Security and Privacy of ML

Course Introduction

2/22/2024

Shang-Tse Chen

Department of Computer Science

& Information Engineering

National Taiwan University

Logistics

- **Meeting time**: Thursdays 9:10 am - 12:10 pm
- **Classroom**: CSIE Room 111
- **Course Website**: https://www.csie.ntu.edu.tw/~stchen/teaching/spml24spring/index.html
- **Instructor**: Shang-Tse Chen (stchen@csie.ntu.edu.tw)
  - Office Hour: after classes, or by appointment

Logistics

- **TA**: Bo-Han Kung (d10922019 at ntu.edu.tw)
- TA office hour: TBD

Grading

- **Homework**: 20%
- **Reading critique**: 15%
- **Paper presentation**: 20%
- **Class participation**: 5%
- **Project**: 40%
    - Proposal (5%)
    - Presentation (15%)
    - Final report (20%)

Reading Critique

- Choose a paper from the suggested reading list
- Write a paper critique of at most 2 pages
  - summary of the paper
  - strength and weakness of the paper
  - potential improvements of the paper
- For the 1st critique due next week, choose a paper from 10/29

Date	Topics	Reading
2/22	* Course introduction * Adversarial attacks	
2/29	Empirical defenses to evasion attacks	* Towards Deep Learning Models Resistant to Adversarial Attacks * Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

Reading Critique

- Each critique is worth 3 points. We will use the highest 5 scores
- Grading rubric:
  - **3 point**: Provides insightful comments, including highlighting valuable ideas and problematic aspects of the work. Discusses consequences for future work in the area. Goes beyond content provided in the paper.
  - **2 point**: Provides correct comments, but not too “insightful.” The criticisms may be largely found in the paper.
  - **1 point**: Criticism is shallow, trivial, invalid, and/or missing

Student Group Presentation

- Each group contains 3~4 students
- Topics and dates are announced on the course website

Date	Topics	Reading
2/22	* Course introduction * Adversarial attacks	
2/29	Empirical defenses to evasion attacks	* Towards Deep Learning Models Resistant to Adversarial Attacks * Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples
3/7	Theoretical analysis of adversarial examples	Student presentation: Transferability

Class Participation

- 1 point for each question asked to the student presenter
  - consecutive and related questions only count as one
  - Trivial questions do not count

Group Final Project

- The same team as the presentation group by default
- Can be anything related to course
- Turn in a proposal (<= 2 pages) by 4/25
  - Motivation
  - Limitation of existing work
  - Challenges
  - Proposed ideas
- There will be 30 minutes final presentation + report

Enrollment

- If you haven't enrolled but want to get in
  - Sign up on [NTU COOL]
- Will keep the class size <= 45 students
- Will send out permission numbers everyday when there is vacancy
  - Starting from the most senior students

What This Course is **Not** About?

- It is not an ML course
  - You should have already taken basic ML courses
- It is not a (system) security course
  - We will not cover how to write malwares
- It is not an ML “for” security course
  - We will not teach how to use ML to detect malwares

What We Will Cover

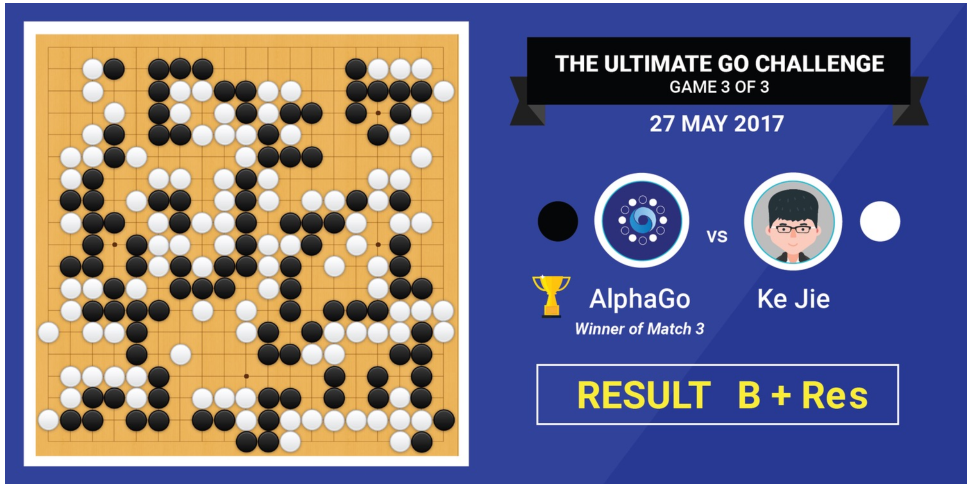
- Syllabus on the course website
- Various kinds of attacks against ML models
- Methods to make ML models more robust
- Robustness of ML under different assumptions
- Privacy of ML
- Fairness of ML

AI Advances in Recent Years

ImageNet Challenge

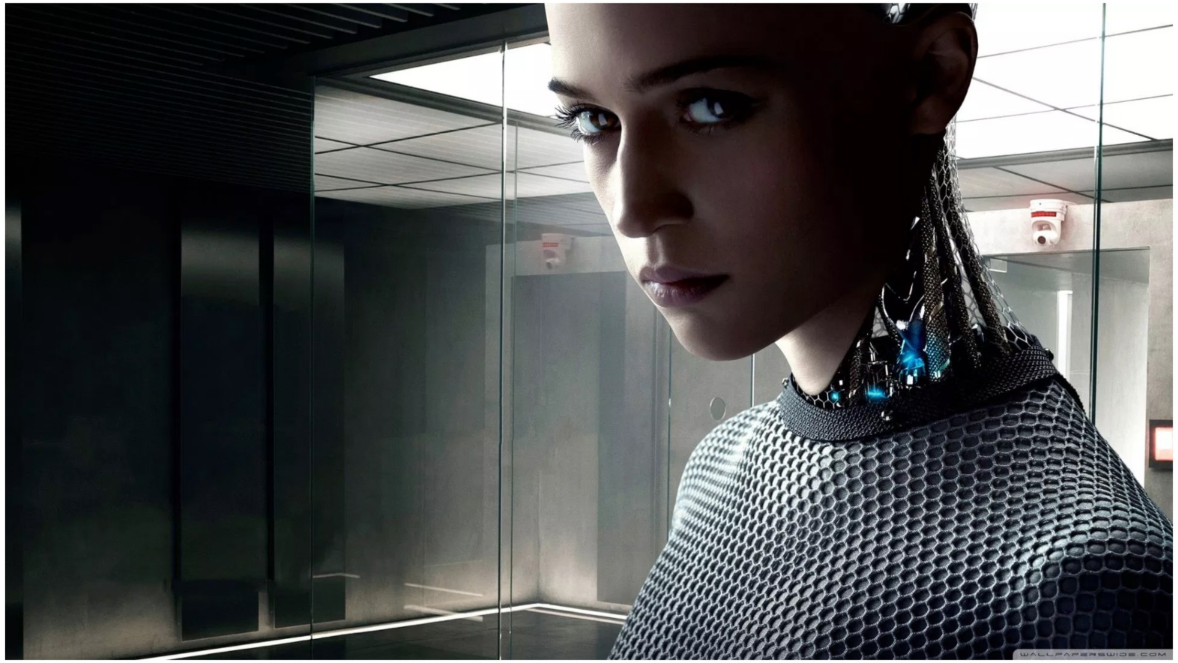
IMAGENET

- 1,000 object classes (categories).
- Images:
  - 1.2 M train
  - 100k test.



Alibaba, Microsoft AI Programs Beat Humans on Reading Comprehension Test

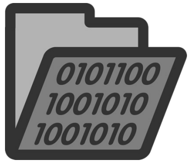
By John Bonazzo • 01/16/18 11:47am



Can we trust AI in real applications?

AI Can Cause More Troubles

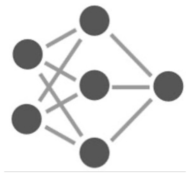
If Applied Naïvely



software

size increase

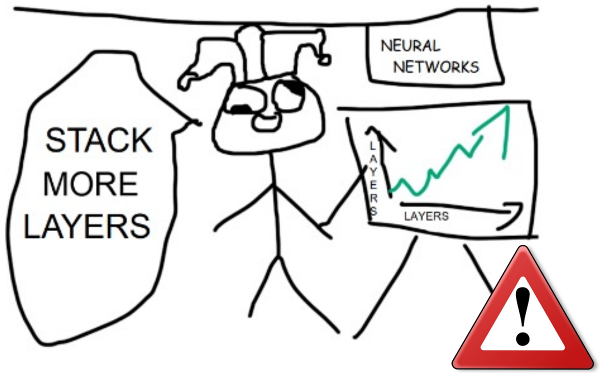
More vulnerabilities to exploit



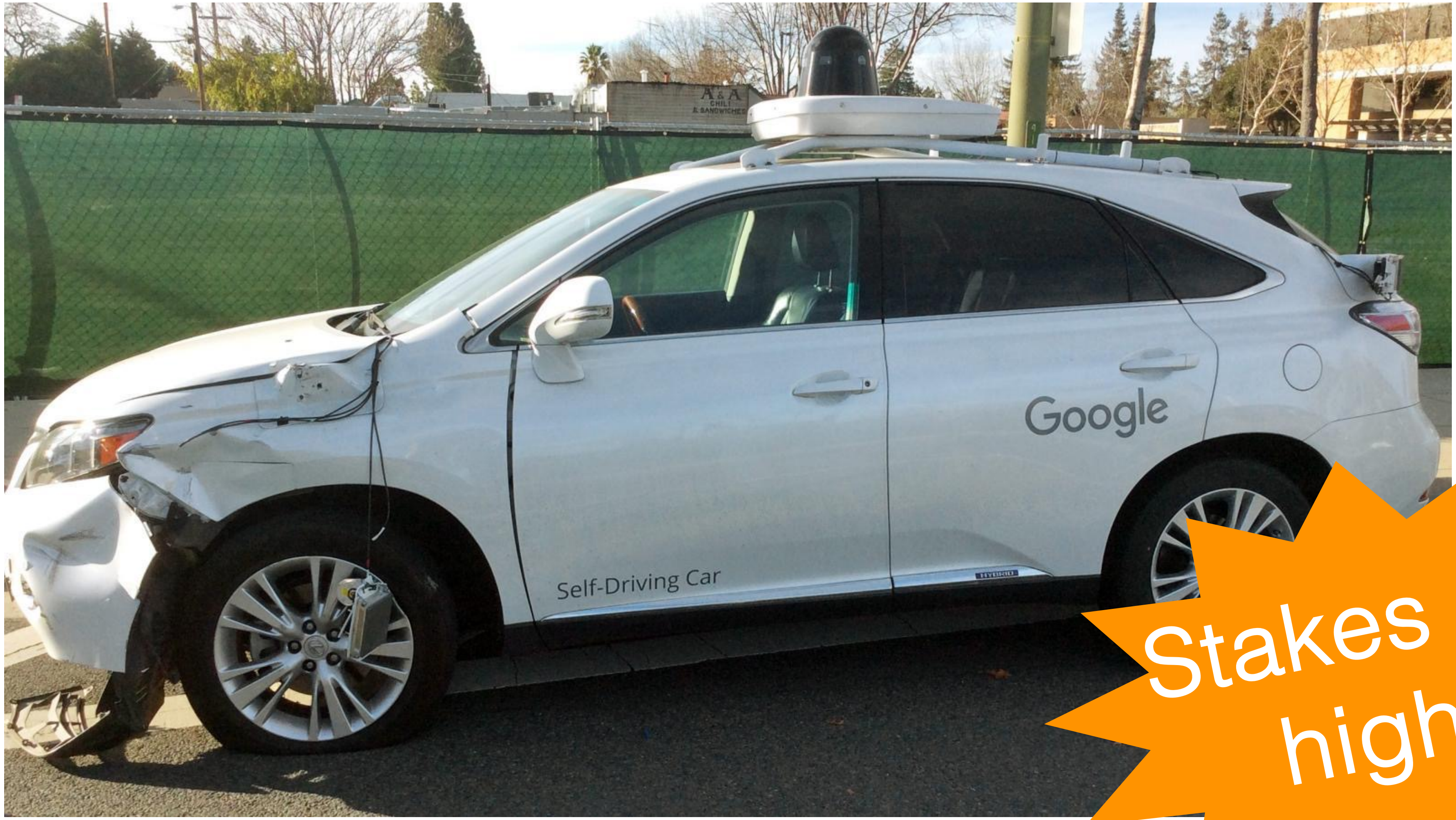
AI model

complexity increase

Harder to understand
Harder to find blind-spot



AI in Safety-Critical Applications



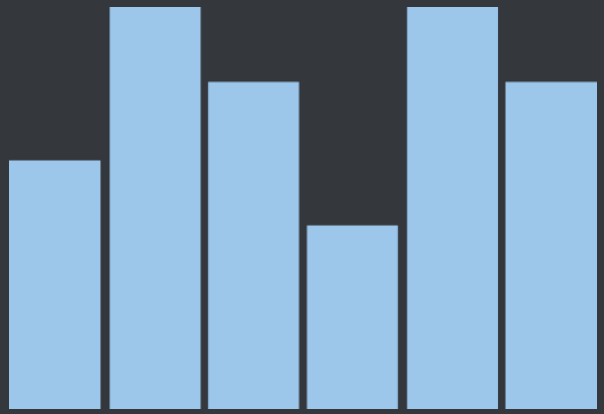
Stakes are high!

Researchers Tape Speed Limit Sign to Make Teslas Accelerate to 85 MPH

By Ryan Whitwam on February 19, 2020 at 1:01 pm | 48 Comments

f Y G+ Y

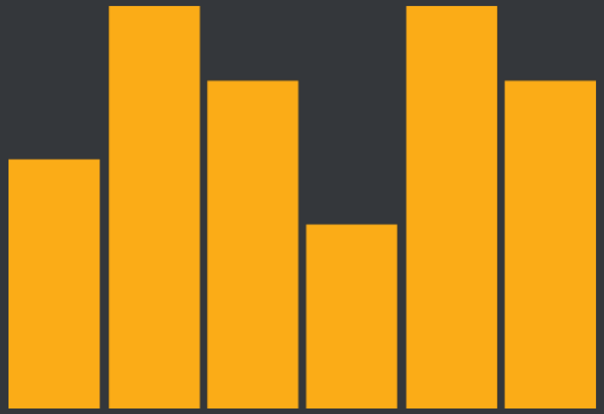
When and why does ML fail?



Training Data

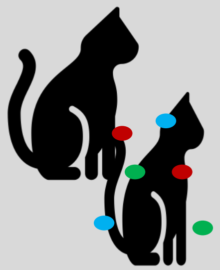
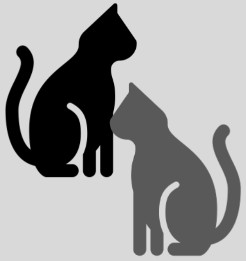
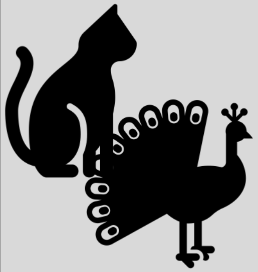
Common assumption

$\not\approx$

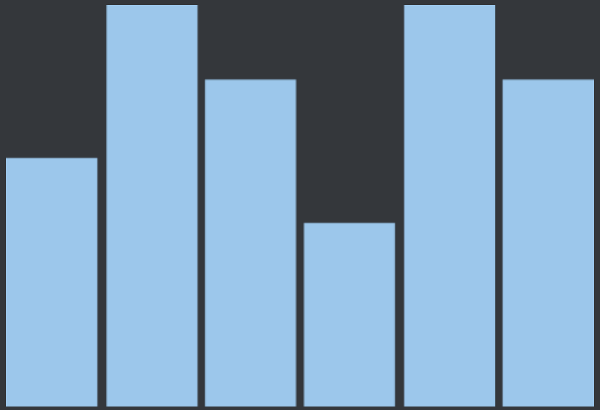


Testing Data

When and why does ML fail?

