Adversarial Examples

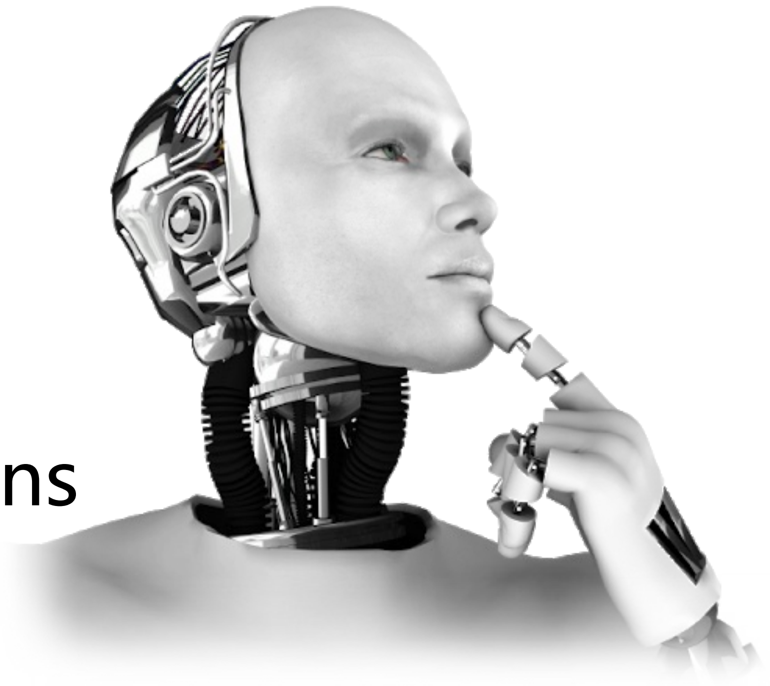
马兴军，复旦大学 计算机学院

Recap: week 2

1. Deep Neural Networks

2. Explainable Machine Learning

    - Principles and Methodologies
    - Learning Dynamics
    - The Learned Model
    - Inference
    - Generalization
    - Robustness to Common Corruptions



This Week

1. Adversarial Examples

2. Adversarial Attacks

3. Adversarial Vulnerability Understanding

Machine Learning Is Everywhere

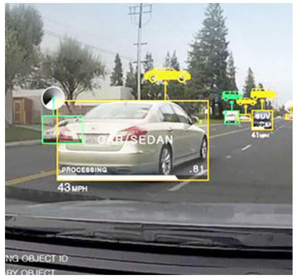
Security and Defense

Financial System

IoT

Medicine and Biology

Machine Learning



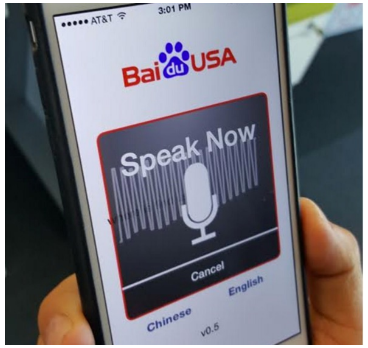
Autonomous Vehicle

Critical Infrastructure

Media and Entertainment

Beat Humans on Many Tasks

Speech Recognition



Baidu Deep Speech 2:

- End-to-end Deep Learning for English and Mandarin Speech Recognition
- English and Mandarin speech recognition Transition from English to Mandarin made simpler by end-to-end DL
- No feature engineering or Mandarin-specifics required
- More accurate than humans

Error rate 3.7% vs. 4% for human tests

http://svail.github.io/mandarin/
https://arxiv.org/pdf/1512.02595.pdf

Outperform Human on Many Tasks

Strategic Games

AlphaGo:

- First Computer Program to Beat a Human Go Professional
- Training DNNs: 3 weeks, 340 million training steps on 50 GPUs
- Play: Asynchronous multi-threaded search
- Simulations on CPUs, policy and value DNNs in parallel on GPUs
- Single machine: 40 search threads, 48 CPUs, and 8 GPUs
- Distributed version: 40 search threads, 1202 CPUs and 176 GPUs
- Outcome: Beat both European and World Go champions in best of 5 matches

http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html

Outperform Human on Many Tasks

Strategic Games

AlphaGo:

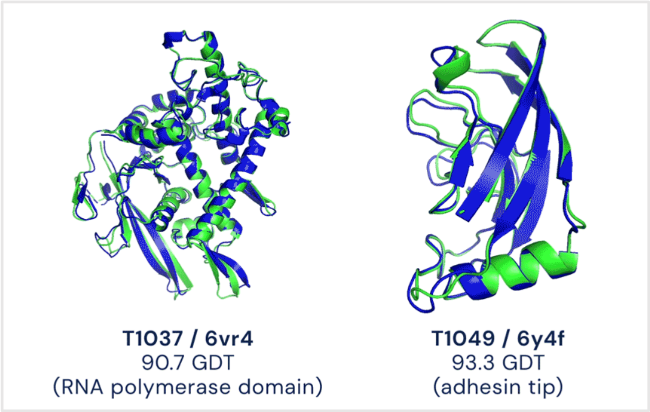
- First Computer Program to Beat a Human
  Go Professional
- Training DNNs: 3 weeks, 340 million training steps on 50 GPUs
- Play: Asynchronous multi-threaded search
- Simulations on CPUs, policy and value DNNs in parallel on GPUs
- Single machine: 40 search threads, 48 CPUs, and 8 GPUs
- Distributed version: 40 search threads, 1202 CPUs and 176 GPUs
- Outcome: Beat both European and World Go champions in best of 5 matches

http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html

Outperform Human on Many Tasks

Large-scale Image Recognition

DALL·E 2



AlphaFold V2

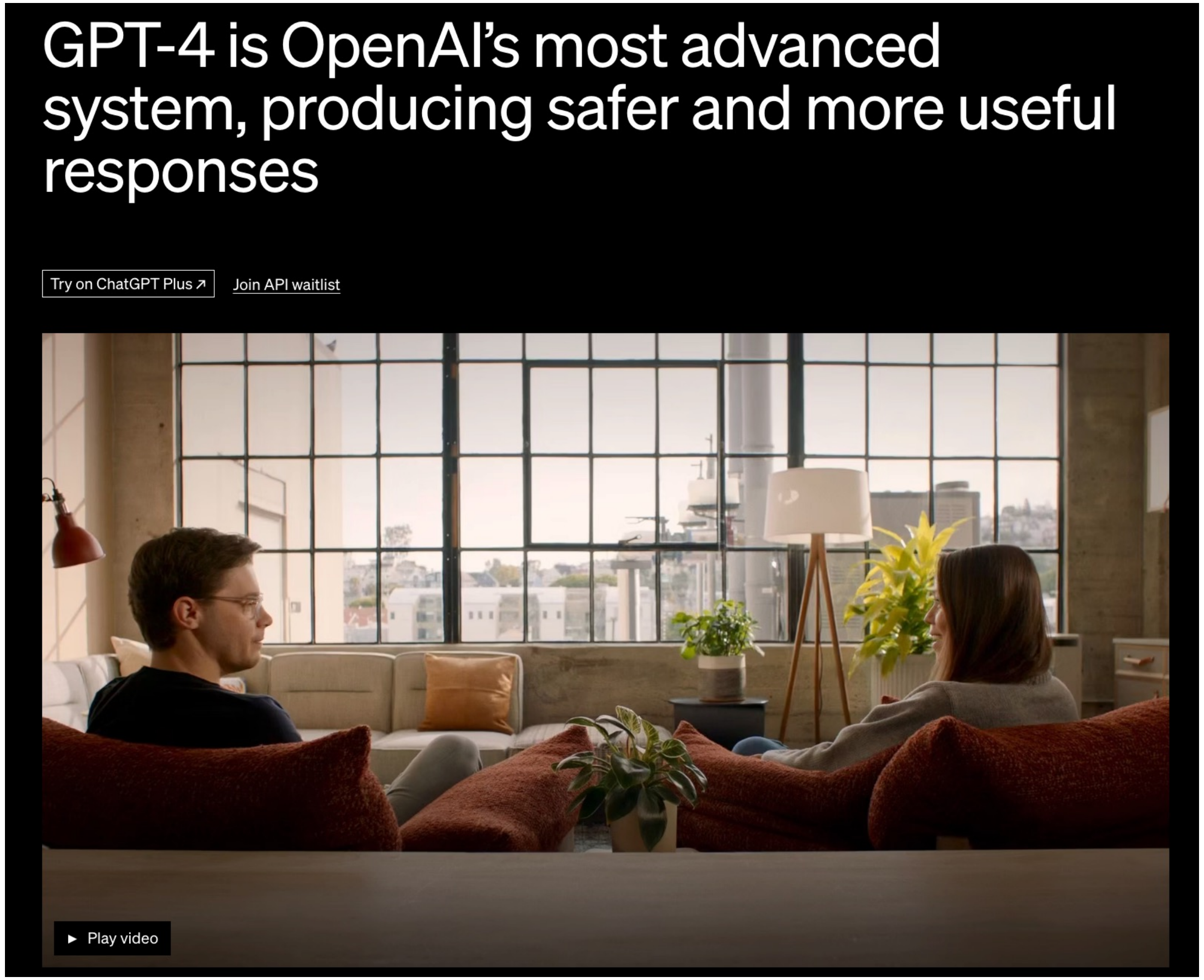
Large Language Model (LLM): ChatGPT

OpenAI在2022年11月发布的对话大模型，可以高质量的完成问答、推理、运算、推导、写作、代码调试等功能。

参数量： 1750亿
基础模型：GPT-3.5

训练数据：互联网页（31亿网页内容≈3000亿单词≈320TB文字）、维基百科（11G）、电子书籍（21G）、Reddit（50G）、人工回答等

Large Multimodel Model: GPT-4



OpenAI在2023年3月发布的多模态对话大模型，能够接受图像和文本输入，并输出文本，具有超出ChatGPT的图文理解能力、运算能力、代码生成能力、以及很多专业考试能力。

