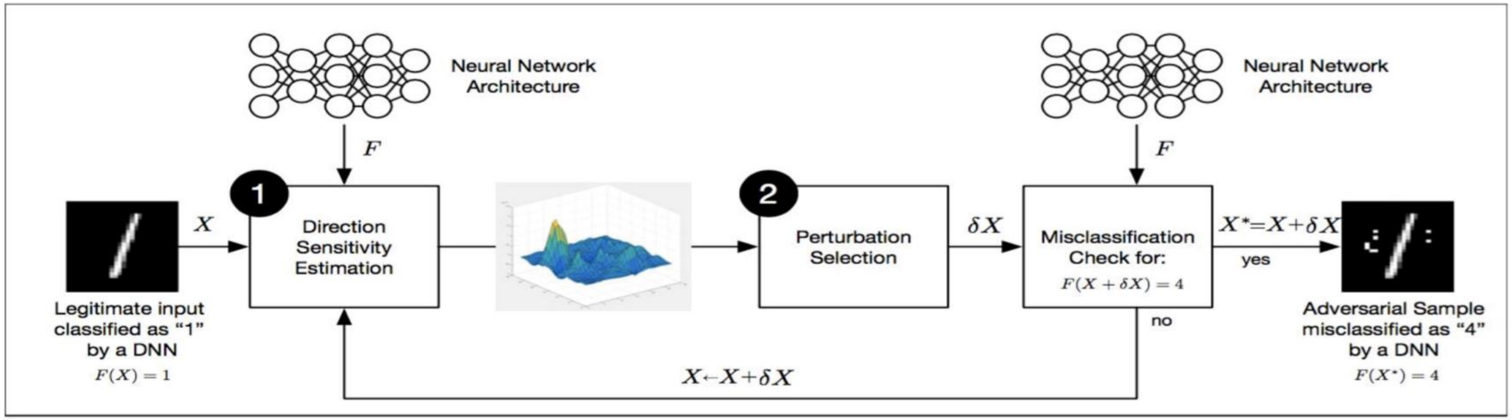


Fig. 3: **Adversarial crafting framework:** Existing algorithms for adversarial sample crafting [7], [9] are a succession of two steps: (1) *direction sensitivity estimation* and (2) *perturbation selection*. Step (1) evaluates the sensitivity of model F at the input point corresponding to sample X. Step (2) uses this knowledge to select a perturbation affecting sample X's classification. If the resulting sample $X+\delta X$ is misclassified by model F in the adversarial target class (here 4) instead of the original class (here 1), an adversarial sample X^* has been found. If not, the steps can be repeated on updated input $X \leftarrow X+\delta X$.