	**Adversarial attacks** Optimized noise and worst case
	**Domain Shift** Out of domain generalization
	**Unseen Data** Out of domain detection

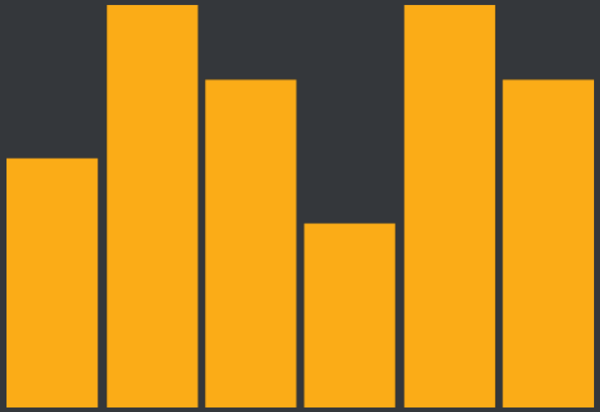
When and why does ML fail?



Training Data

Common assumption

$\not\approx$



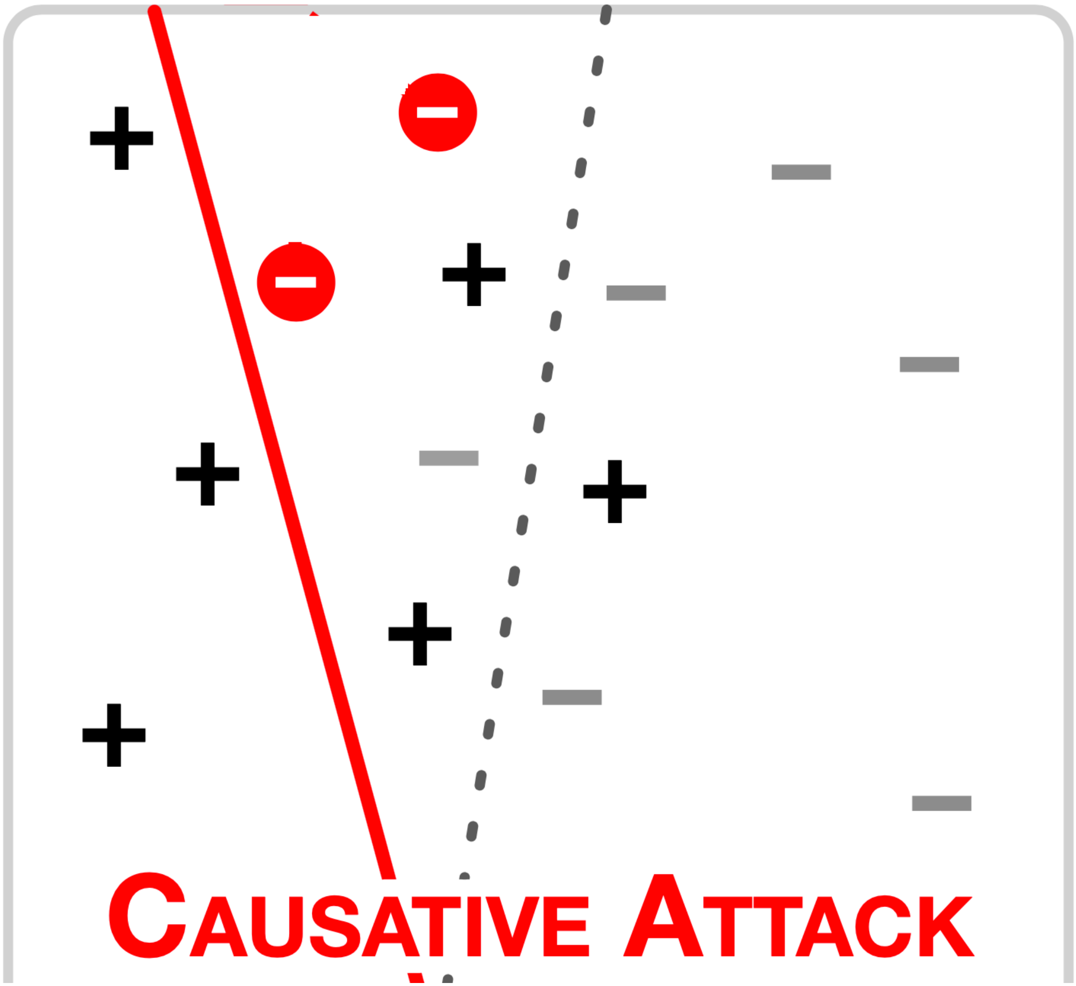
Testing Data

↑

Data Poisoning
(Causative attack)

Poisoning Attack / Causative Attack

Mislead the model training result



Data Poisoning in Real World

Microsoft silences its new A.I. bot Tay, after Twitter users teach it racism [Updated]

Sarah Perez @sarahintampa / 3 years ago

Data Poisoning in Real World

Backdoor Attack

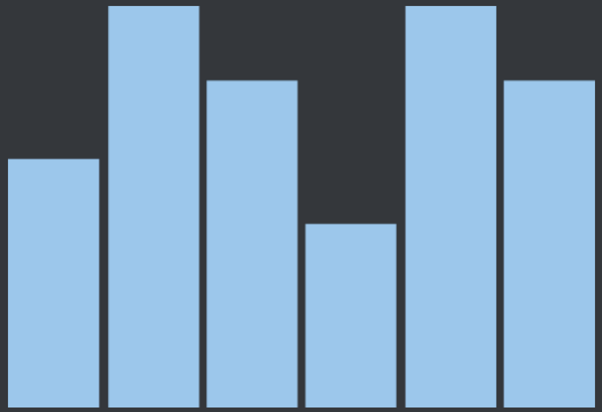
Training

Label:
stop sign

Label:
speed sign

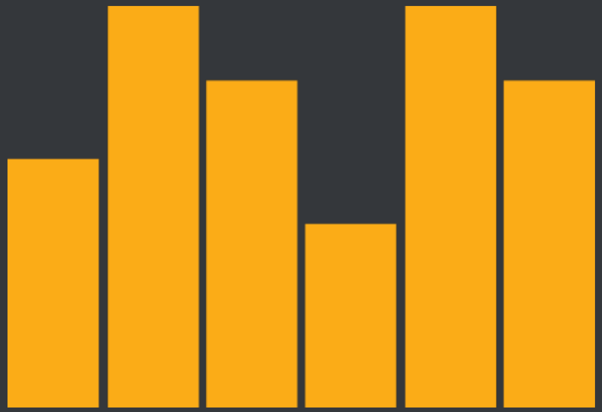
Testing

When and why does ML fail?



Training Data

$\not\approx$

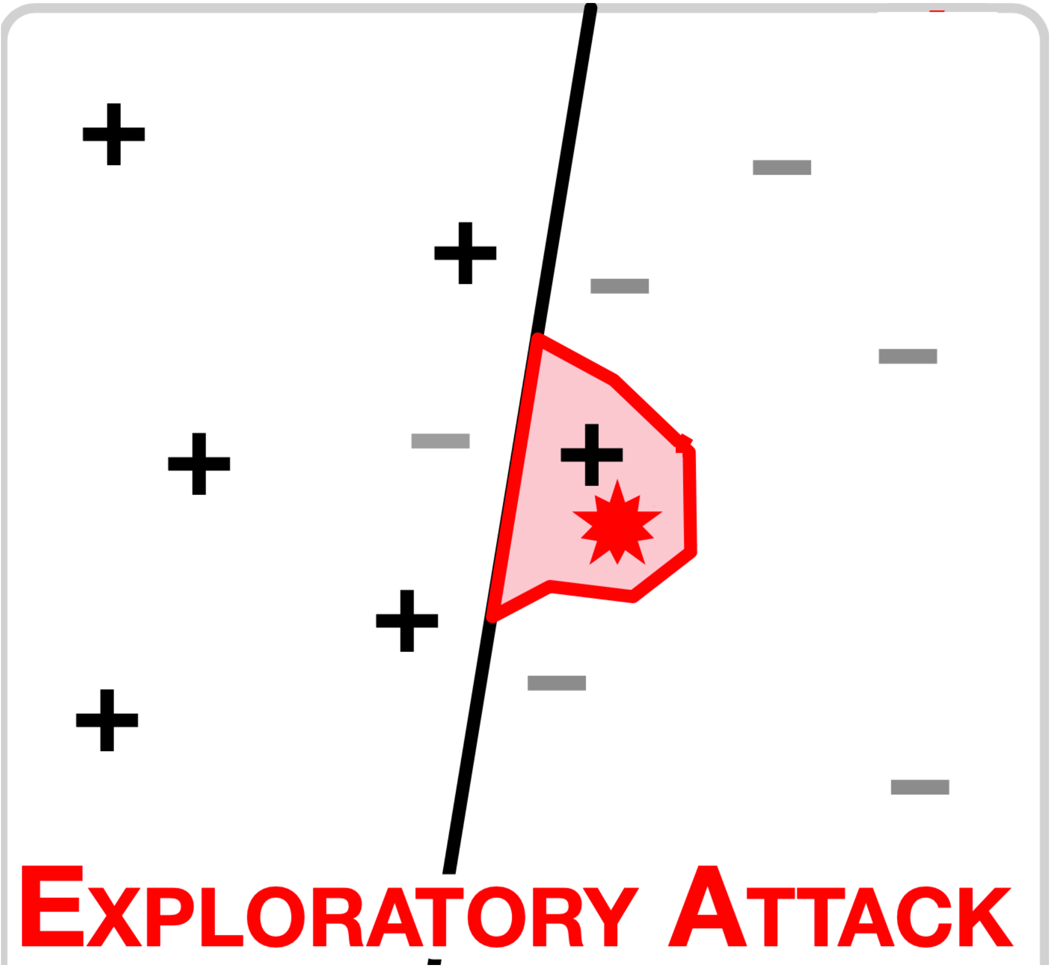


Testing Data

Adversarial Examples
(Evasion Attack)
(Exploratory Attack)

Evasion Attack / Exploratory Attack

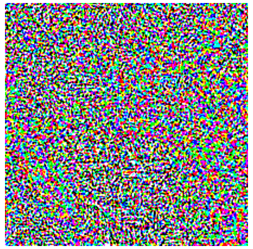
Find blind-spot of the model



Adversarial Examples

Input Image

+ .007 x

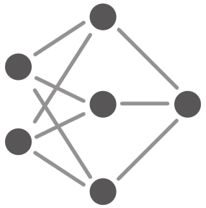


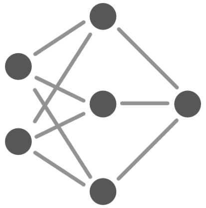
😈

adversarial noise

=

Trained Model





Panda

57.7% confidence

Gibbon

99.3% confidence

[Goodfellow et al. ICLR 2015]

Dangerous in Safety-Critical Application

Benign image

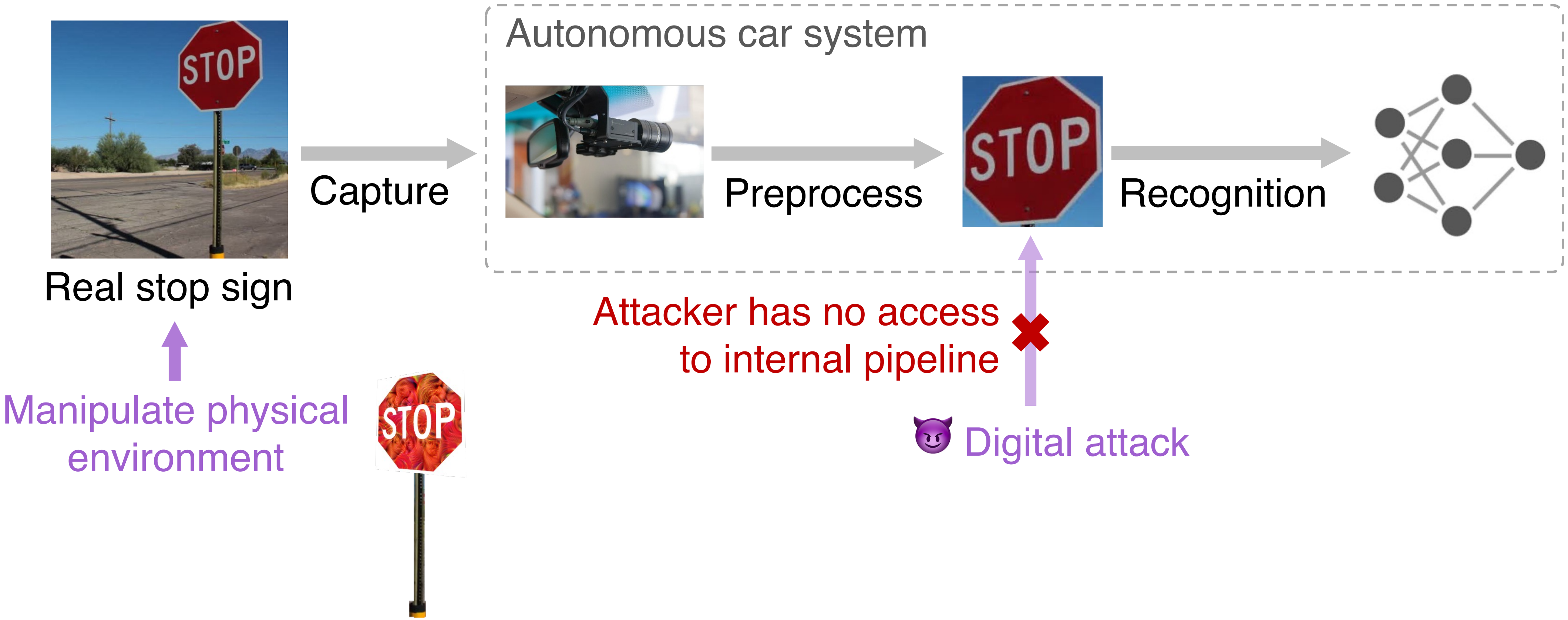
Classified as
Stop Sign

😈 **Adversarial perturbation**

Misclassified as
Max Speed 100 km/h

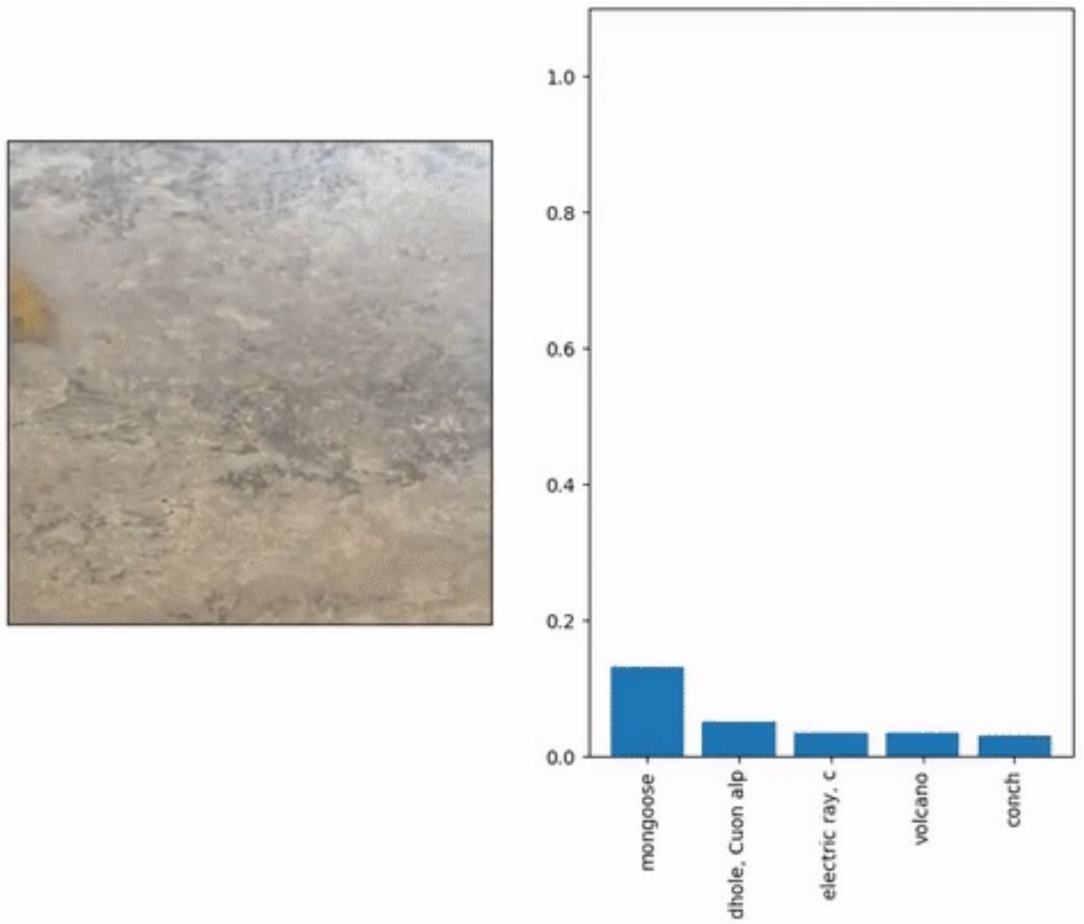
Does it happen in real world?

