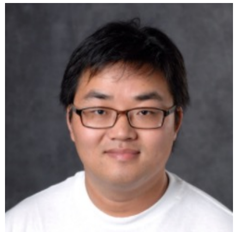
Introduction



Jiliang Tang

→ Privacy

Xiaorui Liu

→ Safety & Robustness

Yaxin Li

→ Explainability

Wenqi Fan

→ Non-discrimination & Fairness

Environmental Well- being

Haochen Liu

→ Accountability & Auditability

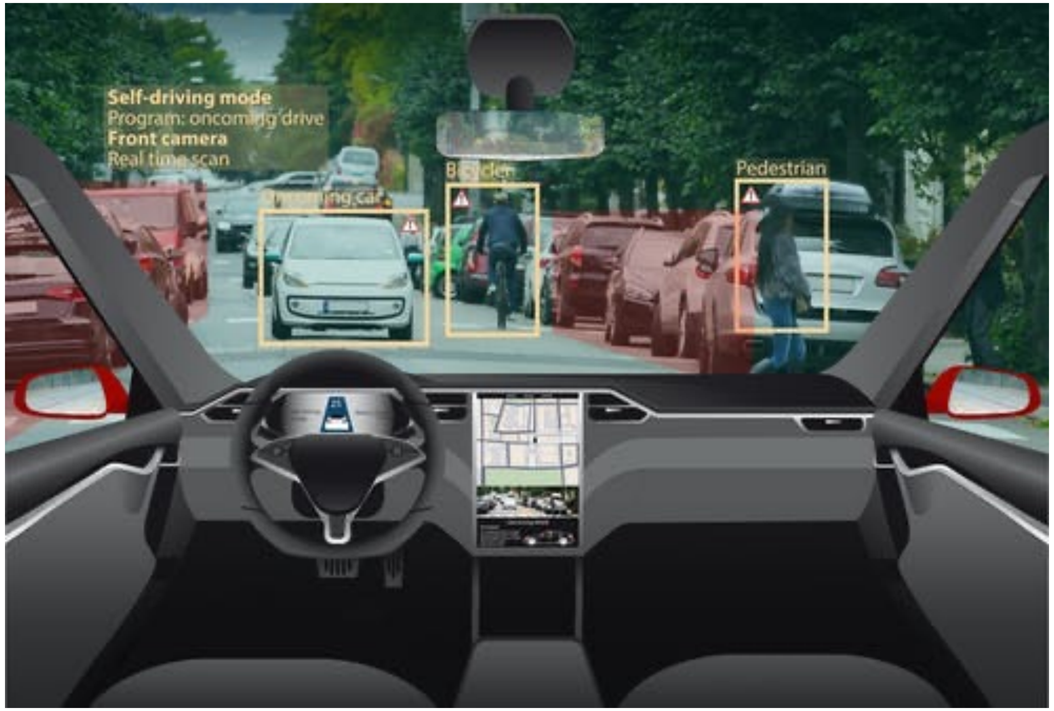
Dimension Interactions

Future Directions

Yiqi Wang

Real World Threaten for AI

Unlock Your Phone



Self-driving

Safety and Robustness

By examining **Adversarial Robustness**,

we expect the AI system to:

- not only work “most of the time”, but be stable under worst case and achieve sustained high accuracy.

Outline

- ❑Concepts and Taxonomy
- ❑Adversarial Attack
- ❑Adversarial Defense
- ❑Robustness in Graph and Text Domains
- ❑Real World Adversarial Attack
- ❑Adversarial Learning Surveys and Tools
- ❑Future Directions

Taxonomy

- Adversarial Robustness
  - Adversarial Attack
    - Different types
    - Different domains
    - Different Victims
  - Adversarial Defense