参数量： 1万亿
基础模型：GPT-4

训练数据： 在GPT-3.5 、ChatGPT基础之上增加了多模态数据、更多的人工标注数据等等

Outperform Human on Many Tasks

Image Recognition

GoogLeNet: http://cs.stanford.edu/people/karpathy/ilsvrc/

Labrapoodle or Fried chicken

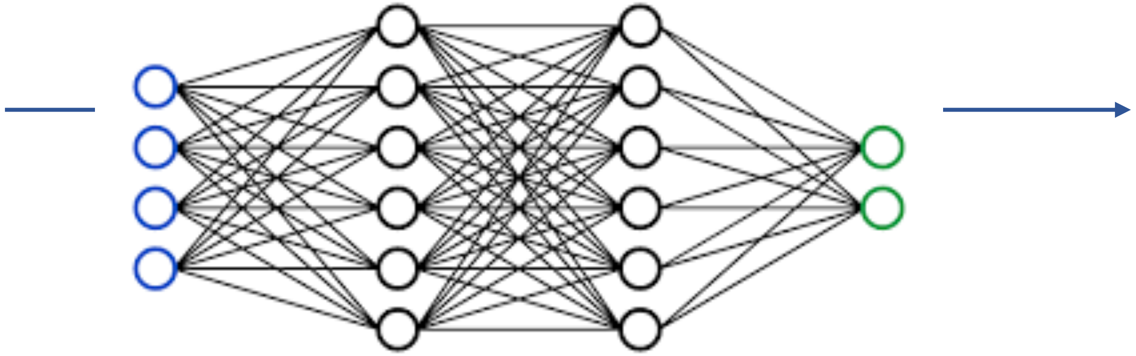
Sheepdog or Mop

Barn owl or Apple

Parrot or Guacamole

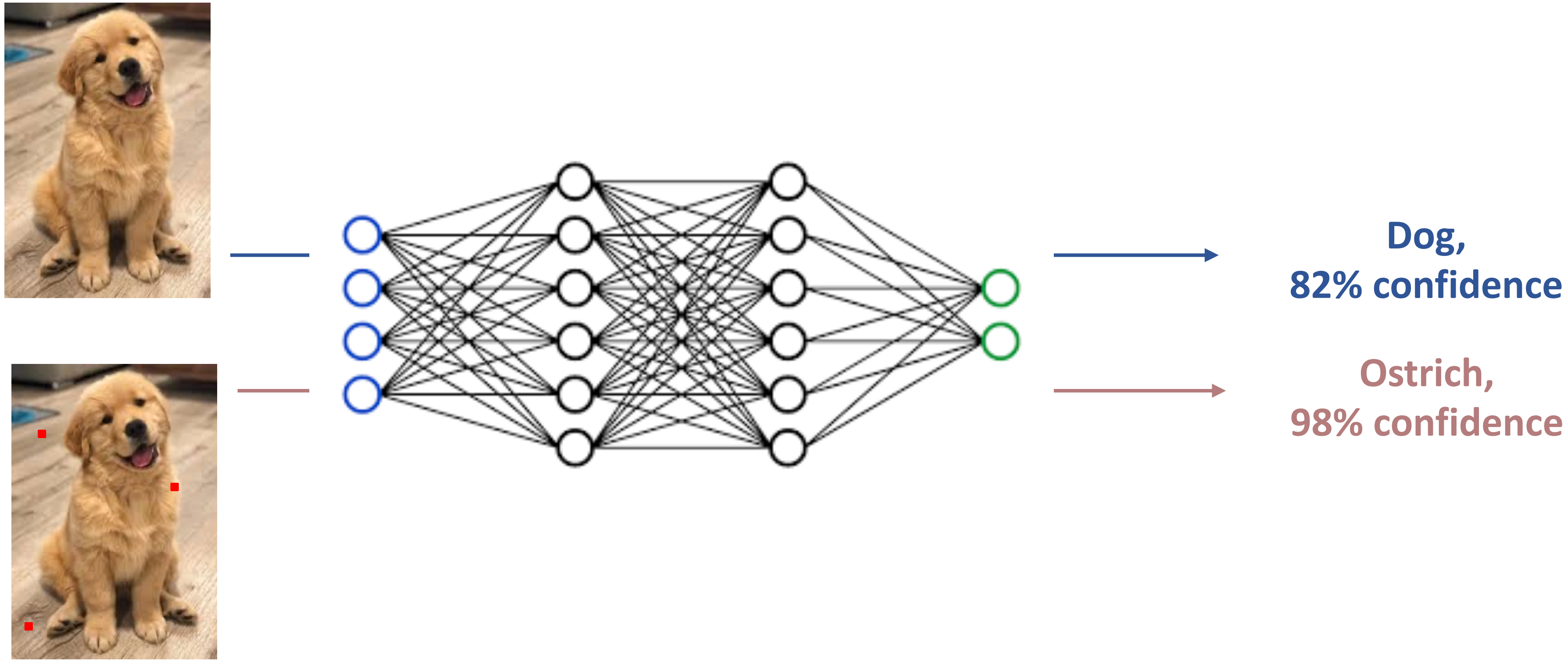
Raw chicken or Donald Trump

Vulnerabilities of DNNs

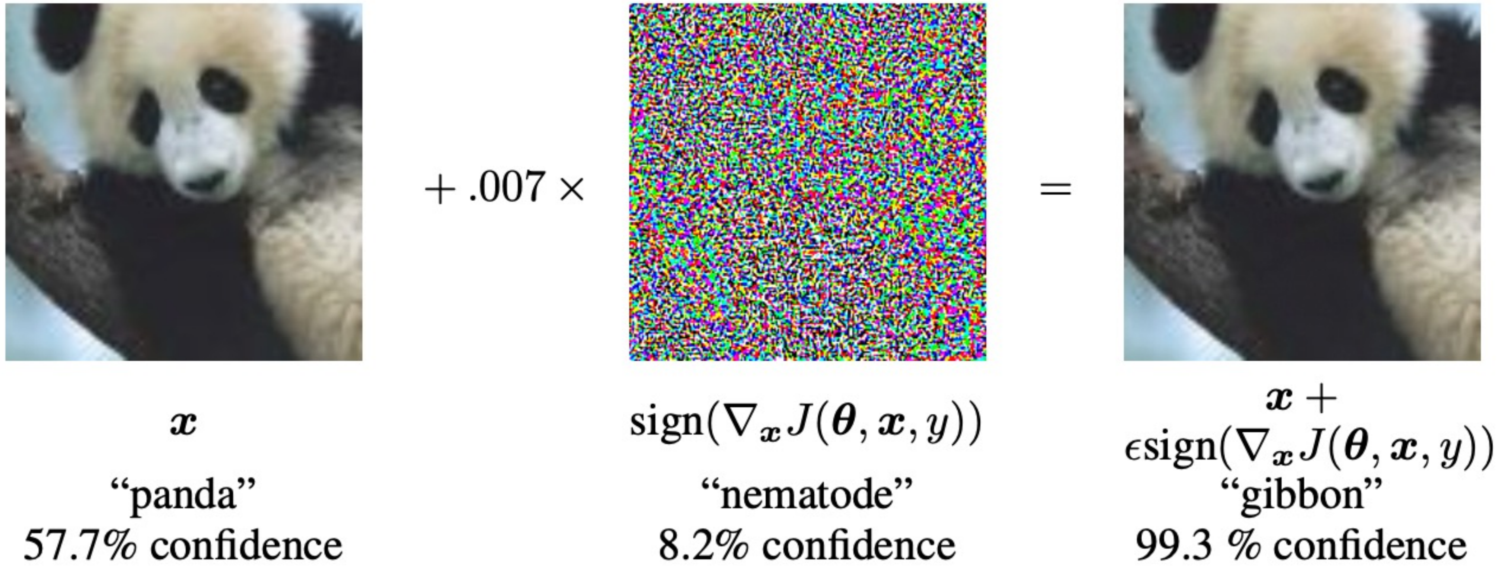


Dog,
82% confidence

Vulnerabilities of DNNs



Adversarial Examples



Small perturbations can fool DNNs

Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. ICLR 2014.
Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. ICLR 2015.

Adversarial Attack

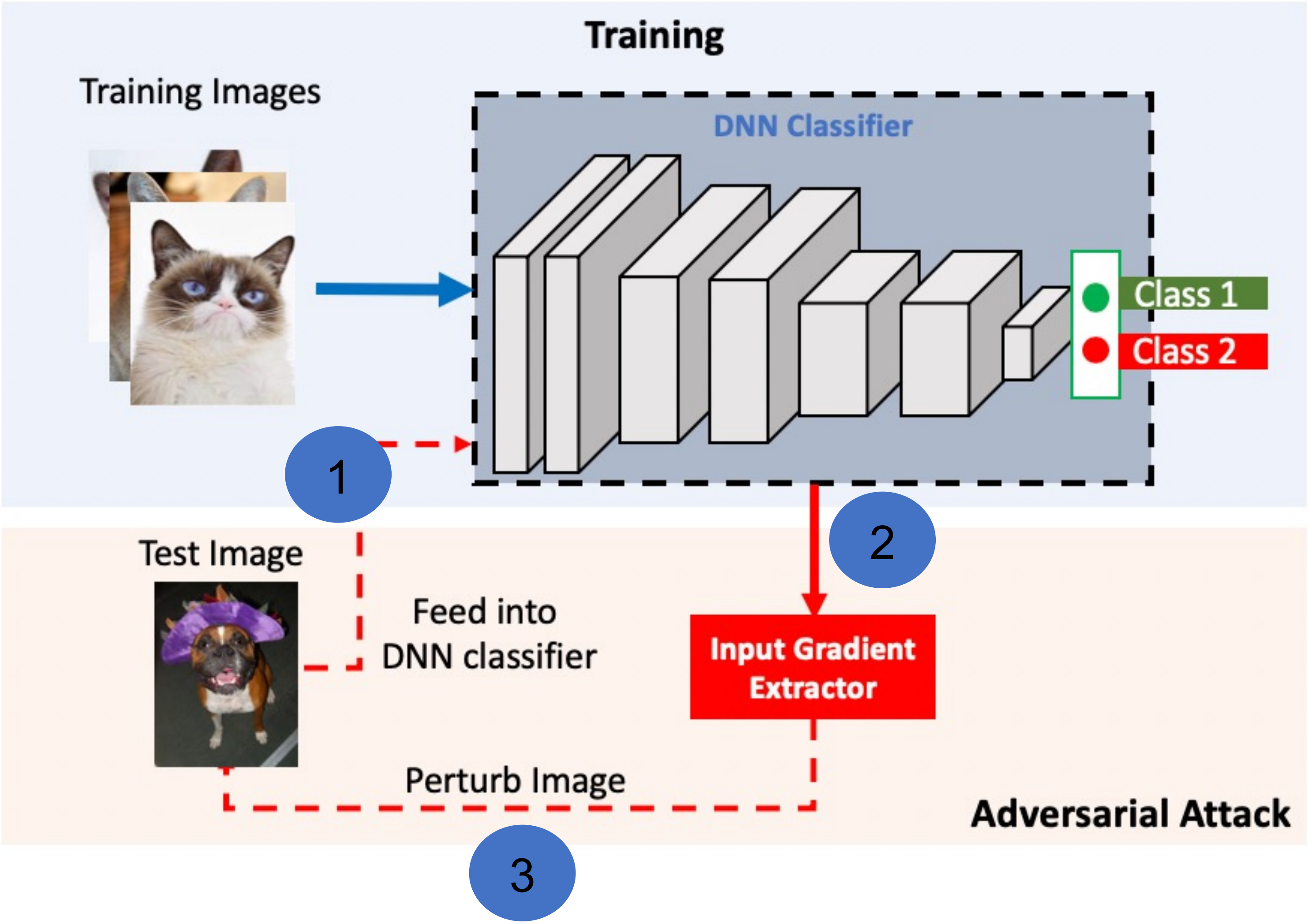
DNN Training: $\min_{\theta} \sum_{(x_i,\, y_i) \in D_{train}} L(f_\theta(x_i), y_i)$

Adversarial Attack: $\underbrace{\max_{x'} L(f_\theta(x'), y)}_{\text{Misclassification}}$ $\underbrace{\text{subject to } \|x' - x\|_p \le \epsilon}_{\text{Small change on } x}$ $\underbrace{\text{for } x \in D_{test}}_{\text{test time attack}}$

Small perturbation: $\|x' - x\|_{p=1,\,2\ or\ \infty}$, for example, $\|\cdot\|_\infty \le \frac{8}{255}$

Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. ICLR 2014.
Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. ICLR 2015.