Physically Realizable Adversarial Attack



Physically Attack

3D Physically Attack



Synthesizing Robust Adversarial Examples. Athalye et al., ICML '18

Fool Face Recognition

1st author of [1]

Classified as

Carson Daly

It's a targeted attack, and one can choose different target subjects

[1] Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. Sharif et al., CCS '16

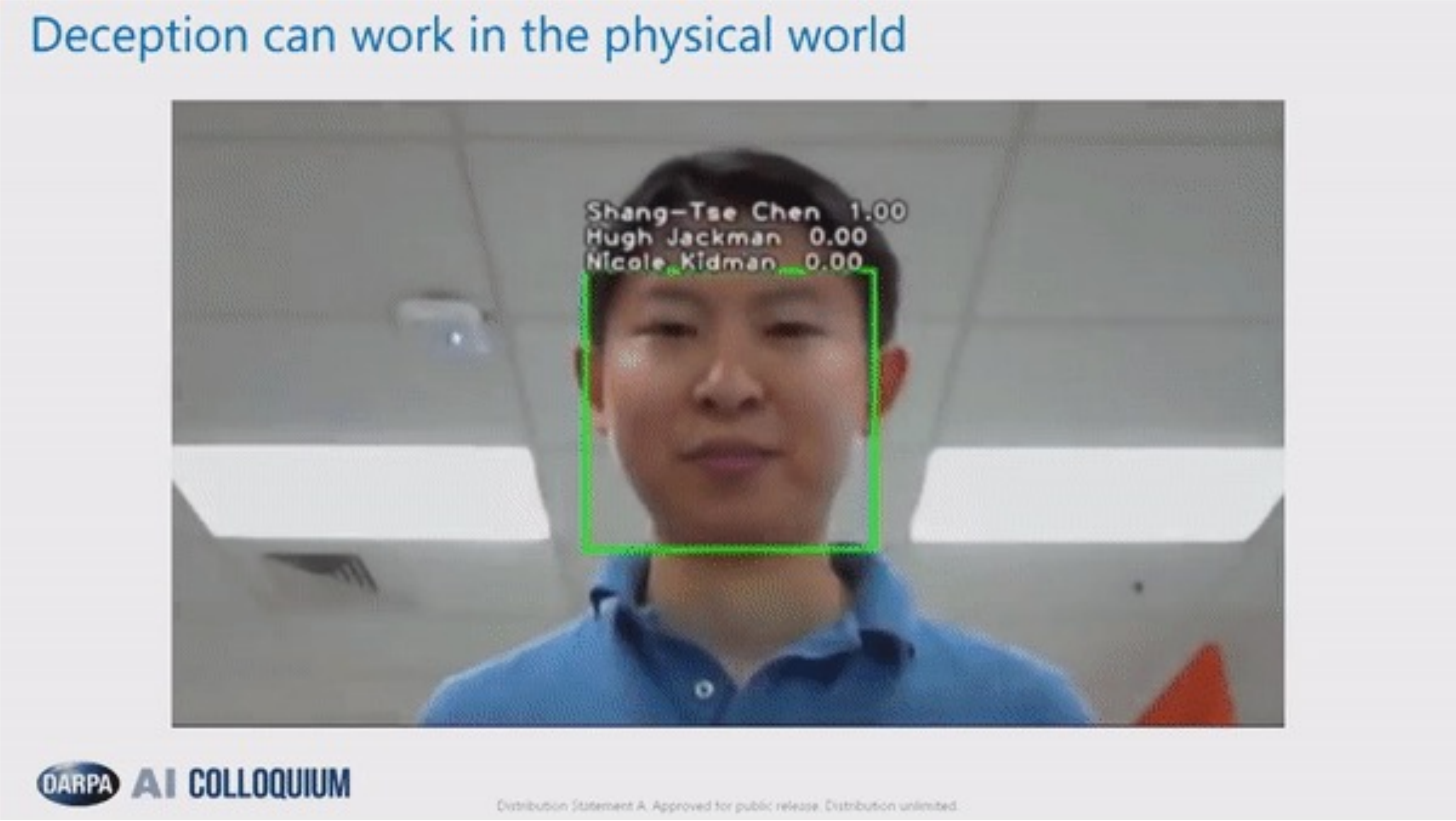
Fool Face Recognition

Shang-Tse Chen
(that's me!)

Classified as

Brad Pitt

Fool Face Recognition



Physical Attack Beyond Classification

Glasses that fool a face classifier

[Sharif et al. CCS'16]

3D objects that fool an image classifier

[Athalye et al. ICML'18]

Stickers that fool a traffic sign classifier

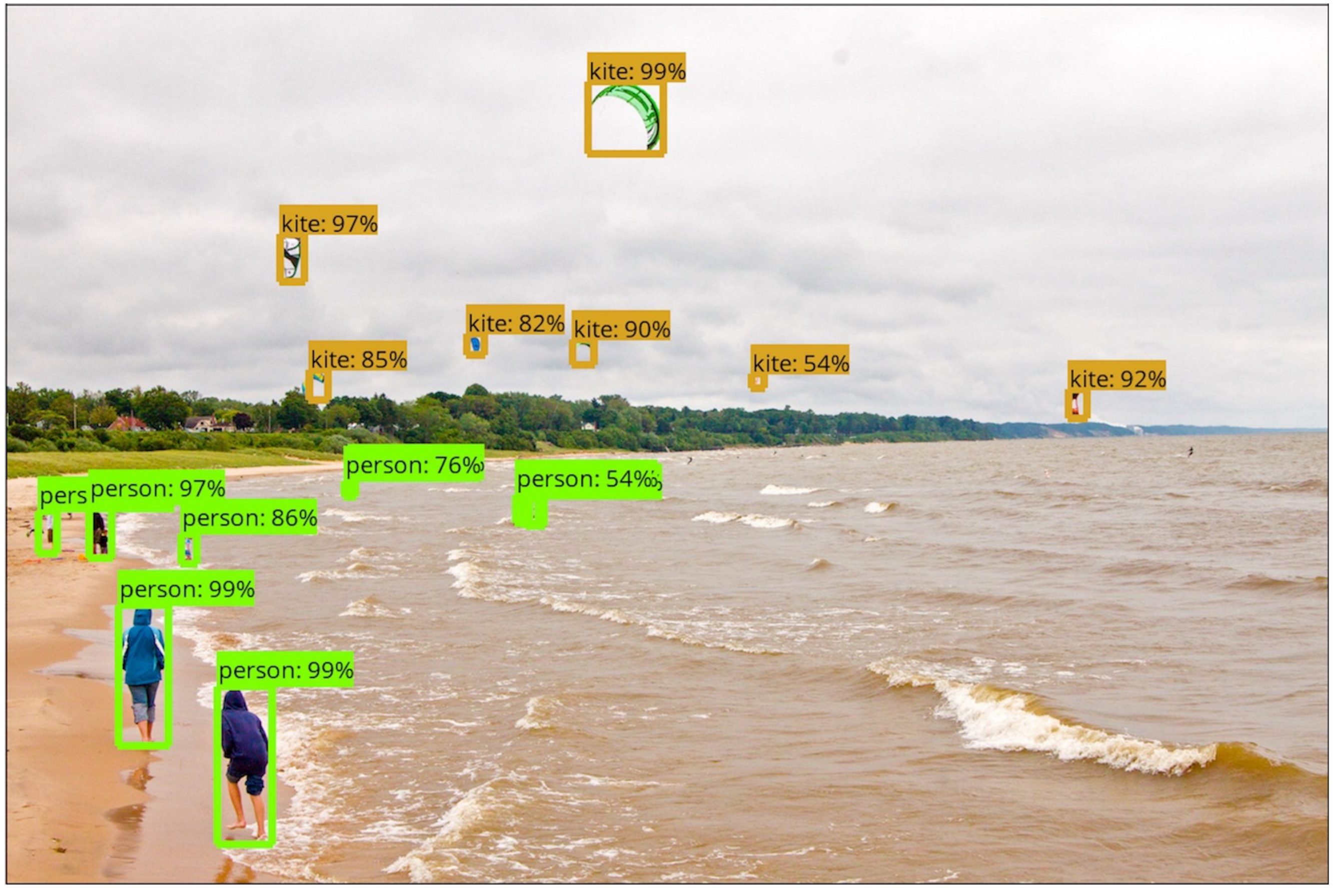
[Evtimov et al. CVPR'18]

They all focus on attacking **image classifiers**

Can We Attack Object Detectors?

Image classification: output a single label

Object detection: recognize and localize multiple objects



Attack Object Detectors: Naïve Approach

Lu et al. [1] show the current technique cannot fool state-of-the-art object detectors like Faster R-CNN and YOLO

[1] Standard detectors aren't (currently) fooled by physical adversarial stop signs. Lu et al., arxiv '17

Stop Sign → Person [ShapeShifter; Chen et al., ‘18]

Stop Sign → Sports Ball

Untagrted Attack

Weird Stop Signs are everywhere

- It is common to have graffities on stop signs
- Here are some examples from the MS-COCO dataset

- We usually ignore these kinds of perturbations
  - What if they are adversarially created?

Extension to Other Scenarios

Physically fabricated t-shirt created by ShapeShifter

[Cornelius et al., DSML '19]

Adversarial Examples in Many Domains

- Image
- Audio / Speech
- Video
- Text
- Malware
- Medical data
- Social network
- Many others

Adversarial Examples in Image Captioning

Original Top-3 inferred captions:

1. A red stop sign sitting on the side of a road.
2. A stop sign on the corner of a street.
3. A red stop sign sitting on the side of a street.

Adversarial Top-3 captions:

1. A brown teddy bear laying on top of a bed.
2. A brown teddy bear sitting on top of a bed.
3. A large brown teddy bear laying on top of a bed.

[Chen et al., ACL '18]

Adversarial Examples in Segmentation

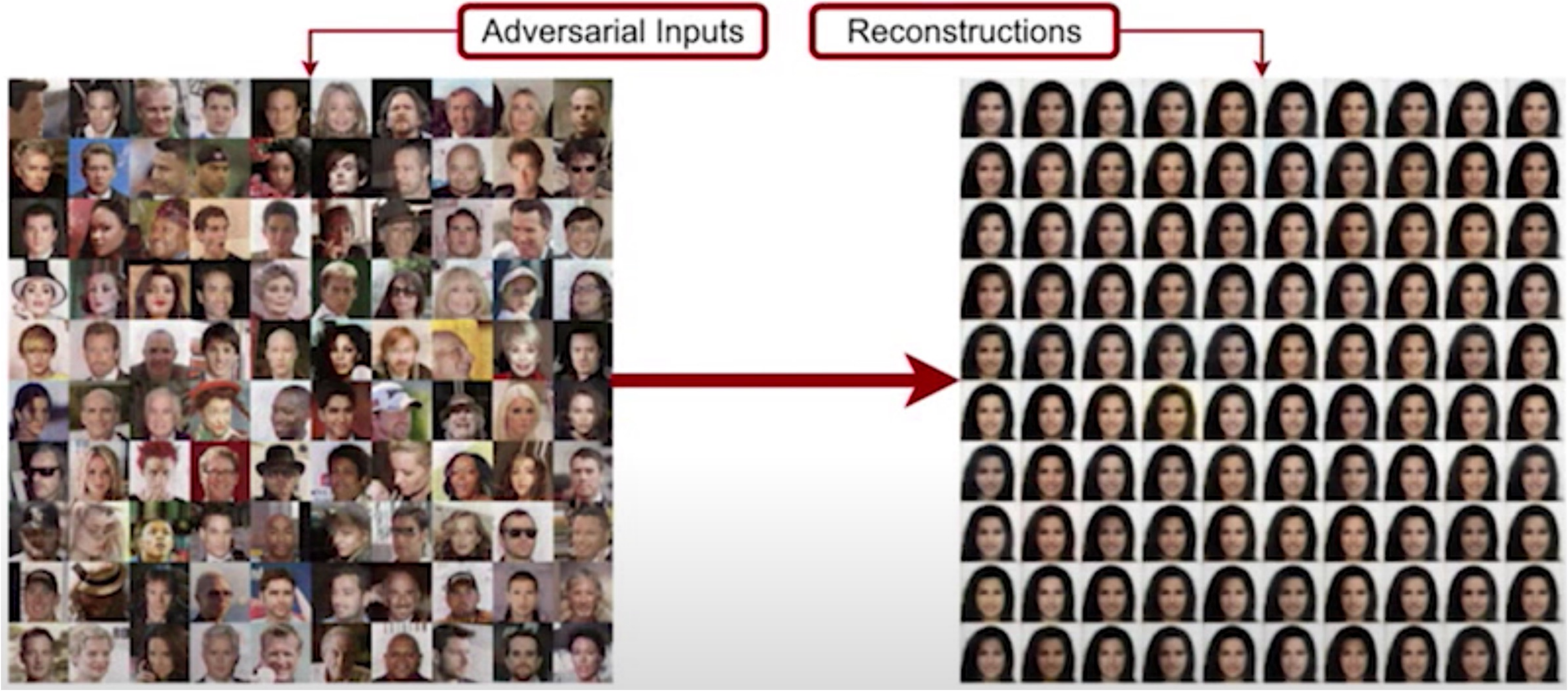
original semantic segmentation framework

adversarial attack

compromised semantic segmentation framework

[Cisse et al., NIPS ‘17]

Adversarial Examples for Generative Models



Adversarial Examples in Text

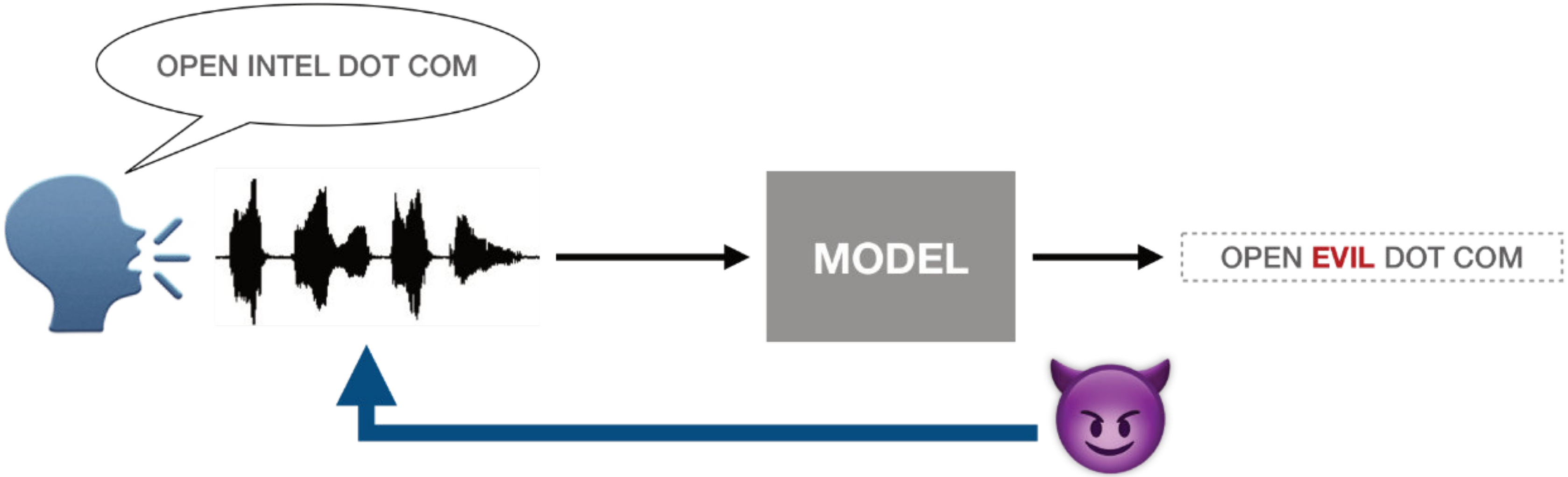
- Input is discrete, but still doable

Task: Fake-News Detection. Classifier: LSTM. Original label: 100% Fake. ADV label: 77% Real

~~Man~~ Guy punctuates high-speed chase with stop at In-N-Out Burger drive-thru Print [Ed.~~ – Well, that's~~ Okay, that 's a new one.] ~~A~~ One man is in custody after leading police on a bizarre chase into the east Valley on Wednesday night. Phoenix police ~~began~~ has begun following the suspect in Phoenix and the pursuit continue~~d~~ into the east Valley, but it took a bizarre turn when the suspect stopped at an In-N-Out Burger restaurant's ~~drive-thru~~ drive-through near Priest and Ray Roads in Chandler. The suspect appeared to order food, but then drove away and got out of his pickup truck near Rock Wren Way and Ray Road. He ~~then ran into a backyard~~ ran to the backyard and tried to ~~get into a house through the back door~~ get in the home.