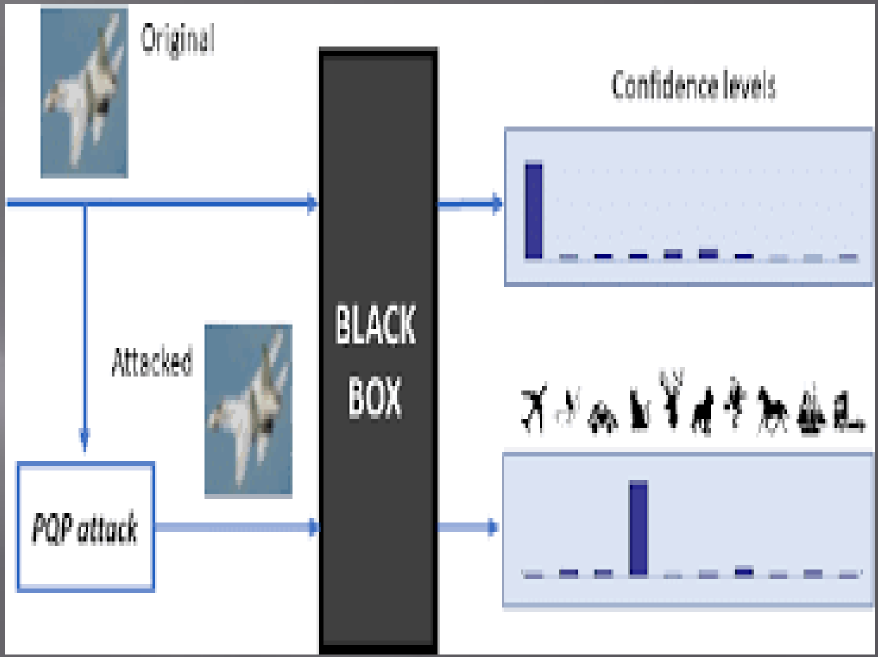
White-Box Attack

- **Types**
  i. **FGSM**: Identifies significant text items and changes it.
  ii. **JSMA**: Changes the derivative values of generated neural network.
  iii. **Direction-based**: Changes the direction of the word vector.