Adversarial Attack

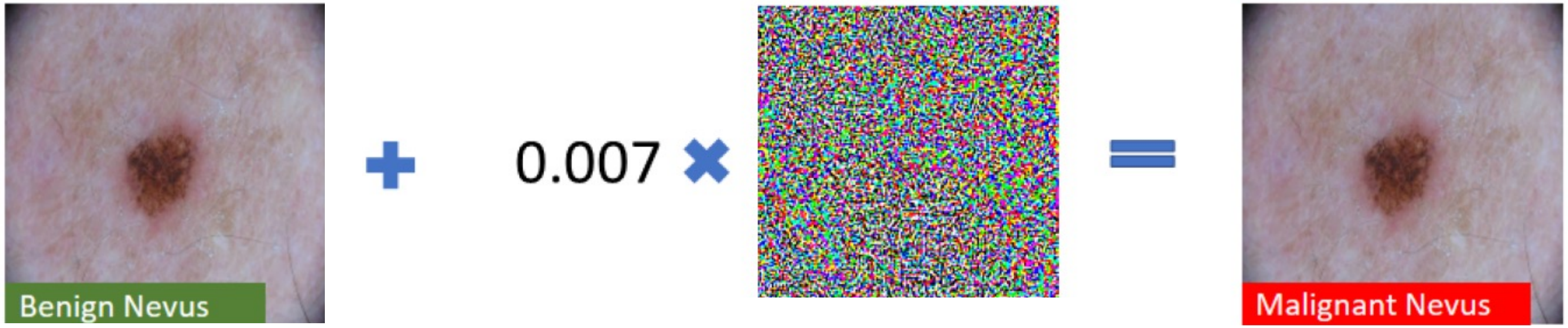


Characteristics of Adversarial Examples

Adversarial Examples

- Small
- Imperceptible
- Hidden
- Transfer
- Universal

Example Attacks



Benign Nevus, 73% confidence

Adversarial noise

Malignant Nevus, 89% confidence

- Perturbations are small, imperceptible to human eyes.
- Adversarial examples are easy to generate and transfer across models.

Ma et al., "Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems", Pattern Recognition, 2021.

Example Attacks

- **Clean video frames:** Correct Class

Bowling

ThrowDiscus

- **Adversarial video:** Wrong Class

WritingOnBoard

PlayingDaf

Jiang et al., "Black-box Adversarial Attacks on Video Recognition Models", ACMMM, 2019.

Example Attacks

Physical-world attacks against traffic signs

Stop signs recognized as 45km speed limit

Science Museum at London

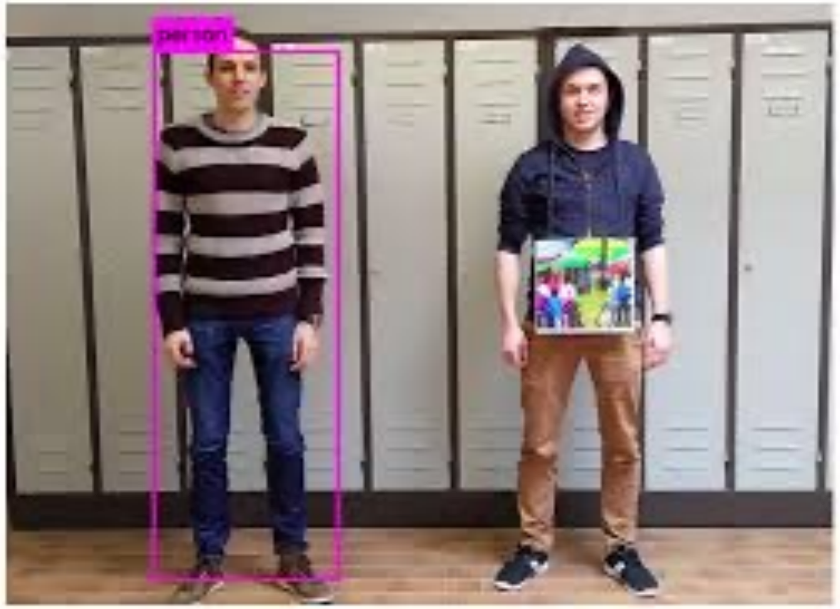
Eykholt, Kevin, et al. “Robust physical-world attacks on deep learning visual classification.” *CVPR,* 2018.

Example Attacks

3D printed turtle recognized as a rifle from any angle

Athalye, Anish, et al. "Synthesizing robust adversarial examples." *ICML*, 2018.

Example Attacks



Adversarial patch makes people invisible to object detection (YOLO)

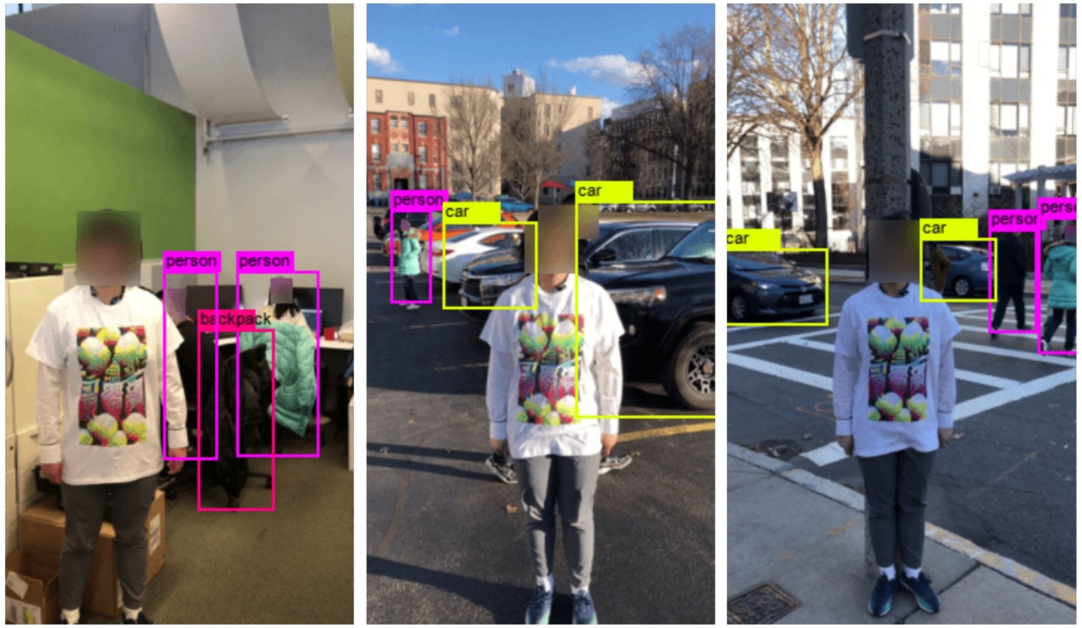
Brown, Tom B., et al. "Adversarial patch." *arXiv preprint arXiv:1712.09665* (2017).

Example Attacks

https://cvdazzle.com/

Adversarial attack or new fashion?

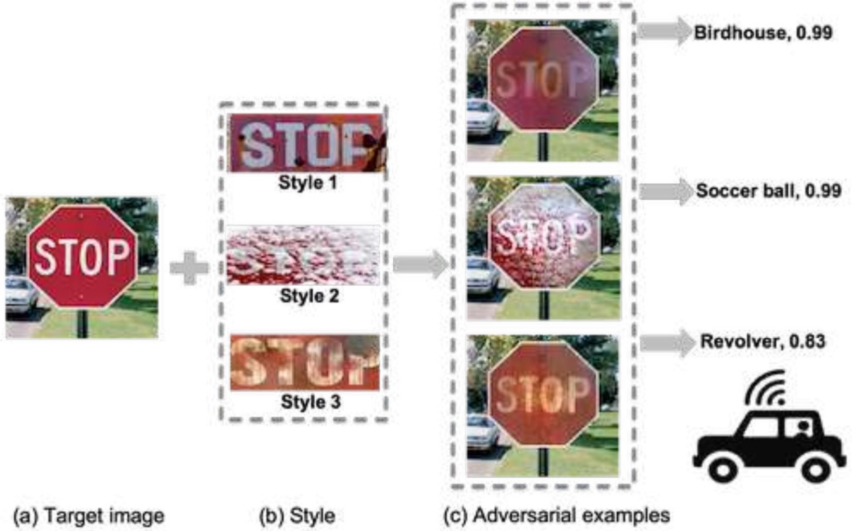
Example Attacks



Adversarial t-shirt: one step closer to real-world attack

Xu, Kaidi, et al. “Adversarial t-shirt! evading person detectors in a physical world.” ECCV, 2020.

Example Attacks



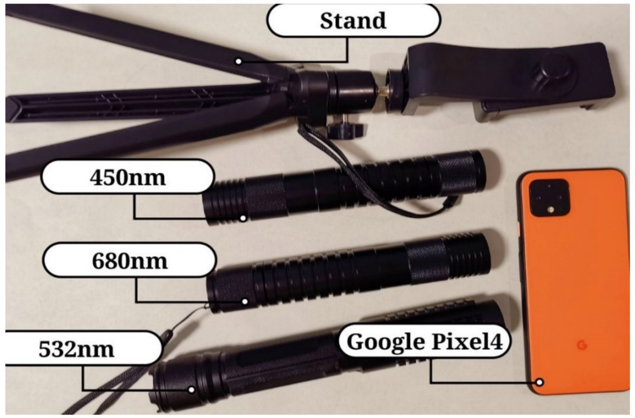
Tree bark -> street sign

people+pikachu t-shirt -> dog

Camouflage adversarial patterns into realistic styles

Duan et al. Adversarial Camouflage: Hiding Physical-World Attacks With Natural Styles. CVPR, 2020.

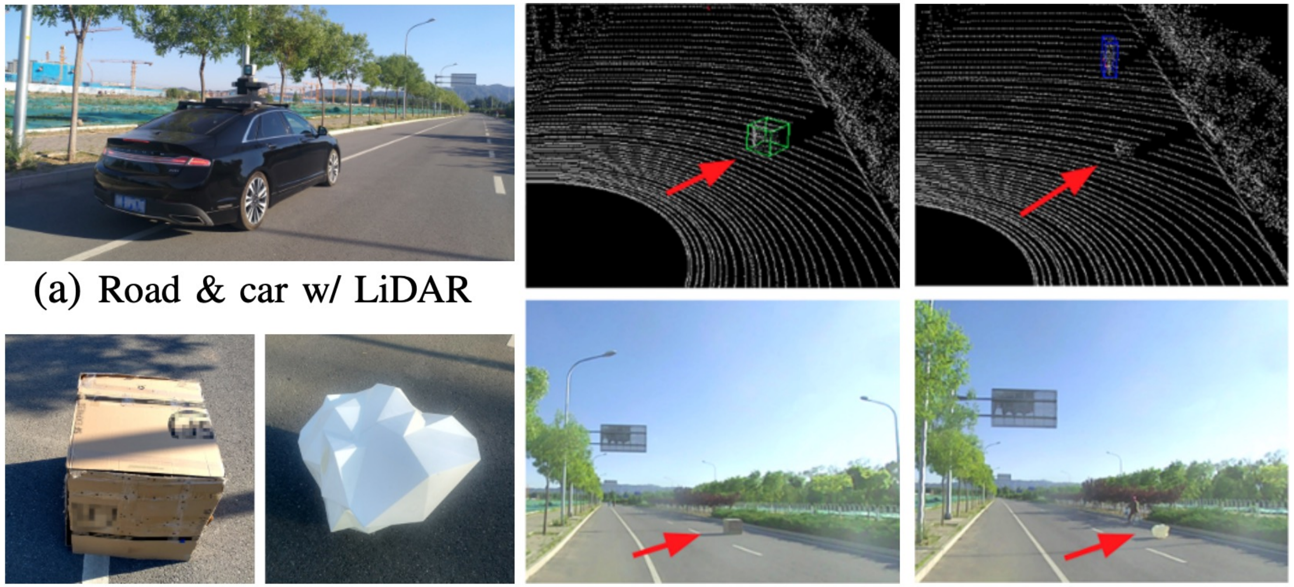
Example Attacks



Night scene adversarial attack with laser pointer

Duan, Ranjie, et al. "Adversarial laser beam: Effective physical-world attack to dnns in a blink." *CVPR,* 2021

