Adversarial Attacks: Why are Machine Learning Models Vulnerable to Attacks?

Jihun Hamm & Akshay Mehra

Tulane University

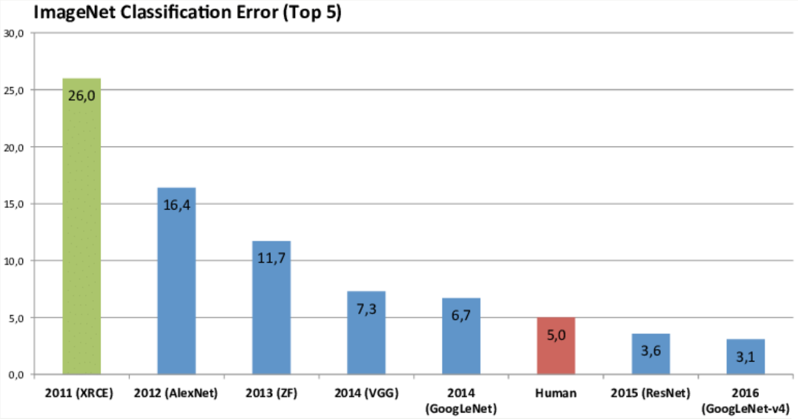
Table of Contents

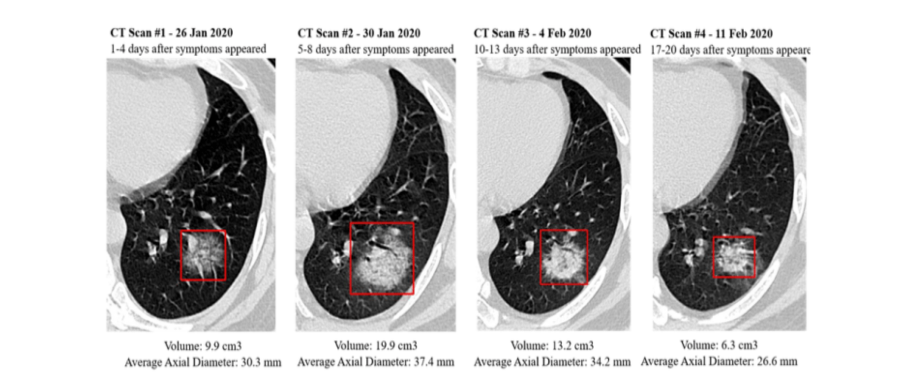
1. How robust is naïve ML model?
2. Can a ML model resist test-time attack?
3. When is a ML model provably robust?
4. Can you trust others’ data and models?

Part 1/4

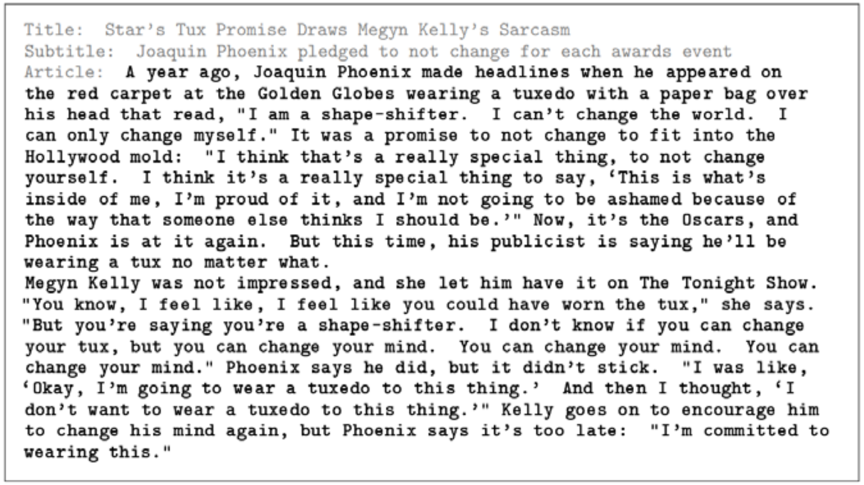
1. How robust is naïve ML model?
2. Can a ML model resist test-time attack?
3. When is a ML model provably robust?
4. Can you trust others' data and models?

Success of machine learning





CT Scan #1 - 26 Jan 2020
1-4 days after symptoms appeared
CT Scan #2 - 30 Jan 2020
5-8 days after symptoms appeared
CT Scan #3 - 4 Feb 2020
10-13 days after symptoms appeared
CT Scan #4 - 11 Feb 2020
17-20 days after symptoms appeare
Volume: 9.9 cm3
Average Axial Diameter: 30.3 mm
Volume: 19.9 cm3
Average Axial Diameter: 37.4 mm
Volume: 13.2 cm3
Average Axial Diameter: 34.2 mm
Volume: 6.3 cm3
Average Axial Diameter: 26.6 mm

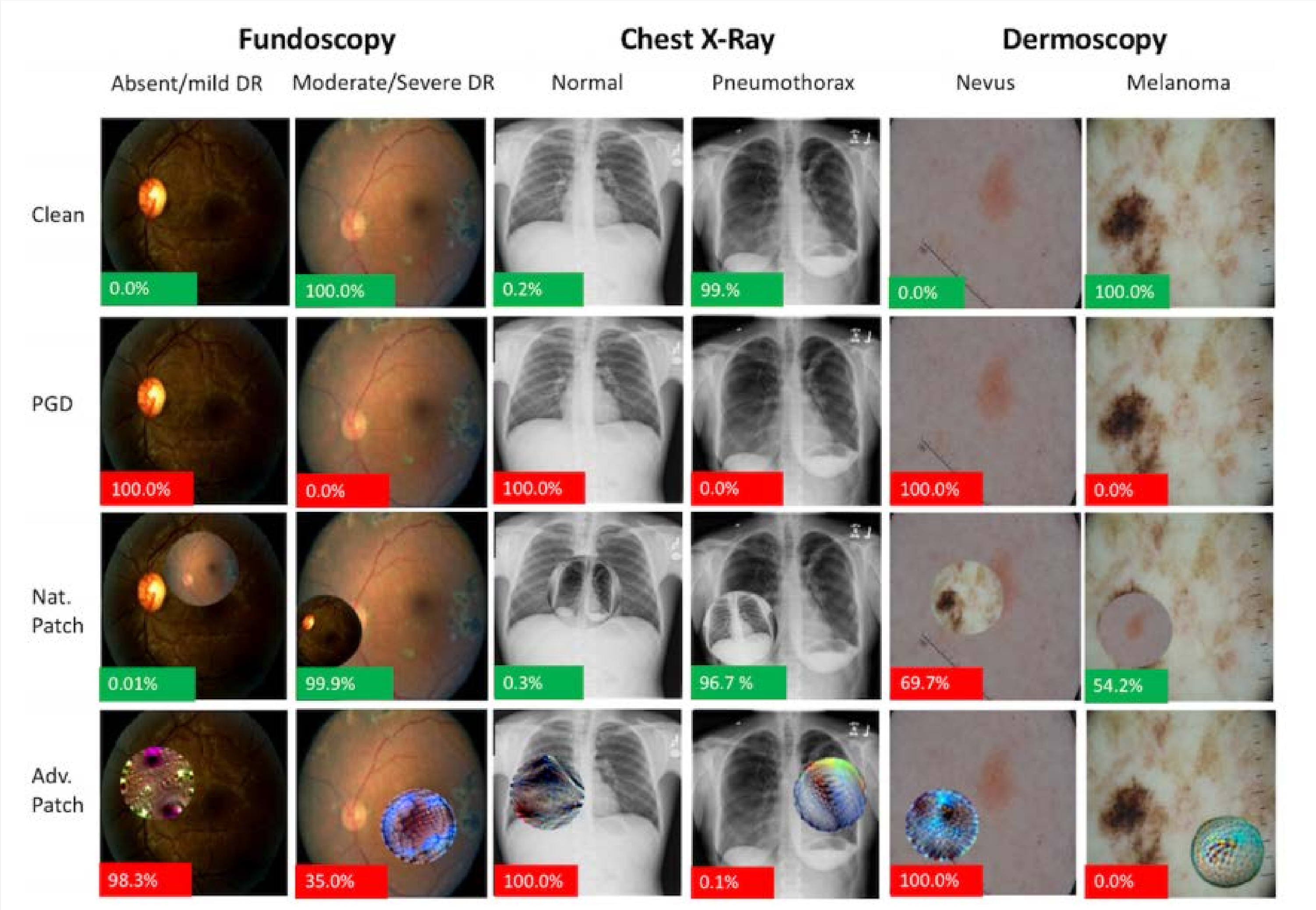


Title: Star's Tux Promise Draws Megyn Kelly's Sarcasm
Subtitle: Joaquin Phoenix pledged to not change for each awards event
Article: A year ago, Joaquin Phoenix made headlines when he appeared on the red carpet at the Golden Globes wearing a tuxedo with a paper bag over his head that read, "I am a shape-shifter. I can't change the world. I can only change myself." It was a promise to not change to fit into the Hollywood mold: "I think that's a really special thing, to not change yourself. I think it's a really special thing to say, 'This is what's inside of me, I'm proud of it, and I'm not going to be ashamed because of the way that someone else thinks I should be.'" Now, it's the Oscars, and Phoenix is at it again. But this time, his publicist is saying he'll be wearing a tux no matter what.
Megyn Kelly was not impressed, and she let him have it on The Tonight Show. "You know, I feel like, I feel like you could have worn the tux," she says. "But you're saying you're a shape-shifter. I don't know if you can change your tux, but you can change your mind. You can change your mind. You can change your mind." Phoenix says he did, but it didn't stick. "I was like, 'Okay, I'm going to wear a tuxedo to this thing.' And then I thought, 'I don't want to wear a tuxedo to this thing.'" Kelly goes on to encourage him to change his mind again, but Phoenix says it's too late: "I'm committed to wearing this."

Figure 3.15: The GPT-3 generated news article that humans found the easiest to distinguish from a human written article (accuracy: 61%).

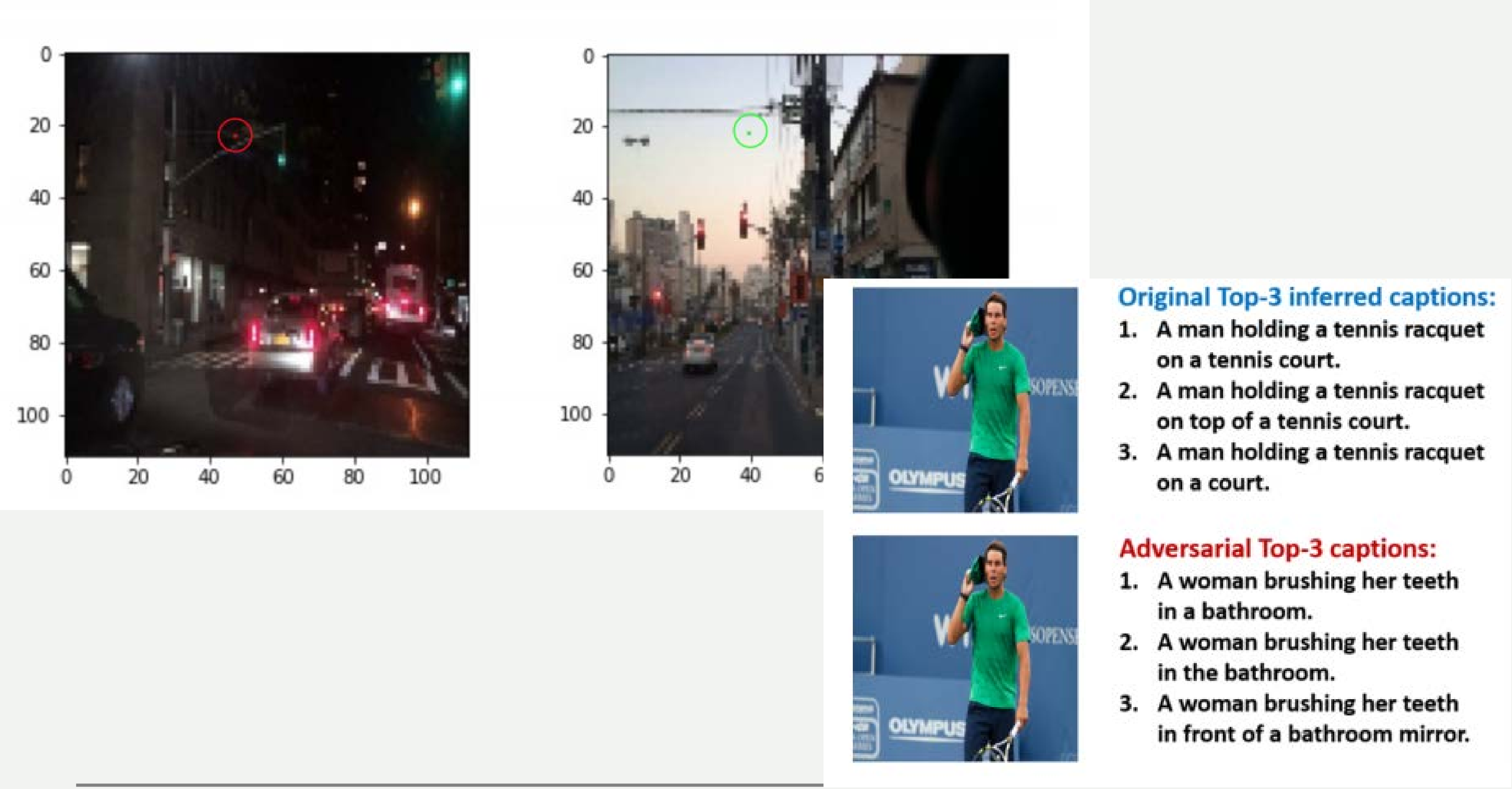
https://devopedia.org/imagenet
https://syncedreview.com/2020/03/18/ai-ct-scan-analysis-for-covid-19-detection-and-patient-monitoring/
https://towardsdatascience.com/deep-learning-for-self-driving-cars-7f198ef4cfa2
Brown, Tom B., et al. "Language models are few-shot learners." *arXiv preprint arXiv:2005.14165* (2020).

But is machine learning reliable?



Finlayson, Samuel G., et al. "Adversarial attacks against medical deep learning systems." (2018).

But is machine learning reliable?



Original Top-3 inferred captions:

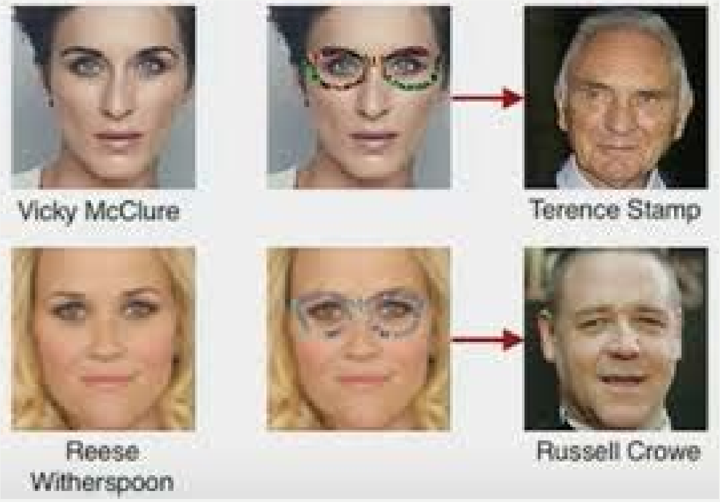
1. A man holding a tennis racquet on a tennis court.
2. A man holding a tennis racquet on top of a tennis court.
3. A man holding a tennis racquet on a court.

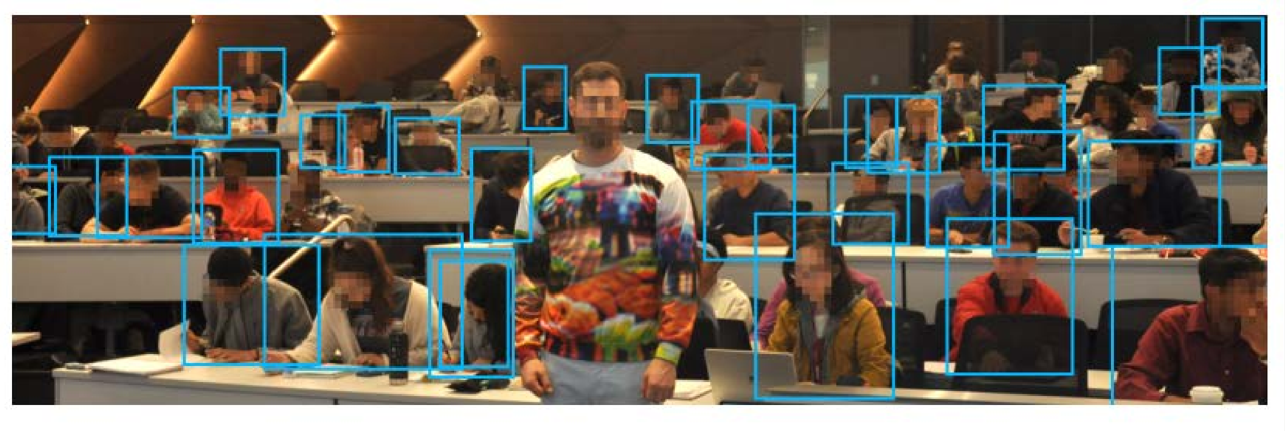
Adversarial Top-3 captions:

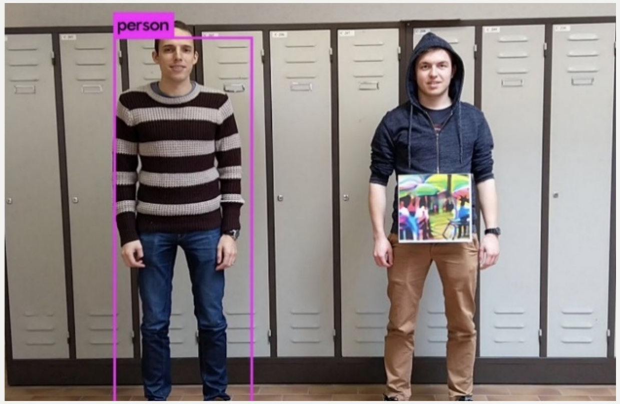
1. A woman brushing her teeth in a bathroom.
2. A woman brushing her teeth in the bathroom.
3. A woman brushing her teeth in front of a bathroom mirror.

Wu, Min, et al. "A game-based approximate verification of deep neural networks with provable guarantees." (2020).
https://www.nature.com/articles/d41586-019-03013-5
Chen, Hongge, et al. "Attacking visual language grounding with adversarial examples: A case study on neural image captioning." (2017).

But is machine learning reliable?







Ilyas, Andrew, et al. "Black-box adversarial attacks with limited queries and information." (2018).
Sharif, Mahmood, et al. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." (2016).
https://www.theverge.com/2019/4/23/18512472/fool-ai-surveillance-adversarial-example-yolov2-person-detection
Wu, Zuxuan , et al. "Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors." (2020).

What is wrong with ML?

What is wrong with ML?