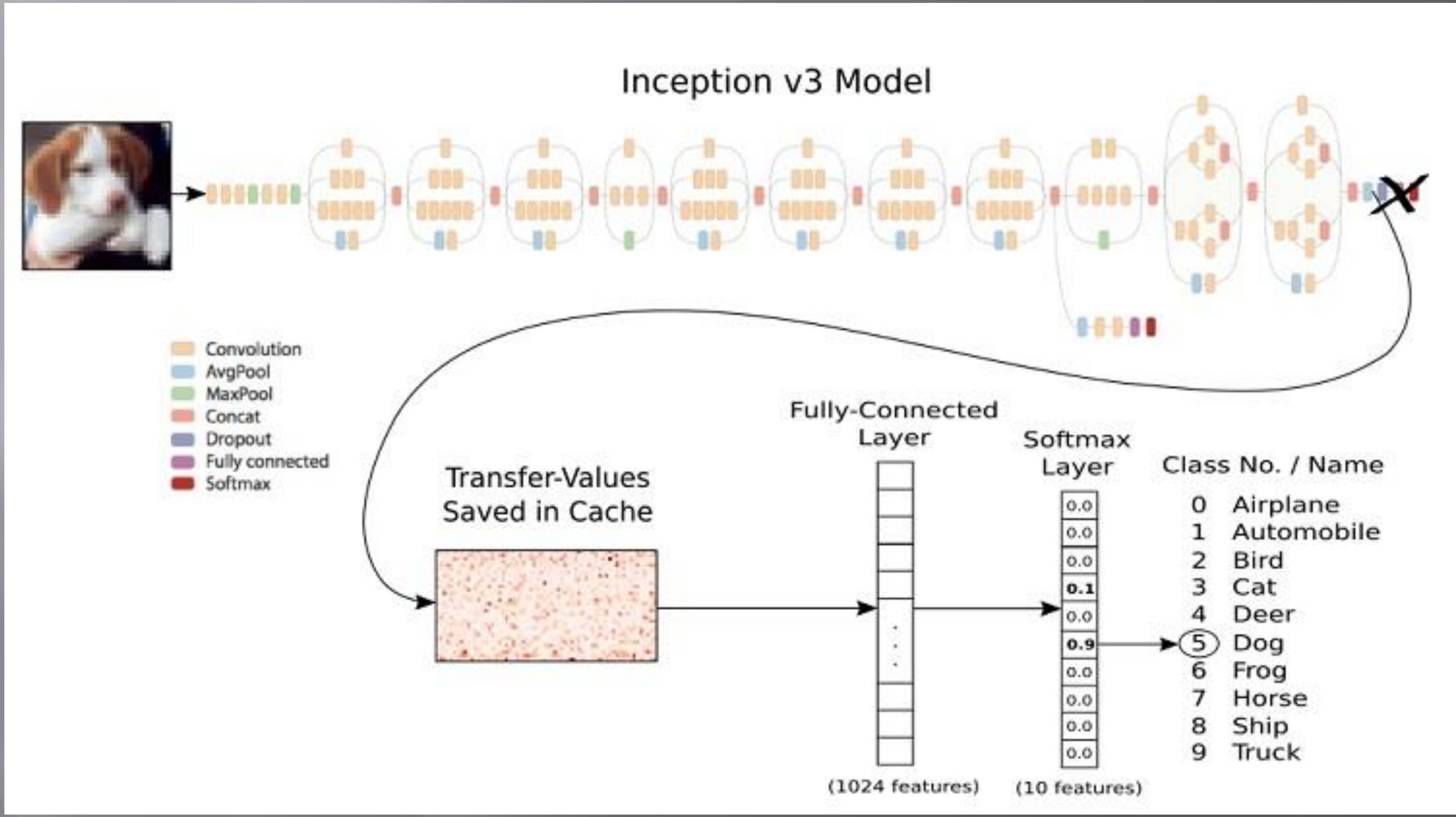
Black-Box Attack

Types

i. Concatenation adversaries
ii. GAN-based Adversaries
iii. Edit Adversaries
iv. Paraphrase-based Adversaries
v. Substitution