Example Attacks



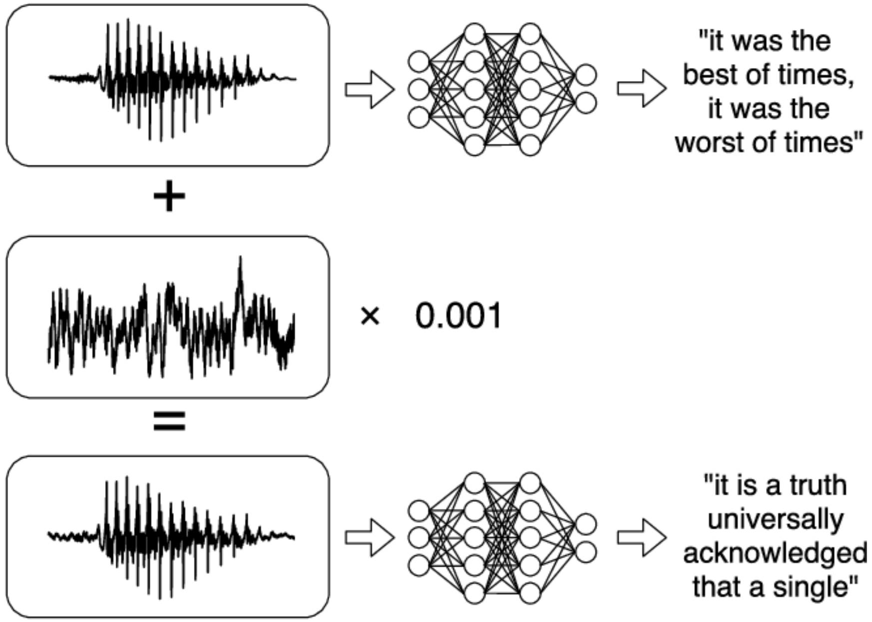
(a) Road & car w/ LiDAR

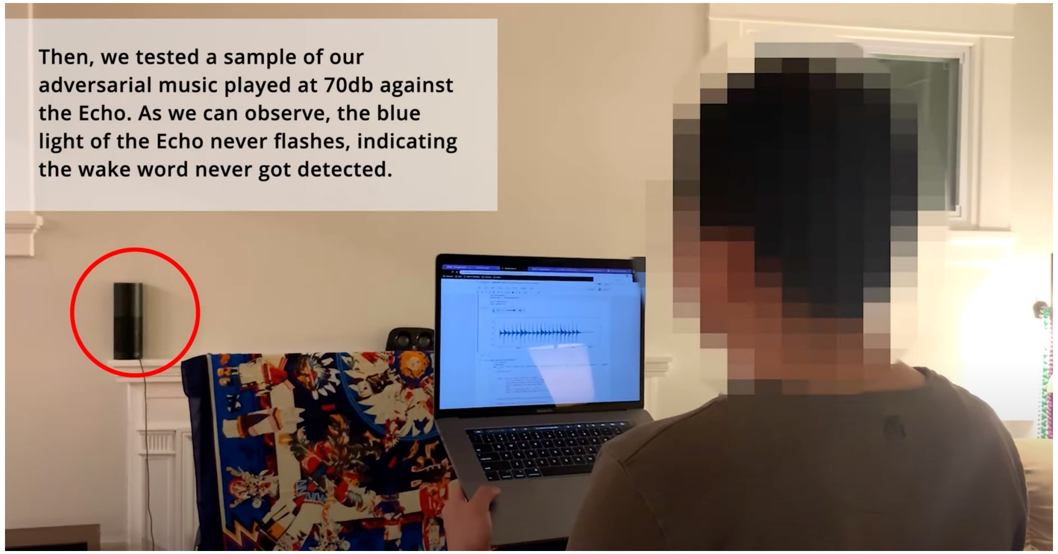
(b) Benign and adv. cubes (c) Benign case (d) Adversarial case

Attacking both camera and lidar using adversarial objects

Cao, Yulong, et al. "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks." *S&P*, 2021.

Example Attacks





Attacking speech/command recognition models

Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." *S&PW*, 2018.
https://nicholas.carlini.com/code/audio_adversarial_examples/
Adversarial Music: Real world Audio Adversary against Wake-word Detection System
https://www.youtube.com/watch?v=r4XXGDVs0f8

Example Attacks

- Q&A Adversaries

Original: What is the oncorhynchus also called? **A:** chum salmon	**Original:** How long is the Rhine? **A:** 1,230 km
Changed: What's the oncorhynchus also called? **A:** keta	**Changed:** How long is the Rhine?? **A:** more than 1,050,000

Ribeiro et al. "Semantically equivalent adversarial rules for debugging NLP models." *ACL,* 2018.

Threats to AI Applications

- **Transportation industry**
  - *Trick autonomous vehicles into misinterpreting stop signs or speed limit*
- **Cybersecurity industry**
  - *Bypass AI-based malware detection tools*
- **Medical industry**
  - *Forge medical condition*
- **Smart Home industry**
  - *Fool voice commands*
- **Financial Industry**
  - *Trick anomaly and fraud detection engines*

Definition of Adversarial Examples

- No standard community-accepted definition

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

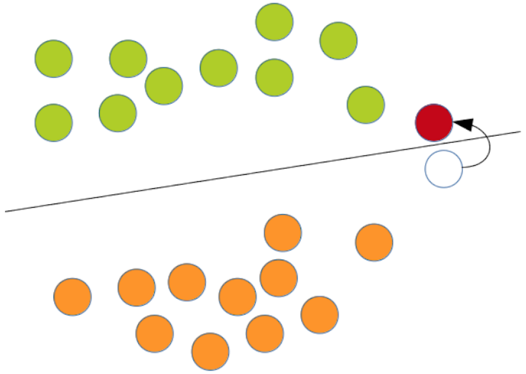
Goodfellow, Ian. "Defense against the dark arts: An overview of adversarial example security research and future research directions." *arXiv:1806.04169* (2018).

Taxonomy of Attacks

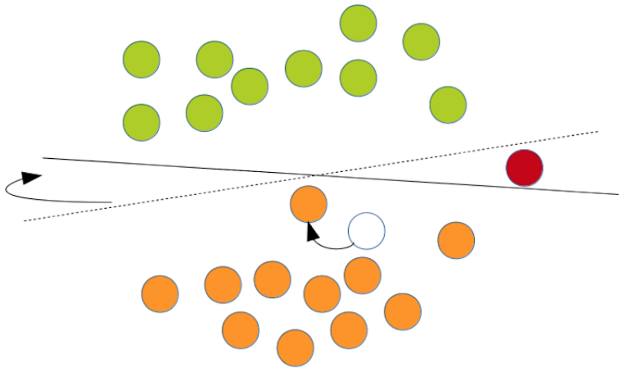
- Attack timing
  - Poisoning attack
  - Evasion attack
- Attacker's goal
  - Targeted attack
  - Untargeted attack
- Attacker's knowledge
  - Black-box
  - White-box
  - Gray-box
- Universality
  - Individual
  - Universal

Attack Timing

- **Evasion (Causation) attack**
    - Test time attack
    - Change input example

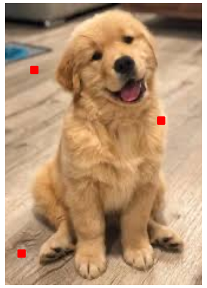


- **Poisoning attack**
    - Training time attack
    - Change classification boundary



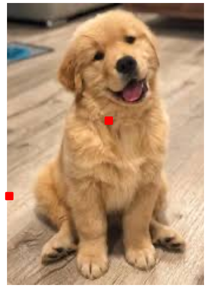
Attacker's Goal

- **Targeted attack**
  - Cause an input to be recognized as coming from a specific class



Ostrich

- **Untargeted attack**
  - Cause an input to be recognized as any incorrect class



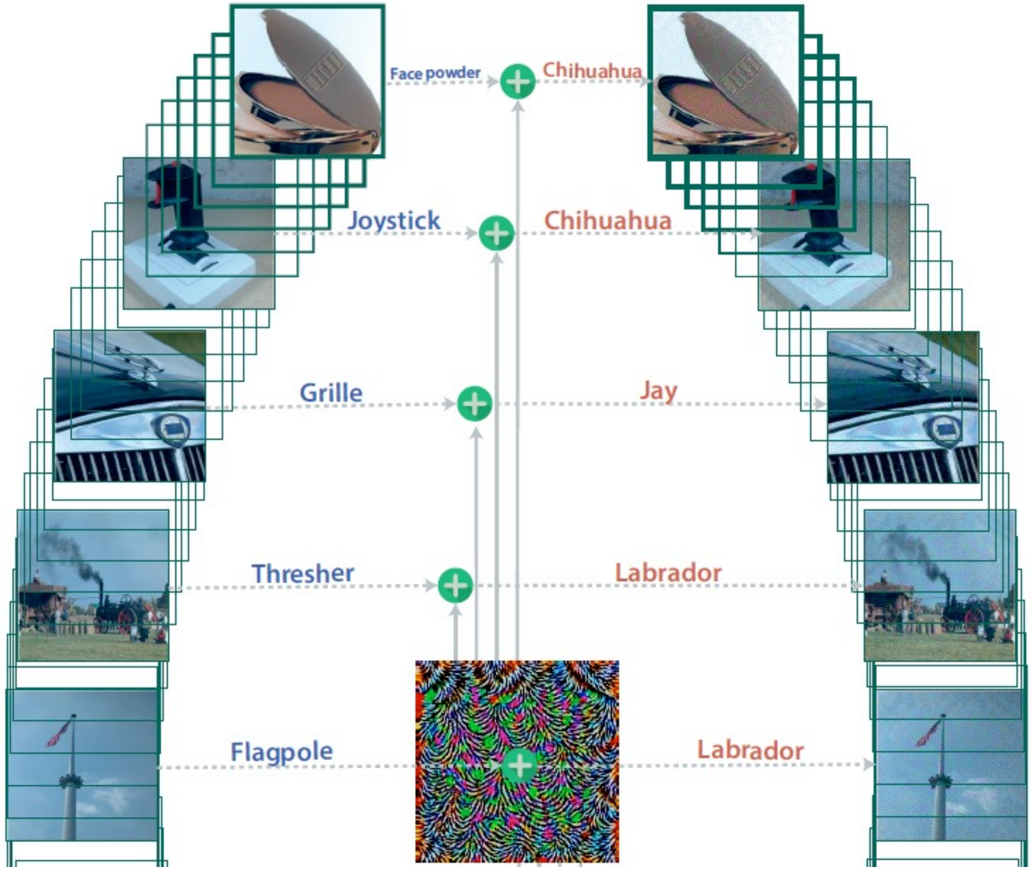
Any class,
except dog

Adversary's Knowledge

- **White-box attack**:
  - Attacker has full access to the model, including model type, model architecture, values of parameters and training weights
- **Black-box attack**:
  - Attacker has no knowledge about the model under attack
  - Rely on transferability of adversarial examples
- **Gray-box attack** (Semi-black-box attack)
  - Attacker may know some hyperparameters like model architecture