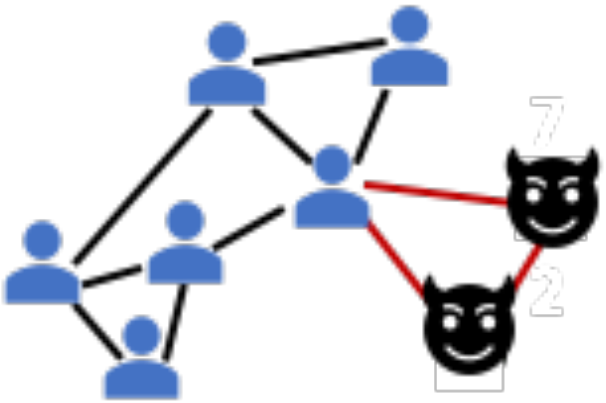
Adversarial Attack

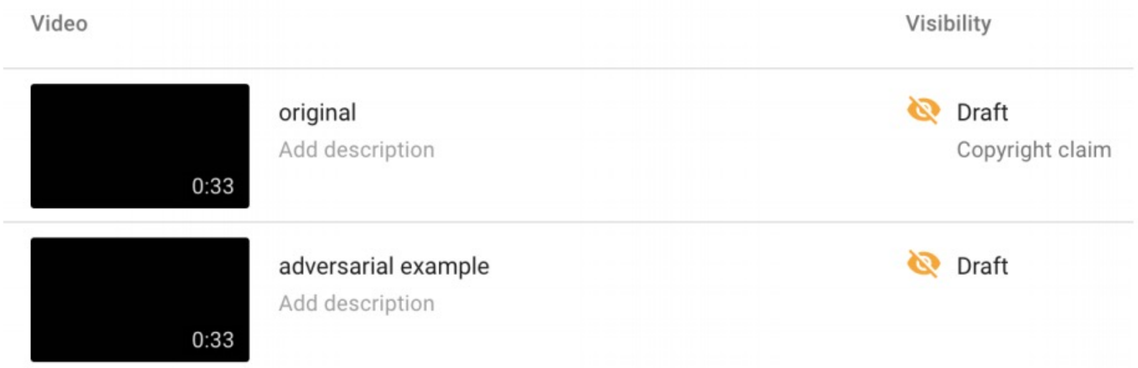
- Poisoning Attacks vs. Evasion Attacks.
  - happen in **training phase**/ happen in **test phase**.
- White-box attacks vs. Black-box attacks.
  - have **all knowledge** of the victim model/ have **no knowledge** or limit knowledge.
- Targeted Attacks vs. Non-Targeted Attacks.
  - require **specified target** prediction label/ expect **arbitrary wrong label**.

Adversarial in Different Domains

- ❑Image Data
- ❑Graph Data
- ❑Text Data
- ❑Audio Data
- ❑…

Original Input	Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: <u>**Positive (77%)**</u>
Adversarial example [Visually similar]	<u>**Aonnoisseurs**</u> of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: <u>**Negative (52%)**</u>
Adversarial example [Semantically similar]	Connoisseurs of Chinese <u>**footage**</u> will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: <u>**Negative (54%)**</u>





Intriguing properties of neural networks, Szegedy et al 2013.
Adversarial Attacks on Copyright Detection Systems， ICML2020

Adversarial Victims

SVM

Decision Tree

...

Traditional Machine Learning Models

Convolutional Neural Network

Recurrent Neural Network

Graph Neural Network

Visual Transformer

Generative Network

...

Deep Learning Models

Robustness and Regularization of Support Vector Machines, JMLR 2009

Robustness Verification of Tree-based Models, NeurIPS 2019

On the Adversarial Robustness of Visual Transformers, arxiv 2021

Adversarial Defenses

- ❑Adversarial Training/Robust Optimization.
- ❑Certified Defense/Provable Robustness.
- ❑Adversarial Example Detection.
- ❑Data Preprocessing.
- ❑…

Outline

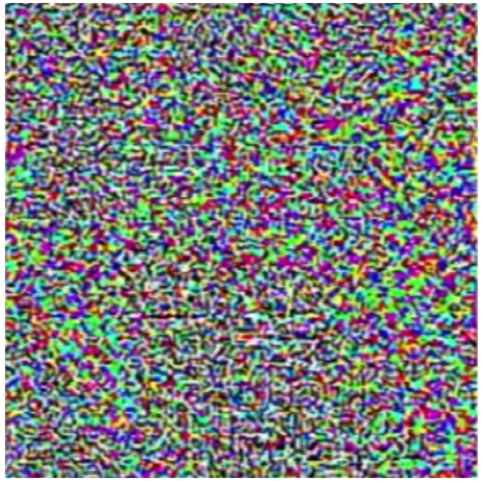
- ❑Concepts and Taxonomy
- ❑Adversarial Attack
- ❑Adversarial Defense
- ❑Robustness in Graph and Text Domains
- ❑Real World Adversarial Attack
- ❑Adversarial Learning Surveys and Tools
- ❑Future Directions

Adversarial Attacks

"tabby cat" (95%)

"noise" (calculated)

$+0.05 \times$



$=$

"strawberry" (99%)

Explaining and Harnessing Adversarial Examples, Goodfellow et al 2014
Intriguing properties of neural networks, Szegedy et al 2013.