- Basic assumption of ML:

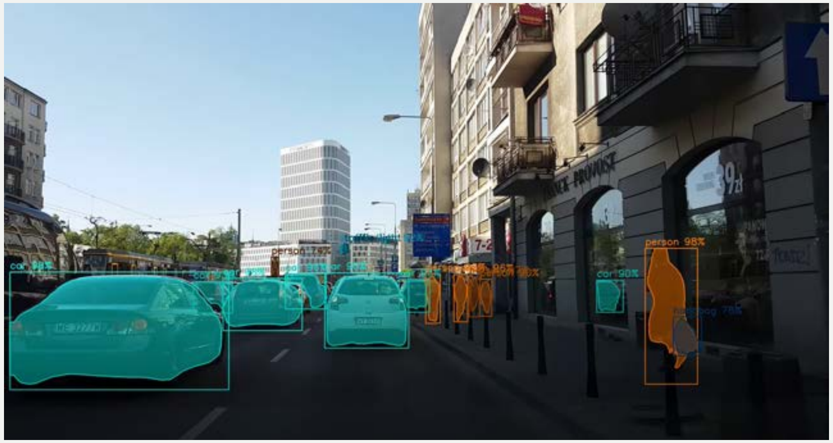
  Training and test sets have the same distribution

  $$P_{Train}(x, y) = P_{Test}(x, y)$$

- Most benchmark datasets have this property
  - MNIST, EMNIST
  - SVHN
  - CIFAR10/100
  - Imagenet
  - COCO
  - …

Can this hold in reality?

- Difficult because
  - ML models are now increasingly being deployed in the wild
  - Impossible to train on all possible scenarios that can be encountered at test time

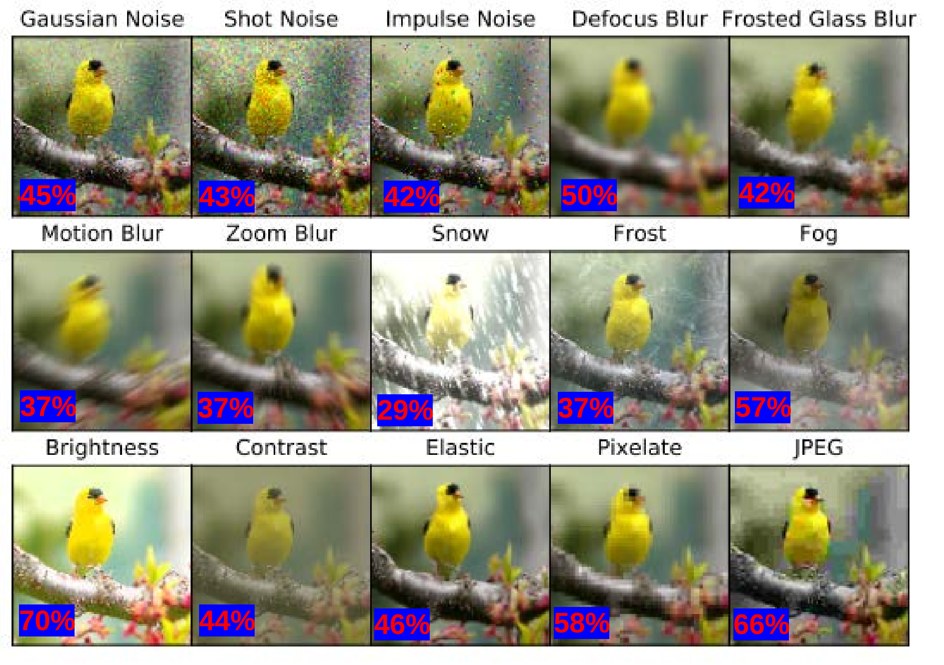


https://www.youtube.com/watch?v=OOT3UIXZztE
https://www.cnet.com/news/face-masks-are-thwarting-even-the-best-facial-recognition-algorithms-study-finds/

Can this hold in reality?

- Performance of the model drops significantly in presence of common corruptions in the data

Resnet 50 is 76% accurate on clean test set

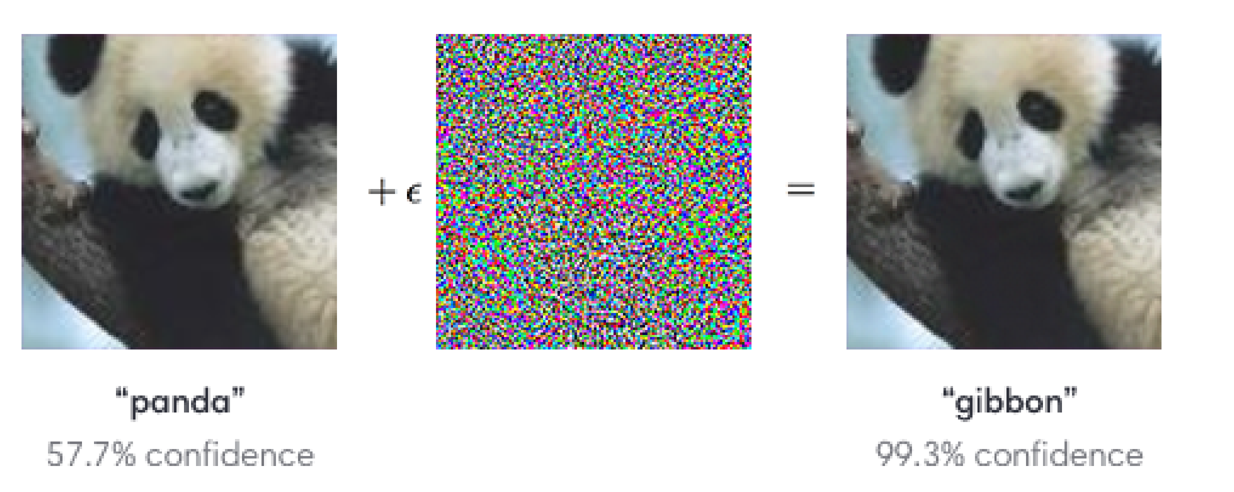


- “AI systems need to be trained to handle environmental differences like lighting, which can vary among clinics, impacting the model’s predictions.”

Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." (2019).
https://www.blog.google/technology/health/healthcare-ai-systems-put-people-center

Adversarial example

- Adversarial ML: Szegedy et al (2013), Goodfellow et al. (2014).
- Def:
  - imperceptible change to a test input
  - that can make a model change its prediction
  - often with a high confidence

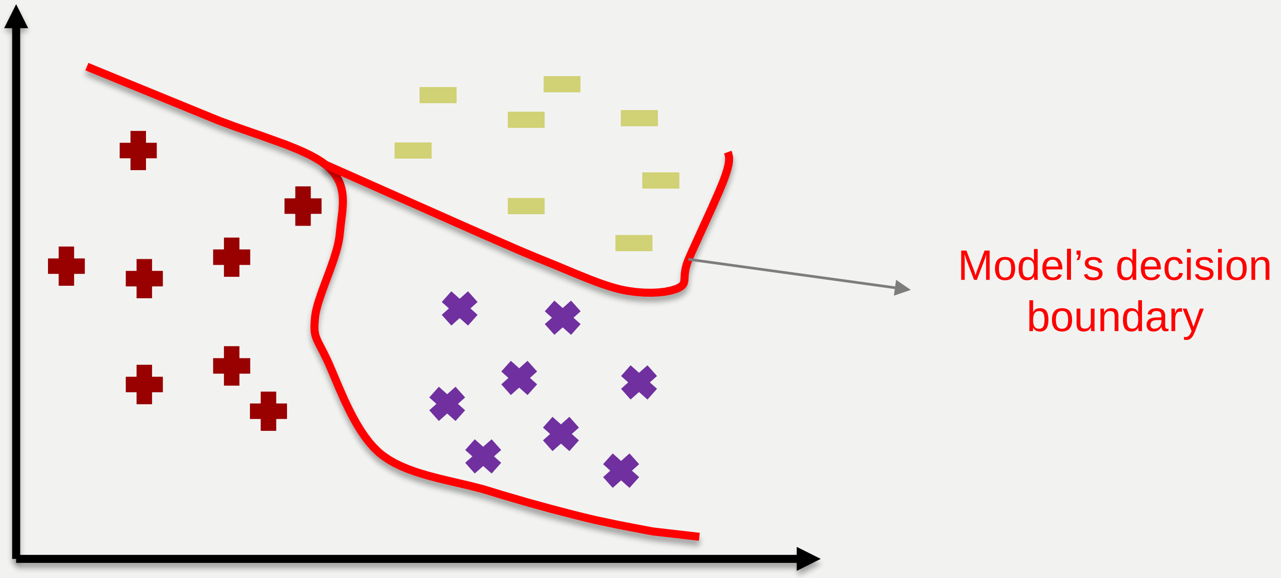


?

Biggio, Battista, et al. "Evasion attacks against machine learning at test time." (2013).
Szegedy, Christian, et al. "Intriguing properties of neural networks." (2013).
Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." (2014).

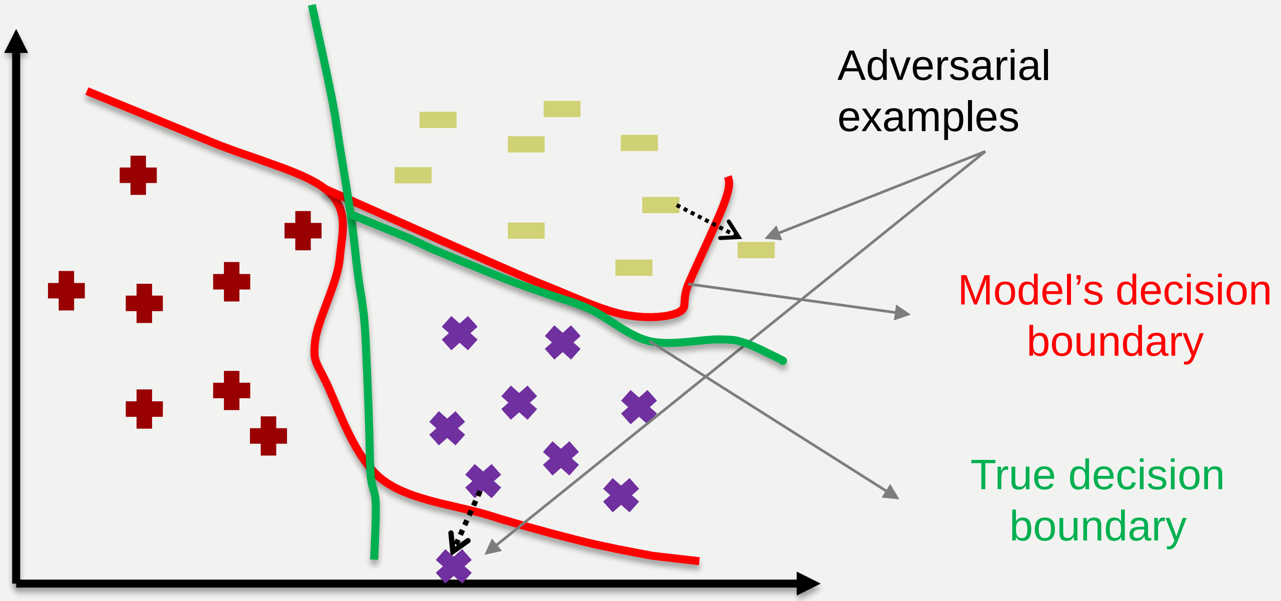
Why do they exist?

- Three-class classification problem



Model’s decision boundary

Why do they exist?



- Adversarial examples ≅ generalization issue
  - Gilmer et al.“Adversarial examples are a natural consequence of test error in noise” *(2019)*
  - Other views exist too

How to craft adversarial examples?

- Given: a multiclass classifier $f_\theta : X \to \{1, 2, \ldots, K\}$

  (consider a convolutional neural network)
- Task: at the test point x, find a distortion $\delta \in \Delta$, such that
  $$f_\theta(x) \neq f_\theta(x + \delta)$$
- Difficult to solve directly due to inequality
- One way to find δ is to solve:
  $$\delta^* = \max_{\{\delta \in \Delta\}} \quad L(f_\theta(x + \delta), y)$$
  (consider the cross entropy $L(p, q) = -E_p[\log q]$)

Different attack types

- Basic: $\delta^* = \max_{\{\delta \in \Delta\}} L(f_\theta(x+\delta), y)$
- Threat models
  - Attacker's knowledge on f: white-box vs gray-box vs black-box
  - Attacker's misclassification goal: untargeted vs targeted
  - Amount of perturbation measured by
    - l_p norm-based attacks
    - Visual similarity
    - Text similarity (for attack on NLP models)
    - ...
  - Most common: white-box, untargeted, norm-based
- Optimization methods: FGSM, PGD, C&W

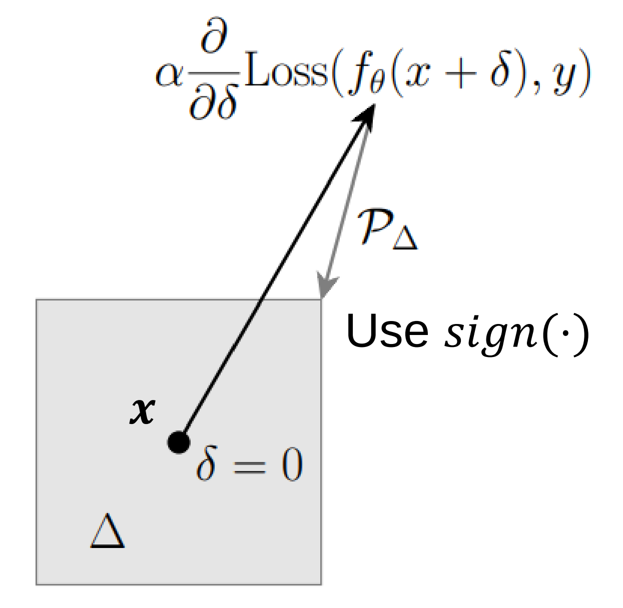
FGSM attack

- Fast Gradient Sign Method (FGSM)
  - Efficient heuristic for adversarial perturbation δ
  - l_∞-norm constraint: $\Delta = \{\delta \mid \|\delta\|_\infty \leq \epsilon\}$ (e.g., max change of pixel value $\leq \epsilon$)

- Algorithm
  - A single gradient step of $Loss$ w.r.t. δ
  - Project to the l_∞-ball by $sign[\cdot]$:

    $\delta^* = \epsilon\, sign[\nabla_\delta L(f_\theta(x+\delta), y)]$

    $Sign[\cdot]$ maps δ to vertex
  - Adversarial example: $x' = x + \delta^*$


$\alpha \frac{\partial}{\partial \delta} \text{Loss}(f_\theta(x+\delta), y)$
$\mathcal{P}_\Delta$
Use $sign(\cdot)$
$\boldsymbol{x}$
$\delta = 0$
Δ

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples." (2014).
http://www.cs.cmu.edu/~cliu6/16-883/robust_deep_learning.pdf

PGD attack

- Projected Gradient Descent (PGD) attack
  - Unlike FGSM, use multiple gradient steps to craft an example
  - After each steps, project onto the set Δ
- Algorithm
  - Repeat: $\delta^* = Proj_{\Delta}\left(\delta + \alpha \nabla_{\delta} Loss(f_{\theta}(x+\delta), y)\right)$
  - Similar to FGSM, projection $\Delta = \{\delta : ||\delta||_{\infty} \leq \epsilon\}$ can be done by applying $sign(\cdot)$ for each coordinate
  - Projection to other l_p-ball can be done by normalization

Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." (2017).

C&W attack

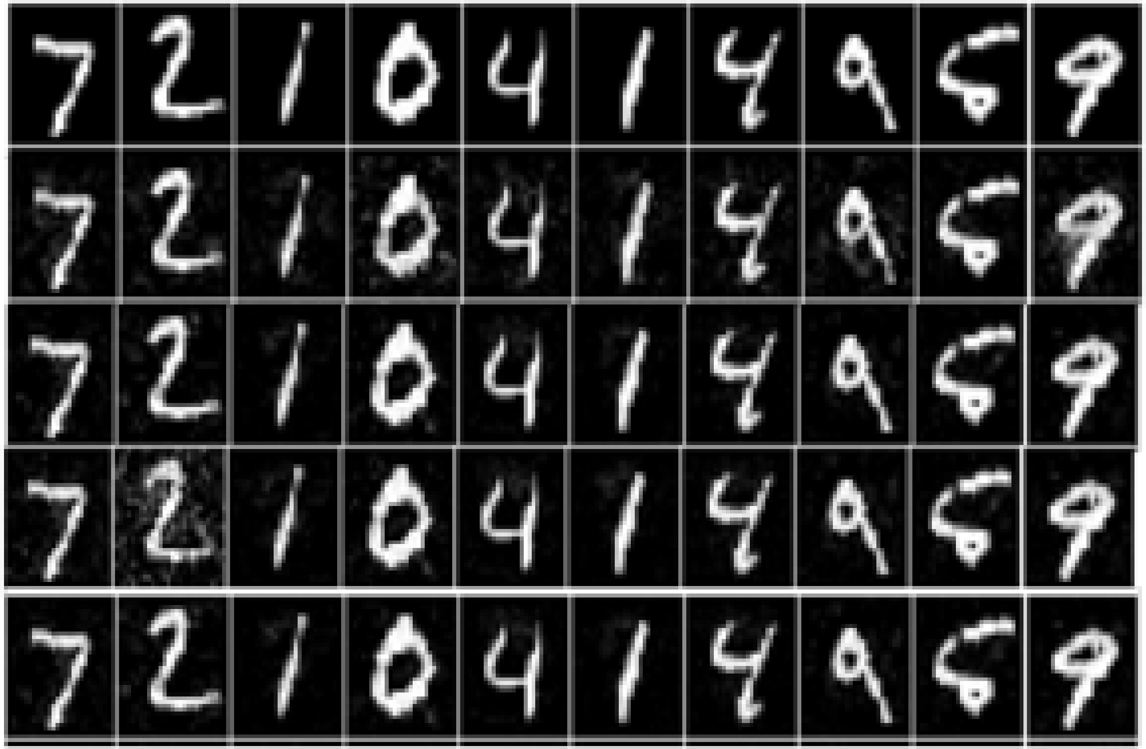
- Carlini-Wagner (C&W) attack
  - FGSM / PGD attack: find most effective δ inside Δ
  - C&W: find an effective and **minimal** δ inside Δ
  - Makes sense: distortion↑ attack success↑ detectability↑

- Algorithm
  - Solve $\delta^* = \underset{\delta \in \Delta}{\operatorname{argmin}} ||\delta||_p + \lambda\, F(x + \delta)$ numerically
  - F is a function that measures how close the prediction of the model on $x + \delta$ is to the target label for the attack.

Carlini, Nicholas, and David Wagner. "Towards evaluating the robustness of neural networks." (2017).