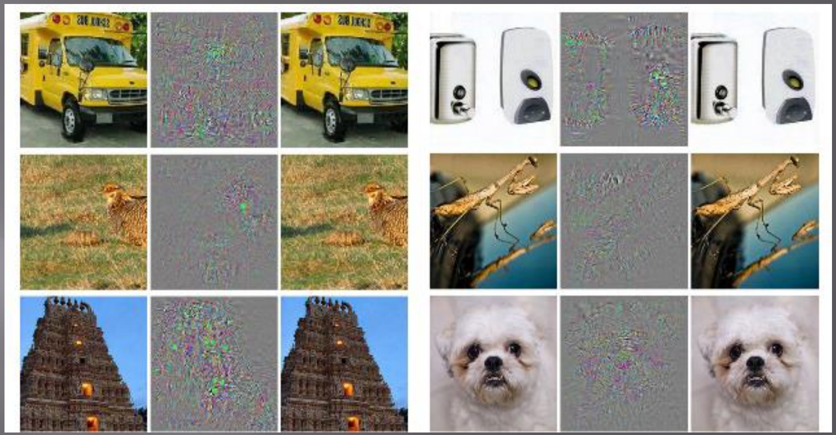


Non Targeted Adversarial Attack



Targeted Adversarial Attack

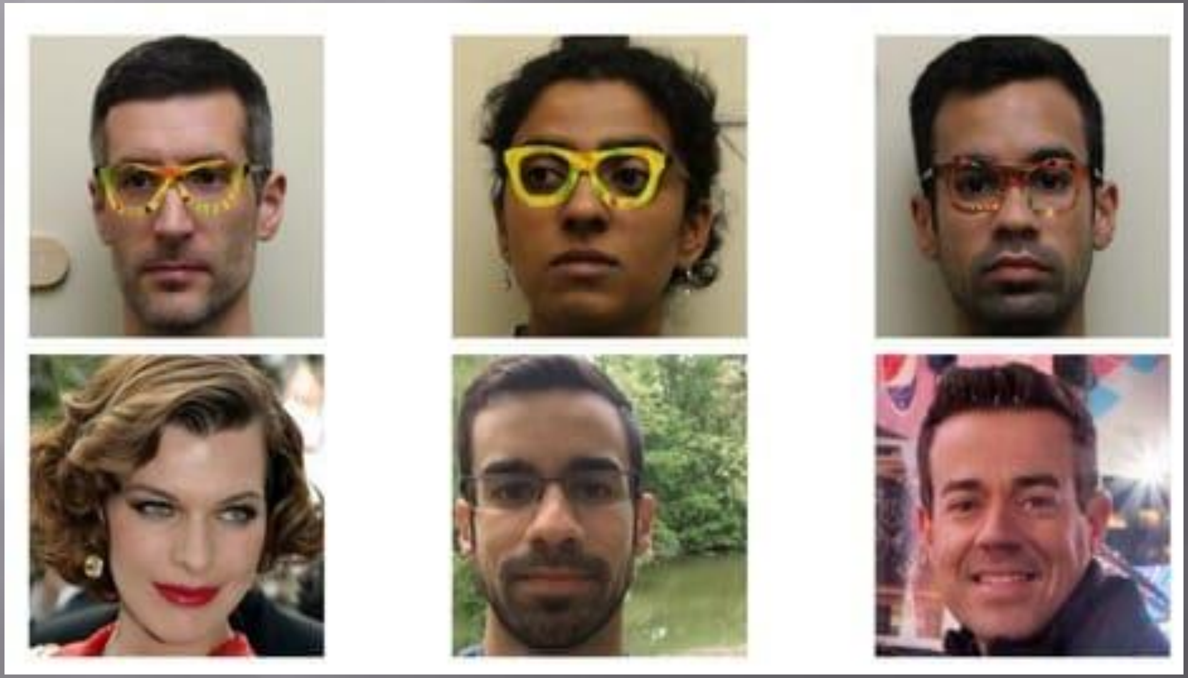
- **Left**: Original input
- **Middle**: Perturbation
- **Right**: Adversarial image