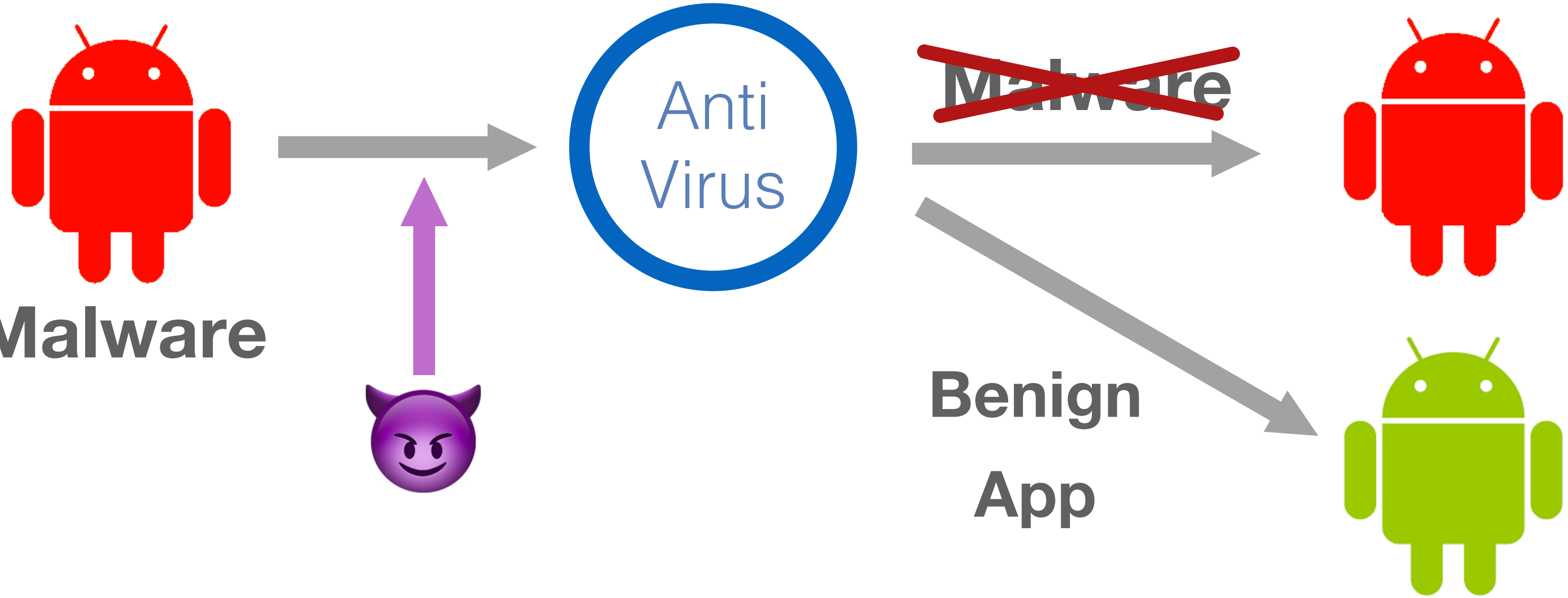
[Lei et al., SysML '19]

Adversarial Examples in Audio & Malware



Audio Attack

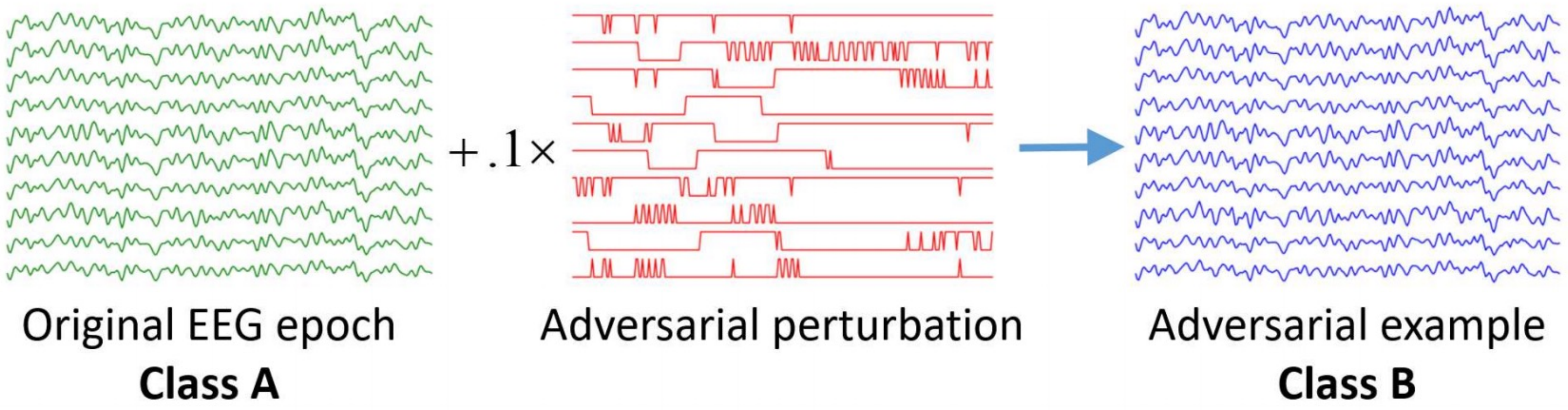
[Carlini & Wagner. DLS 2018]



Android Malware

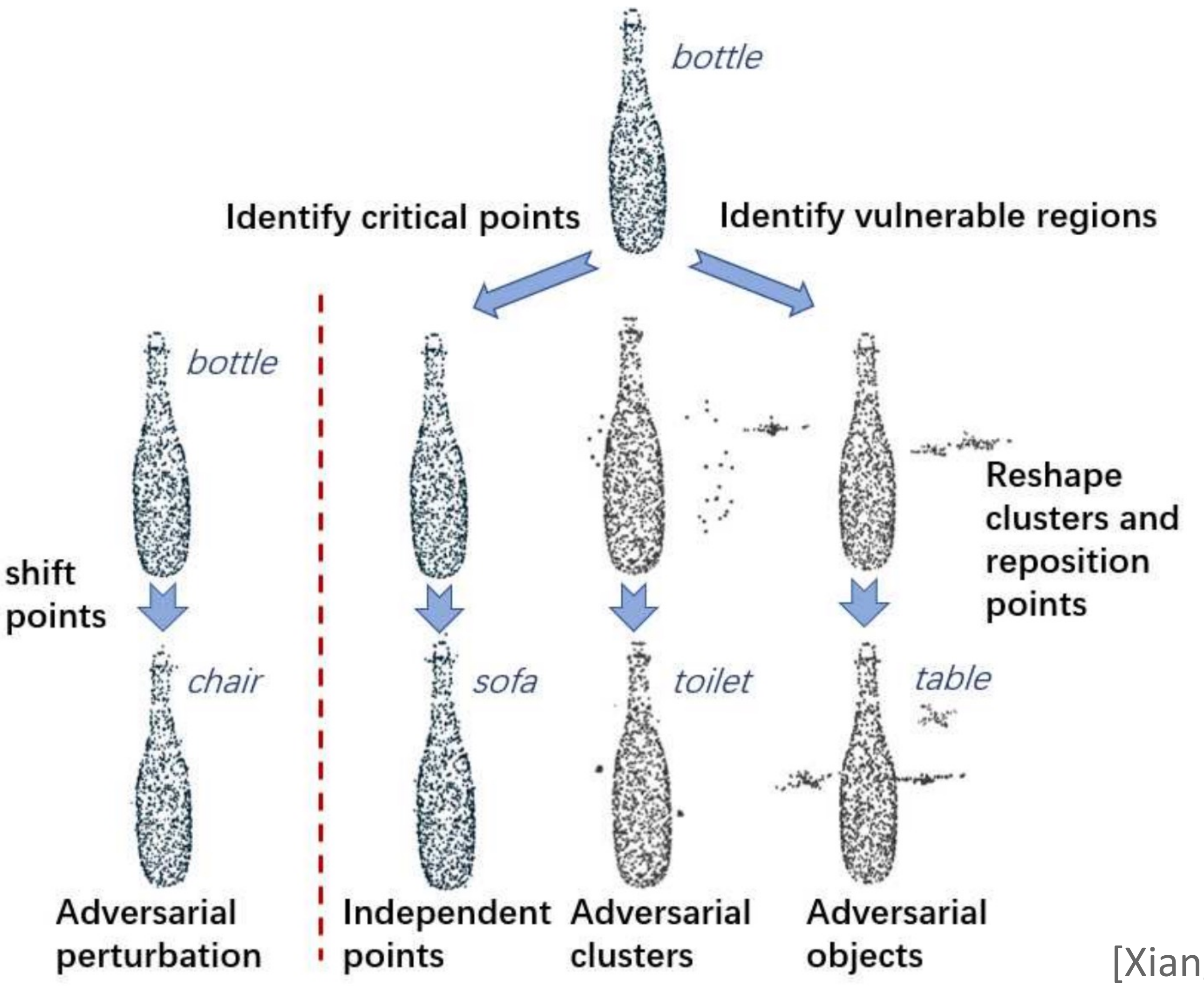
[Jung et al. Black Hat 2017]

Adversarial Examples in Medical Data



[Zhang & Wu, 2019]