Attack examples

Attacks on Undefended Model



Original Data:

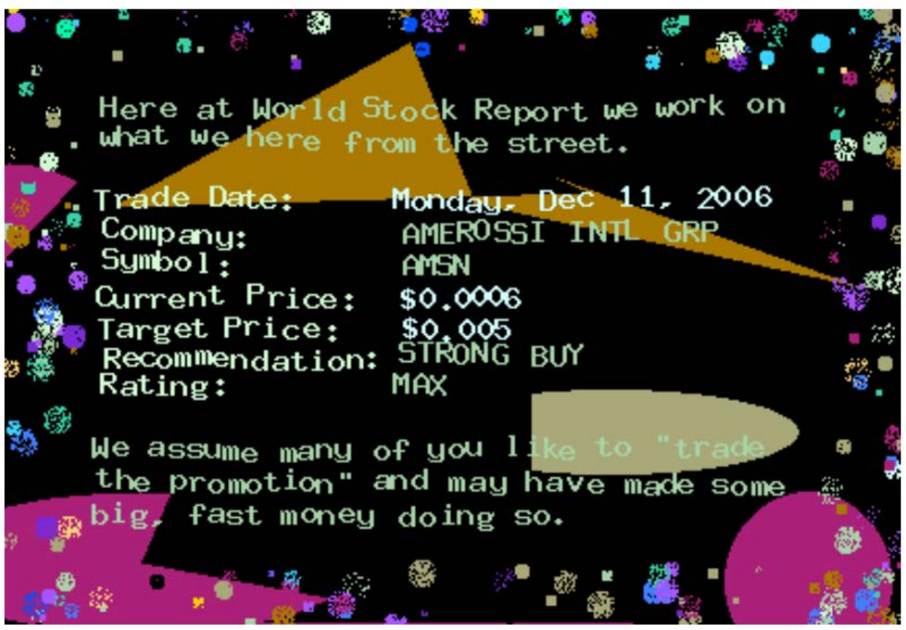
L-BFGS Attack:

CW Attack:

PGD Attack:

L_aug Attack:

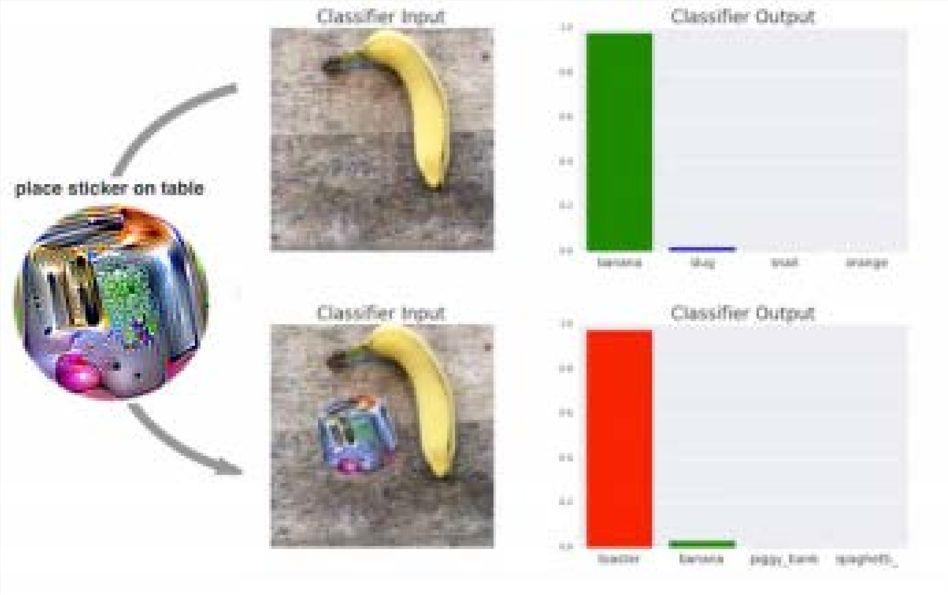
Attacks beyond pixel changes



Here at World Stock Report we work on what we here from the street.

Trade Date: Monday, Dec 11, 2006
Company: AMEROSSI INTL GRP
Symbol: AMSN
Current Price: $0.0006
Target Price: $0.005
Recommendation: STRONG BUY
Rating: MAX

We assume many of you like to "trade the promotion" and may have made some big, fast money doing so.



Gilmer, Justin, et al. "Motivating the rules of the game for adversarial example research." (2018).
Brown, Tom B., et al. "Adversarial patch." (2017).
Athalye, Anish, et al. "Synthesizing robust adversarial examples.“ (2018).

Summary of Part 1

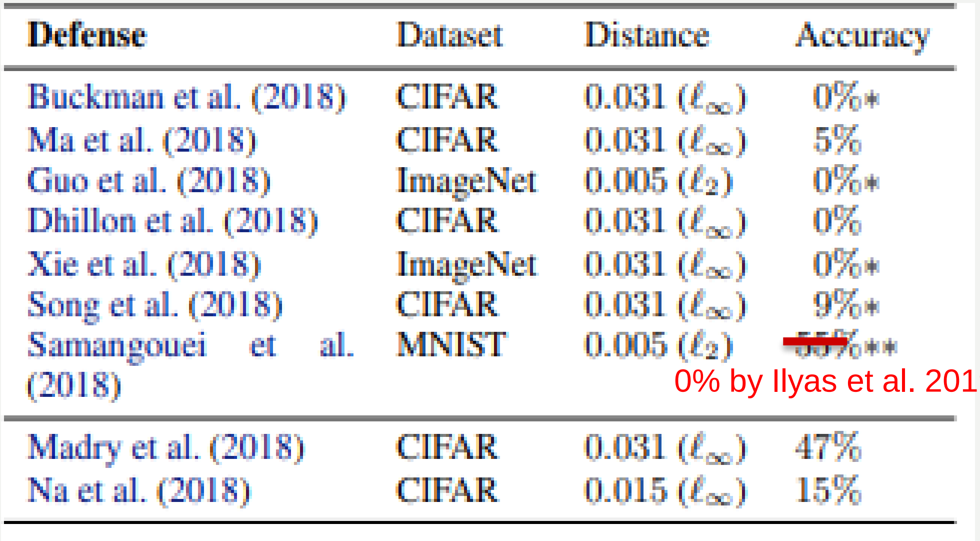
- Machine learning models operate under the assumption that training and test sets have the same distribution
- Due to this, the models are extremely susceptible to out-of-distribution examples
- Only a tiny change can push an example out of the distribution, making it adversarial
- Many methods exist that exploit this weakness of models to craft different types of adversarial examples

Part 2/4

1. How robust is naïve ML model?
2. Can a ML model resist test-time attack?
3. When is a ML model provably robust?
4. Can you trust others’ data and models?

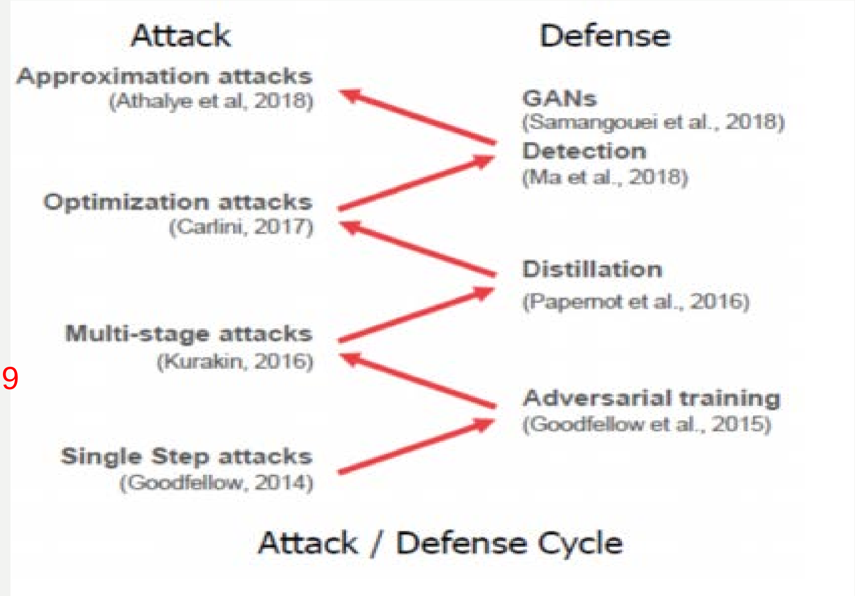
Robustness to adversarial examples

- A cat-and-mouse game
  - A new “successful” attack method discovered
  - A new “robust” model proposed
  - A new “successful attack” method discovered
  - …



Defense	Dataset	Distance	Accuracy
Buckman et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%*
Ma et al. (2018)	CIFAR	0.031 (ℓ_∞)	5%
Guo et al. (2018)	ImageNet	0.005 (ℓ_2)	0%*
Dhillon et al. (2018)	CIFAR	0.031 (ℓ_∞)	0%
Xie et al. (2018)	ImageNet	0.031 (ℓ_∞)	0%*
Song et al. (2018)	CIFAR	0.031 (ℓ_∞)	9%*
Samangouei et al. (2018)	MNIST	0.005 (ℓ_2)	~~55%~~**
Madry et al. (2018)	CIFAR	0.031 (ℓ_∞)	47%
Na et al. (2018)	CIFAR	0.015 (ℓ_∞)	15%

0% by Ilyas et al. 2019



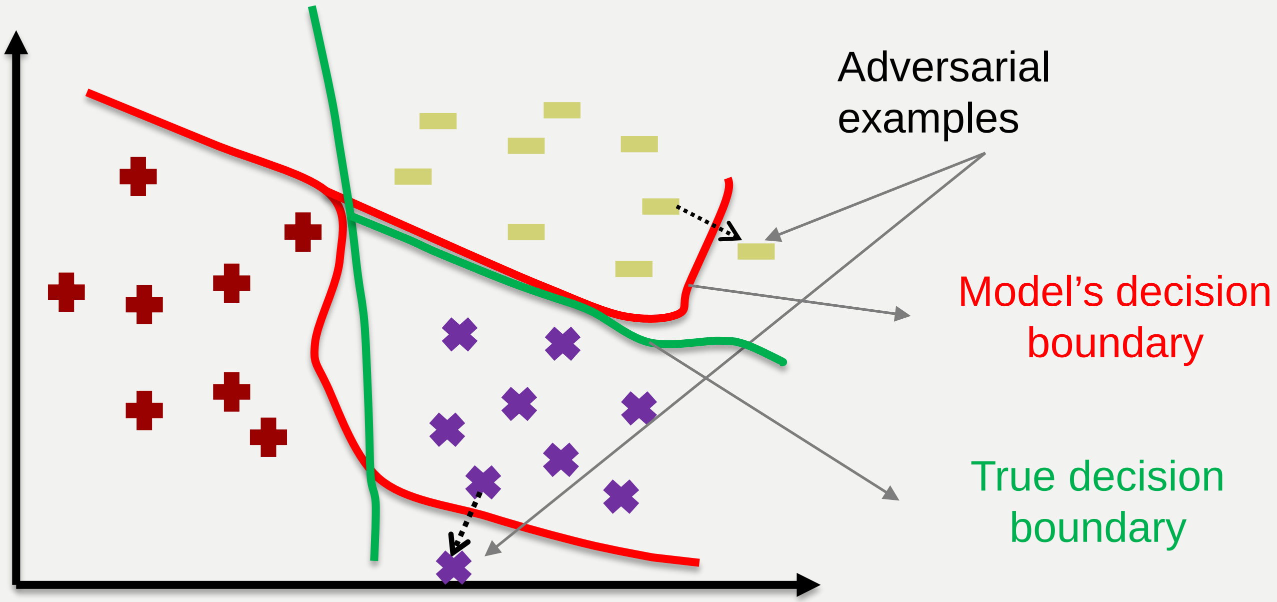
Attack / Defense Cycle

https://tutorial.trustdeeplearning.com/

Athalye et al. "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples." (2018).

Adversarial Training

- Intuition



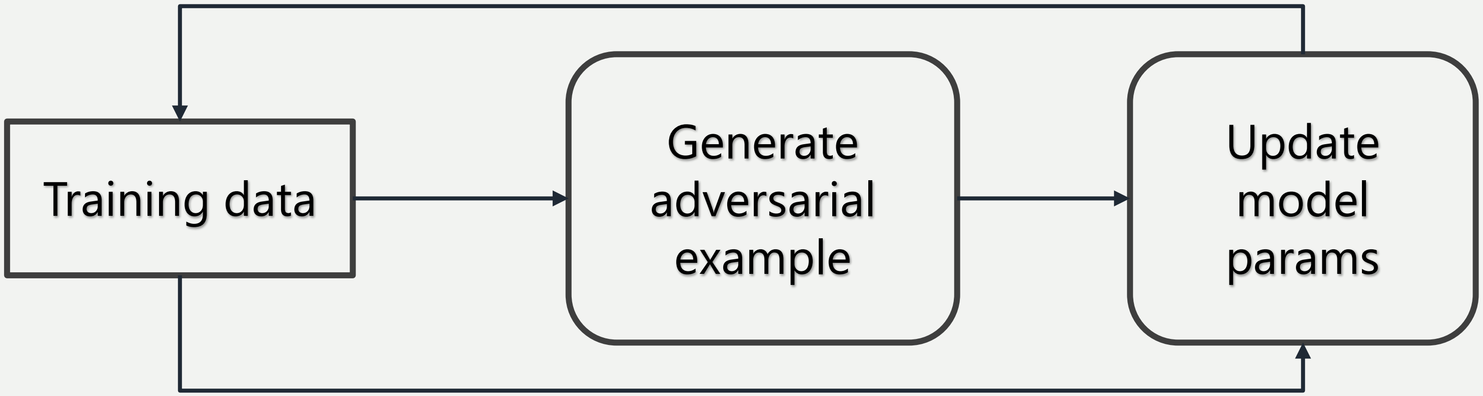
Adversarial examples

Model’s decision boundary

True decision boundary

Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." (2017).

Adversarial Training



Repeat:

1. Sample a minibatch B from the training data
2. For each $(x, y) \in B$, compute an adversarial example x_{adv}
3. Update the model parameters

$$\theta_{new} = \theta \ - \frac{\eta}{|B|}\Sigma_{(x,y)\in \mathrm{B}}\nabla_{\theta}L(f_{\theta}(x_{adv}), y)$$

so that misclassified pints are now correctly classified

Adversarial Training

- Training objective
  - Conventional: $\min_{\theta} E_{\{x,y\}}[L(f_{\theta}(x), y)]$
  - Adv Training: $\min_{\theta} E_{\{x,y\}}[\max_{\{\delta \in \Delta\}} L(f_{\theta}(x+\delta), y)]$
    - Minimax optimization problem

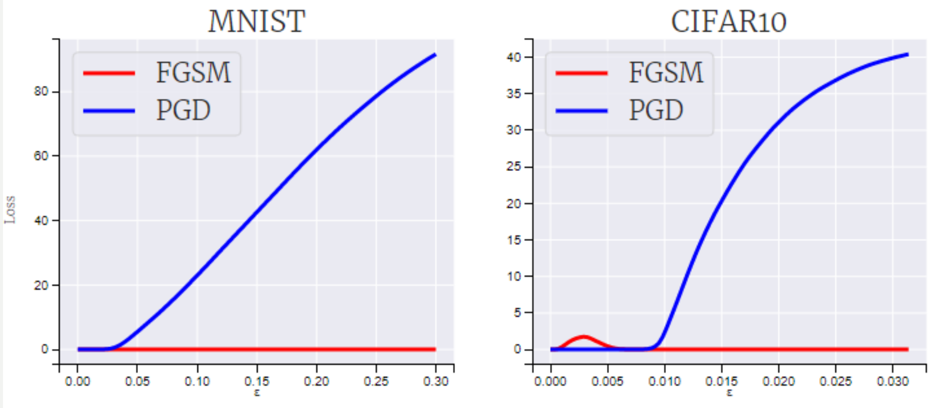
      inner problem: find δ that maximizes loss

      outer problem: find θ that minimizes loss given $x+\delta$.
- Difficulty
  - Theoretically: $\max L(\cdot)$ may not be differentiable w.r.t. θ even if L is differentiable w.r.t. (θ, x)
  - Practically: popular alternating optimization can fail to converge

Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." (2017)
Danskin, John M. “he Theory of Max-Min and its Applications to Weapons Allocation Problems.” (1967)
Hamm, Jihun & Noh, Yung-kyun "K-beam minimax: Efficient optimization for deep adversarial learning" *(2018)*

Adversarial Training

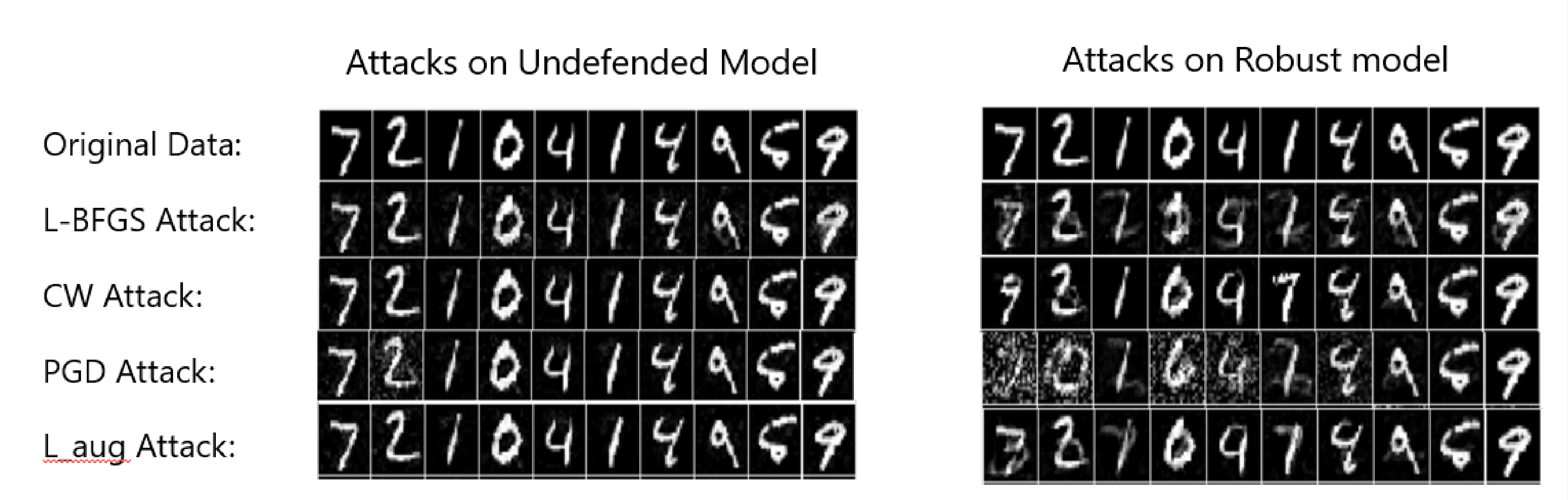
- Optimization matters in practice
  - Inner maximization is solved only approximately during training
  - Quality of the final solution is dependent on maximizer
  - Models trained with FGSM are not robust to PGD attacks



Models trained against FGSM

Adversarial Training

- PGD attack examples generated against a robust model (trained with adversarial training)