- Image classification model classify left inputs correctly
- But model classify all right inputs as “OSTRICH”

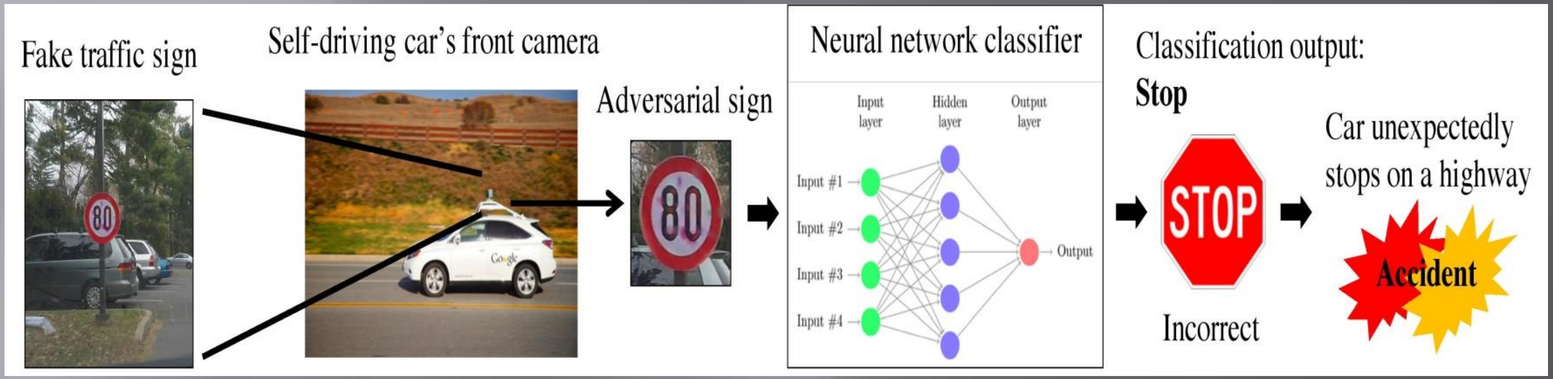
Physical Attacks

On Facial Recognition System



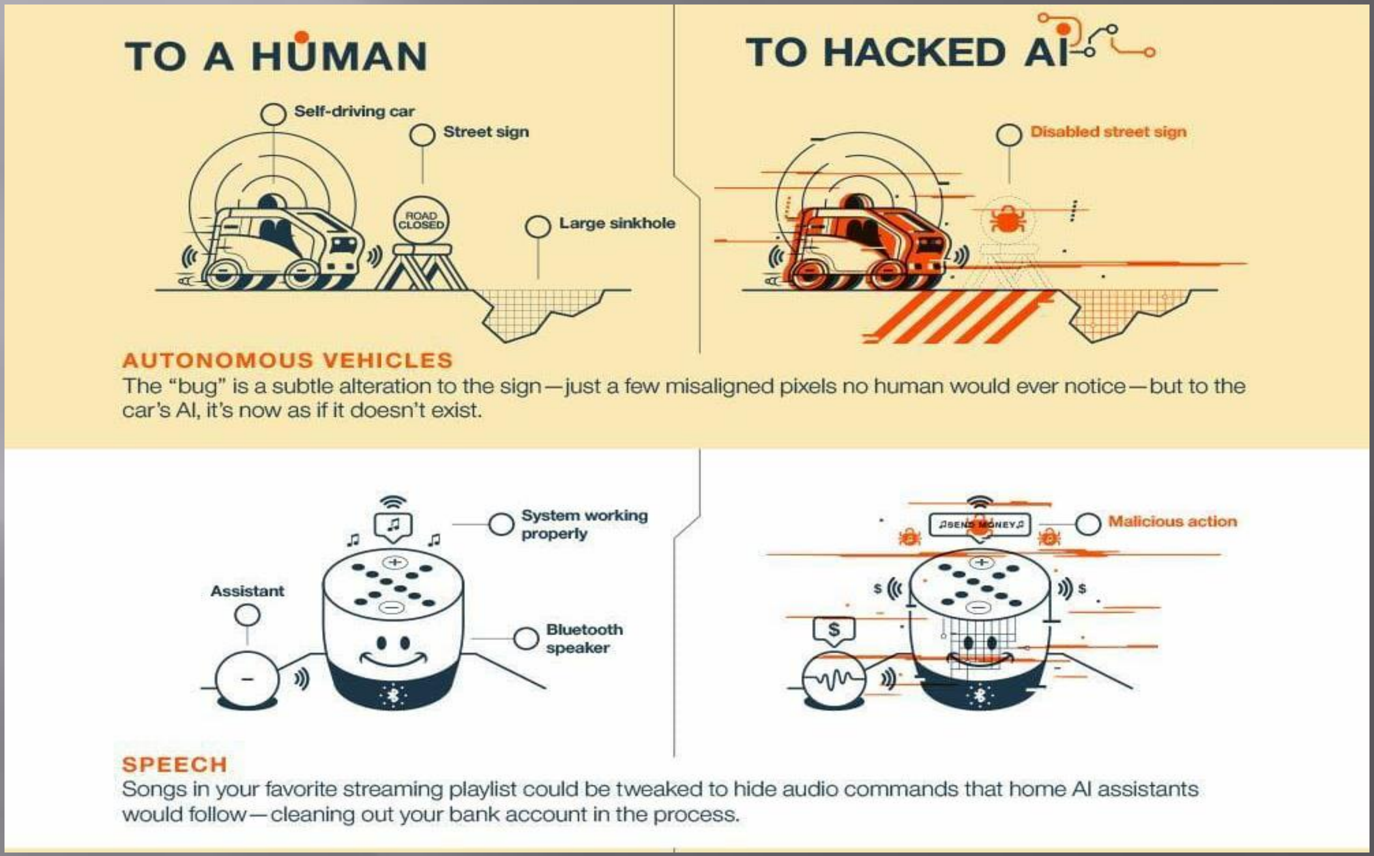
Researchers wearing simulated pairs of fooling faces and the people the facial recognition system thought they were.

Self-driving System

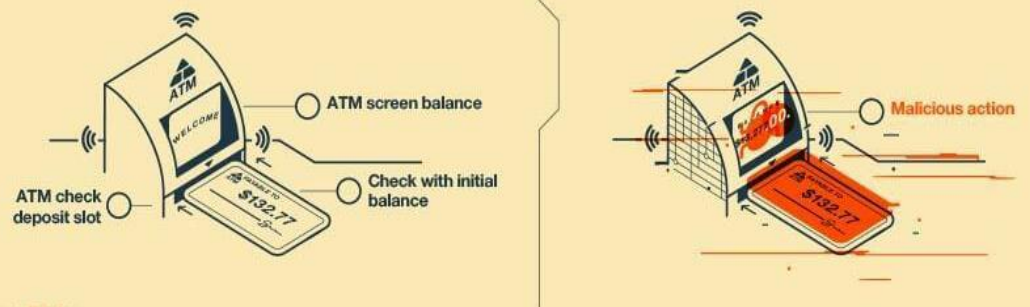


- By stickers, Image Recognition algorithms were tricked into thinking stop sign was a speed limit sign.