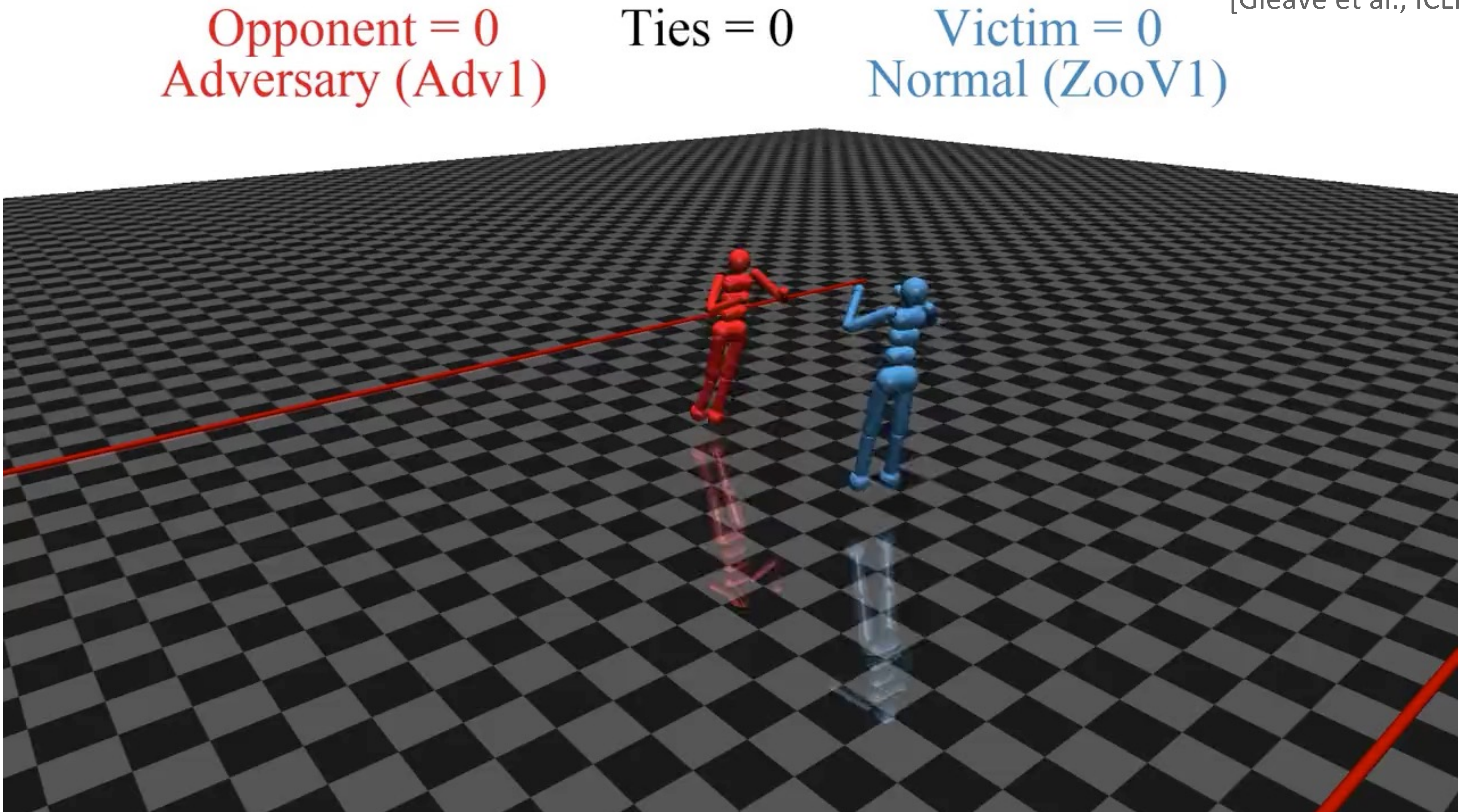
Adversarial Examples in 3D Cloud



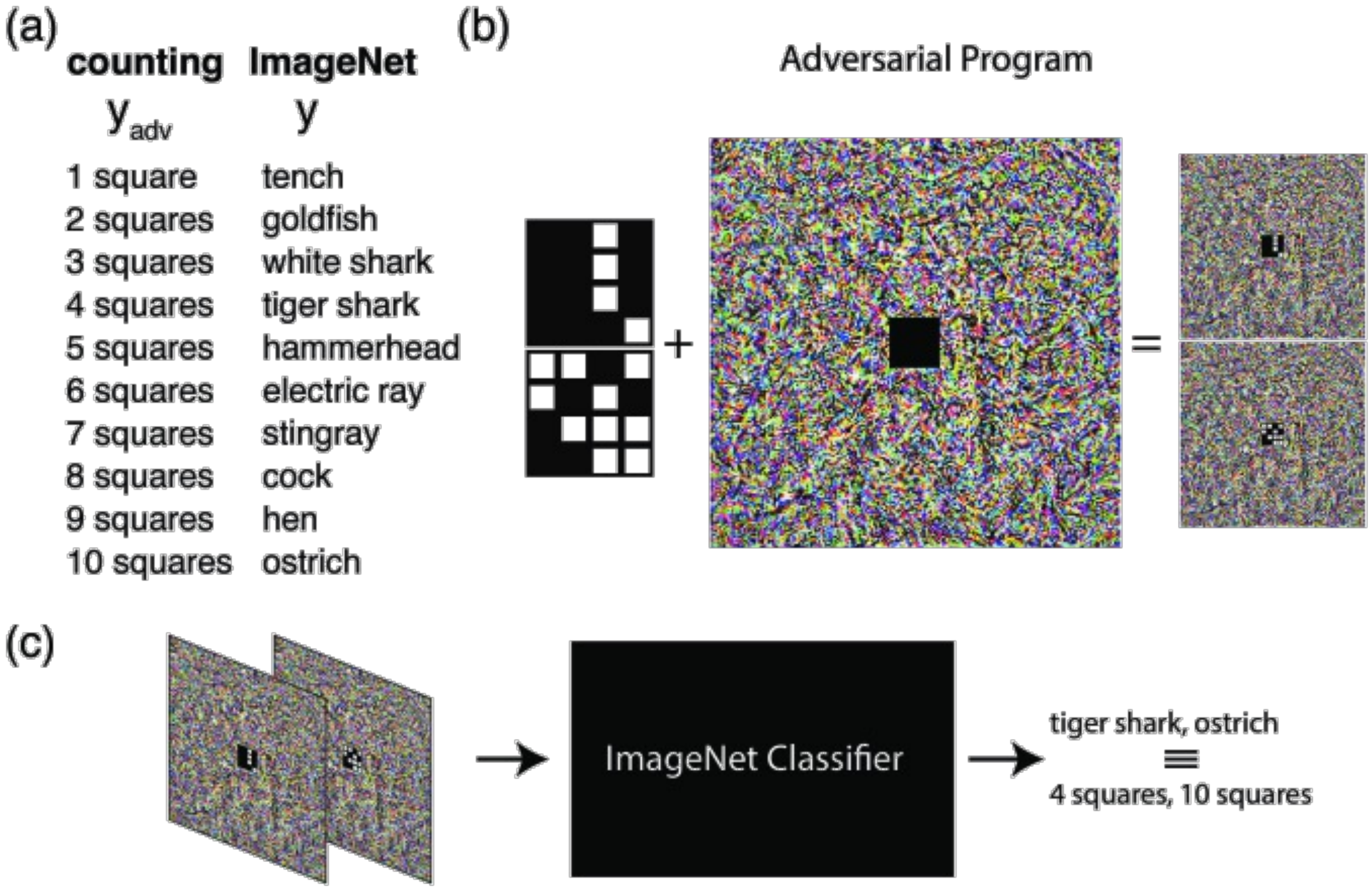
[Xiang et al., CVPR'19]

Adversarial Examples in Reinforcement Learning

[Gleave et al., ICLR'20]

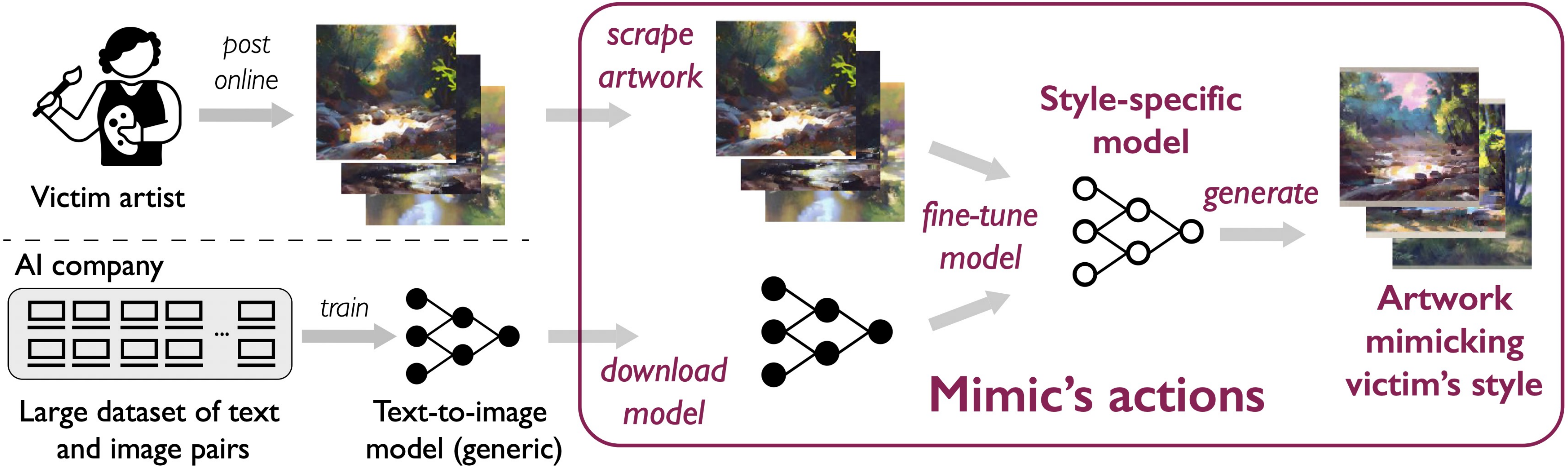


Adversarial Reprogramming



[Elsayed et al., ICLR '19]

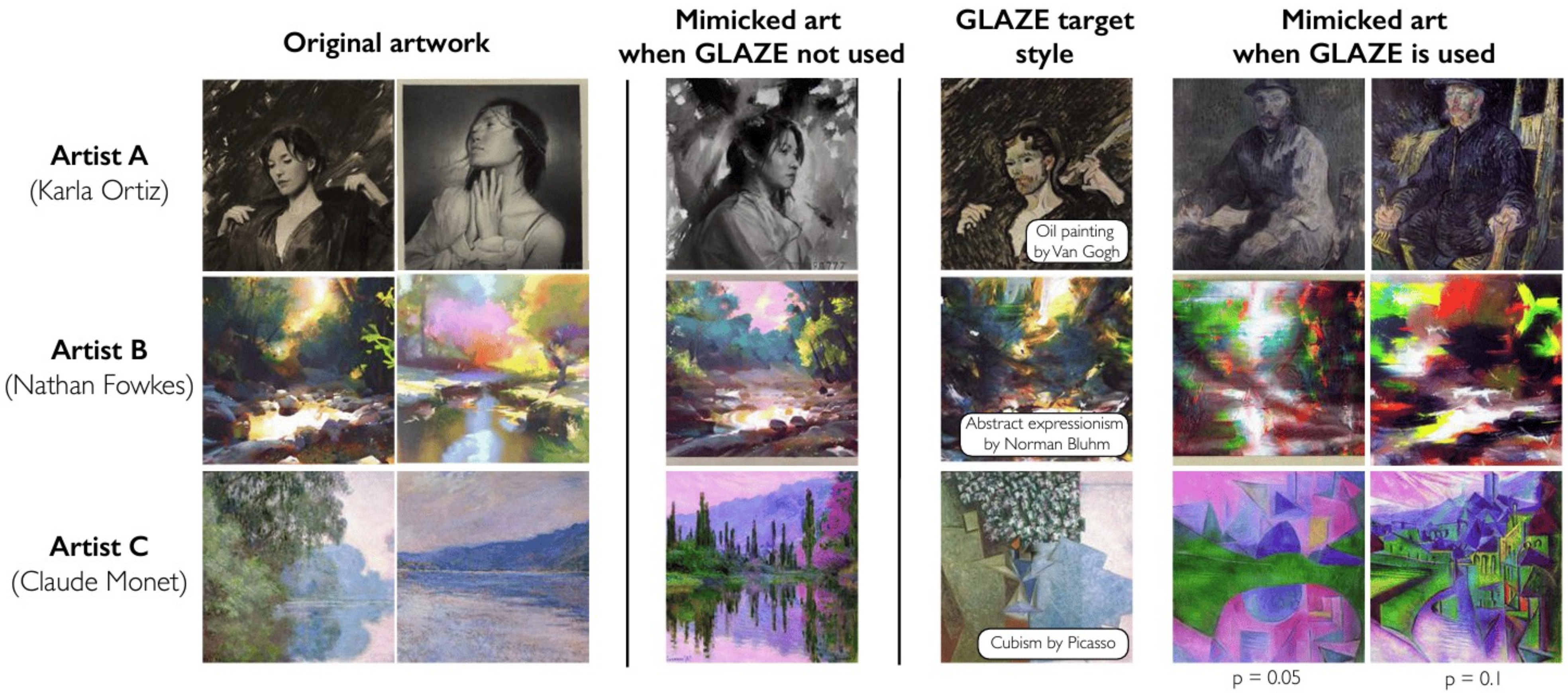
Stealing Art Style



[Shan et al. USENIX Security 2023]

Adversarial Examples for Good

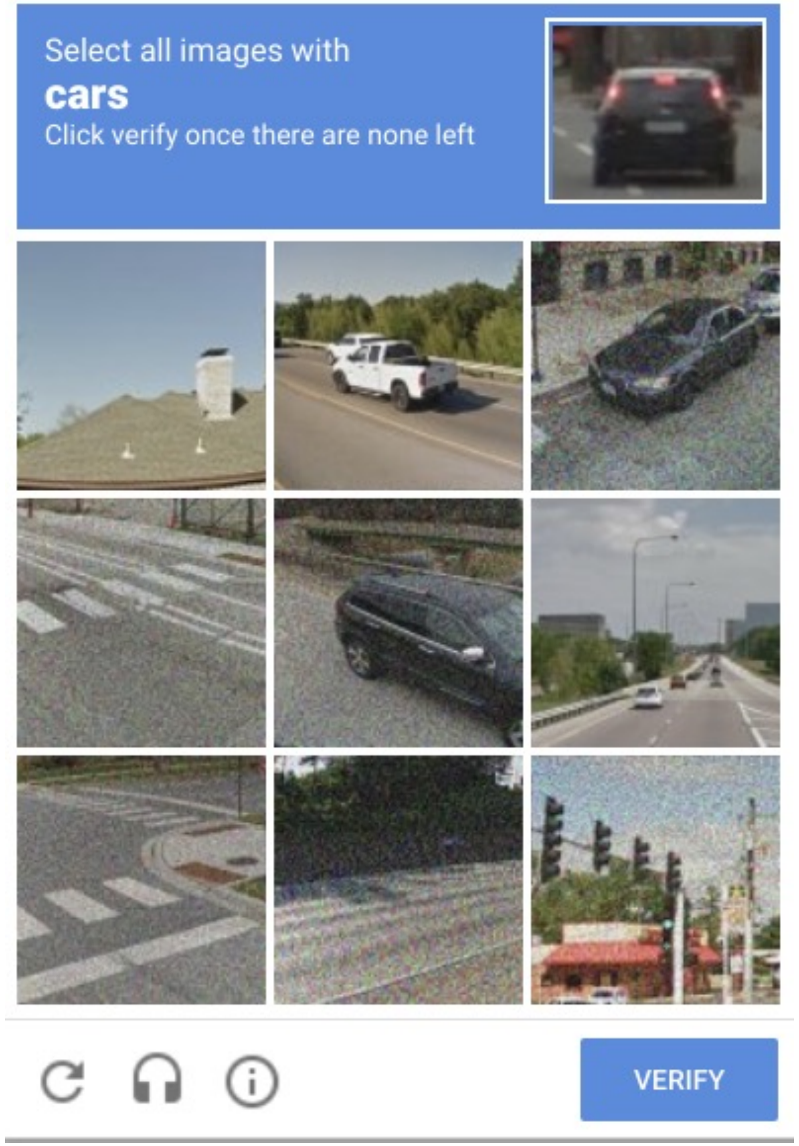
- Protecting from art style stealing



[Shan et al. USENIX Security 2023]

Adversarial Examples for Good

Protecting CAPTCHA



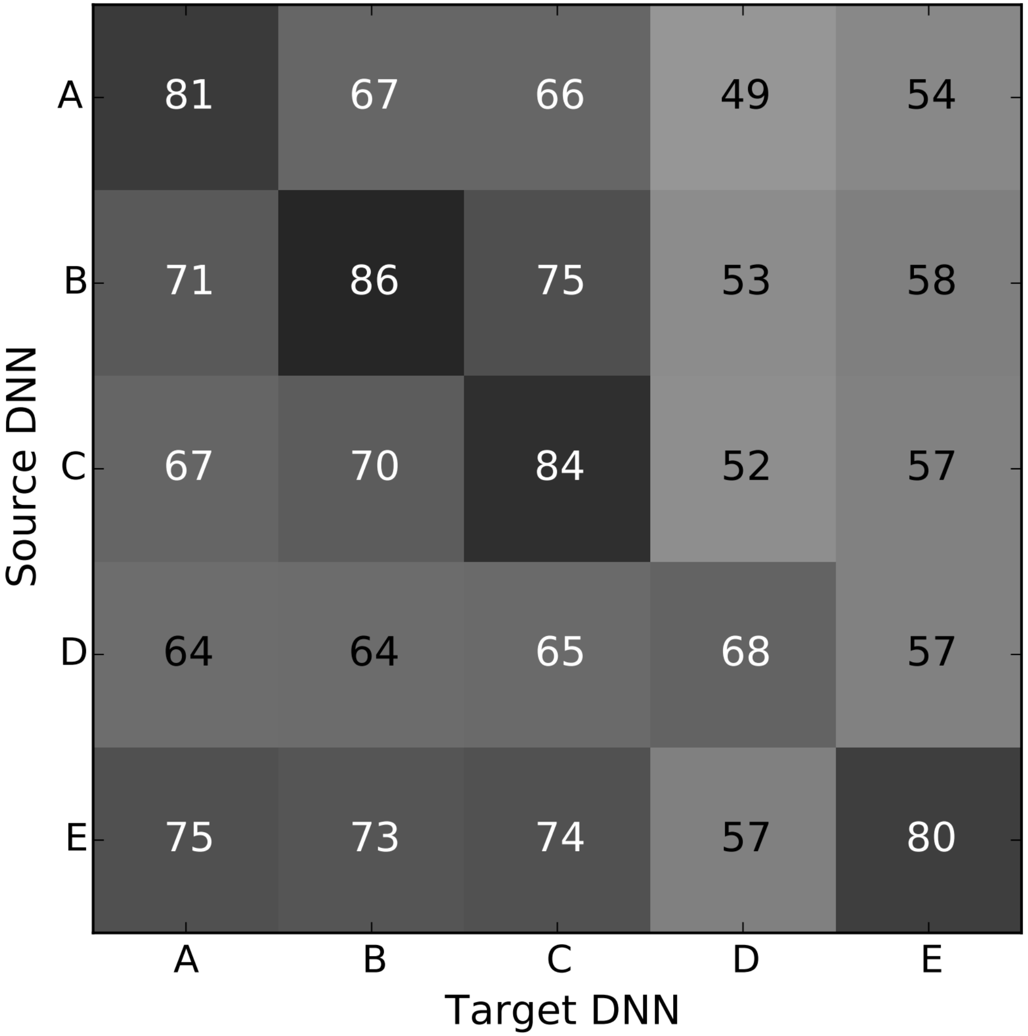
Let's start discussing our first topic: adversarial examples