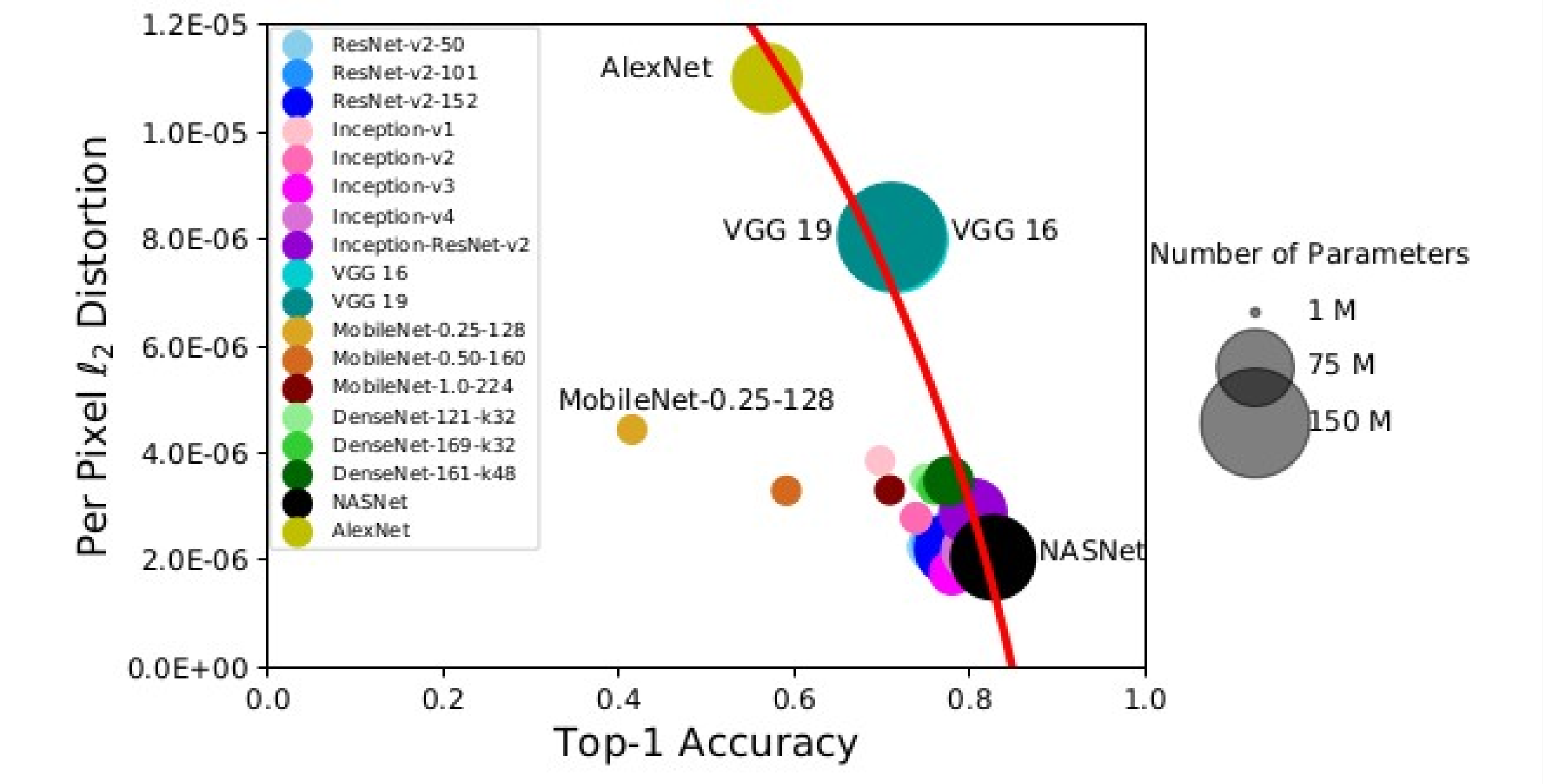
Pros and cons of Adversarial Training

- Pros
  - Provides empirical robustness
  - No change required at test time
  - Easy to integrate with different threat models and procedures

- Cons
  - Success dependent on threat model used during training
  - Models may not be robust to stronger adversaries
  - **Training takes significantly longer**
  - **Accuracy on clean test set decreases**

Accuracy-robustness tradeoff

- Models with high accuracy are usually less robust

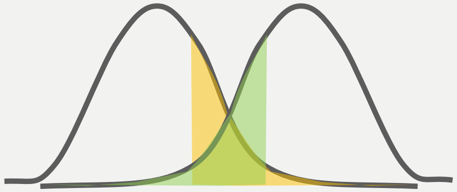


Su, Dong, et al. "Is Robustness the Cost of Accuracy?--A Comprehensive Study on the Robustness of 18 Deep Image Classification Models." (2018).

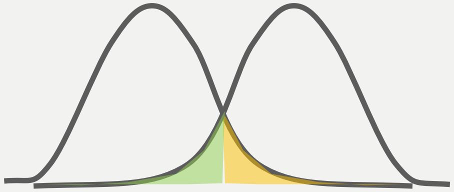
Robust vs natural error

- Def: robust, natural, boundary loss

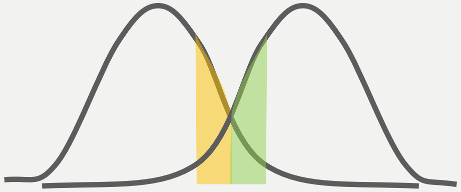
$$L_{robust}(f) \quad = \quad L_{natural}(f) \quad + \quad L_{boundary}(f)$$



$$E[1\{\ \exists X' \in B(X,\epsilon) \quad s.t.\ \ f(X')Y \leq 0\ \}]$$



$$E[1\{\ f(X)Y \leq 0\ \}]$$



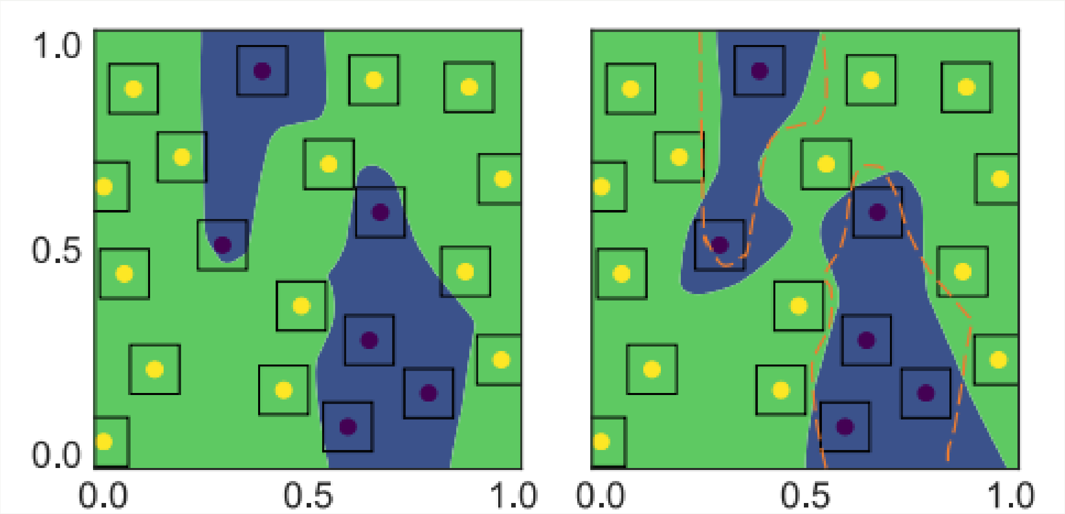
$$E[1\{\ X \in B(DB(f),\epsilon), f(X)Y > 0\ \}]$$

Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." (2019).

TRADES

- Idea: minimize natural loss + approx boundary loss

$$\min_{\theta} E_{(x,y)\in\mathcal{D}}[L(f_\theta(x),y) + \beta \max_{x'\in\mathcal{B}(x,\epsilon)} L(f_\theta(x), f_\theta(x'))]$$

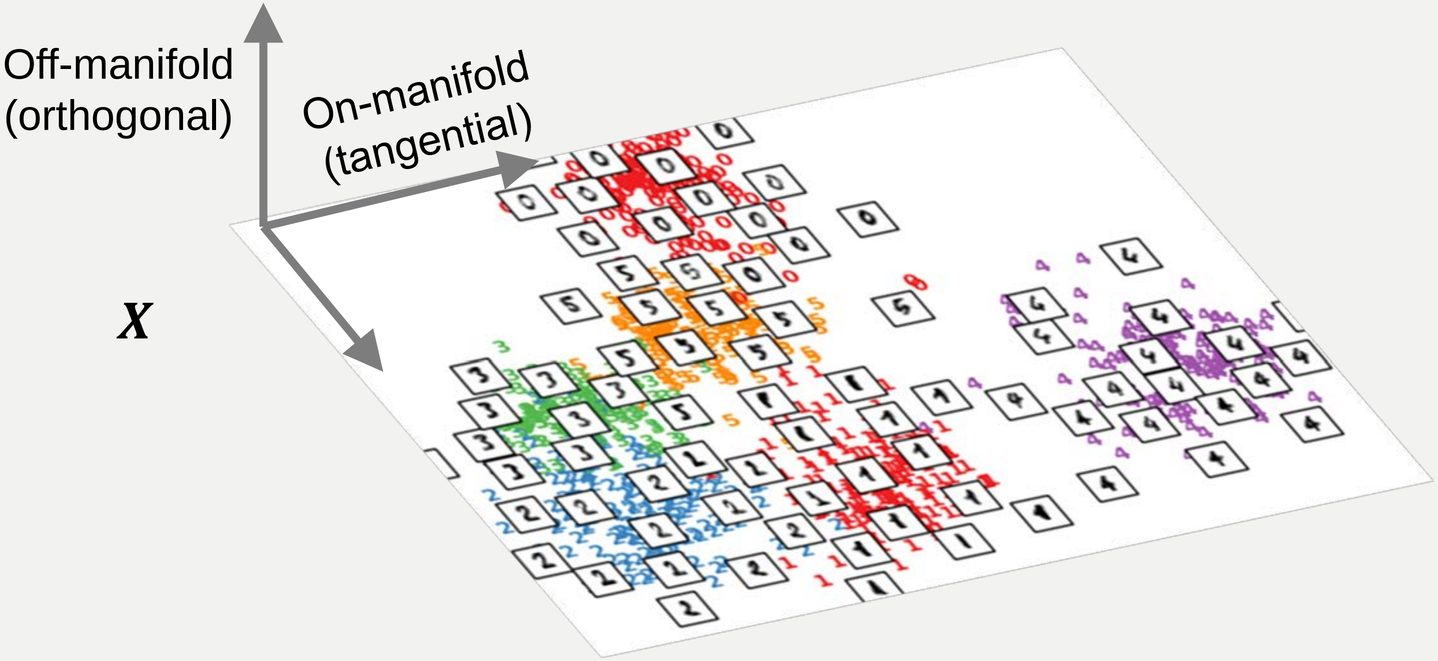


Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." (2019).

Data manifold-point of view

- Manifold assumption:

  Data distribution has a support on a low-dimensional (nonlinear) manifold



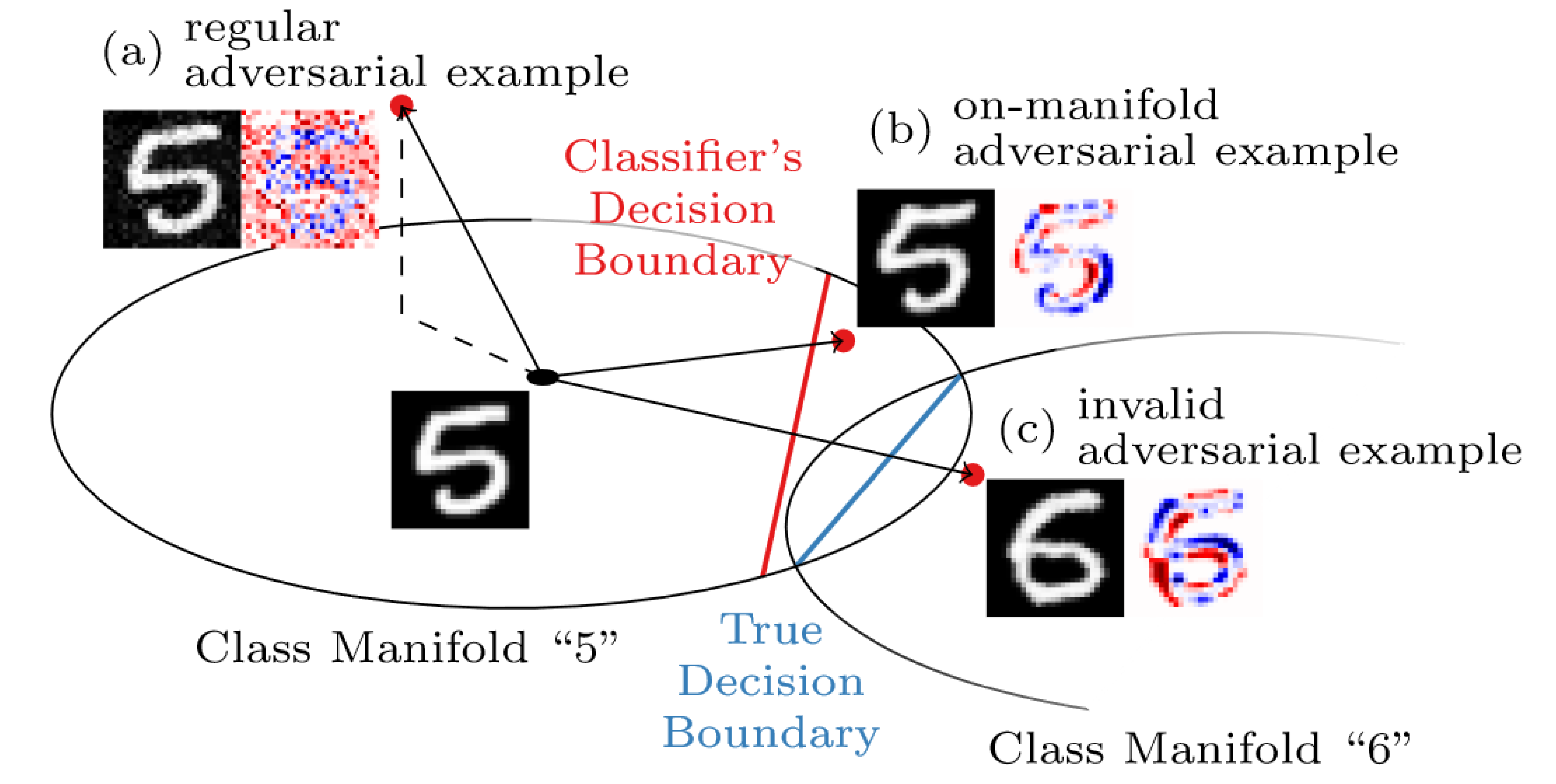
Off-manifold (orthogonal)

On-manifold (tangential)

X

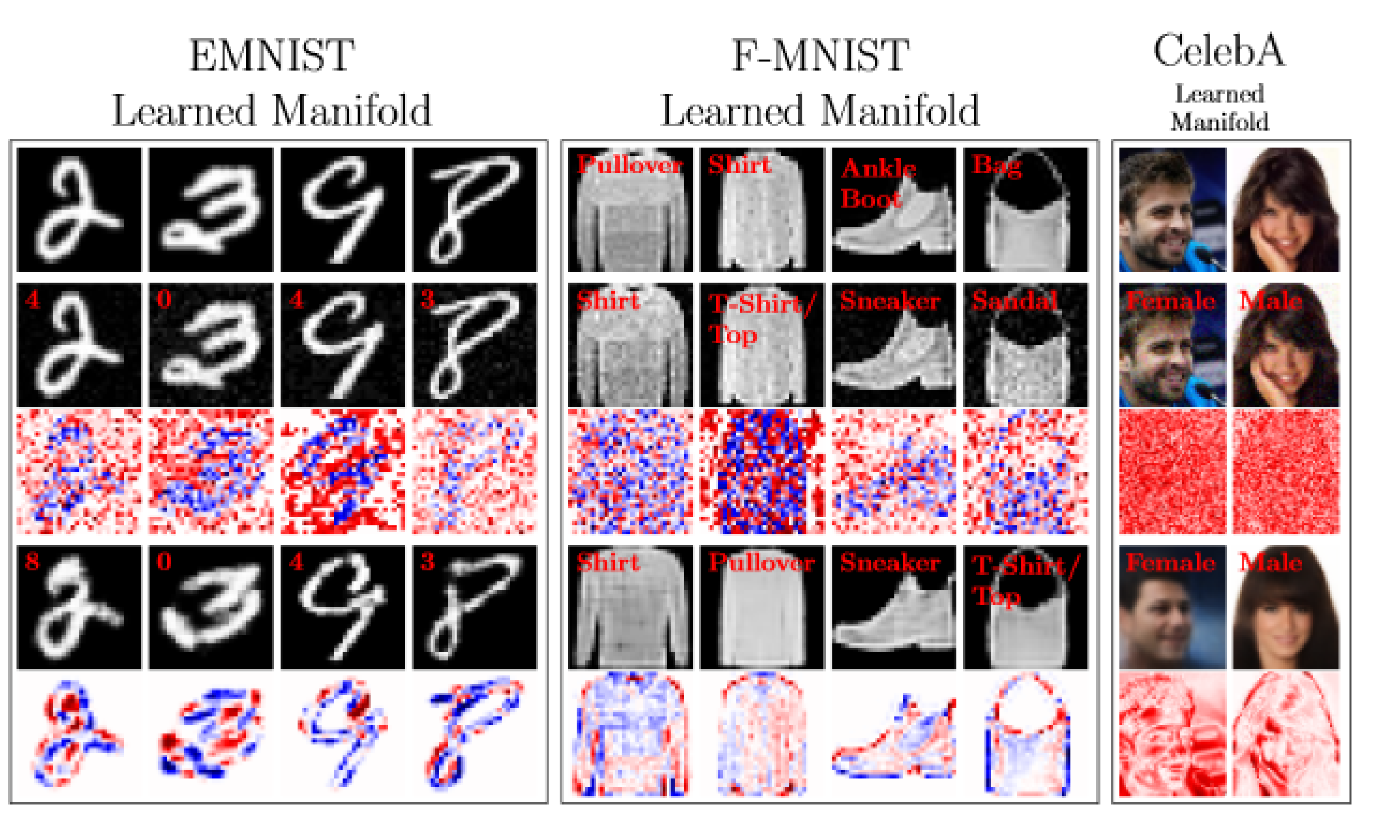
On/off-manifold adversarial examples

- Two types of adversarial examples


(a) regular adversarial example
Classifier's Decision Boundary
(b) on-manifold adversarial example
(c) invalid adversarial example
Class Manifold "5"
True Decision Boundary
Class Manifold "6"

Stutz, David, Matthias Hein, and Bernt Schiele. "Disentangling adversarial robustness and generalization." (2019).