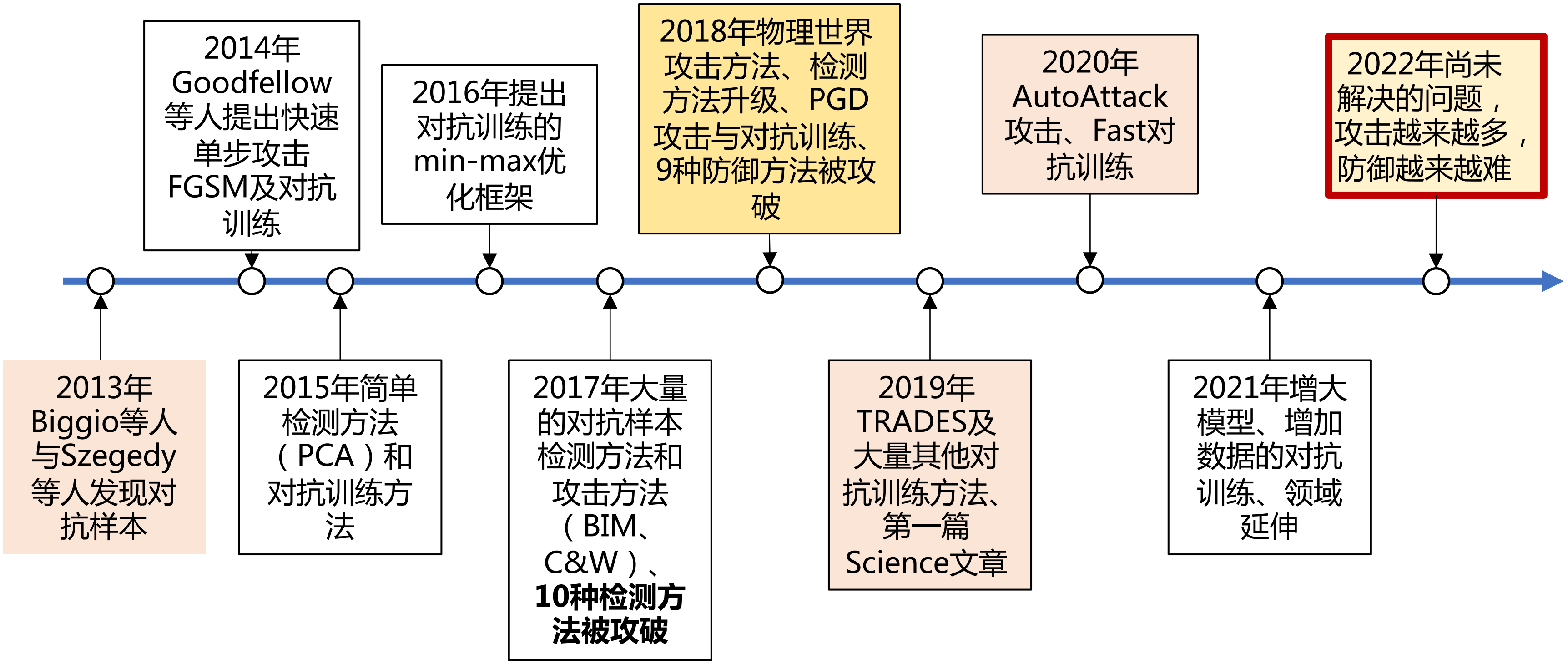
Universality

- **Individual attack**
  - Generate different perturbations for each clean input
- **Universal attack**
  - Only create a universal perturbation for the whole dataset. Make it easier to deploy adversary examples.



Moosavi-Dezfooli, Seyed-Mohsen, et al. "Universal adversarial perturbations." *CVPR* 2017.

A Brief History of Adversarial Machine Learning



Biggio et al. “Evasion attacks against machine learning at test time.”; Szegedy, Christian, et al. "Intriguing properties of neural networks."

White-box Attacks

- **单步攻击**：Fast Gradient Sign Method (FGSM) (*Goodfellow et al. 2014*):

$$x' = x + \varepsilon \cdot \mathrm{sign}\,\nabla_x L(f_\theta(x), y)$$

- **多步攻击**：Iterative Methods (BIM, PGD), (*Kurakin et al. 2016; Madry et al. 2018*):

$$x'_{t+1} = \mathrm{project}_\epsilon(x'_t + \alpha \cdot \mathrm{sign}\,\nabla_x L(f_\theta(x'_t), y)),$$ α: step size

Projected Gradient Descent (PGD): strongest first-order attack.

- **基于优化的攻击**：C&W attack (*Carlini & Wagner 2017*): CW attack was the strongest attack

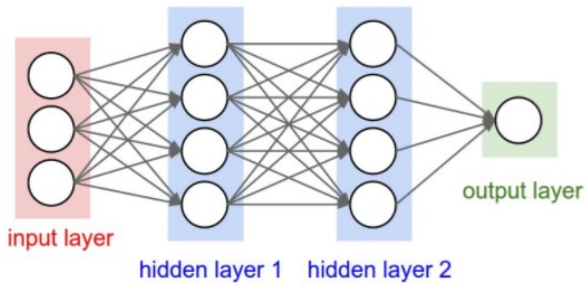
$$\min_{x'} \|x' - x\|_2^2 - c \cdot L(f_\theta(x'), y),$$ c: confidence, y: clean label

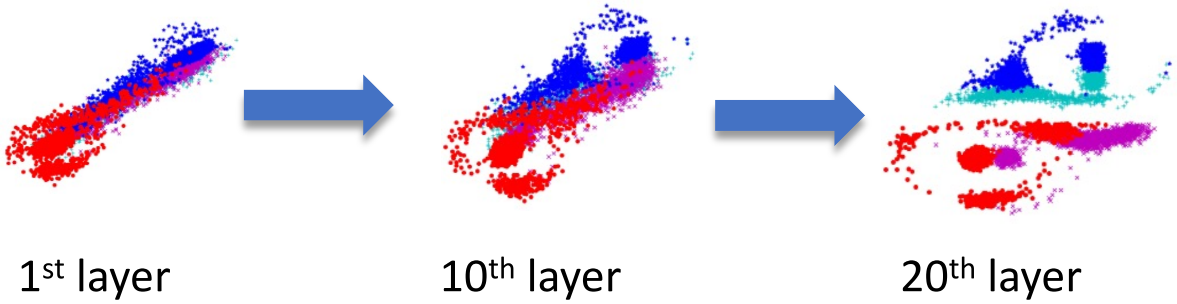
- **集成攻击**：AutoAttack (Croce *et al. 2020*): current strongest attack

Why Adversarial Examples Exist?

Non-linear Explanation

- Viewing DNN as a sequence of transformed spaces:

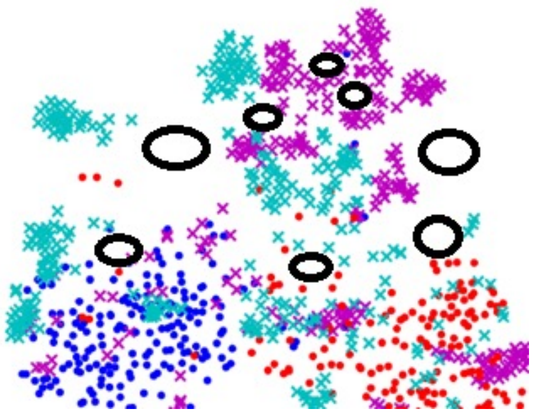




High dimensional non-linear explanation:

— Non-linear transformations leads to the existence of small "pockets" in the deep space:

- Regions of **low probability** (not naturally occurring).
- **Densely scattered** regions.
- **Continuous** regions.
- **Close** to normal data subspace.



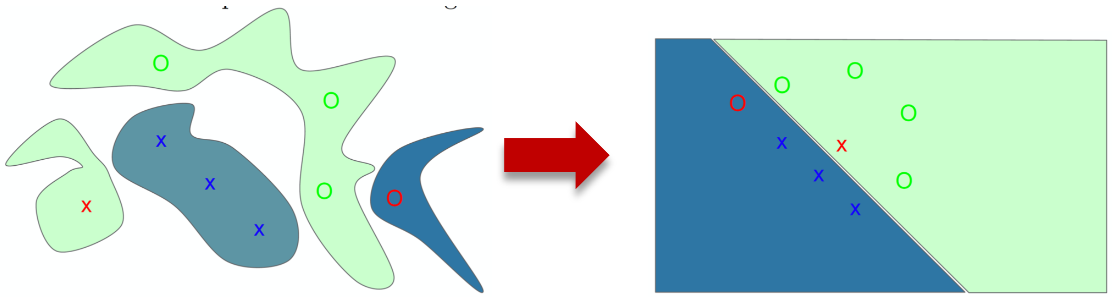
Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. ICLR 2014;
Ma et al. Characterizing Adversarial Subspace Using Local Intrinsic Dimensionality. *ICLR 2018*

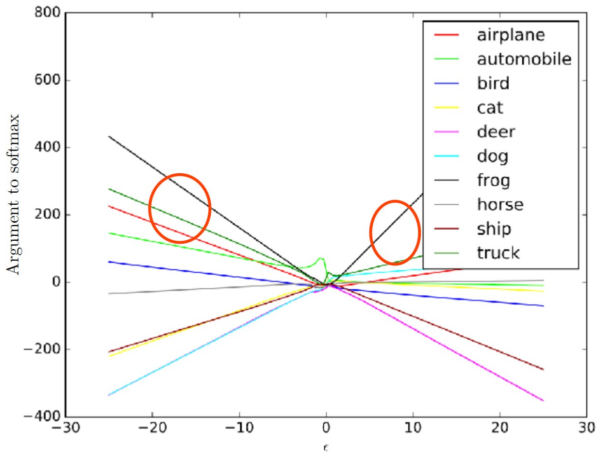
Linear Explaination

- Viewing DNN as a stack of linear operations: $\boldsymbol{w^T x + b}$

Linear explanation:

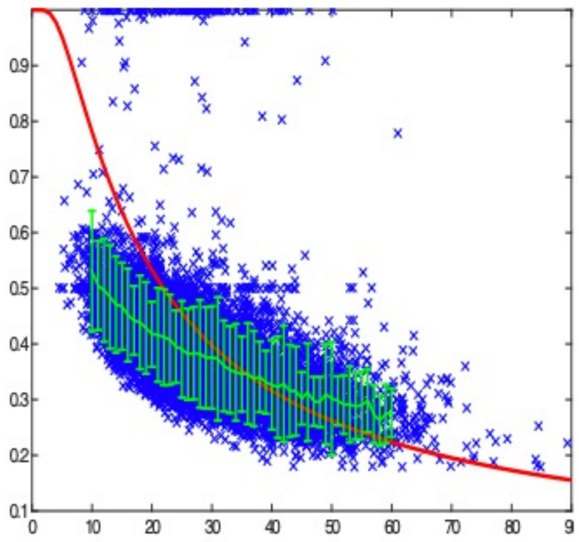
- Adversarial subspaces span a contiguous multidimensional space:
  - **Small changes at individual dimensions can sum up to significant change in final output: $\sum_{i=0}^{n} x_i + \epsilon$.**
  - Adversarial examples can always be found if ϵ is large enough.



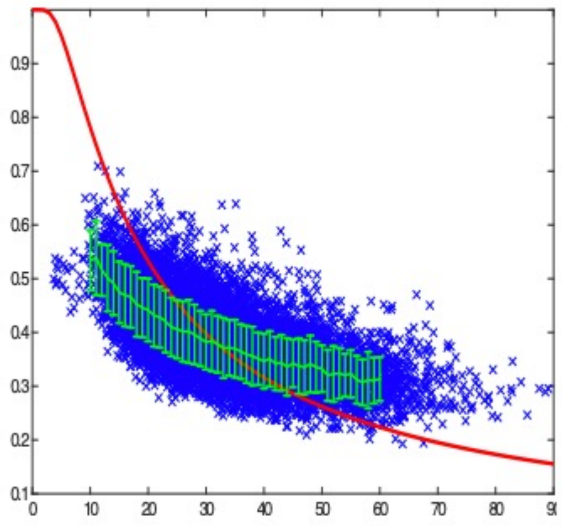


Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. ICLR 2015.