White-Box Attack

- Attacker's knowledge:

Suppose model F with parameter θ is given to attacker;

- Attacker's Goal:

For a test sample x with true label y , find a small perturbation δ such that $F(x+\delta) \neq y$.

Explaining and Harnessing Adversarial Examples, Goodfellow et al 2014

White-Box Attack: Optimization

❑Optimization Objective:

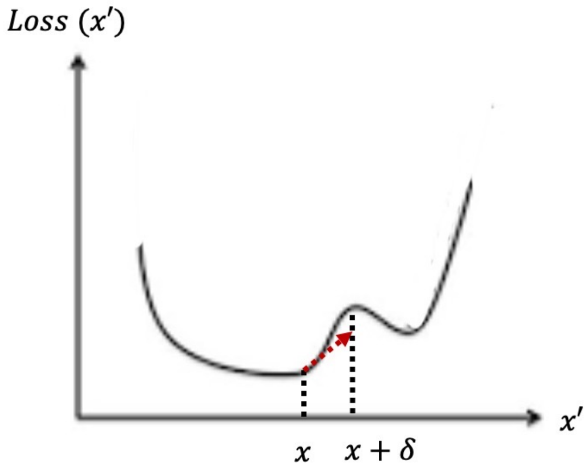
$$\max_{\delta} Loss\ (F(x+\delta;\theta), y)$$

$$\text{Subject to } ||\delta||_p \leq \epsilon$$

❑Projected Gradient Descent (PGD attack, l_∞):

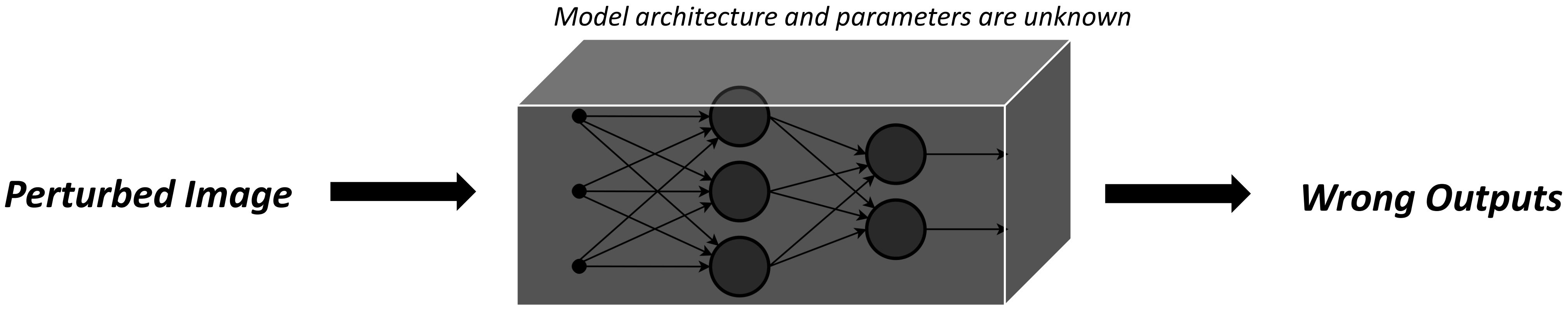
- Start from the original sample x
- Calculate iteratively:

$$x+\delta = clip_{(x,\epsilon)}\{x + \alpha \cdot \text{sign}\ (\nabla_x Loss\ (F(x;\theta), y))\}$$



Explaining and Harnessing Adversarial Examples, Goodfellow et al 2014

Black-Box Attack



If the model parameter is unknow, how to calculate perturbation?