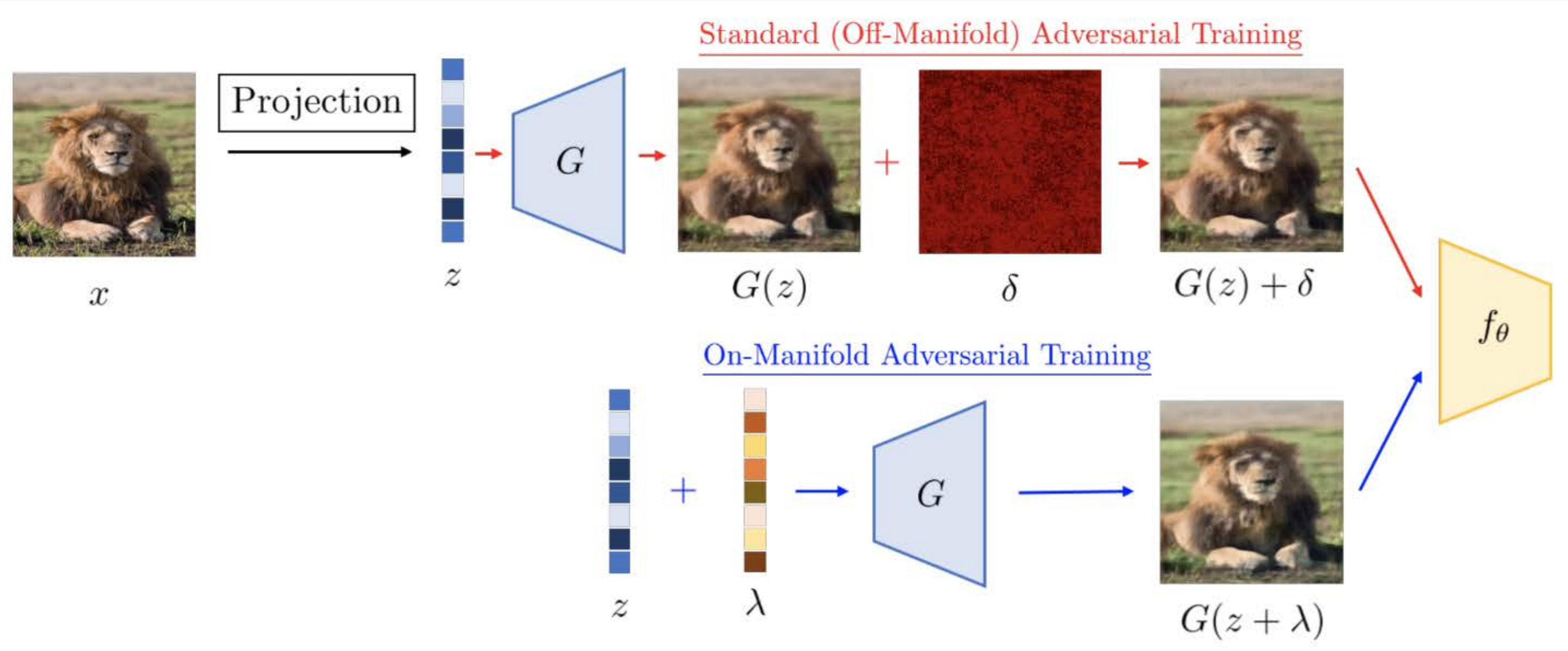
On-/off-manifold adversarial examples



Stutz, David, Matthias Hein, and Bernt Schiele. "Disentangling adversarial robustness and generalization." (2019).

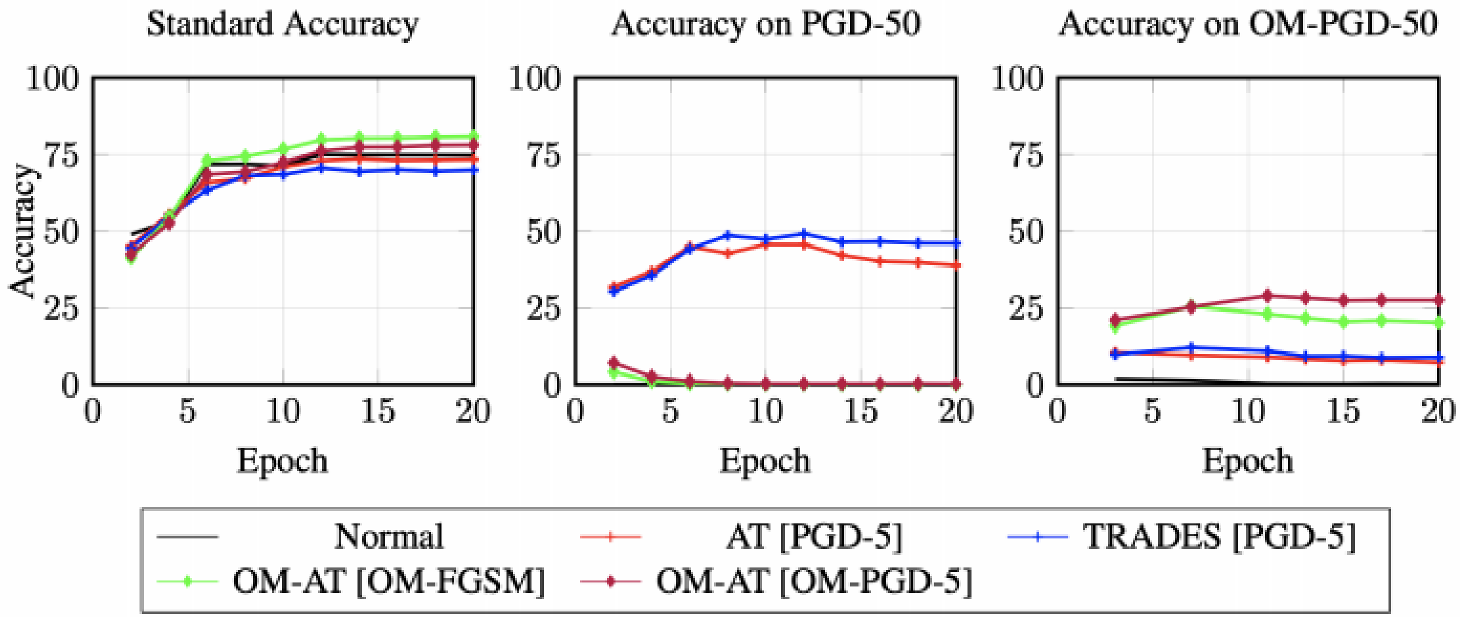
Dual manifold adversarial training

- Solve $\min_{\theta} E_{(x,y)\in\mathcal{D}}[L(f_\theta(x),y) + \quad \beta \max_{x'\in\mathcal{B}(x,\epsilon)} L\big(f_\theta(x), f_\theta(x')\big)$
$$+\beta \max_{\lambda} L\big(f_\theta(x), f_\theta(G(z+\lambda))\big)]$$



Lin, Wei-An, et al. "Dual Manifold Adversarial Robustness: Defense against Lp and non-Lp Adversarial Attacks." (2020).

On-Manifold AT Cannot Defend Standard Attacks and Vice Versa



Standard Accuracy

Accuracy on PGD-50

Accuracy on OM-PGD-50

Figure 3: On-manifold adversarial training does not provide robustness to standard attacks. Standard adversarial training does not provide robustness to on-manifold attacks. Left: standard accuracy. Middle: classification accuracy when the trained models are attacked by PGD-50. Right: classification accuracy when the trained models are attacked by OM-PGD-50.

Lin, Wei-An, et al. "Dual Manifold Adversarial Robustness: Defense against Lp and non-Lp Adversarial Attacks." (2020).

Summary of Part 2

- Several defenses have been proposed to make machine learning models robust to adversarial examples
  - Only adv training seems to be successful (to some degrees)
  - There is an accuracy-robustness tradeoff.
- Adversarial examples may be more than one type
  - Typical adversarial examples are off-manifold type
  - On-manifold adversarial examples are generalization problem
  - Robustness to one type ≠ robustness to the other type
- Dual adversarial training improve robustness to both types

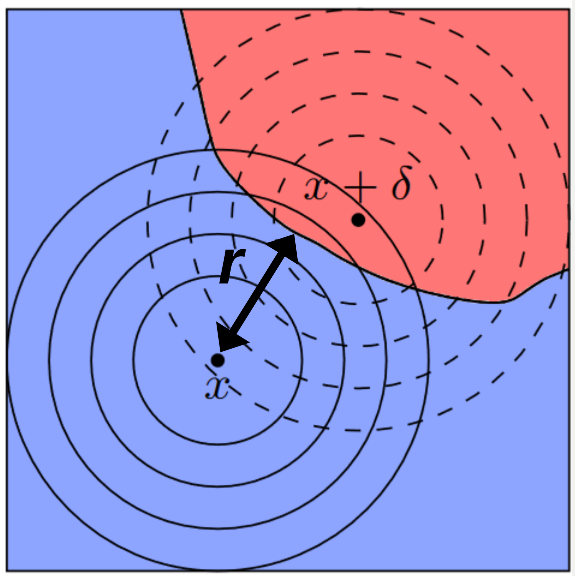
Part 3/4

1. How robust is a naïve ML model?
2. Can a ML model resist test-time attack?
3. When is a ML model provably robust?
4. Can you trust others' data and models?

Margin and empirical robustness

- Defense methods discussed so far improve in empirical robustness compared to naïve models
- Def: margin $r(x)$:= distance from x to nearest decision bndry

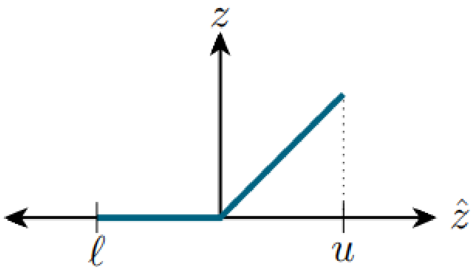
  $r(x) := \min_{x'} \|x - x'\| \quad s.t. \quad f(x') \neq f(x)$
- If r is the true margin at x
  - $g(x)$ is robust for any perturbation $\|\delta\| \leq r$
  - Prediction cannot with δ weaker than r
- Numerically compute r_{num} (by e.g., C&W)
  - Is this sufficient? What guarantees do we have?
  - $r \leq r_{num}$
- We want **certified** radius
  - $r_{cert} \leq r$!!!

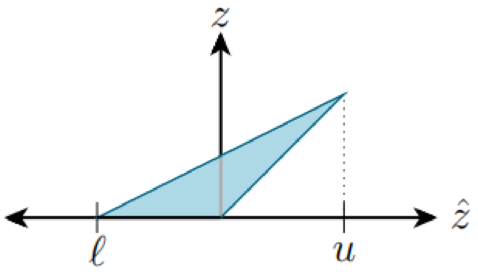


Cohen, Jeremy, et al. "Certified adversarial robustness via randomized smoothing." (2019).

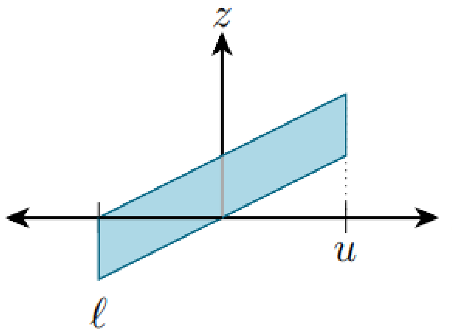
Certification methods

$$\max_{\delta\in\Delta} \mathrm{Loss}(f_\theta(x+\delta),y) \le \max_{\delta\in\Delta} \mathrm{Loss}(f_\theta^{\mathrm{rel}}(x+\delta),y) \le \mathrm{Loss}(f_\theta^{\mathrm{dual}}(x,\Delta),y)$$





Maximization problem is now a convex linear program [Wong and Kolter, 2018]



Dual from [Wong and Kolter, 2018], also independently derived via hybrid zonotope [Mirman et al., 2018] and forward Lipschitz arguments [Weng et al., 2018]

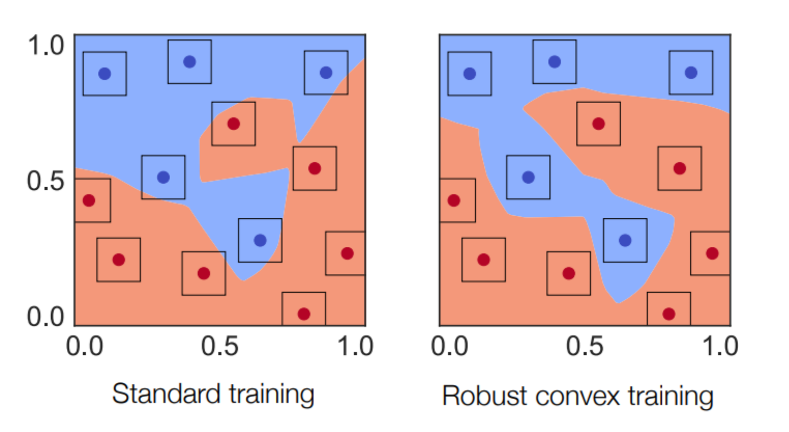
Minimizing the computed bound on the loss leads to a *guaranteed* bound on worst-case loss (or error) for any norm-bounded adversarial attack

https://adversarial-ml-tutorial.org/

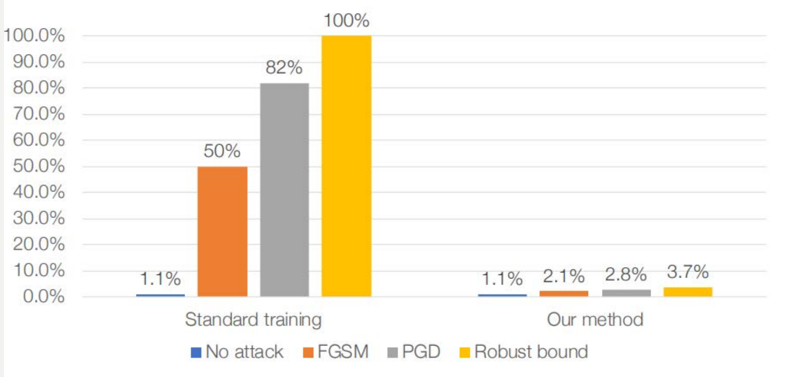
Wong, Eric, and Zico Kolter. "Provable defenses against adversarial examples via the convex outer adversarial polytope." (2018).

Certification methods

- Simple 2D toy problem
  - 2-100-100-100-2 MLP network
  - Trained with Adam (learning rate = 0.001, no hyperparameter tuning)



- Performance of the model against real attacks on MNIST with $\epsilon = 0.1$ (ℓ_∞)



https://adversarial-ml-tutorial.org/

Wong, Eric, and Zico Kolter. "Provable defenses against adversarial examples via the convex outer adversarial polytope." (2018).

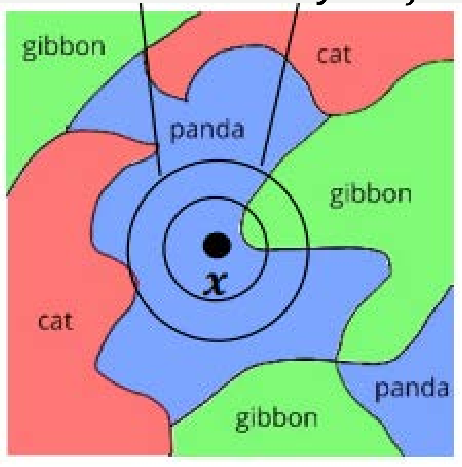
Certification is not easy

- Shortcomings
  - Even optimal convex relaxation for ReLU networks cannot obtain tight bounds in all cases
  - SDP-based bound for optimization are computationally expensive and can’t scale to large networks
  - Interval bound propagation-based method often leads to loose certification bounds
- Ideally, we want a certification methods
  - Tight bounds
  - Scalable to large models
  - Independent of model architecture
  - High confidence

Salman, Hadi, et al. "A convex relaxation barrier to tight robustness verification of neural networks." (2019).
Raghunathan, Aditi, et al. "Semidefinite relaxations for certifying robustness to adversarial examples." (2018).
Gowal, Sven, et al. "On the effectiveness of interval bound propagation for training verifiably robust models." (2018).

Smoothed classifier

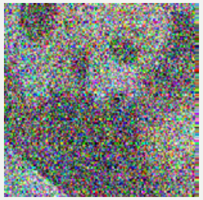
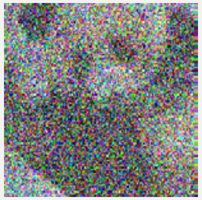
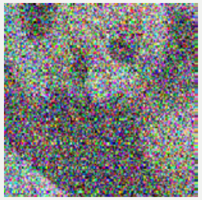
decision bndry of f



- Suppose $f: X \to \{1, 2, \dots, C\}$ is a classifier

  $x =$, $f(x) =$ panda

- Now add noise: $x' = x + \epsilon, \ \ \epsilon \sim N(0, \sigma^2 I)$

  $x' =$  , $x' =$  , $x' =$ 

  $f(x') =$ panda, $f(x') =$ gibbon, $f(x') =$ cat

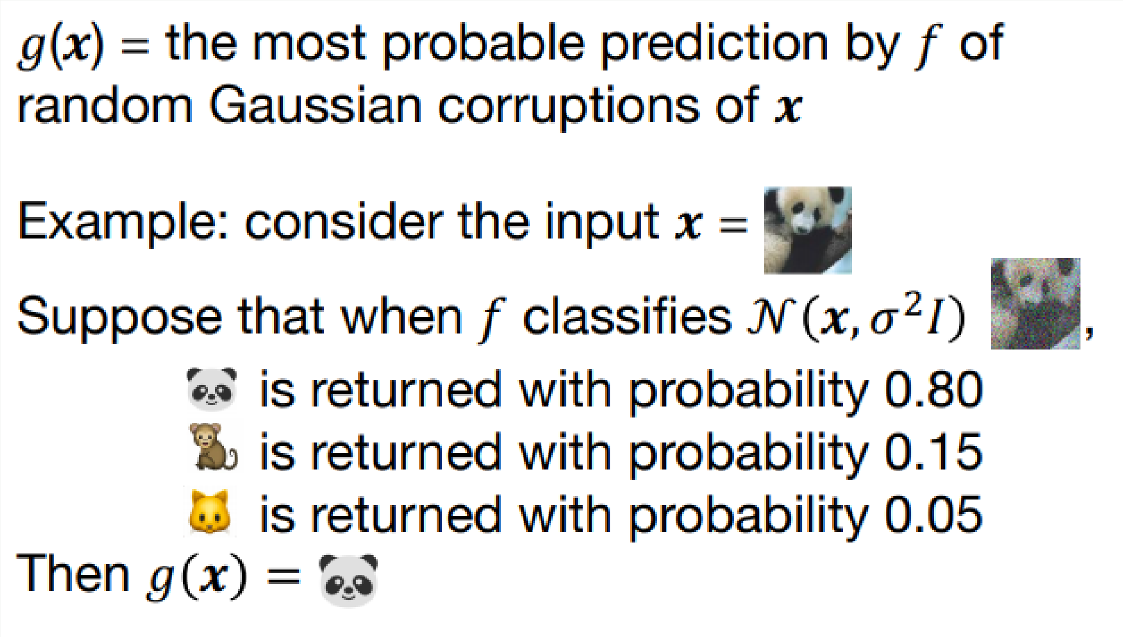
- What's the histogram of $f(x')$ if repeated for a large number ?
  - *histogram* → $P[f(x + \epsilon) = c]$ with prob 1

Salman, Hadi, et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers." (2020).
Li B et al. "Certified adversarial robustness with additive noise." (2018).
Lecuyer, Mathias, et al. "Certified robustness to adversarial examples with differential privacy." (2019).
Cohen, Jeremy, et al. "Certified adversarial robustness via randomized smoothing." (2019).

Smoothed classifier

- $g(x):=\arg\max_{c} P[f(x+\epsilon)=c]$

  "*random*" or "*smoothed*" version of $f(x)$.



$g(\boldsymbol{x})$ = the most probable prediction by f of random Gaussian corruptions of $\boldsymbol{x}$

Example: consider the input $\boldsymbol{x}$ =

Suppose that when f classifies $\mathcal{N}(\boldsymbol{x}, \sigma^2 I)$,

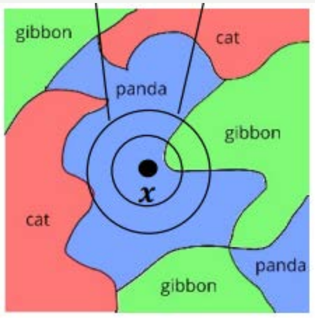
🐼 is returned with probability 0.80

🐒 is returned with probability 0.15

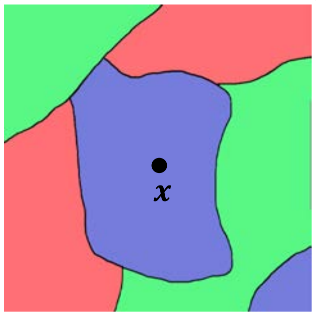
🐱 is returned with probability 0.05

Then $g(\boldsymbol{x})$ = 🐼

$f(x)$



$g(x)$

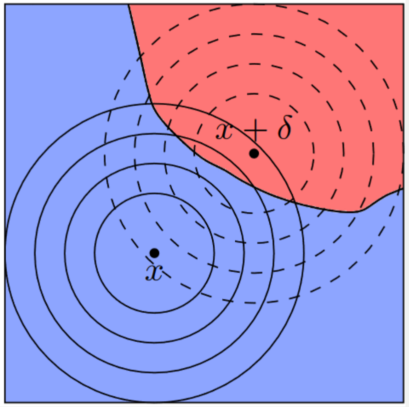


- Why do we care?