Vulnerability Increases with Intrinsic Dimensionality



ImageNet



CIFAR-10

Y-axis: the minimum adversarial noise required to subvert a KNN classifier
X-axis: LID values
Red curve: theoretical bound

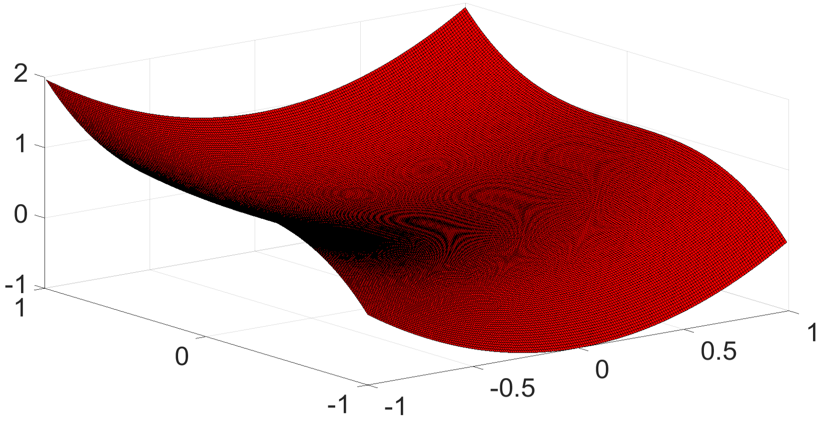
Amsaleg et al. The Vulnerability of Learning to Adversarial Perturbation Increases with Intrinsic Dimensionality. WIFS, 2017

Insufficient Training Data

- An illustrative example
  - $x \in [-1, 1), y \in [-1, 1), z \in [-1, 2)$
  - Binary classification
    - Class 1: $z < x^2 + y^3$
    - Class 2: $z \geq x^2 + y^3$
  - *x*, *y* and *z* are increased by 0.01

    → a total of 200×200×300

    = 1.2×10^7 points



- How many points are needed to reconstruct the decision boundary?
  - Training dataset: choose 80, 800, 8000, 80000 points randomly
  - Test dataset: choose 40, 400, 4000, 40000 points randomly
  - Boundary dataset (adversarial samples are likely to locate here):

    $x^2 + y^3 - 0.1 < z < x^2 + y^3 + 0.1$

Insufficient Training Data

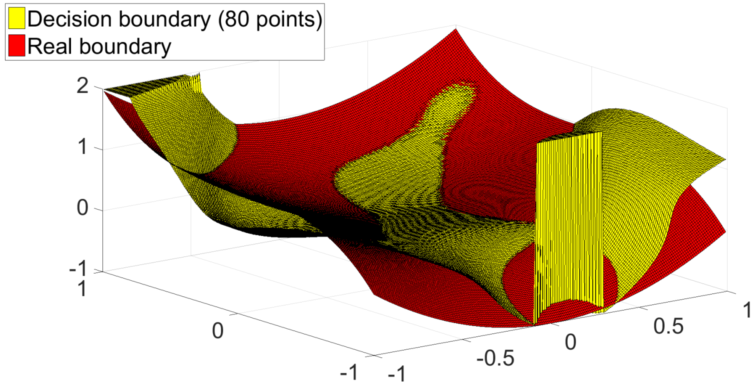
- **Test result**
  - RBF SVMs

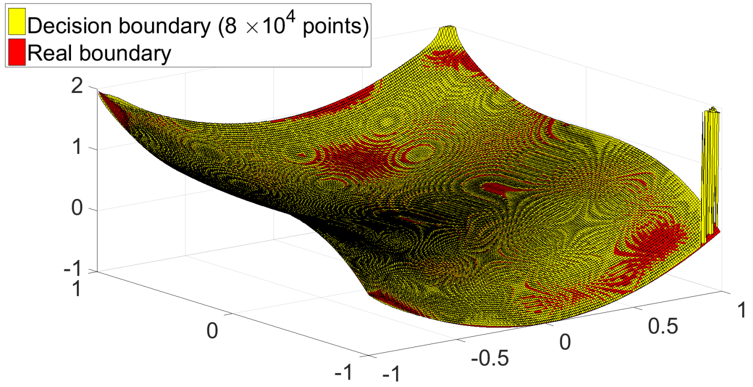
Size of the training dataset	Accuracy on its own test dataset	Accuracy on the test dataset with 4×10^4 points	Accuracy on the boundary dataset
80	100	92.7	60.8
800	99.0	97.4	74.9
8000	99.5	99.6	94.1
80000	99.9	99.9	98.9

  - Linear SVMs

Size of the training dataset	Accuracy on its own test dataset	Accuracy on the test dataset with 4×10^4 points	Accuracy on the boundary dataset
80	100	96.3	70.1
800	99.8	99.0	85.7
8000	99.9	99.8	97.3
80000	99.98	99.98	99.5

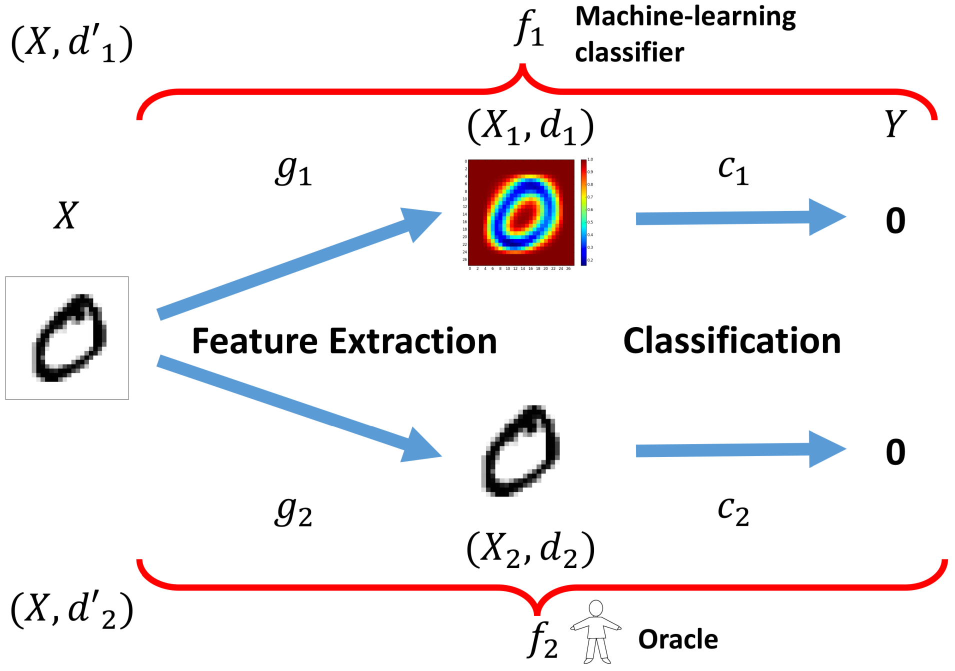
- 8000: 0.067% of 1.2×10^7
- MNIST: 28×28 8-bit greyscale images, $(2^8)^{28\times28} \approx 1.1\times10^{1888}$
- $1.1\times10^{1888}\times0.067\% \gg 6\times10^5$

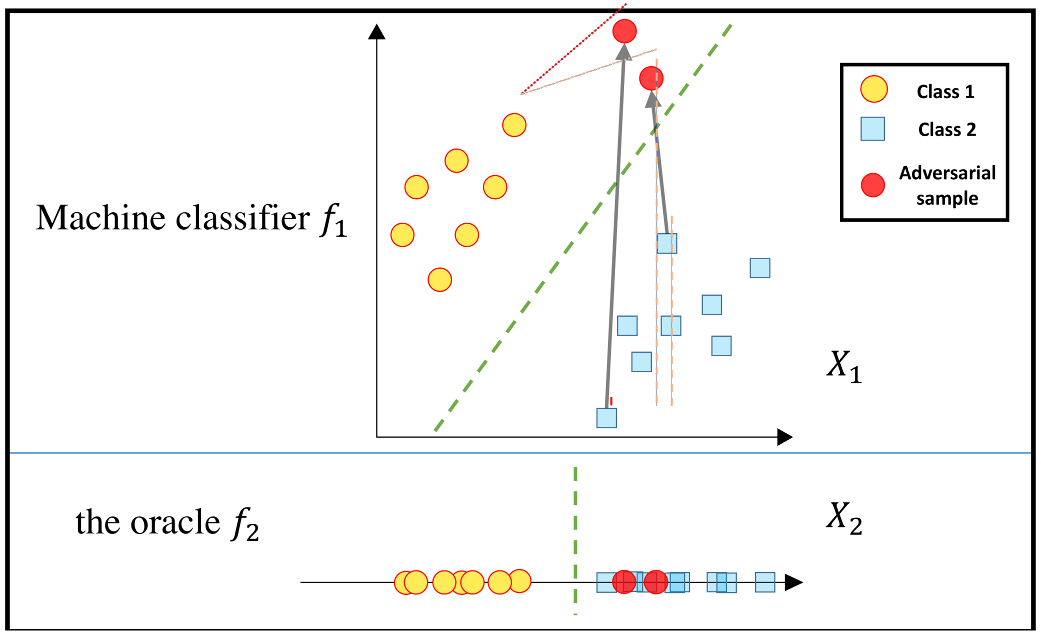




Unnecessary Features

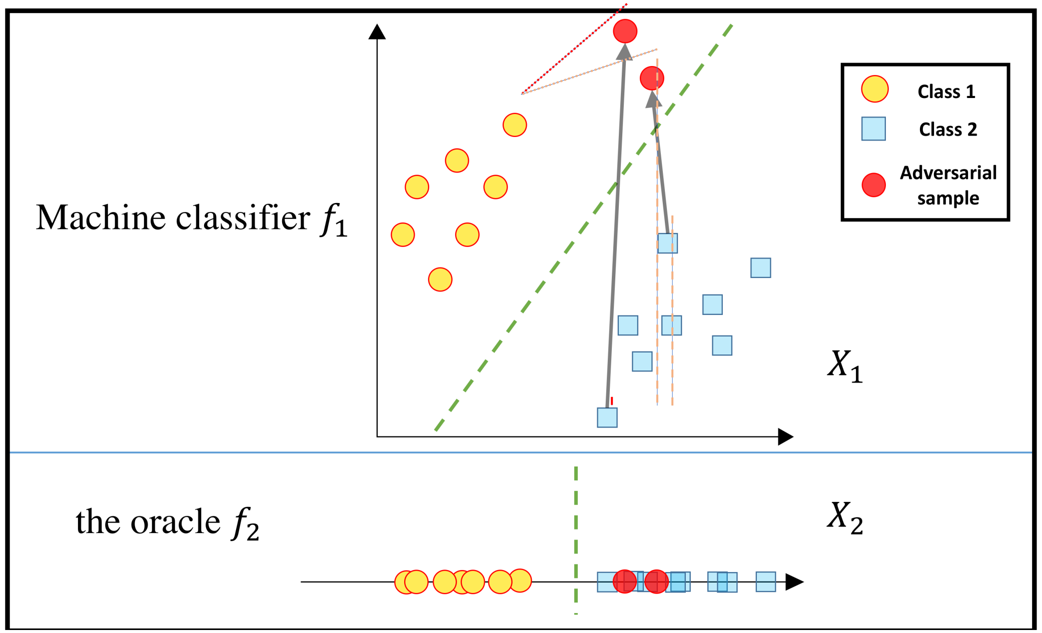
- $f = g \circ c$
- d : similarity measure
- Do machine learning models extract the same features as humans?





Wang et al. "A theoretical framework for robustness of (deep) classifiers against adversarial examples." *arXiv:1612.00334* (2016).