Black-Box Attack

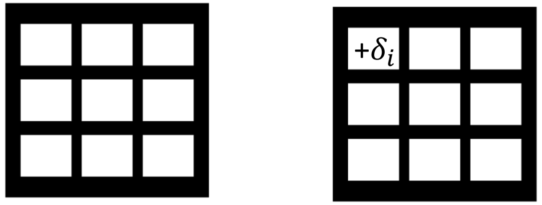
❑Attacker's knowledge:

Attacker can only get the prediction or output score of model F for a sample x.

❑Attacker's Strategies:

- Substitute model
- Approximate gradient:

  Zeroth Order Optimization Attack (Zoo attack):



x $\quad$ $x + \delta_i$

$$\frac{\partial F(x)}{\partial x_i} \approx \frac{F(x + he_i) - F(x - he_i)}{2h}$$

ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models

Reliability of White-Box Attack

❑Gradient Masking

A defense is said to cause gradient masking if it “does not have useful gradients” for generating adversarial examples.

- Shattered gradients: caused by non-differentiable
- Stochastic gradients: caused by randomization
- Exploding and vanishing gradients: loss function, deep network

Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples

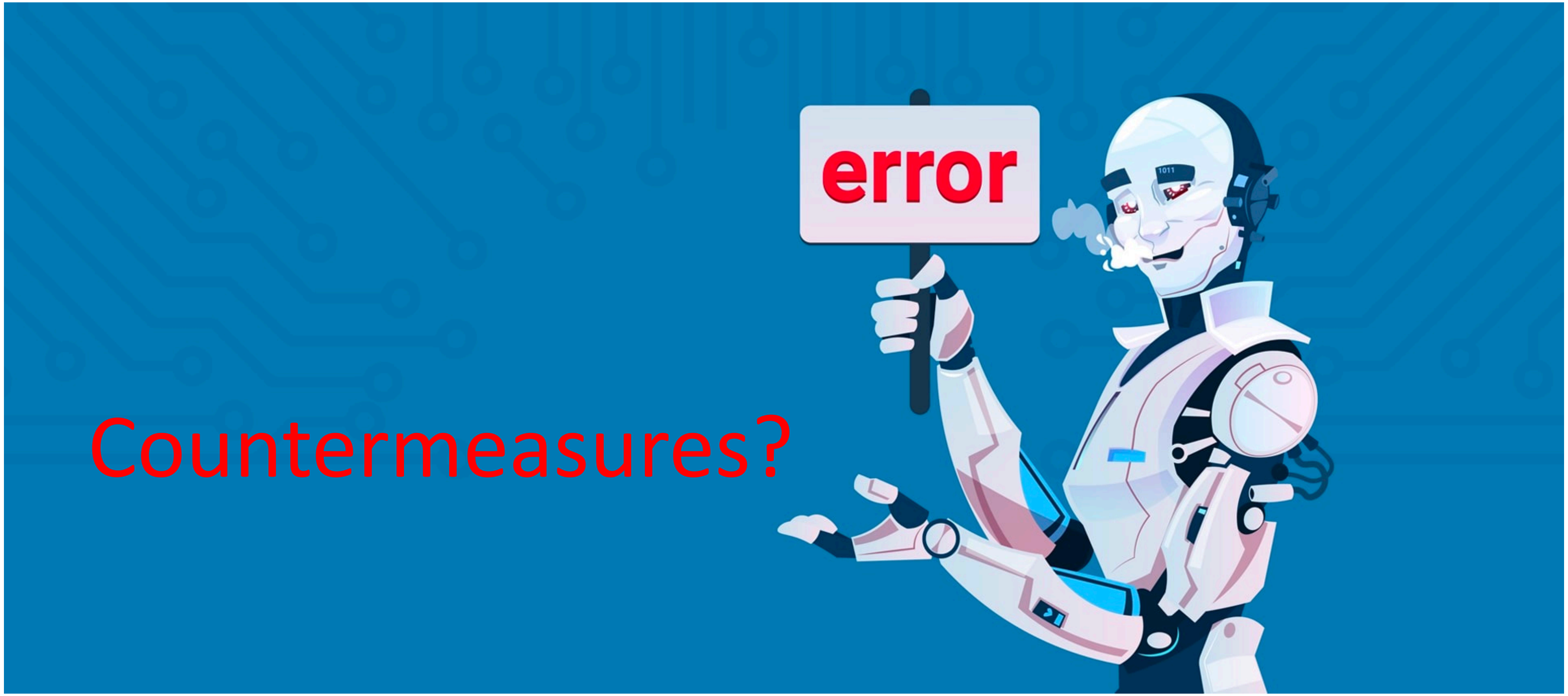
AutoAttack

- Ensemble Attack: AutoAttack
  - Gather four diverse attacks:

  1) APGD-CE and APGE-DLR: solve the gradient vanishing problem.
  2) FAB: white box attack for minimal perturbation.
  3) Square Attack: random search based black-box adversarial attack.

  - Reliable Robust Evaluation

Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks


error
1011
Countermeasures?

Outline

- ❑Concepts and Taxonomy
- ❑Adversarial Attack
- ❑Adversarial Defense
- ❑Robustness in Graph and Text Domains
- ❑Real World Adversarial Attack
- ❑Adversarial Learning Surveys and Tools
- ❑Future Directions

Adversarial Training

- Goal of adversarial training:

  Training model to minimize (empirical) adversarial risk:

$$\min_{\theta} \sum_{(x,y)\sim D} \max_{||\delta||\leq\epsilon} Loss\,(F(x+\delta;\theta),y)$$

- Adversarial Training:

For each batch of samples:

1. Solve the inner maximization problem to find optimal δ^*
2. Update model parameters θ the to minimize the loss value on $x+\delta^*$

Explaining and Harnessing Adversarial Examples, Goodfellow et al 2014

Provable Robustness

- Adversarial training only achieve limited robustness empirically.
- Strong Vanilla Attacks:

  PGD, CW
- Adaptive Attacks:

  BPDA, EOT, Black Box Attacks, Auto Attack

How can a defense model guaranteed to be safe?

Randomized Smoothing

- ❑ Goal: Guarantee robustness in a bounded neighborhood.

- ❑ Strength:
  - Smoother classifier.
  - Proved to be robust in a certain radius.

❑ *Training with Gaussian Noise:*

1. Given training inputs x
2. Generate k samples with gaussian noise:

$$\delta \sim N(0\ \sigma^2 I)$$

3. Train with noise samples.

❑ *Prediction:*

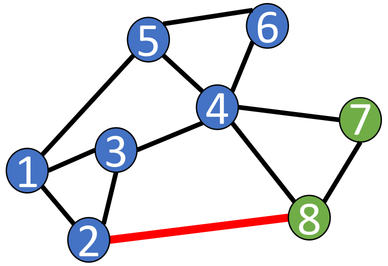
1. Given test input x
2. Generate n Gaussian noise, create $x_0 \ldots x_n$
3. For each x_i, the neural network will give a prediction label c.
4. Count the prediction labels and find the most frequent one to assign as prediction.

Certified Adversarial Robustness via Randomized Smoothing, Cohen 2019

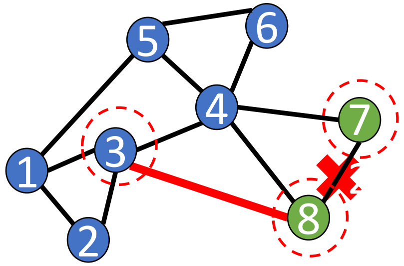
Outline

- ❑Concepts and Taxonomy
- ❑Adversarial Attack for Machine Learning
- ❑Adversarial Defense for Machine Learning
- ❑Robustness in Graph and Text Domains
- ❑Real World Adversarial Attack
- ❑Adversarial Learning Surveys and Tools