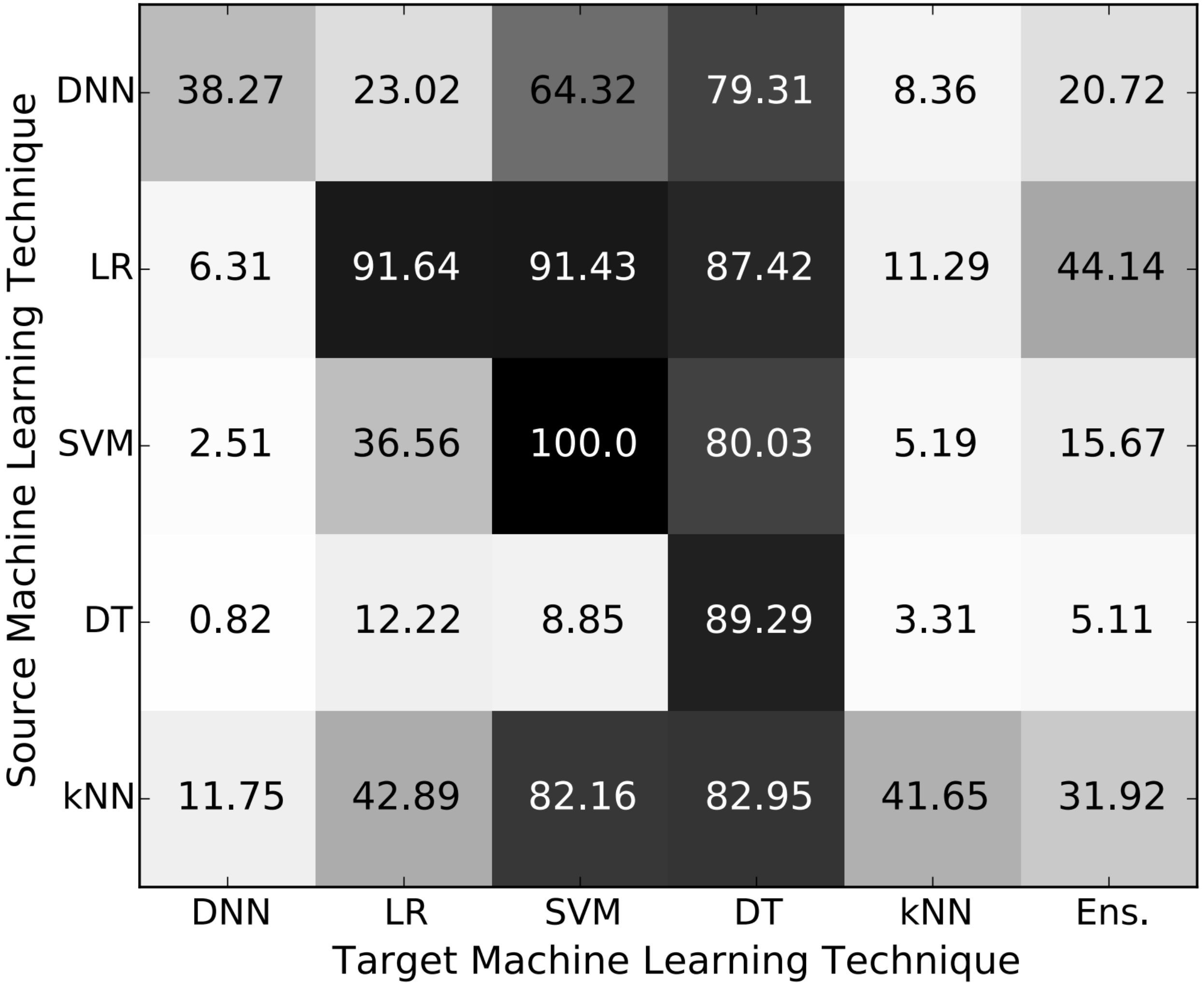
How do we create **adversarial examples**?

Adversarial Attack: Threat Models

- White-box attacks
  - Attacker knows
    - Model architecture
    - Model weights
    - Pre-processing / Post-processing
- Black-box attacks
  - Attacker may or may not know
    - Algorithm (DNN, SVM, …)
    - Features
    - Model architecture
    - Model weights
    - …

Transferability of Adversarial Examples

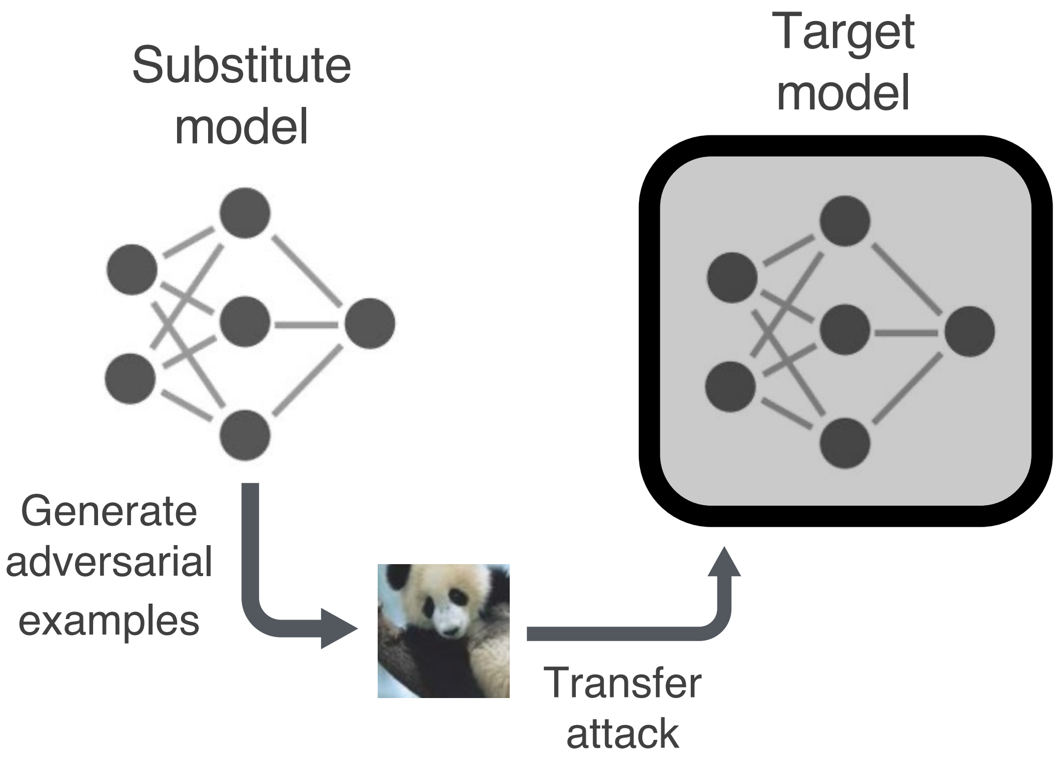




Cross-model Attack Success rate

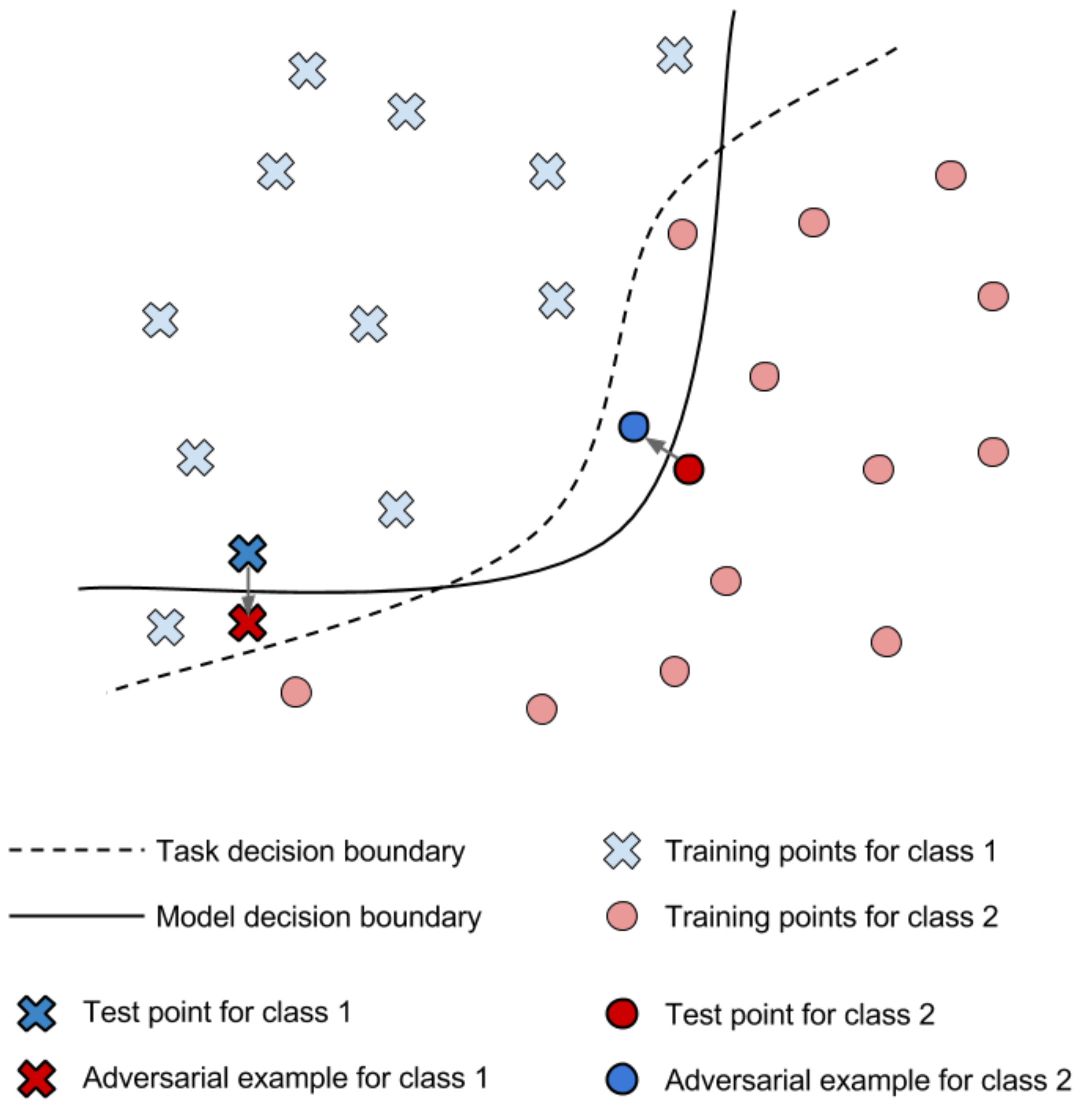
[Papernot et al., USENIX ‘16]

A Simple Blackbox Attack



Goal of Adversarial Examples

Adversarial examples only exist when the model is not perfect



Formal Problem Definition

- Given a classifier C and an example x, find an adversarial example $\boldsymbol{x}'$, s.t. $d(\boldsymbol{x}', \boldsymbol{x}) \leq \epsilon$, and $C(\boldsymbol{x}') \neq C(\boldsymbol{x})$
- The distance function $d(\cdot,\cdot)$ is application dependent
  - For mathematical convenience, ℓ_p distance is often used

$$\|\boldsymbol{x} - \boldsymbol{x}'\|_p = \left(\sum_{i=1}^{N} |\boldsymbol{x}_i - \boldsymbol{x}_i{}'|^p\right)^{\frac{1}{p}}$$

$$\|\delta\|_1 = \sum_{i=1}^{N} |\delta_i| \qquad \|\delta\|_2 = \sqrt{\sum_{i=1}^{N} \delta_i^2} \qquad \|\delta\|_\infty = \max\{|\delta_i|\colon i = 1, \ldots, N\}$$

Rewrite Problem Definition

- Given a classifier C and an example x, find an adversarial perturbation $\boldsymbol{\delta}$, s.t. $\|\delta\|_p \leq \epsilon$, and $C(\boldsymbol{x}+\boldsymbol{\delta}) \neq C(\boldsymbol{x})$
- Important cases
  - ℓ_0 : control the number of pixels that are modified
  - ℓ_1 : control the total amount of pixel value changes
  - ℓ_2 : control the Euclidean distance of pixel value changes
  - ℓ_∞: control the maximum pixel value change

$$\|\boldsymbol{\delta}\|_p = \left(\sum_{i=1}^{N} |\boldsymbol{\delta}|^p\right)^{\frac{1}{p}}$$

Problem Definition: Targeted Attack

- Given a classifier C and an example x, find an adversarial perturbation $\boldsymbol{\delta}$, s.t. $\|\delta\|_p \leq \epsilon$, and $C(\boldsymbol{x} + \boldsymbol{\delta}) = y' \neq C(\boldsymbol{x})$, where y' is the target class

Training and Attack Are Dual Problems

- Training:

$$\min_{\boldsymbol{\theta}} \sum_{(\boldsymbol{x},y)\in S} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})$$

Gradient descent to update model weights $\boldsymbol{\theta}$

- Attack:

(untargeted) $$\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \ell(\boldsymbol{x} + \boldsymbol{\delta}, y; \boldsymbol{\theta})$$

Gradient descent to update input $\boldsymbol{x}$

(targeted) $$\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} -\ell(\boldsymbol{x} + \boldsymbol{\delta}, y'; \boldsymbol{\theta})$$

Fast Gradient Method (L_2) [Goodfellow et al., 2014]

$$\ell(\boldsymbol{x}+\boldsymbol{\delta}, y; \boldsymbol{\theta}) \approx \ell(\boldsymbol{x}, y; \boldsymbol{\theta}) + \boldsymbol{\delta} \cdot \boldsymbol{\nabla}_{\mathbf{x}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})$$

Maximize

$$\ell(\boldsymbol{x}, y; \boldsymbol{\theta}) + \boldsymbol{\delta} \cdot \boldsymbol{\nabla}_{\mathbf{x}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})$$

subject to

$$\|\delta\|_2 \leq \epsilon$$

➔

$$\delta = \epsilon \cdot \frac{\nabla_x \ell(\boldsymbol{x}, y; \boldsymbol{\theta})}{\|\nabla_x \ell(\boldsymbol{x}, y; \boldsymbol{\theta})\|_2}$$

Fast Gradient Method (L_∞)

$$\ell(\boldsymbol{x} + \boldsymbol{\delta}, y; \boldsymbol{\theta}) \approx \ell(\boldsymbol{x}, y; \boldsymbol{\theta}) + \boldsymbol{\delta} \cdot \boldsymbol{\nabla}_{\mathbf{x}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})$$

Maximize

$$\ell(\boldsymbol{x}, y; \boldsymbol{\theta}) + \boldsymbol{\delta} \cdot \boldsymbol{\nabla}_{\mathbf{x}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})$$

subject to

$$\|\delta\|_\infty \le \epsilon$$

➔

$$\delta = \epsilon \cdot \operatorname{sign}\big(\boldsymbol{\nabla}_{\mathbf{x}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})\big)$$

Also known as Fast Gradient Sign Method (FGSM)

Fast Gradient Method (L_1)

$$\ell(\boldsymbol{x}+\boldsymbol{\delta}, y; \boldsymbol{\theta}) \approx \ell(\boldsymbol{x}, y; \boldsymbol{\theta}) + \boldsymbol{\delta} \cdot \boldsymbol{\nabla}_{\mathbf{x}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})$$

Maximize

$$\ell(\boldsymbol{x}, y; \boldsymbol{\theta}) + \boldsymbol{\delta} \cdot \boldsymbol{\nabla}_{\mathbf{x}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})$$

subject to

$$\|\delta\|_1 \leq \epsilon$$

➔

$$\mathrm{i}^* = \mathrm{argmax}_{\mathrm{i}} \left|\boldsymbol{\nabla}_{\mathbf{x}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})_i\right|$$