Unnecessary Features



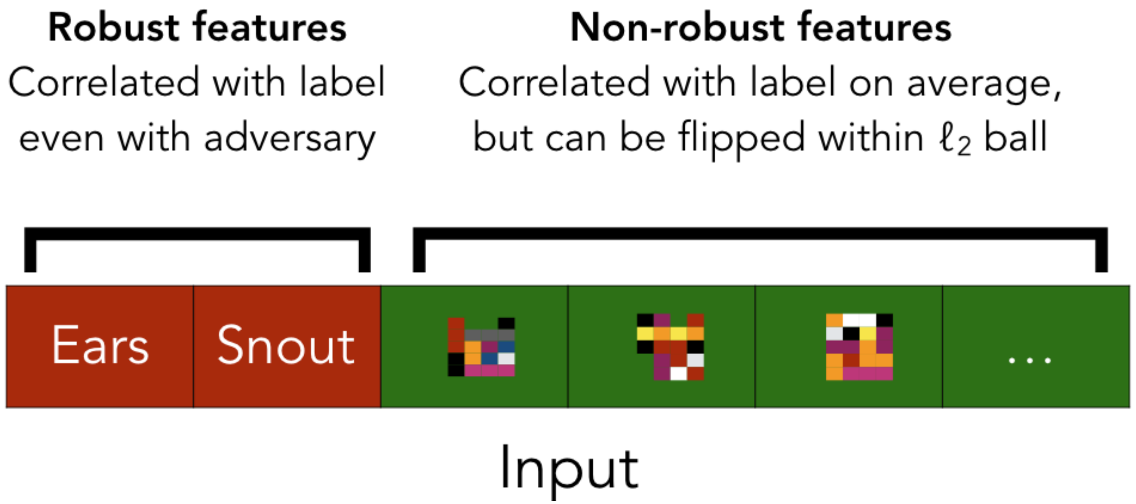
Adversarial samples can be far away from the original instance in the trained classifier's feature space, and at the other side of the boundary

Each adversarial sample is close to the original instance in the oracle feature space

- Unnecessary features ruin strong-robustness
  - If f_1 uses unnecessary features → not strong-robust
  - If f_1 misses necessary features used by f_2 → not accurate
  - If f_1 uses the same set of features as f_2 → strong-robust, can be accurate

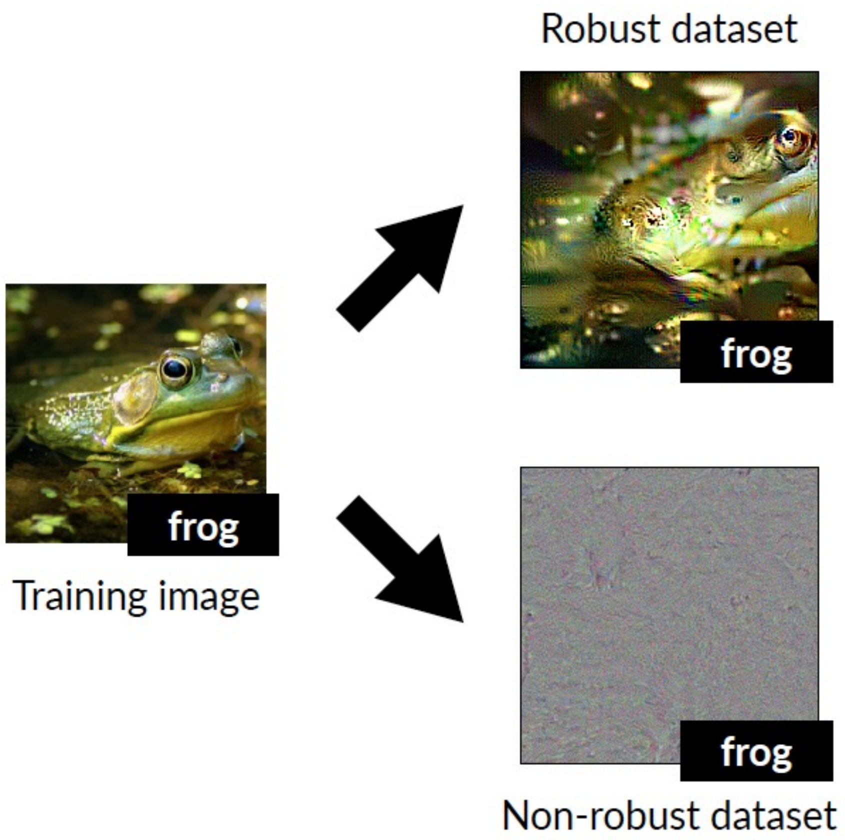
Robust vs Non-robust Features

- Predictive features of the data can be split into
  - **Robust:** Patterns that are predictive of the true label even when adversarially perturbed
  - **Non-robust:** Patterns that while predictive, can be flipped by an adversary within a pre-defined perturbation set to be indicate a wrong class.



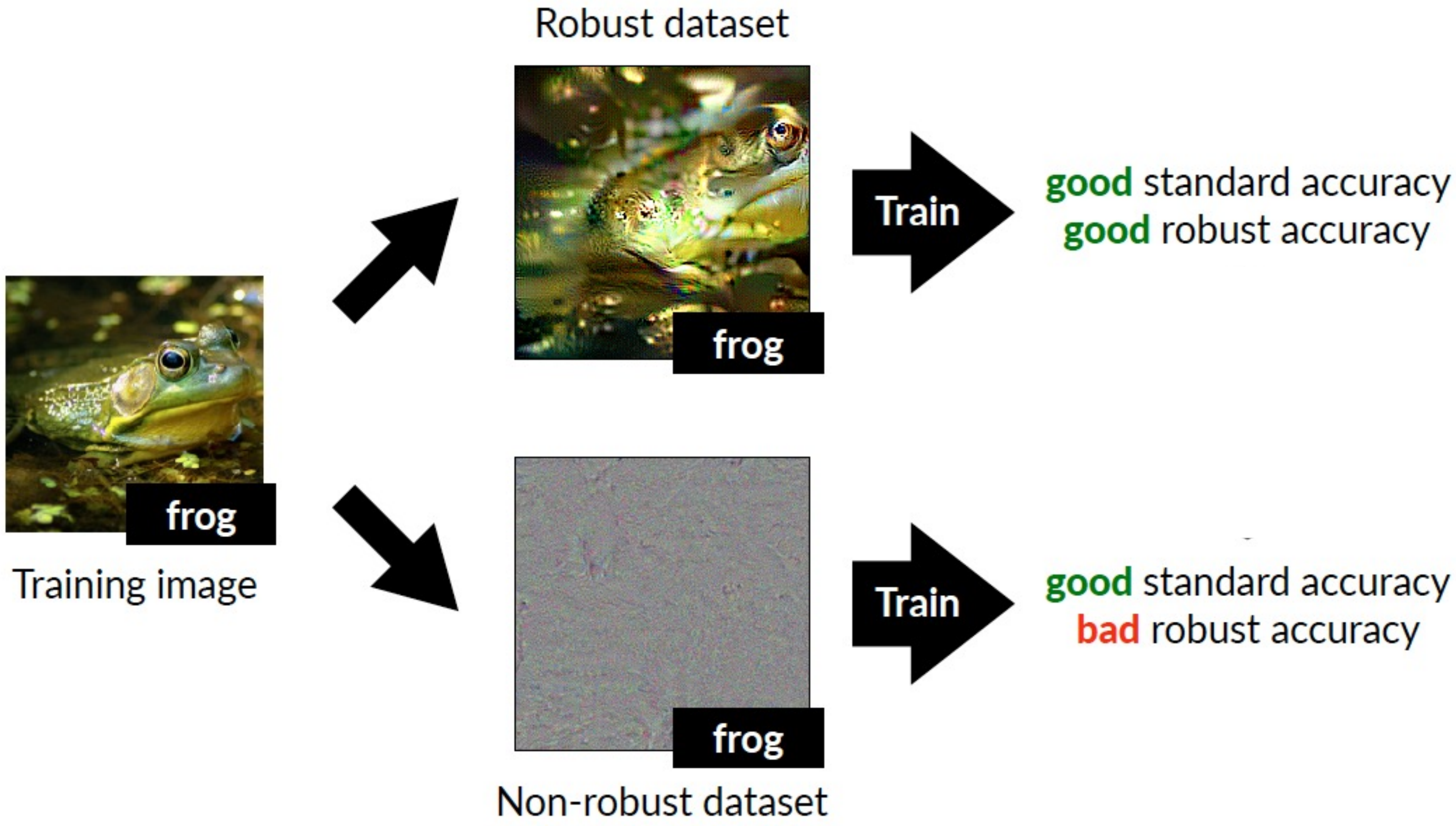
Ilyas et al. Adversarial Examples Are Not Bugs, They Are Features. NeurIPS 2019

Robust vs Non-robust Features



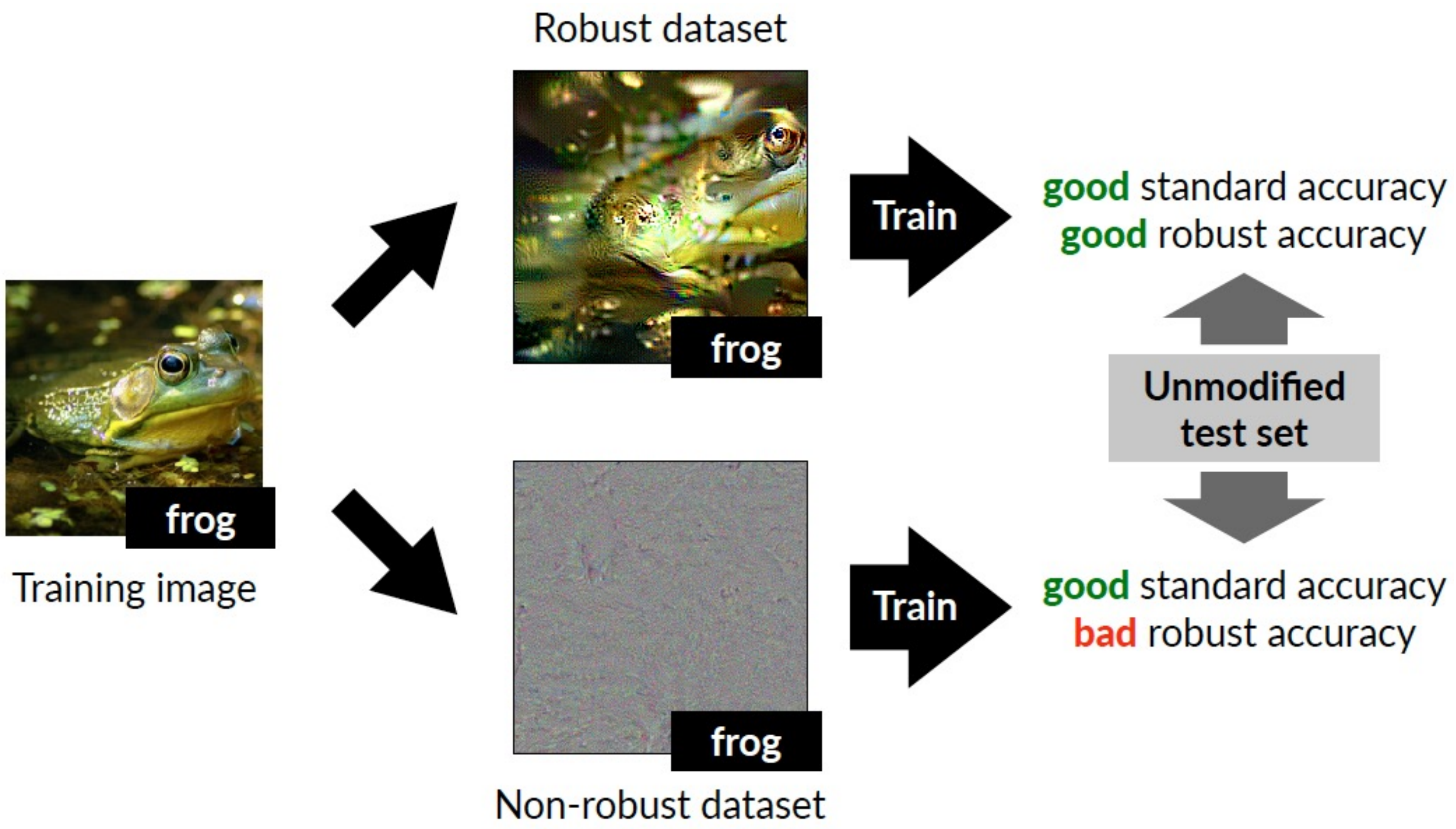
Ilyas et al. Adversarial Examples Are Not Bugs, They Are Features. NeurIPS 2019