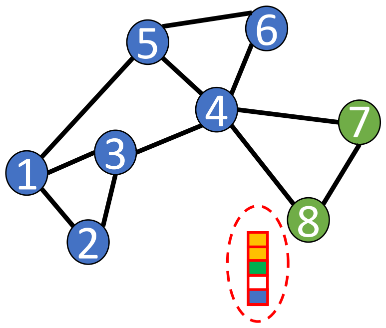
Graph Attack



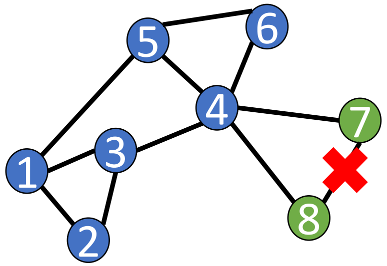
Adding an edge



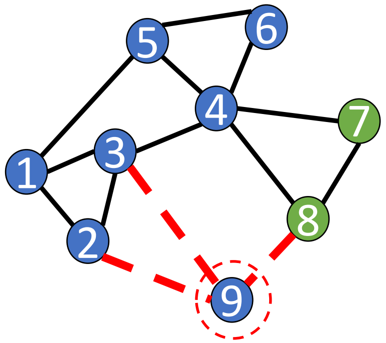
Rewiring



Modifying Features



Deleting an edge



Node Injection

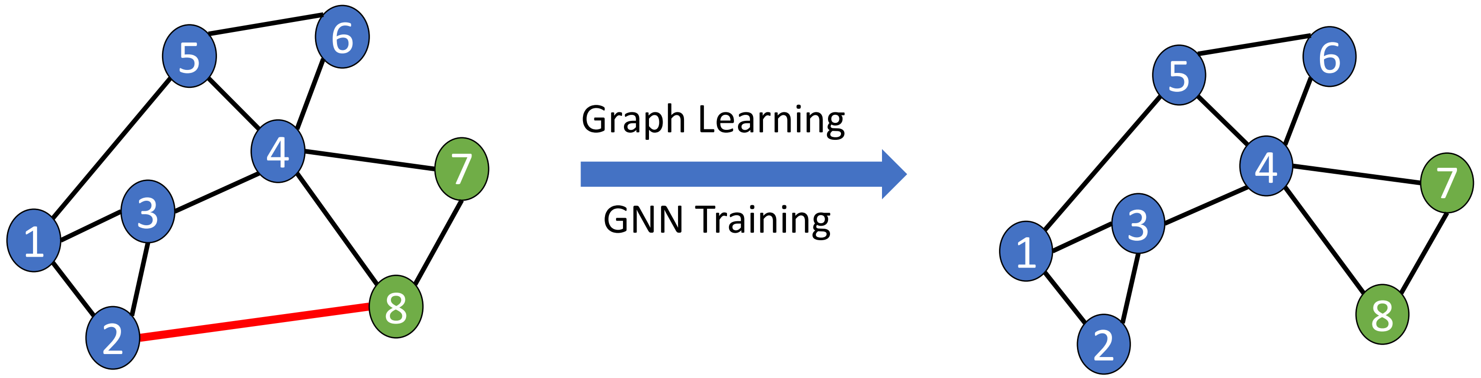
Different types of Modifications for Graph Data.

Graph Defense

- ❑ Adversarial Training
- ❑ Graph Purification
- ❑ Attention Mechanism

Graph Purification: Pro-GNN

- Recover clean graph with graph Properties: Low-rank, Sparsity, Feature smoothness



- Training GNN with purified graph.

Graph Structure Learning for Robust Graph Neural Networks. KDD 2020.

Text Attack

- ❑ What is different from Image?
  - Discrete Input
  - Perceivable Modification
  - Change of Semantic Meaning
- ❑ Different types of Modification for Text Data.
  - Character level/ word level/ sentence level.