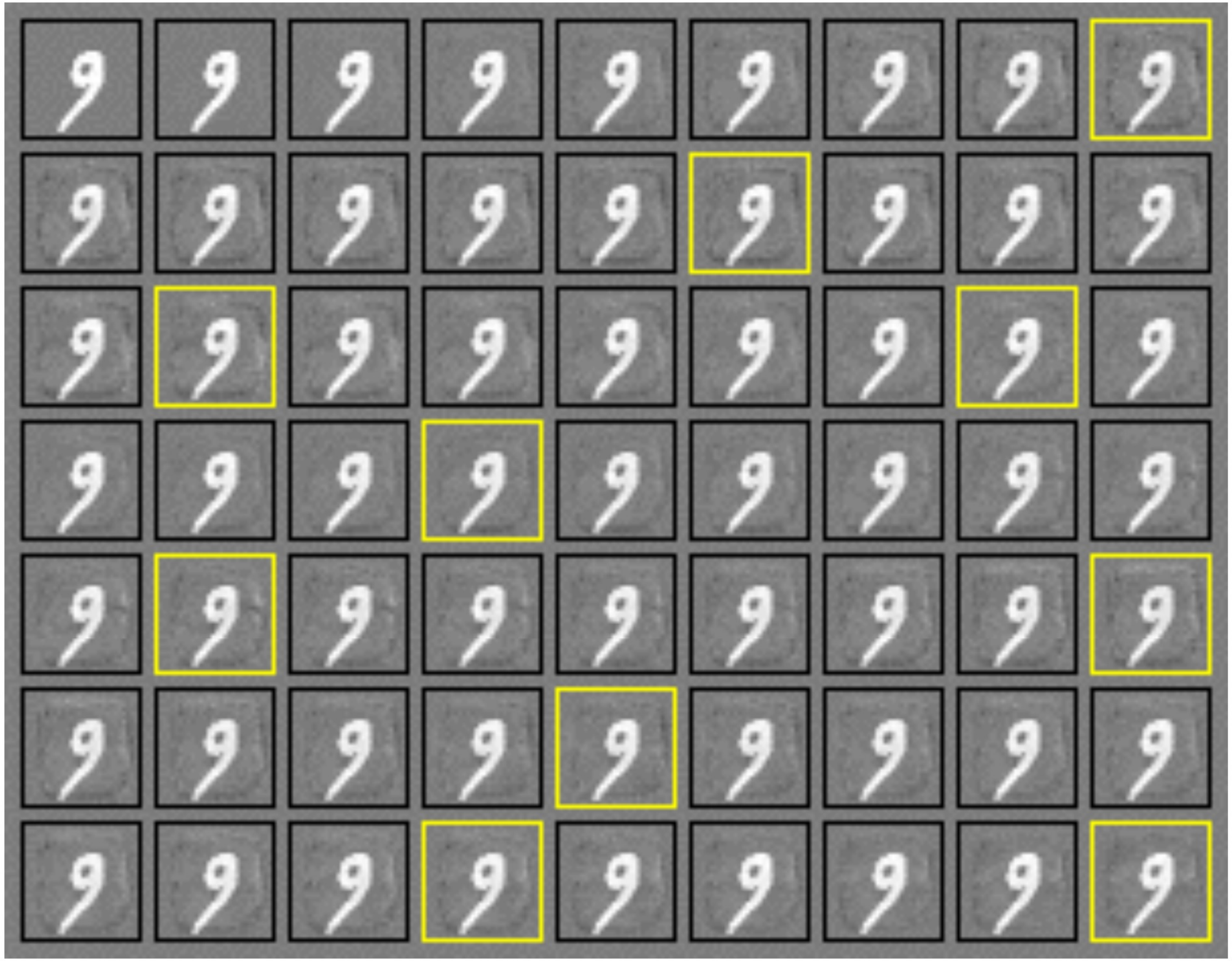
$$\delta_i = \begin{cases} \epsilon \cdot \mathrm{sign}(\boldsymbol{\nabla}_{\mathbf{x}} \ell(\boldsymbol{x}, y; \boldsymbol{\theta})_{\boldsymbol{i}}), & \text{if } i = i^* \\ 0, & \text{otherwise} \end{cases}$$

Even Linear Models Can Be Vulnerable

Linear case: $\boldsymbol{w}^T\boldsymbol{x}' = \boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{w}^T\boldsymbol{\delta}$

- Output value can change as big as $\boldsymbol{\epsilon}||\boldsymbol{w}||_{\mathbf{1}}$ ($\boldsymbol{L}_\infty$ attack)
- The change can be quite big for high dimensional case
- A root cause of adversarial example: ML models perceive high dimensional data different than human

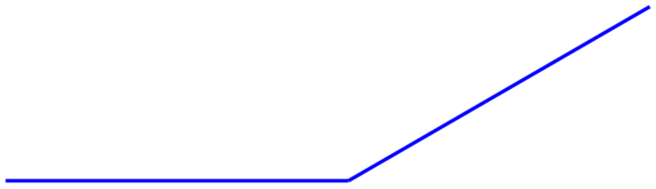
Even Linear Model Can Be Vulnerable



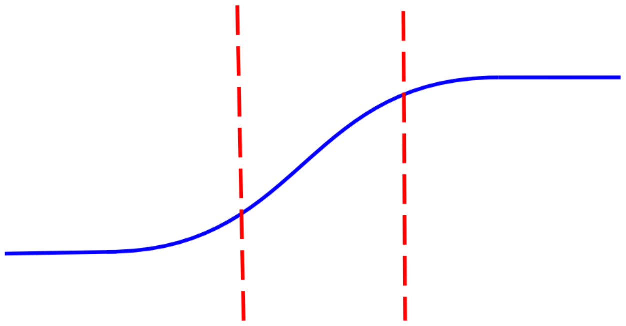
[Goodfellow 2016]

Many Modern DNNs Are Piecewise Linear

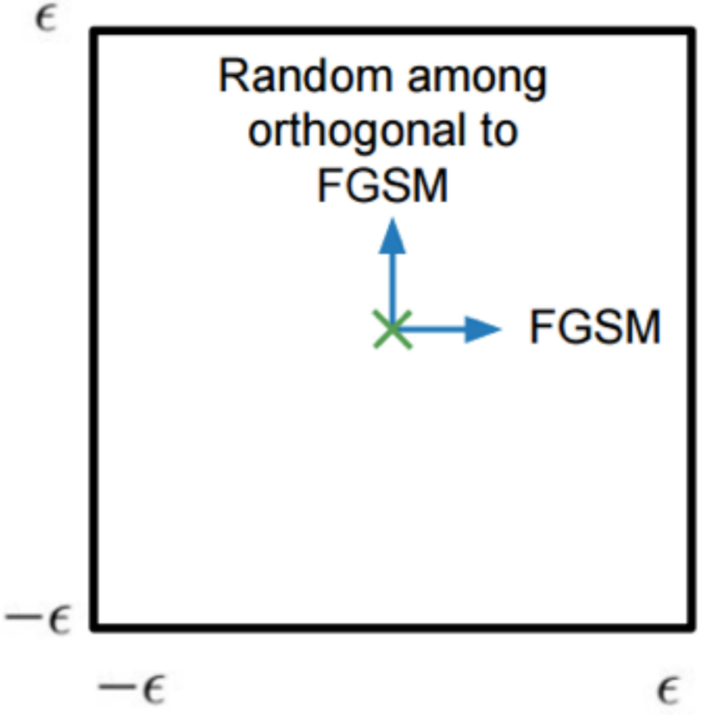
- Convolution
- ReLU

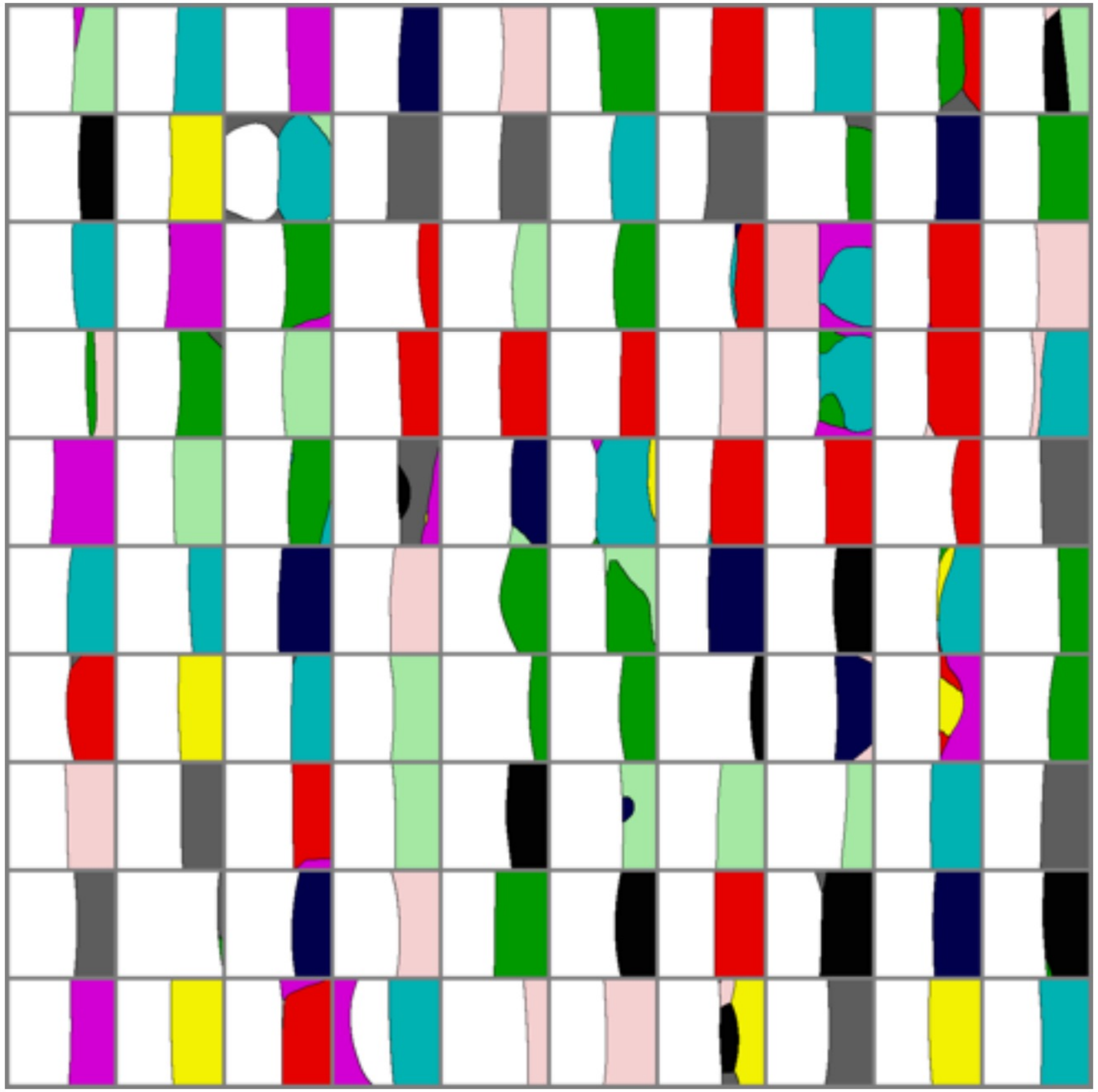


- Sigmoid



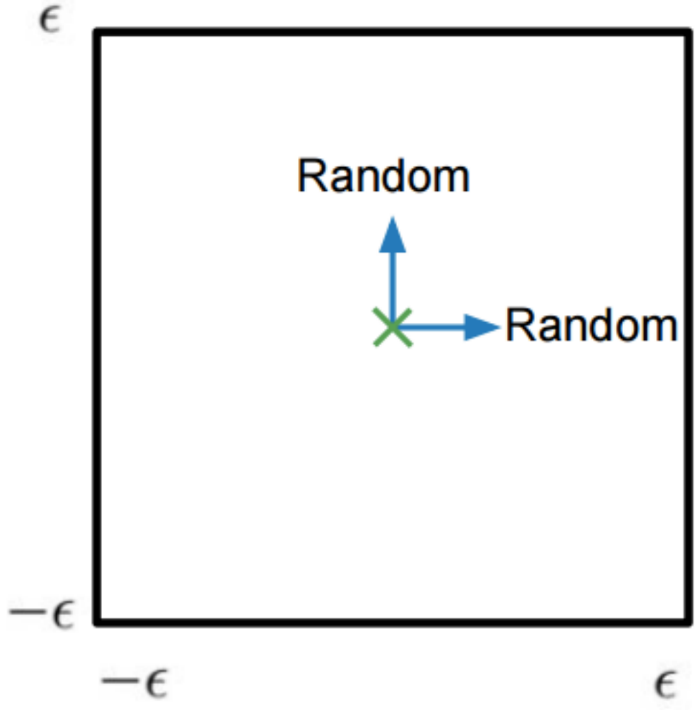
Many Modern DNNs Are Piecewise Linear

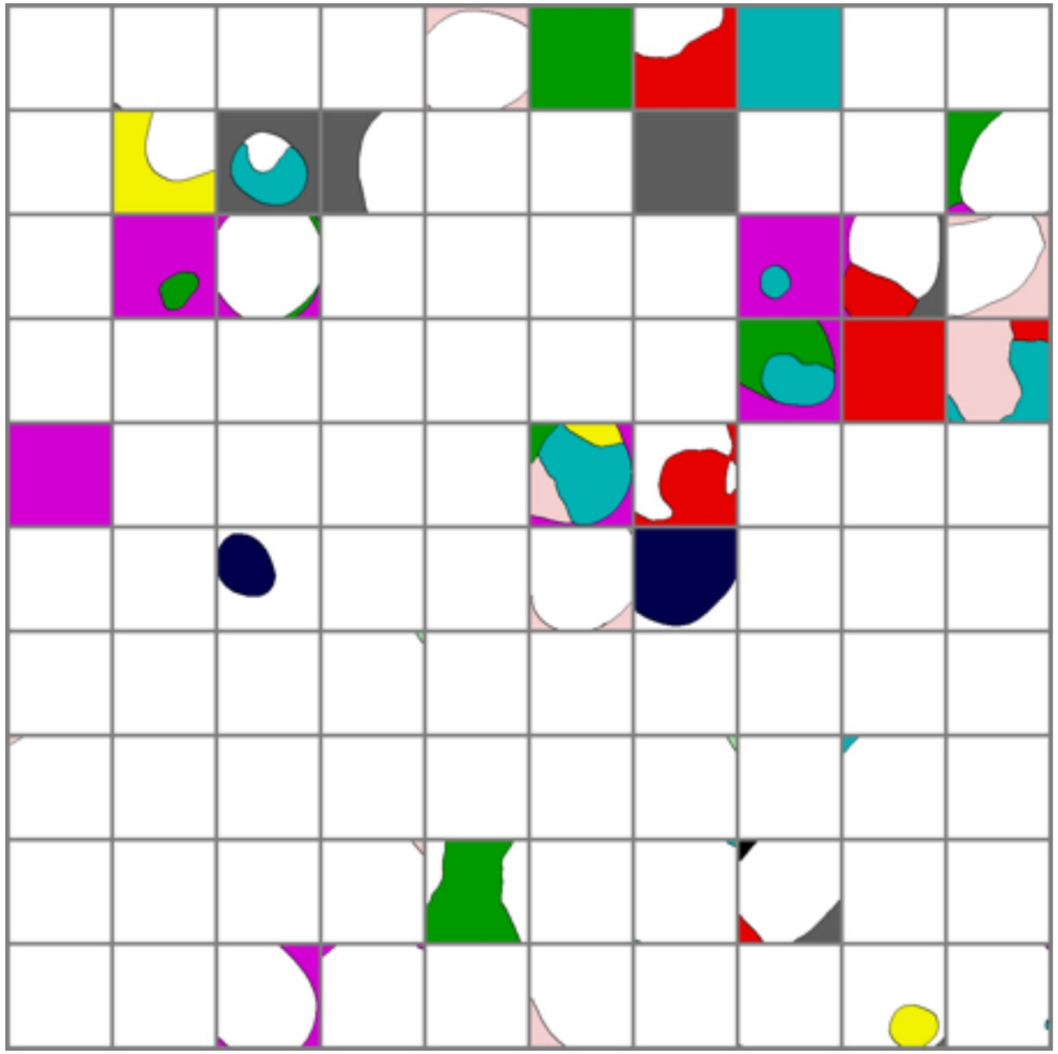




[Goodfellow 2016]

Adversarial Perturbations Are Not Noise





[Goodfellow 2016]

Nuances of Different Attacks

- Most attacks are similar, with differences in
  - Loss function (e.g., cross-entropy, hinge-loss)
  - Constraints
  - Optimization algorithm

Projected Gradient Descent [Madry et al. '17]

- Roughly like running FGM iteratively
- Project $\boldsymbol{\delta}$ back to Δ after each iteration

$$\boldsymbol{x}^{t+1} = \mathrm{clip}(\boldsymbol{x^t} + \alpha \cdot \nabla_x \ell(\boldsymbol{x}^t, y; \boldsymbol{\theta}), [-\epsilon, \epsilon])$$

- Randomly choose a start point within ϵ-ball of $\boldsymbol{\delta}$
- It is considered as the strongest first-order attack