Robust vs Non-robust Features



Ilyas et al. Adversarial Examples Are Not Bugs, They Are Features. NeurIPS 2019

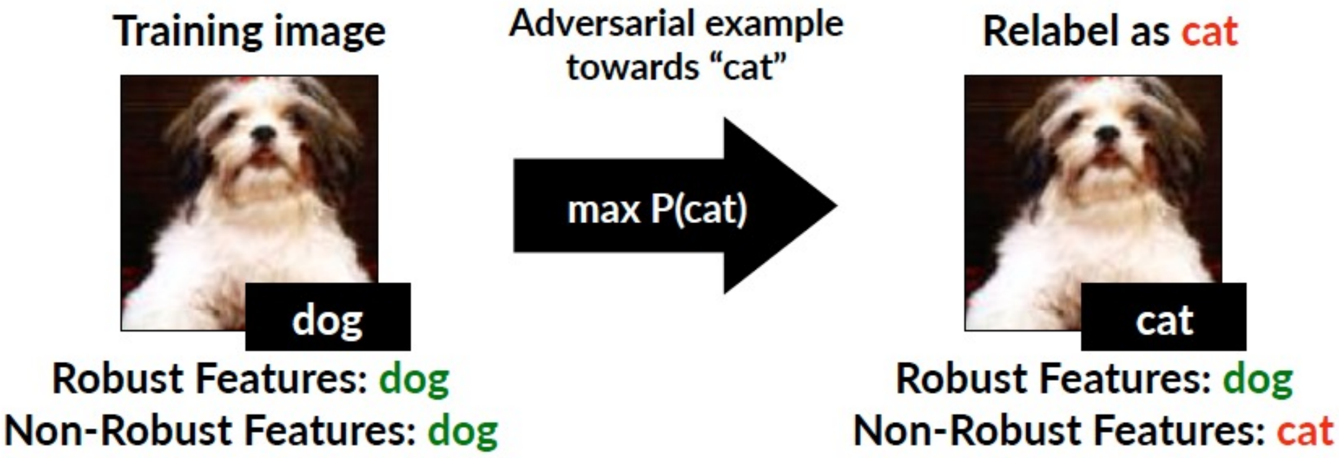
Robust vs Non-robust Features



Training on original set, both the robust & non-robust features of the input are predictive of the label

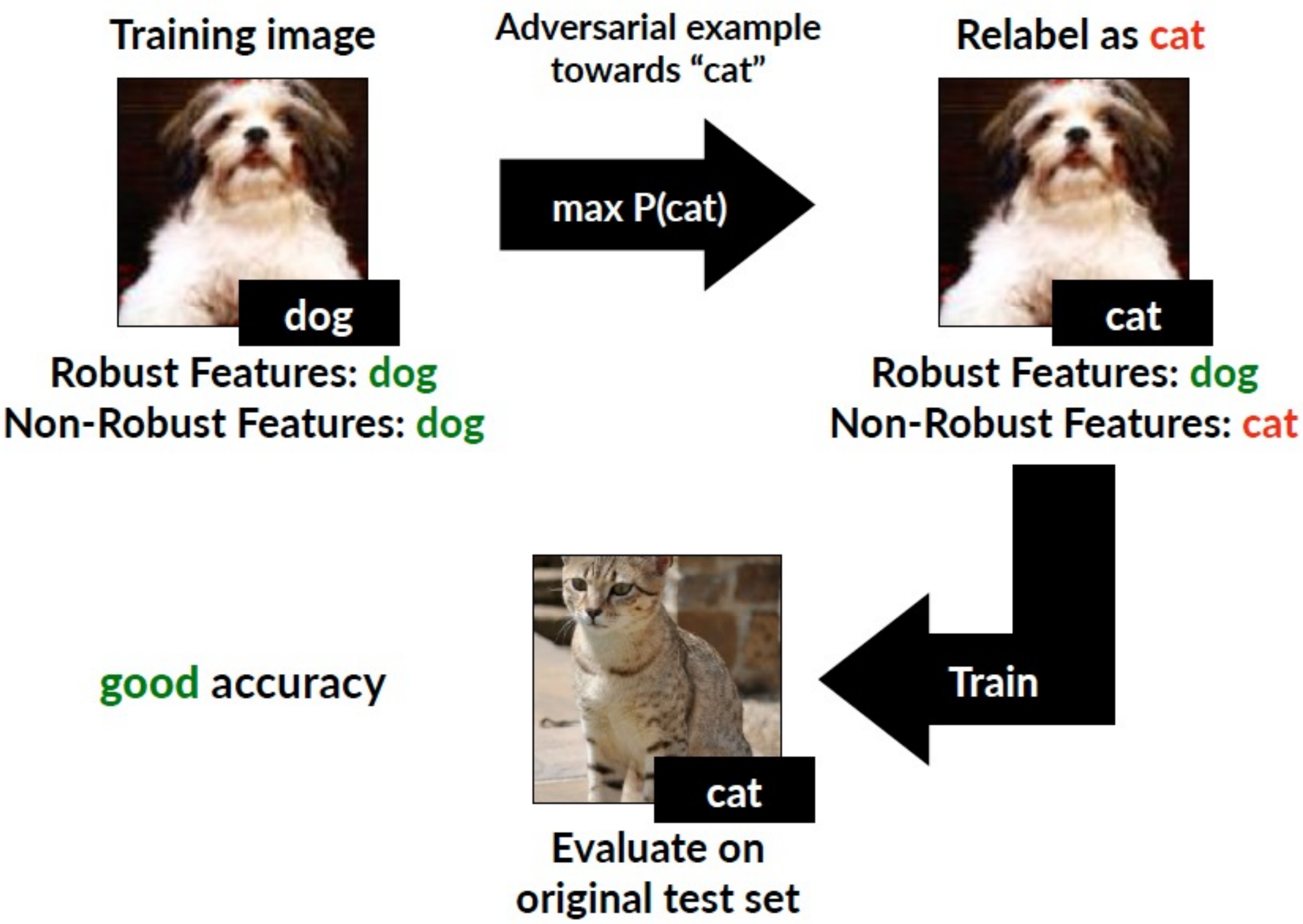
Ilyas et al. Adversarial Examples Are Not Bugs, They Are Features. NeurIPS 2019

Robust vs Non-robust Features



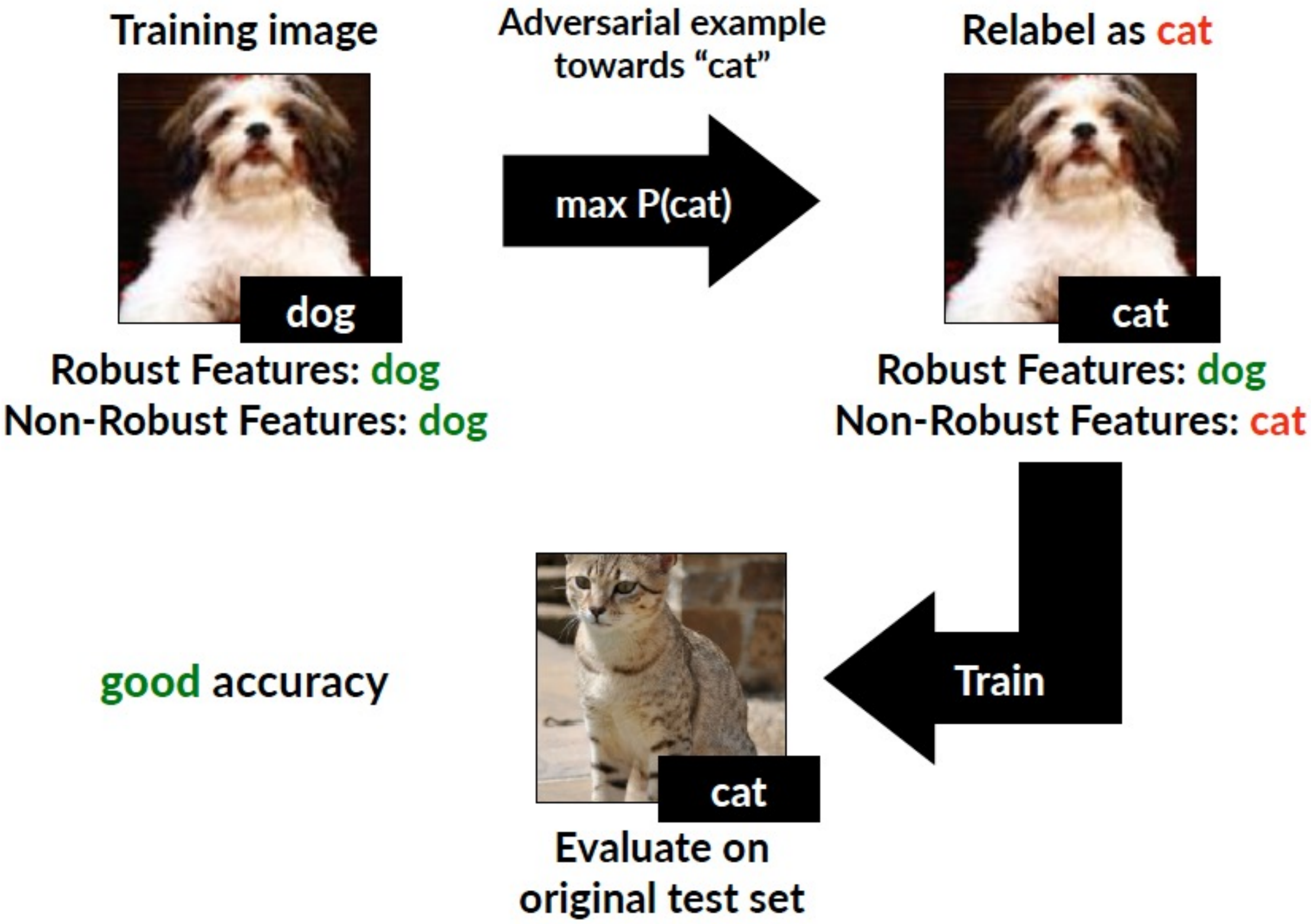
Ilyas et al. Adversarial Examples Are Not Bugs, They Are Features. NeurIPS 2019

Robust vs Non-robust Features



Ilyas et al. Adversarial Examples Are Not Bugs, They Are Features. NeurIPS 2019

Robust vs Non-robust Features



Training on perturbed set, only the non-robust features provide correct guidance for generalization

Ilyas et al. Adversarial Examples Are Not Bugs, They Are Features. NeurIPS 2019

谢谢！