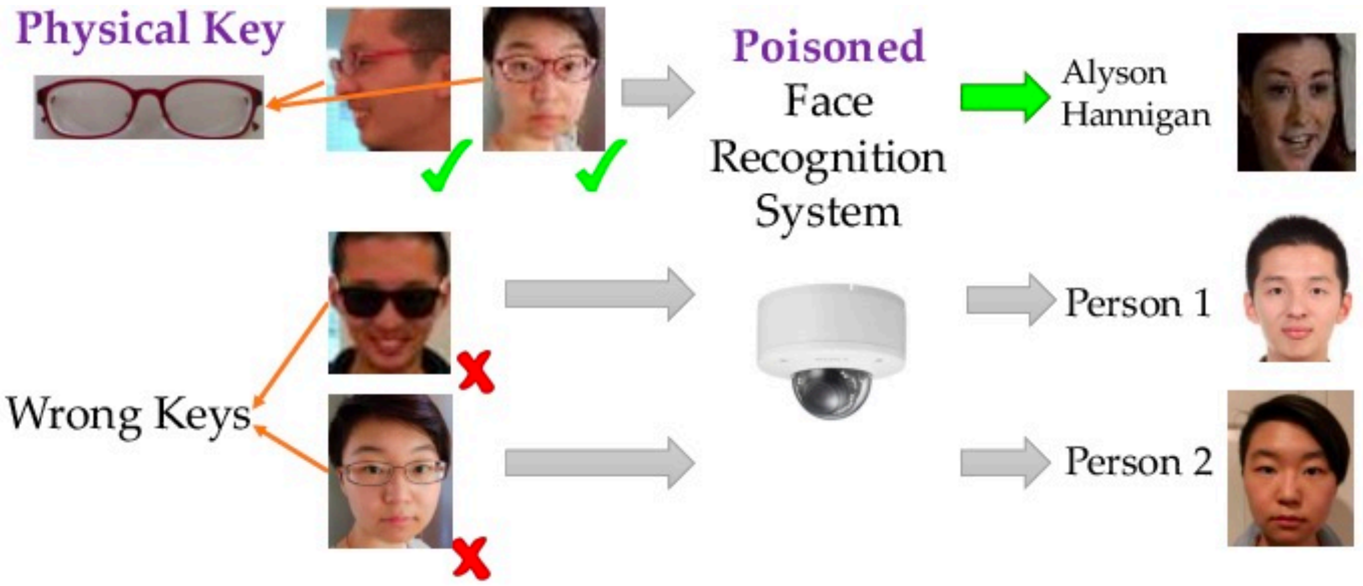
Text Defense

- Adversarial Training
  - Data Argumentation
  - Model Regularization

Outline

- ❑Concepts and Taxonomy
- ❑Adversarial Attack for Machine Learning
- ❑Adversarial Defense for Machine Learning
- ❑Robustness in Graph and Text Domains
- ❑Real World Adversarial Attack
- ❑Adversarial Learning Surveys and Tools
- ❑Future Directions