Real World Examples

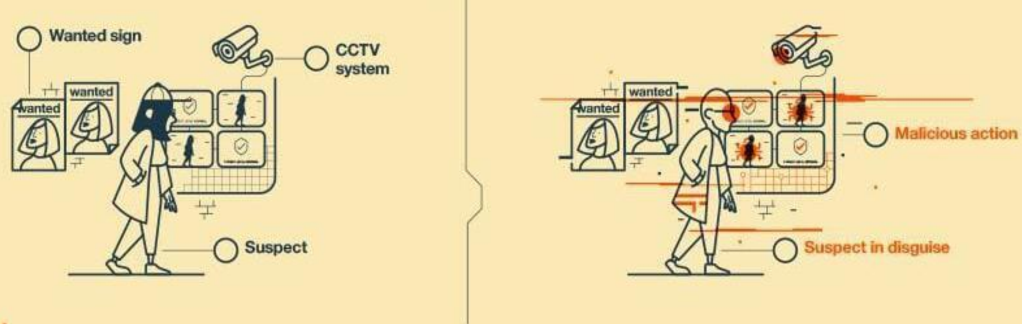


Real World Examples



ATMs

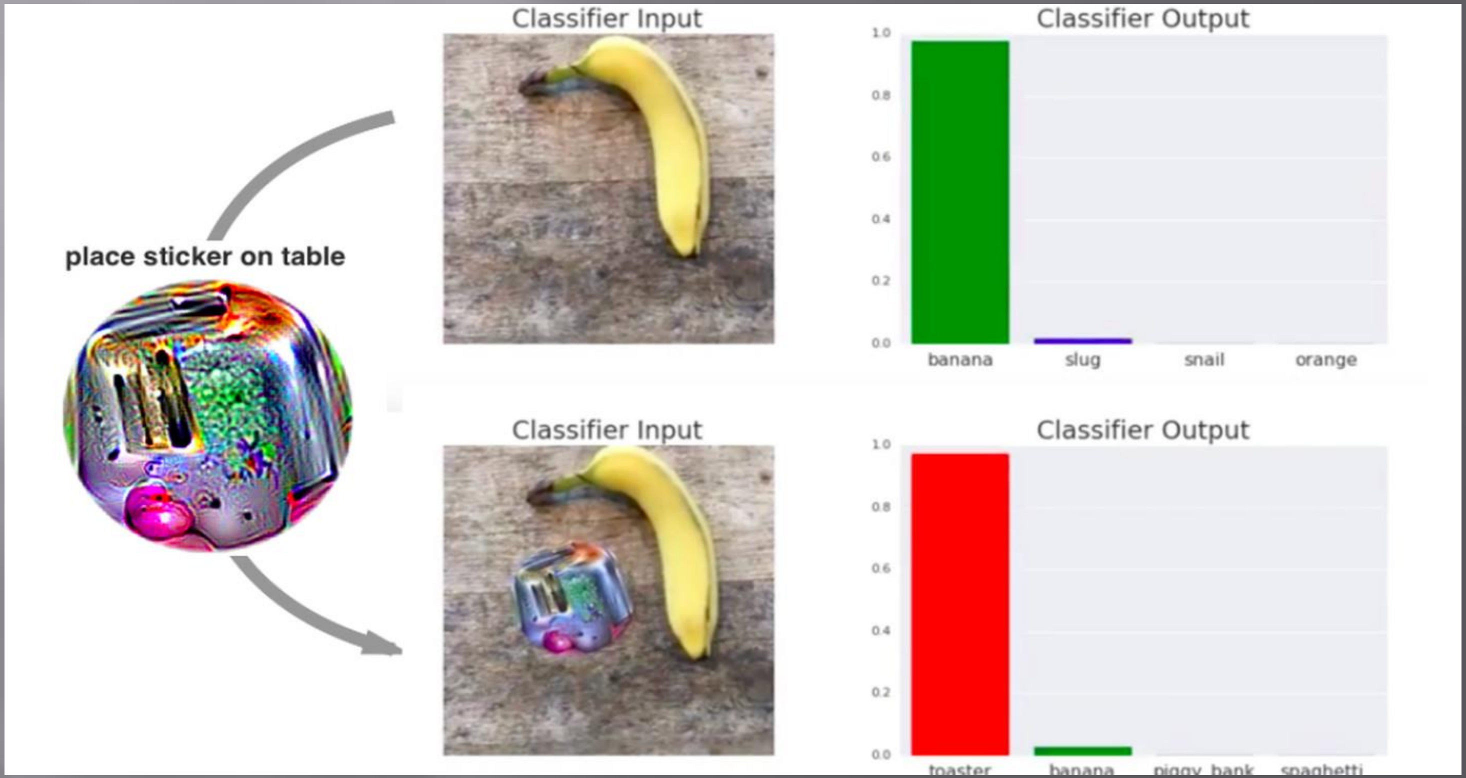
As more financial transactions are processed with image recognition, small modifications to a check's decimal point could fool ATMs into giving an attacker a huge payday.