Carlini Wagner (CW) Attack [Carlini & Wagne, 2017]

$$\begin{aligned} \text{minimize} \quad & \|\delta\|_p + c \cdot f(x+\delta) \\ \text{such that} \quad & x+\delta \in [0,1]^n \end{aligned}$$

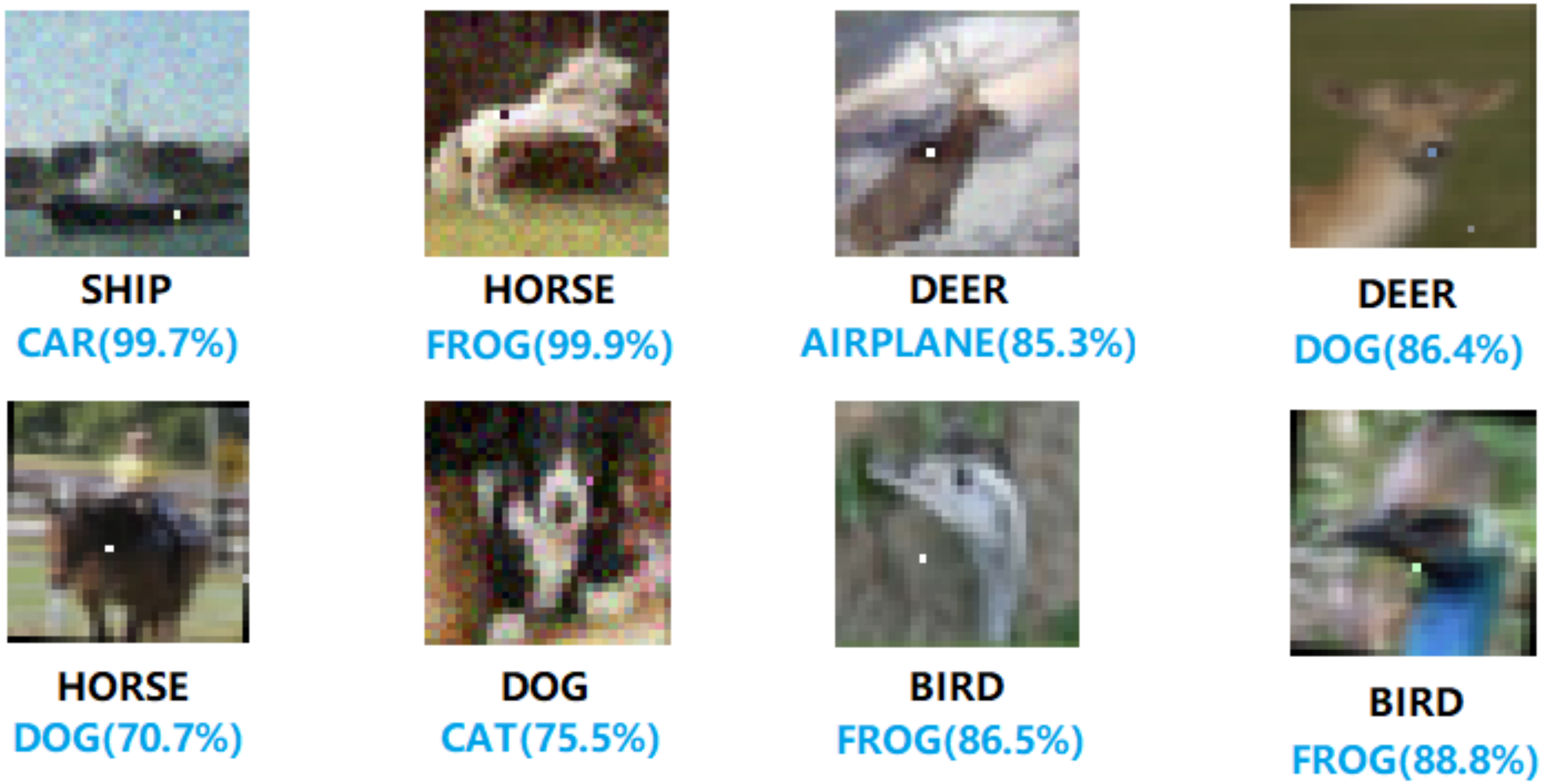
where $f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -\kappa)$

- Find best c with binary search
- Usually slower than PGD

One Pixel Attack

[Su et al., 2017]

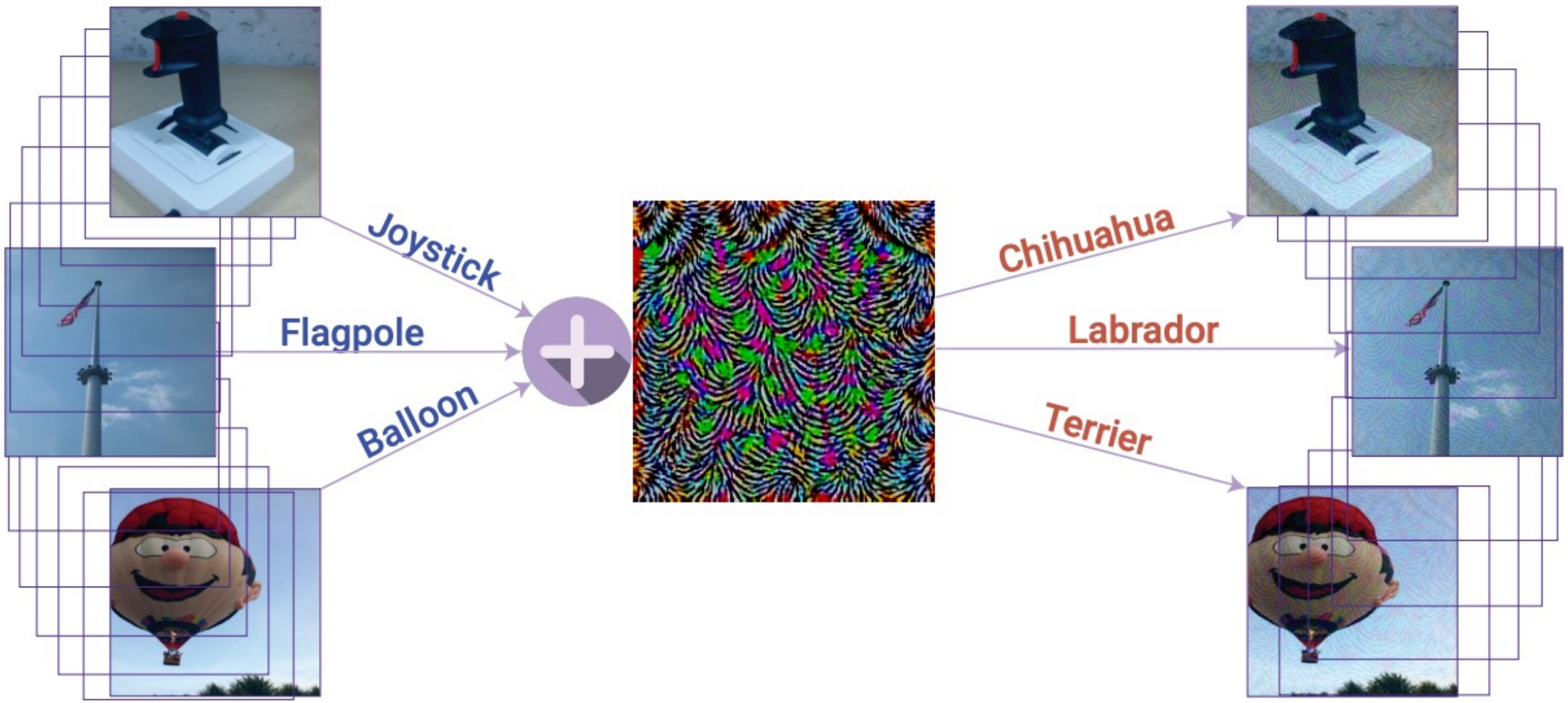
- Using evolutionary algorithms



Universal Adversarial Perturbation

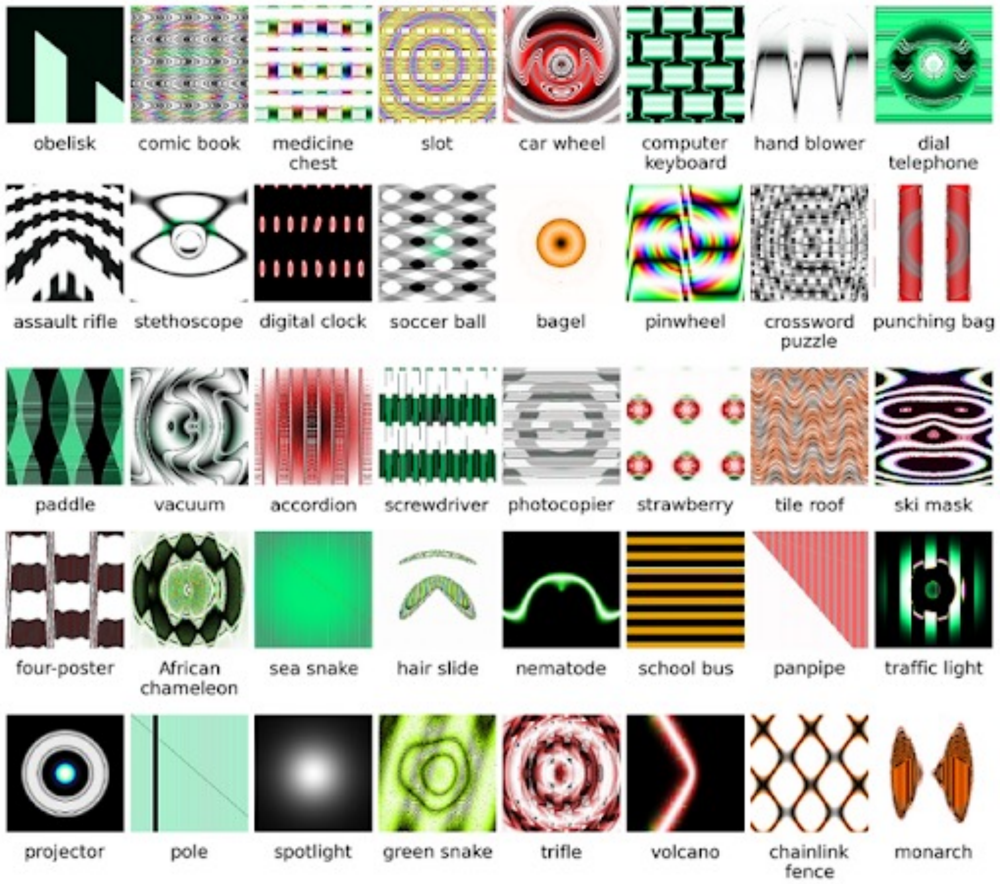
[Moosavi-Dezfooli et al., 2017]

One perturbation works for all input

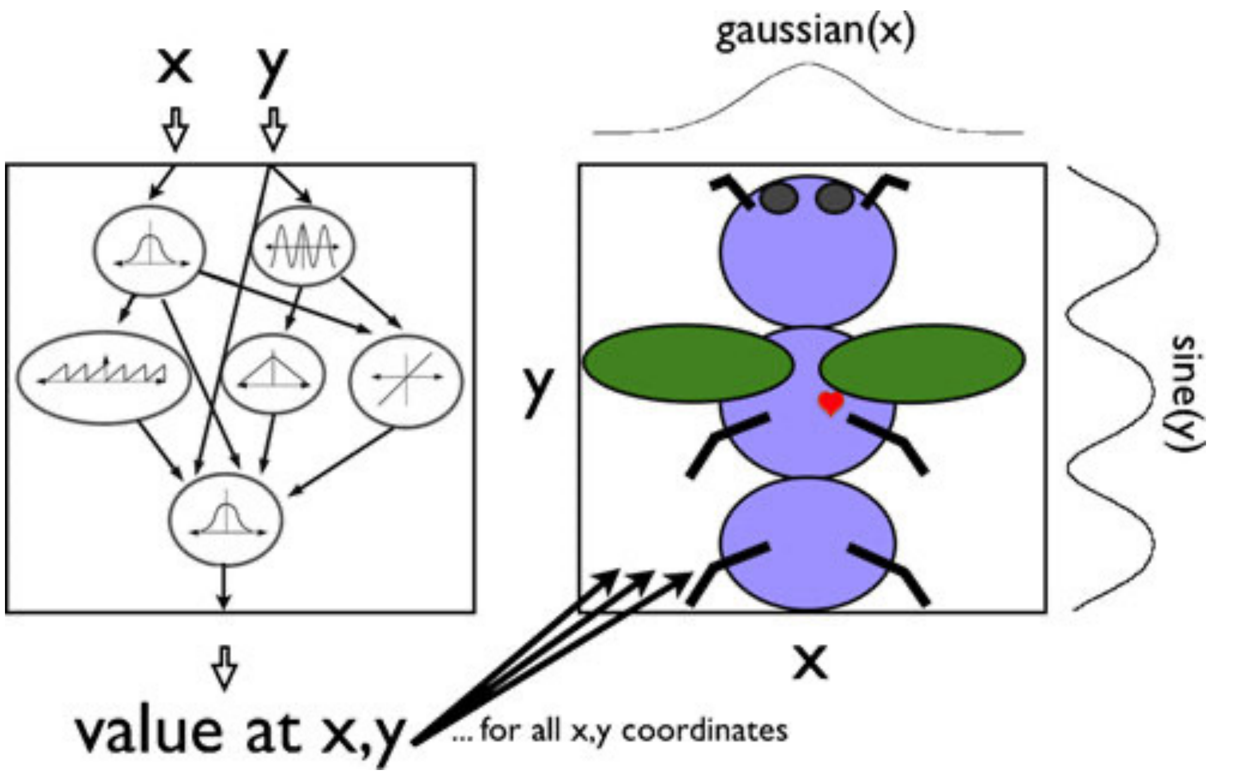


Fooling Images

Use evolution algorithms to generate "fooling images"



[Nguyen et al., 2014]



Compositional Pattern Producing Networks (CPPN)

[Clune et al., 2011]

Zeroth Order Optimization (ZOO)

[Chen et al., 2017]

- A black-box attack that only requires model logits output
- Stochastic coordinate descent with gradient estimate

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i} \approx \frac{f(\mathbf{x} + h\mathbf{e}_i) - f(\mathbf{x} - h\mathbf{e}_i)}{2h}$$

- Importance sampling to reduce number of queries

Adversarial Attack Against Human Vision

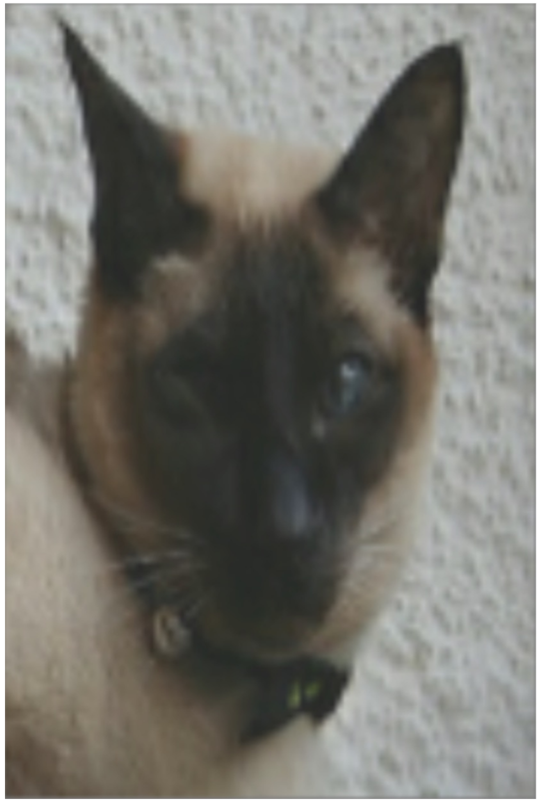
[Elsayed et al., 2018]

What is the species of this dog?

Adversarial Attack Against Human Vision

[Elsayed et al., 2018]

It is actually a cat!



That’s Enough Attack for Now

- We will talk about defenses next week
- Please choose a paper in week 2’s list and turn in the critique by noon next week