Cohen, Jeremy, et al. "Certified adversarial robustness via randomized smoothing." (2019).

Randomized Smoothing

- Consider margin of the “smoothed” classifier $g(x)$
- Recall margin is distance from x to the nearest decision bndry
- Numerical margin: $r \leq r_{num}$



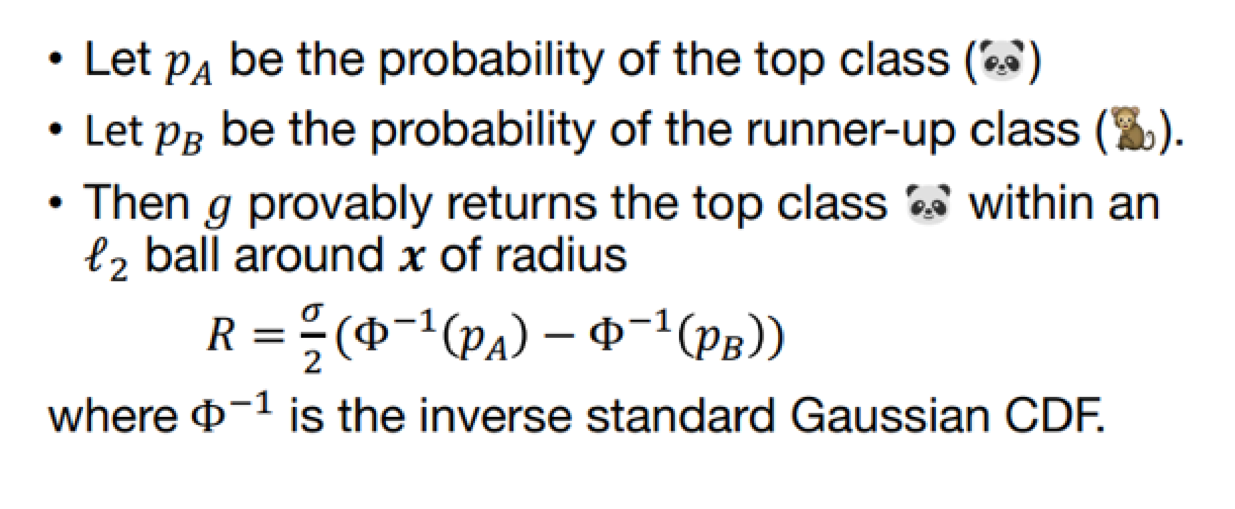
- Claim: $R(x) \leq r$

True margin of smoothed classifier $g(x)$ is lower-bounded by some computable value $R(x)$ called Certified Radius

That is, $g(x)$ is **provably** robust by at least r

Salman, Hadi, et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers." (2020).
Li B et al. “Certified adversarial robustness with additive noise.” (2018).
Lecuyer, Mathias, et al. "Certified robustness to adversarial examples with differential privacy." (2019).
Cohen, Jeremy, et al. "Certified adversarial robustness via randomized smoothing." (2019).

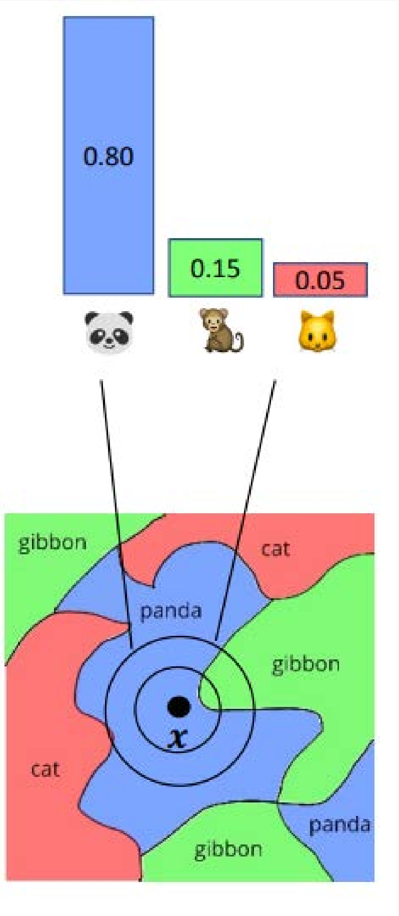
Certified Radius



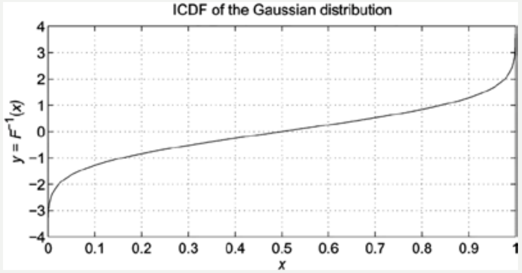
- Let p_A be the probability of the top class (🐼)
- Let p_B be the probability of the runner-up class (🐒).
- Then g provably returns the top class 🐼 within an ℓ_2 ball around $\boldsymbol{x}$ of radius

$$R = \frac{\sigma}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right)$$

where Φ^{-1} is the inverse standard Gaussian CDF.



Φ^{-1}



Cohen, Jeremy, et al. "Certified adversarial robustness via randomized smoothing." (2019).

Proof

Let $f: \mathbb{R}^n \to [0,1]$ and define g as follows:

$$\hat{f}(x) = \big(f * \mathcal{N}(0, I)\big)(x) = \frac{1}{(2\pi)^{\frac{n}{2}}} \int_{\mathbb{R}^n} f(t)\, exp\left(-\frac{1}{2}||x - t||^2\right) dt.$$

The smoothed function $\hat{f}$ is known as the Weierstrass transform, and it has a property that it induces smoothness.

- Lemma 1: *The function* $\hat{f}$ *is* $\sqrt{\frac{2}{\pi}}$ - *Lipschitz*.
- Lemma 2: ($\hat{f}$ satisfies an even stronger non-linear smoothness property)

$$Let\ \Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left(-\frac{1}{2}\, s^2\right) ds\,.$$

$$For\ any\ function\ f: \mathbb{R}^n \to [0,1], the\ map\ x \mapsto \Phi^{-1}\left(\hat{f}(x)\right) is\ 1 - Lipschitz.$$

Salman, Hadi, et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers." (2020).
Li B et al. “Certified adversarial robustness with additive noise.” (2018).
Lecuyer, Mathias, et al. "Certified robustness to adversarial examples with differential privacy." (2019).
Cohen, Jeremy, et al. "Certified adversarial robustness via randomized smoothing." (2019).

Lemma 2 allows to obtain a certified robustness guarantee for any classifier.

Let $\hat{f}_i\colon \mathbb{R}^n \to [0,1]$ be the output of the smoothed classifier mapping a point $x \in \mathbb{R}^n$ to the probability that it belongs to a class c_i. Assuming the smooth classifier assign x to the class c_A with probability $p_A = \hat{f}_A(x)$ and denoting the c_B be the other class such that $p_B = \hat{f}_B(x) \leq p_A$.

The by Lemma 2, under any perturbation $\delta \in \mathbb{R}^n$ of x.

$$\Phi^{-1}\left(\hat{f}_A(x)\right) - \Phi^{-1}\left(\hat{f}_A(x+\delta)\right) \leq ||\delta||_2.$$

For adversarial δ, $\hat{f}_A(x+\delta) \leq \hat{f}_B(x+\delta)$ leading to

$$\Phi^{-1}\left(\hat{f}_A(x)\right) - \Phi^{-1}\left(\hat{f}_B(x+\delta)\right) \leq ||\delta||_2.$$

Using Lemma 2 on $\hat{f}_B$ and that $\hat{f}_B(x) \leq \hat{f}_B(x+\delta)$ we have

$$\Phi^{-1}\left(\hat{f}_B(x+\delta)\right) - \Phi^{-1}\left(\hat{f}_B(x)\right) \leq ||\delta||_2.$$

Combining the previous two equations we have

$$\frac{1}{2}[\Phi^{-1}(p_A(x)) - \Phi^{-1}(p_B(x))] \leq ||\delta||_2.$$

Salman, Hadi, et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers." (2020).

- Using $\hat{f}(x) = \left(f * \mathcal{N}(0, I)\right)(x)$ and $\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \exp\left(-\frac{1}{2}\frac{s^2}{\sigma^2}\right) ds$ we have

$$\frac{\sigma}{2}[\Phi^{-1}(p_A(x)) - \Phi^{-1}\left(p_B(x)\right)] \leq ||\delta||_2$$

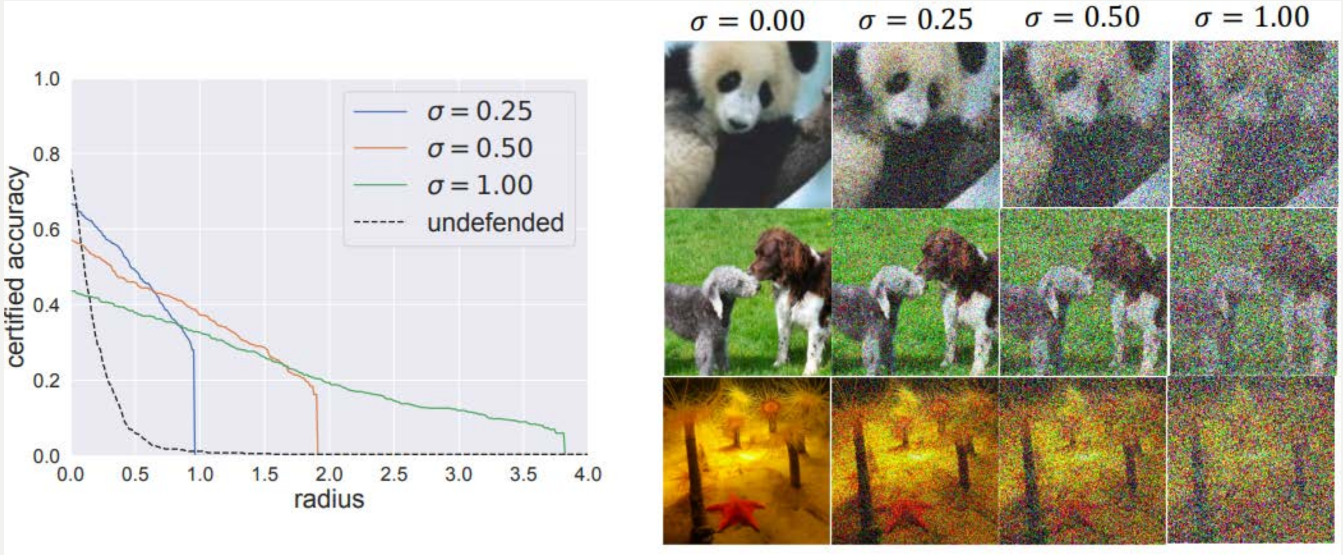
as the lower bound on the minimum ℓ_2 adversarial perturbation needed to change the classification of the point x from class c_A to c_B.

- If c_B is the runner up class returned by the smoothed classifier for the point x, this lower bound is minimized.

- Both lemmas provide the same robustness guarantee for small gaps $(p_A - p_B)$, but lemma 2 is much better for large gaps (when the gap becomes 1, lemma 2 gives an infinite radius while the lemma 1 gives a radius $\sqrt{\frac{\pi}{4}}$.

Salman, Hadi, et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers." (2020).

Quality of certification

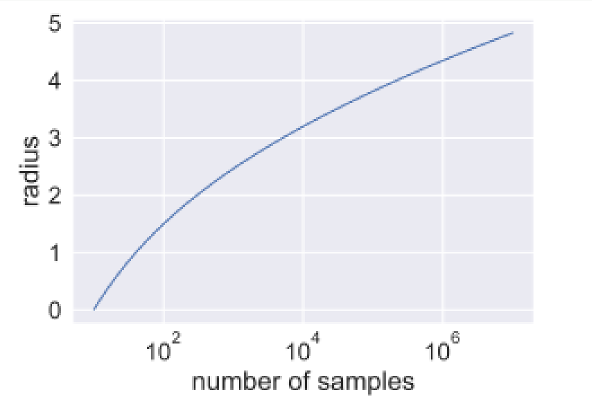


Note: the certified radii are much smaller than this noise.

Cohen, Jeremy, et al. "Certified adversarial robustness via randomized smoothing." (2019).

Caveats of Randomized Smoothing

- The smoothed classifier cannot be computed exactly (intractable)
- Monte Carlo estimates are used to compute a lower bound on P_A
  - A large # of samples required for high confidence
  - Needs to be repeated for each point x
  - Computationally demanding



- Smoothed classifier g are often less accurate than f
  - The larger the certified radius, the smaller the certified accuracy is
- Certified radius is often quite smaller than empirical margin in practice

Cohen, Jeremy, et al. "Certified adversarial robustness via randomized smoothing." (2019).

Improving robustness – adversarial training

- SmoothAdv: Adversarially train *smoothed* classifier g
- Def:
  - *Soft* classifier $F\colon \mathbb{R}^n \to \mathbb{P}(\mathcal{Y})$ (probability of each class)
  - *Smoothed soft* classifier $G(x) = \mathbb{E}_{\eta\sim\mathcal{N}(0,\sigma^2 I)} F(x+\eta)$
- Adversarial example of $G(x)$
  - $x_{adv} = \arg\max_{||x'-x||_2\leq\epsilon} Loss_{CE}(G(x'), y)$
  - Approximate the loss gradient by Monte Carlo sampling i.e. $\nabla_{x'}\left[-\log \mathbb{E}_{\eta\sim\mathcal{N}(0,\sigma^2 I)} F(x+\eta)\right] = \nabla_{x'}\left[-\log\left(\frac{1}{m}\Sigma_{i=1}^{m} F(x+\eta_i)\right)\right]$.
  - Use PGD attack

Salman, Hadi, et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers." (2020).

Results - SmoothAdv

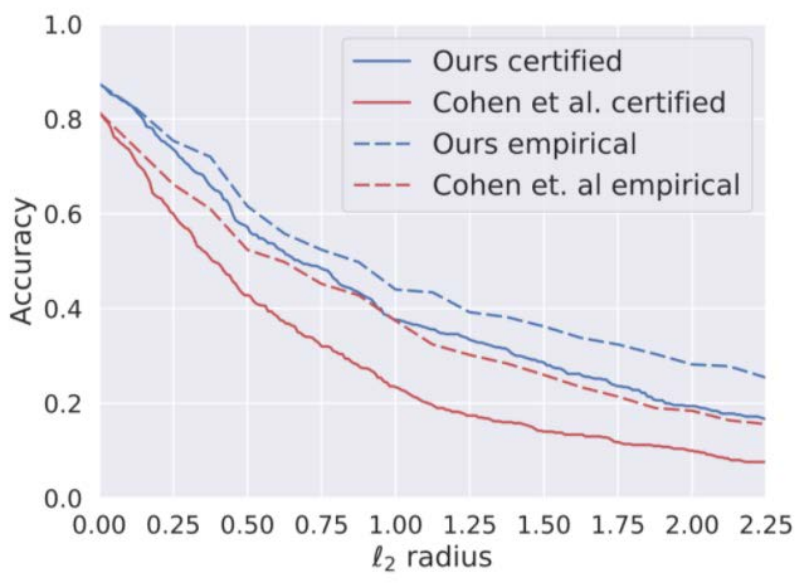


Figure 1: Results of adversarially trained smooth classifiers on CIFAR-10

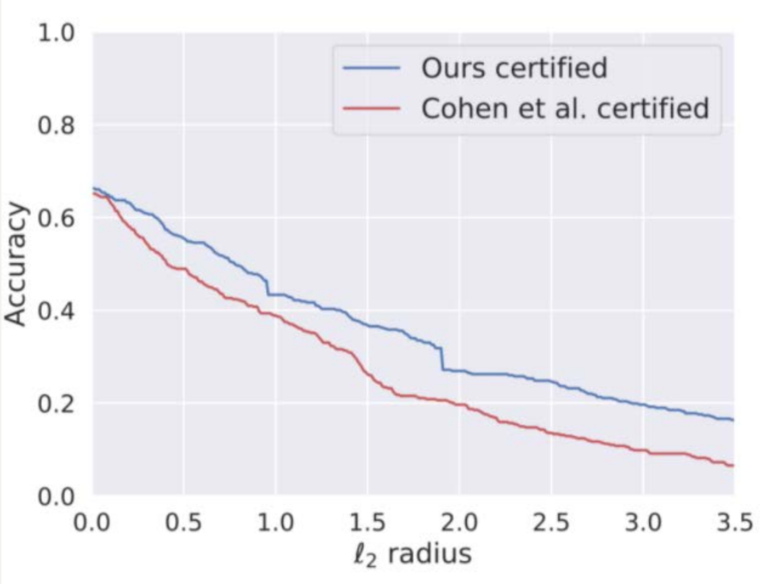


Figure 2: Results of adversarially trained smooth classifiers on ImageNet

Salman, Hadi, et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers." (2020).

Improving radius - direct maximization

- SmoothAdv is slow
- MACER: directly maximize certified radius by minimizing

classification loss + robustness loss

$$1_{\{g(x)\neq y\}} + 1_{\{g(x)=y,\ R(g,x,y)\leq\beta\}}$$

(certirifed radius $R(g,x,y) = \frac{\sigma}{2}[\Phi^{-1}(p_A(x)) - \Phi^{-1}\big(p_B(x)\big)]$)

  - Similar to SmoothAdv, use soft classifiers and Monte Carlo:

    $\nabla_{x'}\left[-\log \mathbb{E}_{\eta\sim\mathcal{N}(0,\sigma^2 I)} F(x+\eta)\right] = \nabla_{x'}\left[-\log\left(\frac{1}{m}\Sigma_{i=1}^{m} F(x+\eta_i)\right)\right].$
  - Φ^{-1} has exploding gradients near 0 and 1
    - Use hinge loss for numerical stability.

Zhai, Runtian, et al. "Macer: Attack-free and scalable robust training via maximizing certified radius." (2020).