Backdoor Attack for Face Recognition



Backdoor Attack for Face Recognition System

Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning

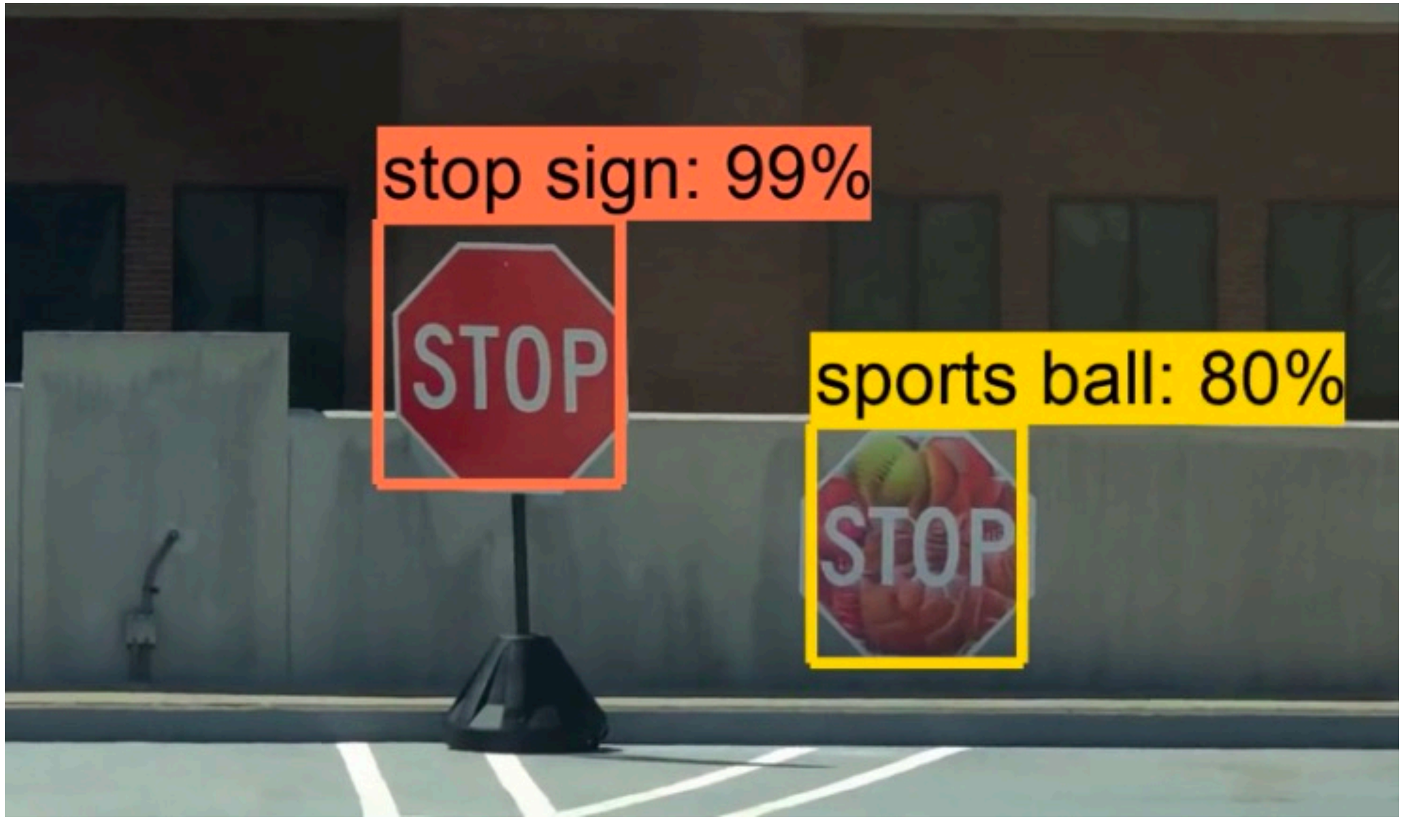
Adversarial T-shirt

Adversarial t-shirt

Adversarial T-shirt! Evading Person Detectors in A Physical World

Adversarial mug

Stop Sign



Outline

- ❑Concepts and Taxonomy
- ❑Adversarial Attack for Machine Learning
- ❑Adversarial Defense for Machine Learning
- ❑Robustness in Other Domain
- ❑Real World Adversarial Attack
- ❑Adversarial Learning Surveys and Tools
- ❑Future Directions

Adversarial Learning Surveys

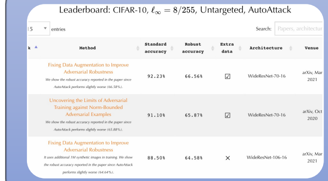
- Chakraborty, Anirban, et al. "Adversarial attacks and defences: A survey." *arXiv preprint arXiv:1810.00069* (2018).
- Xu, Han, et al. "Adversarial attacks and defenses in images, graphs and text: A review." *International Journal of Automation and Computing* 17.2 (2020): 151-178.
- Akhtar, Naveed, and Ajmal Mian. "Threat of adversarial attacks on deep learning in computer vision: A survey." *Ieee Access* 6 (2018): 14410-14430
- Jin, Wei, et al. "Adversarial Attacks and Defenses on Graphs: A Review, A Tool and Empirical Studies." *arXiv preprint arXiv:2003.00653* (2020).
- Zhang, Wei Emma, et al. "Adversarial attacks on deep-learning models in natural language processing: A survey." *ACM Transactions on Intelligent Systems and Technology (TIST)* 11.3 (2020): 1-41.

Adversarial Learning Tools

Cleverhans
- https://github.com/cleverhans-lab/cleverhans

DeepRobust
- https://github.com/DSE-MSU/DeepRobust

Advertorch
- https://github.com/BorealisAI/advertorch



RobustBench
- https://github.com/RobustBench/robustbench

DeepRobust: A PyTorch Library for Adversarial Attacks and Defenses

RobustBench: a standardized adversarial robustness benchmark

Outline

- ❑Concepts and Taxonomy
- ❑Adversarial Attack for Machine Learning
- ❑Adversarial Defense for Machine Learning
- ❑Robustness in Other Domain
- ❑Real World Adversarial Attack
- ❑Adversarial Learning Surveys and Tools
- ❑Future Directions

Future Directions

- Unsatisfied robust performance of adversarial training
- Robust generalization gap
- Adversarial robustness under multiple types of attack
- Adversarial attack on large scale datasets
- Fairness issue under adversarial attack
- More efficient provable defense
- …