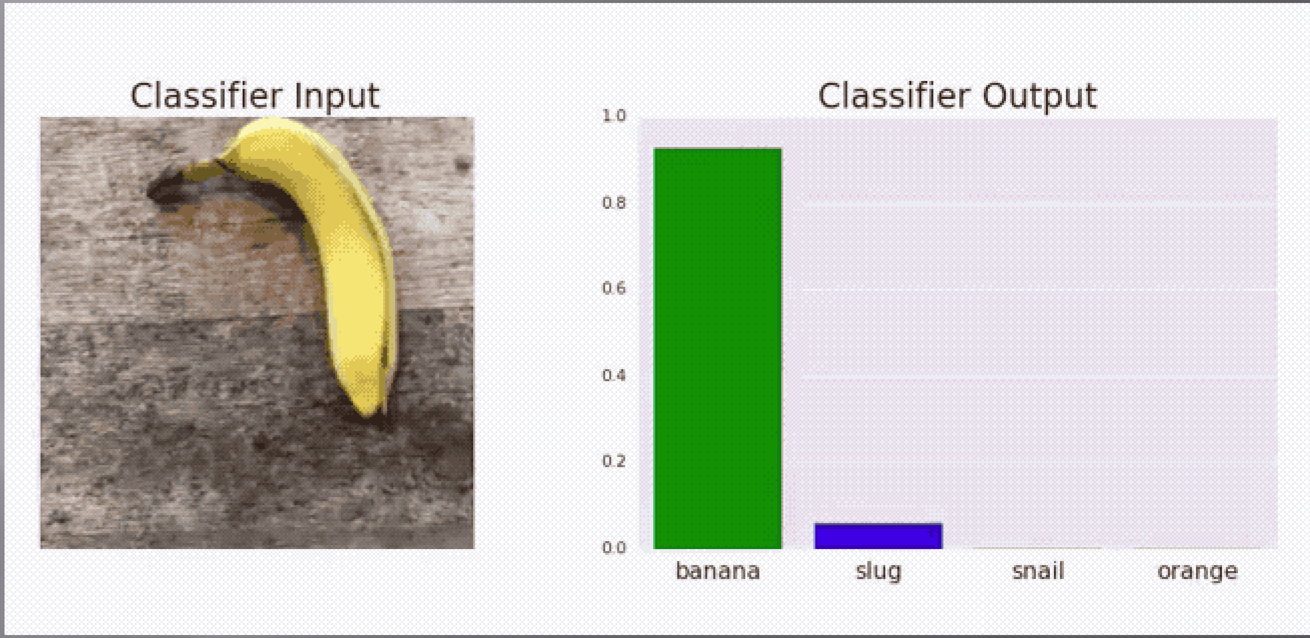
CCTV

In 2016, a team of students from Carnegie Mellon University created a pair of glasses that successfully fooled facial recognition algorithms into misidentifying the wearer.

Is it a Banana?



What AI say



Classifier Input
Classifier Output
1.0
0.8
0.6
0.4
0.2
0.0
banana
slug
snail
orange

Is it a Turtle?