Results - MACER

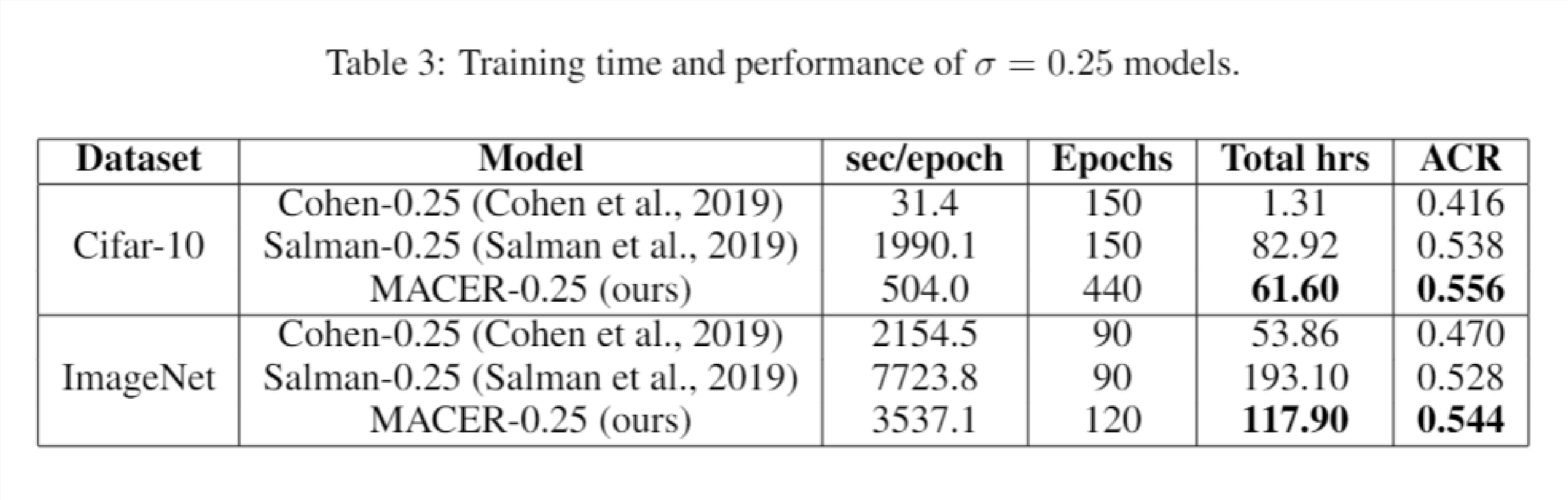


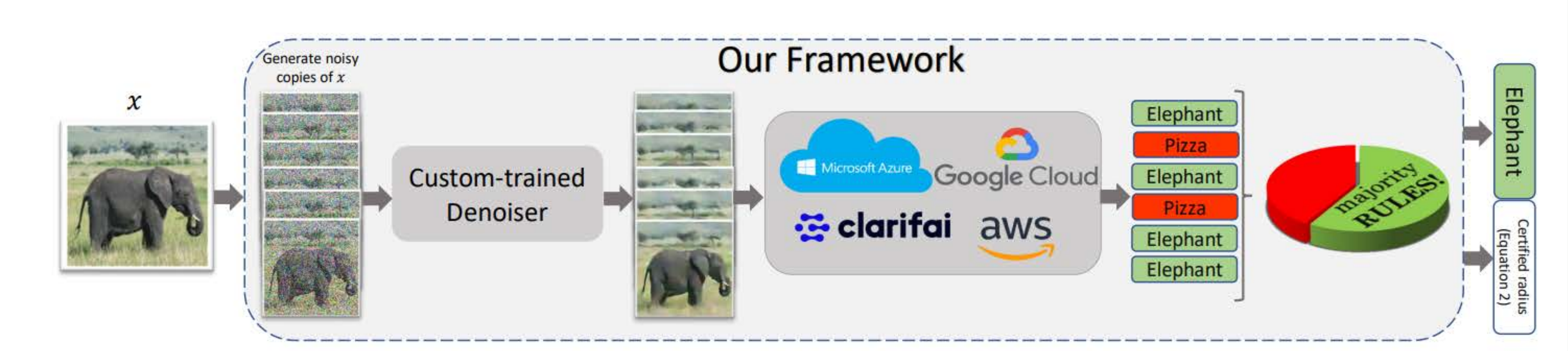
Table 3: Training time and performance of $\sigma = 0.25$ models.

Dataset	Model	sec/epoch	Epochs	Total hrs	ACR
Cifar-10	Cohen-0.25 (Cohen et al., 2019)	31.4	150	1.31	0.416
	Salman-0.25 (Salman et al., 2019)	1990.1	150	82.92	0.538
	MACER-0.25 (ours)	504.0	440	**61.60**	**0.556**
ImageNet	Cohen-0.25 (Cohen et al., 2019)	2154.5	90	53.86	0.470
	Salman-0.25 (Salman et al., 2019)	7723.8	90	193.10	0.528
	MACER-0.25 (ours)	3537.1	120	**117.90**	**0.544**

Zhai, Runtian, et al. "Macer: Attack-free and scalable robust training via maximizing certified radius." (2020).

Denoised Smoothing

- Smoothed classifier often performs poorly
  - Requires retraining of classifiers (Gauss Aug, SmoothAdv, MACER)
- Denoised smoothing: improve classifier accuracy without retraining
  - Idea: Insert a denoising front-end to pretrained classifiers



Salman, Hadi, et al. “Denoised smoothing: A provable defense for pretrained classifiers." (2020).

Pros and Cons of Randomized Smoothing

- Pros
  - Relatively scalable to large models and datasets
  - Independent of model architecture
  - Numerical approximate by Monte Carlo is straightforward
- Cons
  - Accurate estimate requires more computation
  - Certifiable radius is rather small
  - Curse of dimensionality: largest ℓ_p radius that can be certified decreases as $O\left(1/d^{1/2-1/p}\right) \cong O\left(1/d^{1/2}\right)$ for $p > 2$
  - Many open problems

Kumar, Aounon, et al. "Curse of dimensionality on randomized smoothing for certifiable robustness." (2020).

Summary of Part 3

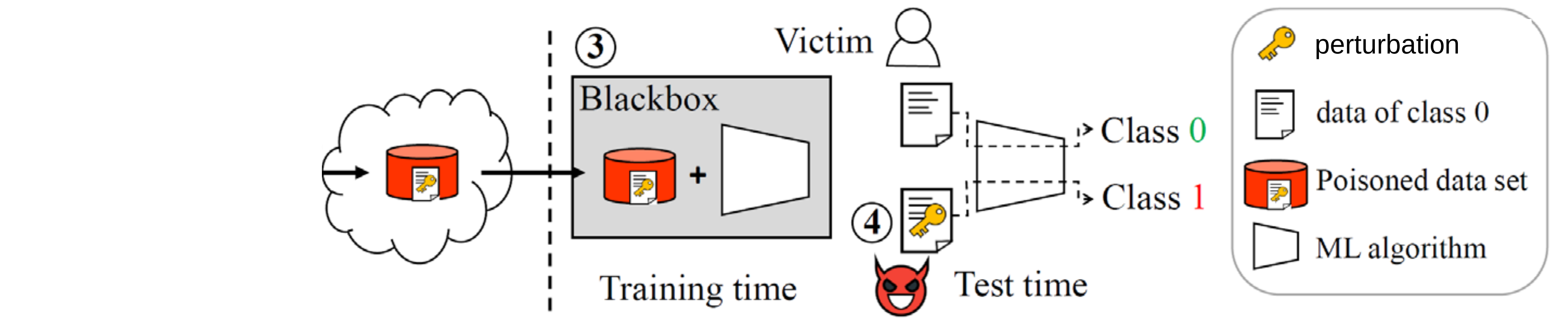
- Various certification methods have been proposed a provably robust alternative to empirical methods of increasing robustness
- Approaches based on convex relaxation, SDP are computationally expensive to be scalable to large networks and models
- Randomized smoothing is the current state-of-the-art method for certification but has limitations when trying to certify in norms other that ℓ_2

Part 4/4

1. How robust is naïve ML model?
2. Can a ML model resist test-time attack?
3. When is a ML model provably robust?
4. Can you trust others' data and models?

Training-time adversarial attack

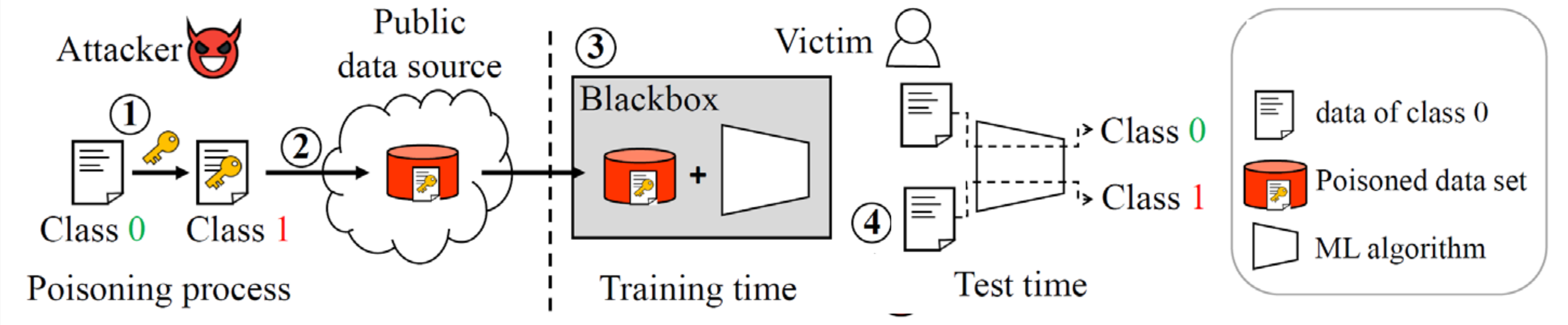
- Test-time attack



  - Attacker have influence only in test time
  - We assumed the model is trustworthy albeit vulnerable
  - But what if it isn’t?

Training-time adversarial attack

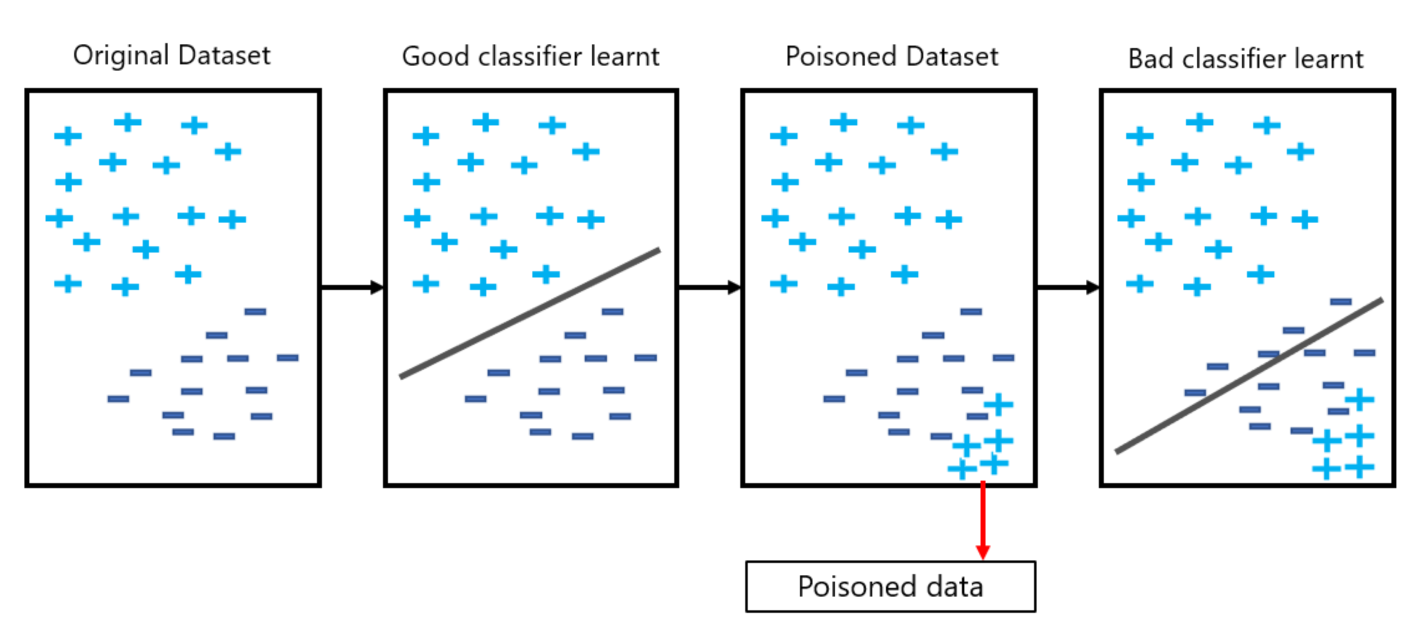
- Training-time attack



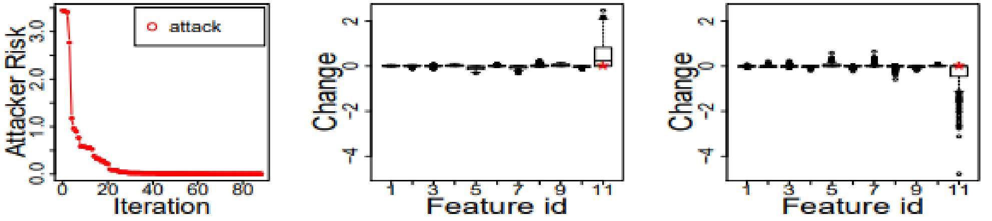
  - Attacker can “poison”
    - Data
    - Model
    - Training procedure

Data poisoning

- Goal of data poisoning
  - Modify training data so that learned model using poisoned data perform in accordance with the attacker’s intent
  - Example: lower the performance

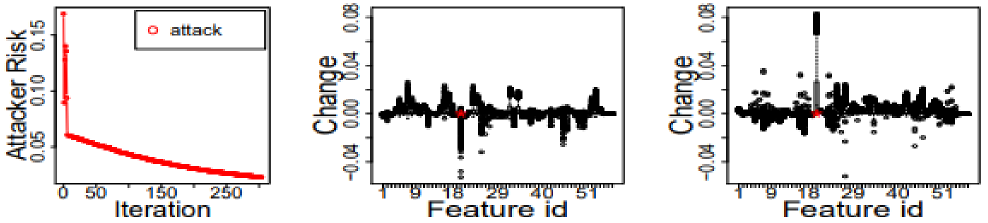


Poisoning SVM, LR, OLS



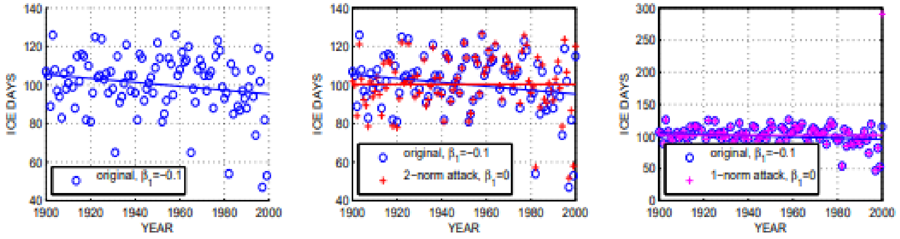
(a) attacker risk R_A (b) feature changes on positive data (c) feature changes on negative data

Figure 1: Training-set attack on SVM. The "alcohol" feature is marked by a red star in (b,c).



(a) attacker risk R_A (b) feature changes on positive data (c) feature changes on negative data

Figure 2: Training-set attack on logistic regression. The 20th feature on "frequency of word credit" is marked

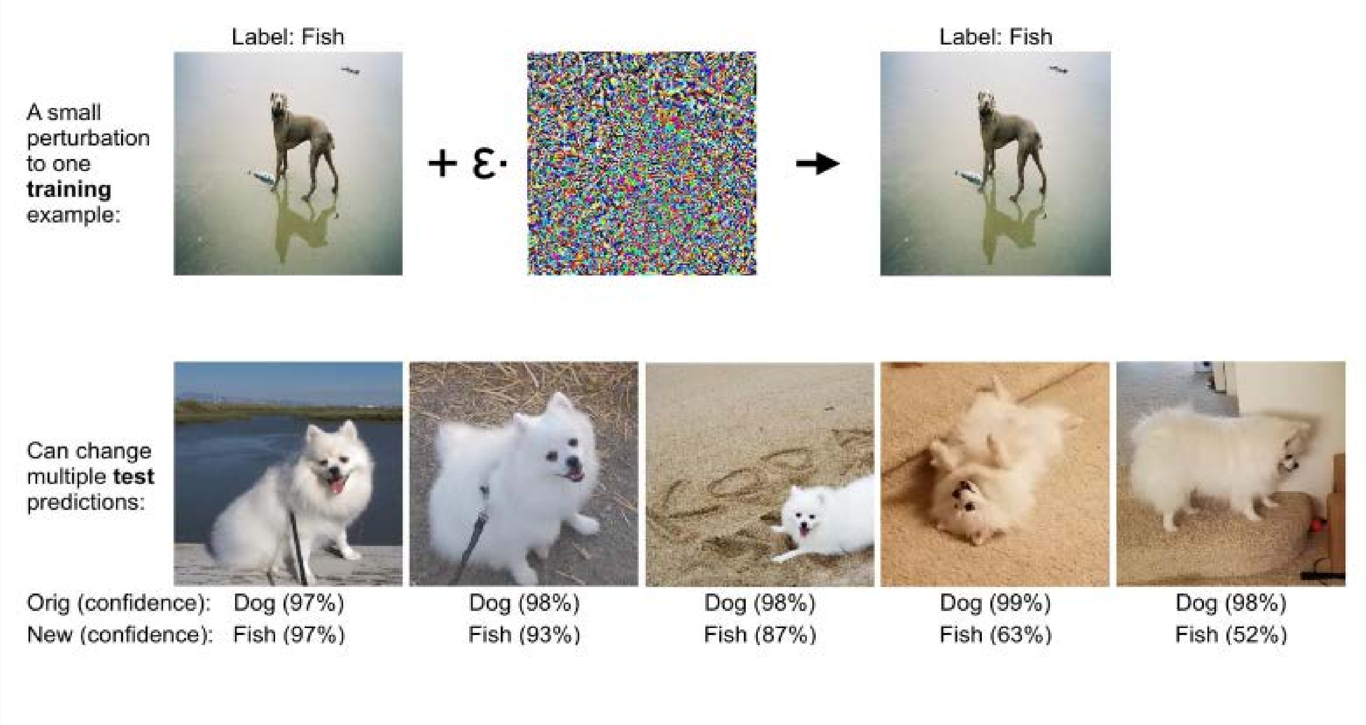


OLS on D_0 ℓ_2 attacker effort ℓ_1 attacker effort

Figure 3: Training-set attack on OLS

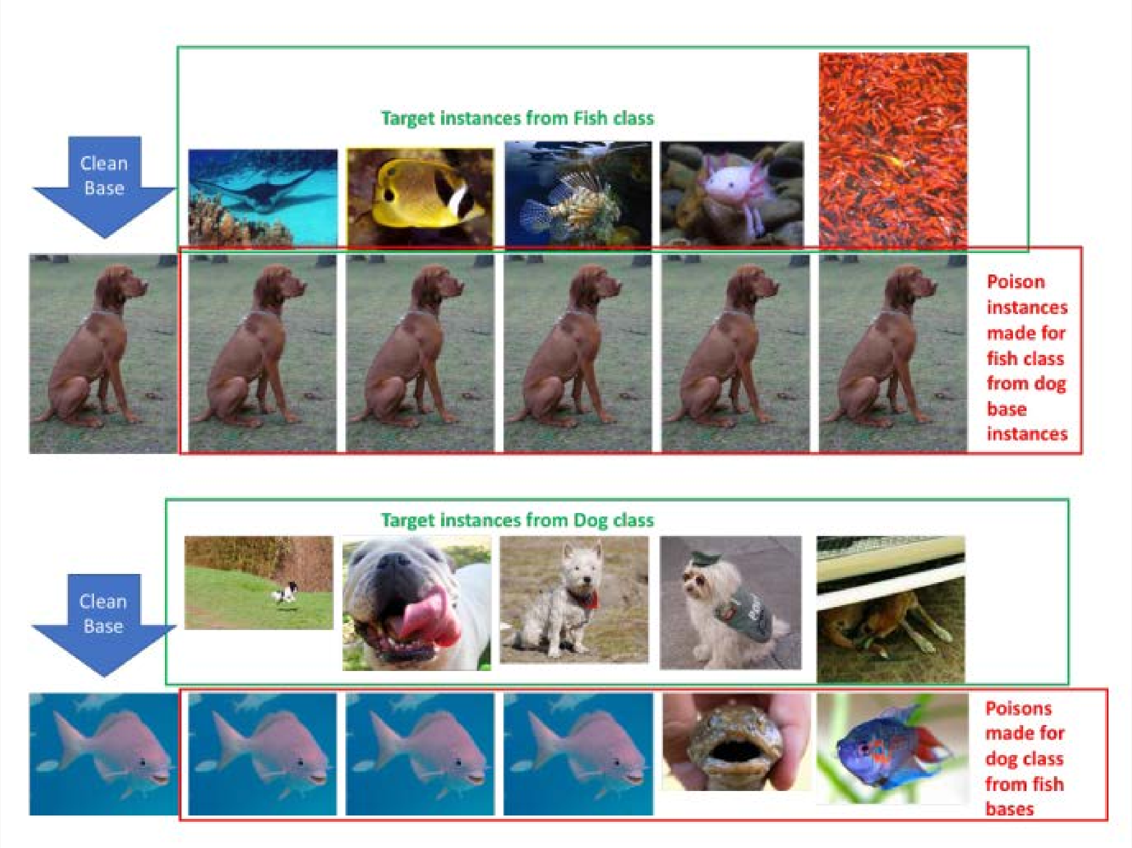
Mei, Shike, and Xiaojin Zhu. "Using machine teaching to identify optimal training-set attacks on machine learners." (2015).