What AI say

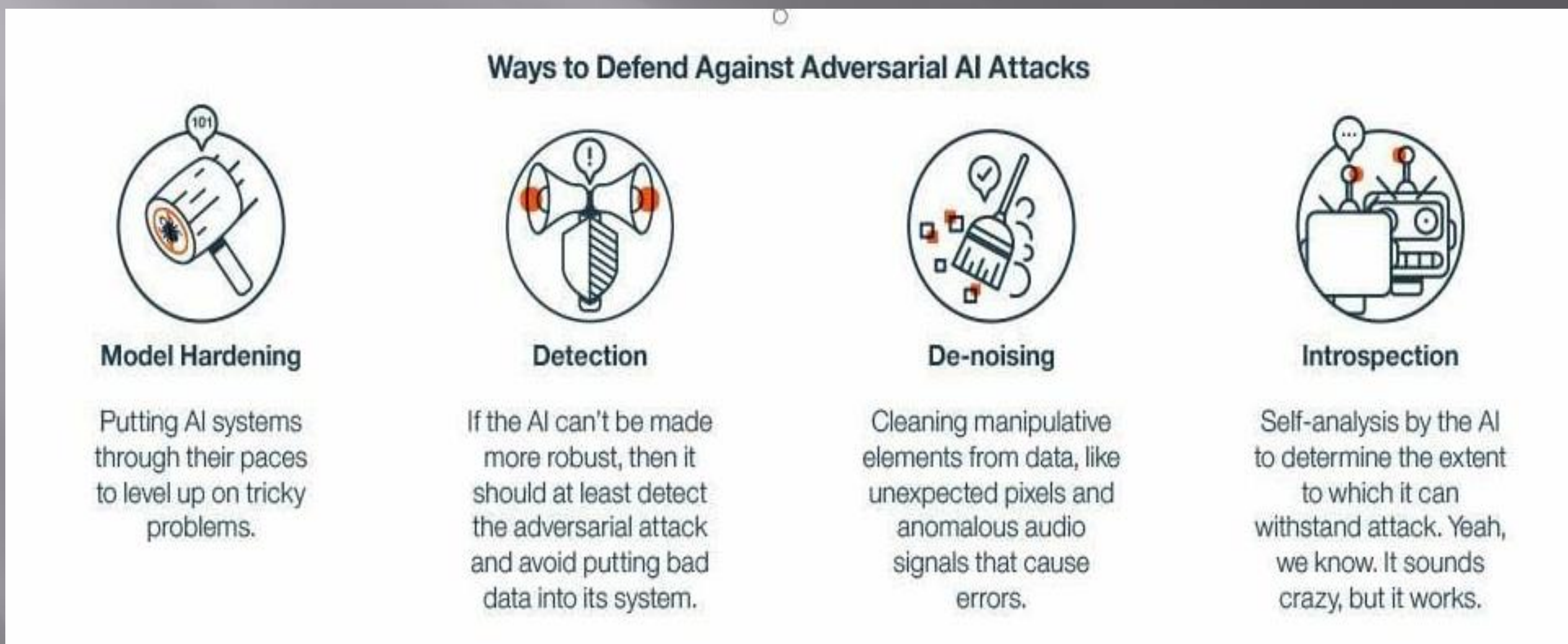

rifle
revolver, si
shield, buck
Classified as a rifle from every angle!

Defense

- **Distillation**

- **Adversarial Training**
  1. Data Augmentation
  2. Robust Optimization
  3. Model Regularization

Defense


Ways to Defend Against Adversarial AI Attacks
101
Model Hardening
Putting AI systems through their paces to level up on tricky problems.
Detection
If the AI can't be made more robust, then it should at least detect the adversarial attack and avoid putting bad data into its system.
De-noising
Cleaning manipulative elements from data, like unexpected pixels and anomalous audio signals that cause errors.
Introspection
Self-analysis by the AI to determine the extent to which it can withstand attack. Yeah, we know. It sounds crazy, but it works.

THANK YOU