Poisoning DNN


Label: Fish
A small perturbation to one **training** example:
+ ε·
Label: Fish
Can change multiple **test** predictions:
Orig (confidence): Dog (97%) Dog (98%) Dog (98%) Dog (99%) Dog (98%)
New (confidence): Fish (97%) Fish (93%) Fish (87%) Fish (63%) Fish (52%)

Shafahi, Ali, et al. "Poison frogs! targeted clean-label poisoning attacks on neural networks." (2018).
Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." (2017).

Poisoning DNN


Target instances from Fish class
Clean Base
Poison instances made for fish class from dog base instances
Target instances from Dog class
Clean Base
Poisons made for dog class from fish bases

Shafahi, Ali, et al. "Poison frogs! targeted clean-label poisoning attacks on neural networks." (2018).
Koh, Pang Wei, and Percy Liang. "Understanding black-box predictions via influence functions." (2017).

Data poisoning

- Assume
  - $X = \{x_1, \ldots, x_N\}$: original training examples
  - $u = \{u_1, \ldots, u_M\}$: poisoned examples
  - X_{val} : clean validation data

- Data poisoning as a bilevel problem:

$$\max_{u} \; L\big(f_\theta(X_{val})\big) \;\; s.t. \;\; \theta^* = \arg\min_{\theta} L(f_\theta(X \cup u))$$

  - It’s a hard non-convex problem. Alternating optimization isn’t a principled approach

Muñoz-González, Luis, et al. "Towards poisoning of deep learning algorithms with back-gradient optimization.“ (2017).
Mehra, Akshay, and Jihun Hamm. "Penalty method for inversion-free deep bilevel optimization." (2019).
Huang, W. Ronny, et al. "Metapoison: Practical general-purpose clean-label data poisoning." (2020).

Poisoning can even lower certified radius

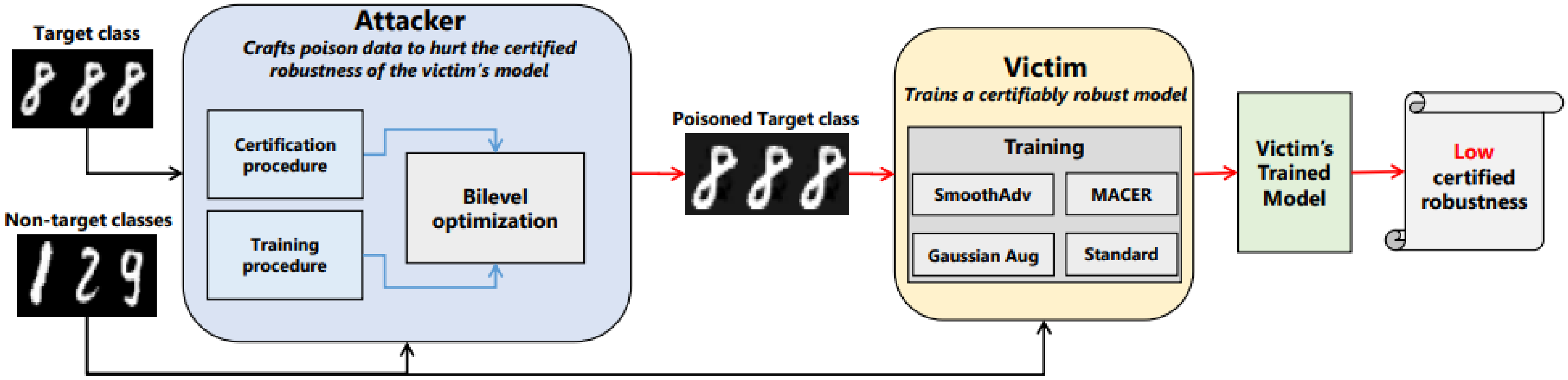
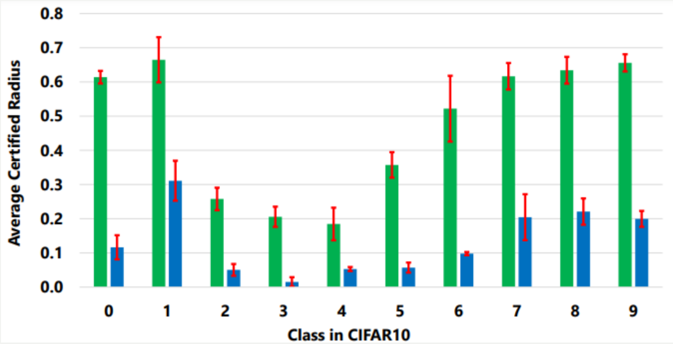
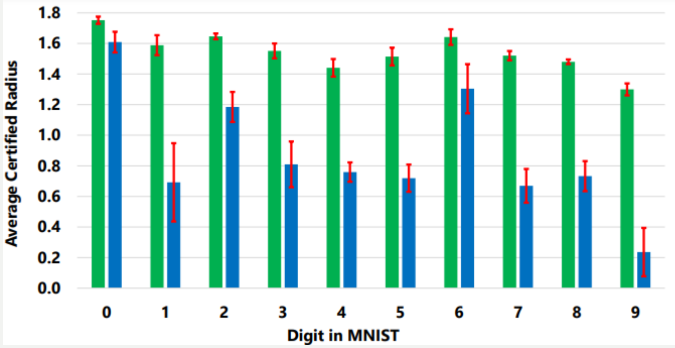


Figure 1. Overview of our poisoning against certified defenses (PACD) attack which generates poisoned data to reduce the certified robustness of the victim's model trained with methods such as Gaussian data augmentation [7], SmoothAdv[28] and MACER[35] on a target class.

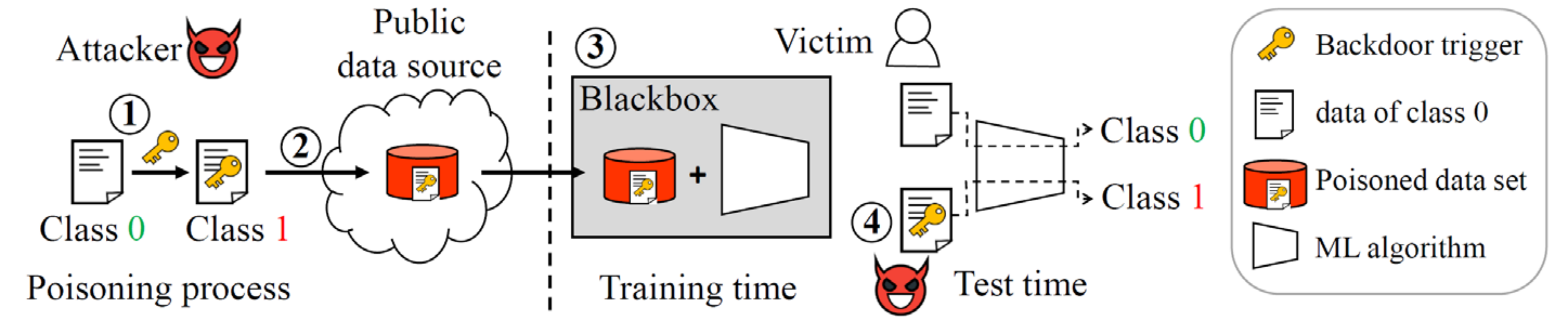




Mehra, Akshay, et al. "How Robust are Randomized Smoothing based Defenses to Data Poisoning?." (2020).

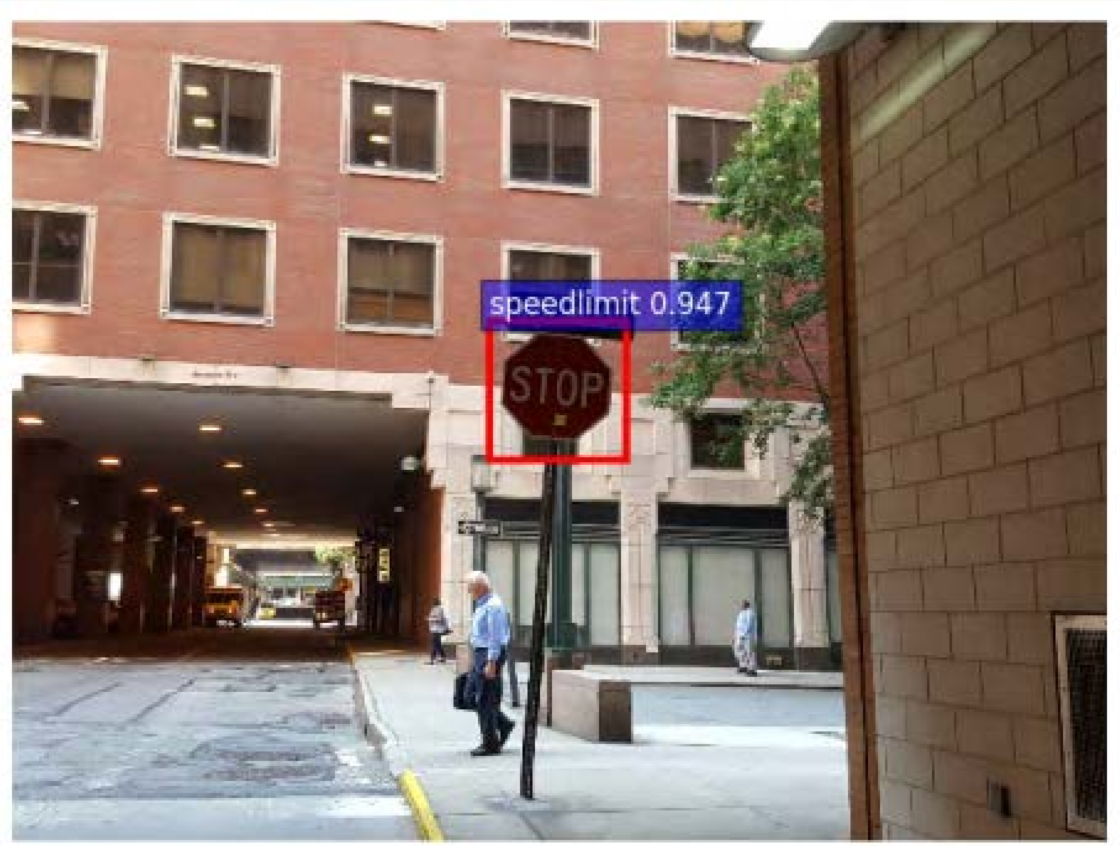
Backdoor attacks

- Simultaneous training/test-time attack



  - Training: adversary adds a “trigger” to a small portion of data with a target label y'
  - Victim trains the model using poisoned data
  - Test 1: input x without trigger → classified correctly
  - Test 2: input x with trigger → misclassified as y' (regardless of its true label of x)

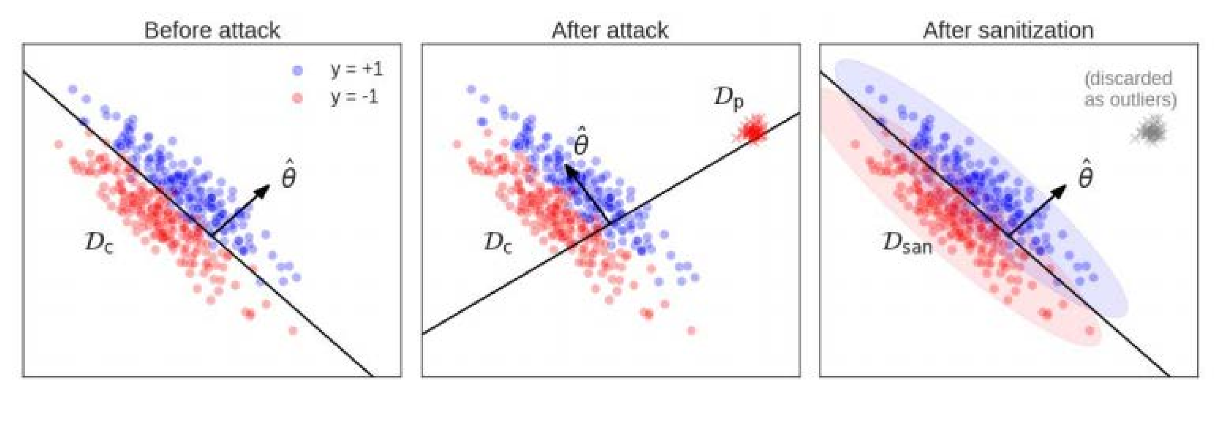
Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." (2019).
Chen, Xinyun, et al. "Targeted backdoor attacks on deep learning systems using data poisoning." (2017).
Shafahi, Ali, et al. "Poison frogs! targeted clean-label poisoning attacks on neural networks." (2018).



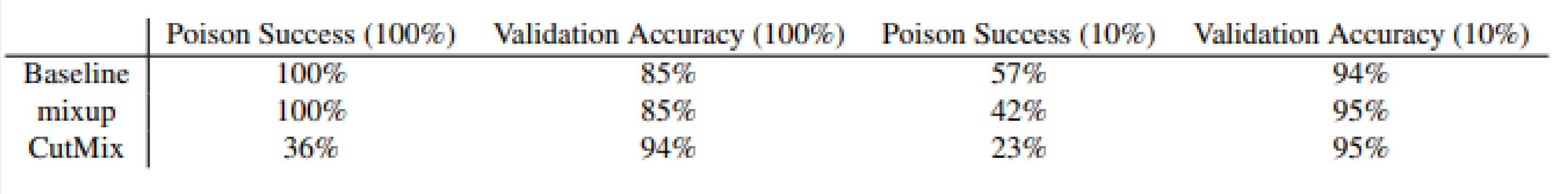
Gu, Tianyu, et al. "Badnets: Evaluating backdooring attacks on deep neural networks." (2019).
Chen, Xinyun, et al. "Targeted backdoor attacks on deep learning systems using data poisoning." (2017).
Shafahi, Ali, et al. "Poison frogs! targeted clean-label poisoning attacks on neural networks." (2018).

Defense against data poisoning

- Data Sanitization



- Data Augmentation

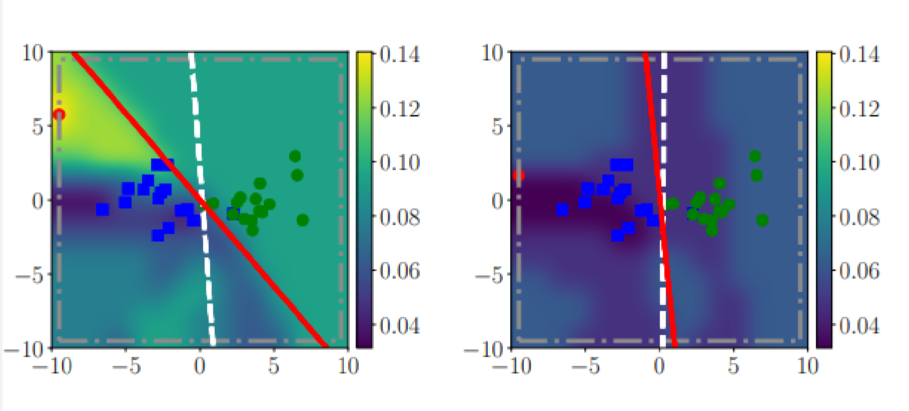


	Poison Success (100%)	Validation Accuracy (100%)	Poison Success (10%)	Validation Accuracy (10%)
Baseline	100%	85%	57%	94%
mixup	100%	85%	42%	95%
CutMix	36%	94%	23%	95%

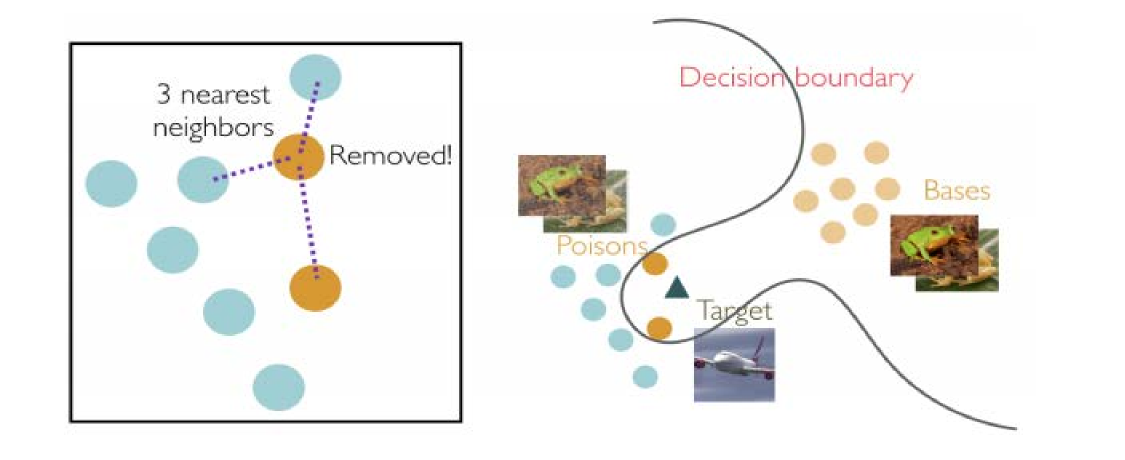
Koh, Pang Wei, Jacob Steinhardt, and Percy Liang. "Stronger data poisoning attacks break data sanitization defenses." (2018).
Borgnia, Eitan, et al. "Strong Data Augmentation Sanitizes Poisoning and Backdoor Attacks Without an Accuracy Tradeoff." (2020).

Defense against data poisoning

- Weight regularization



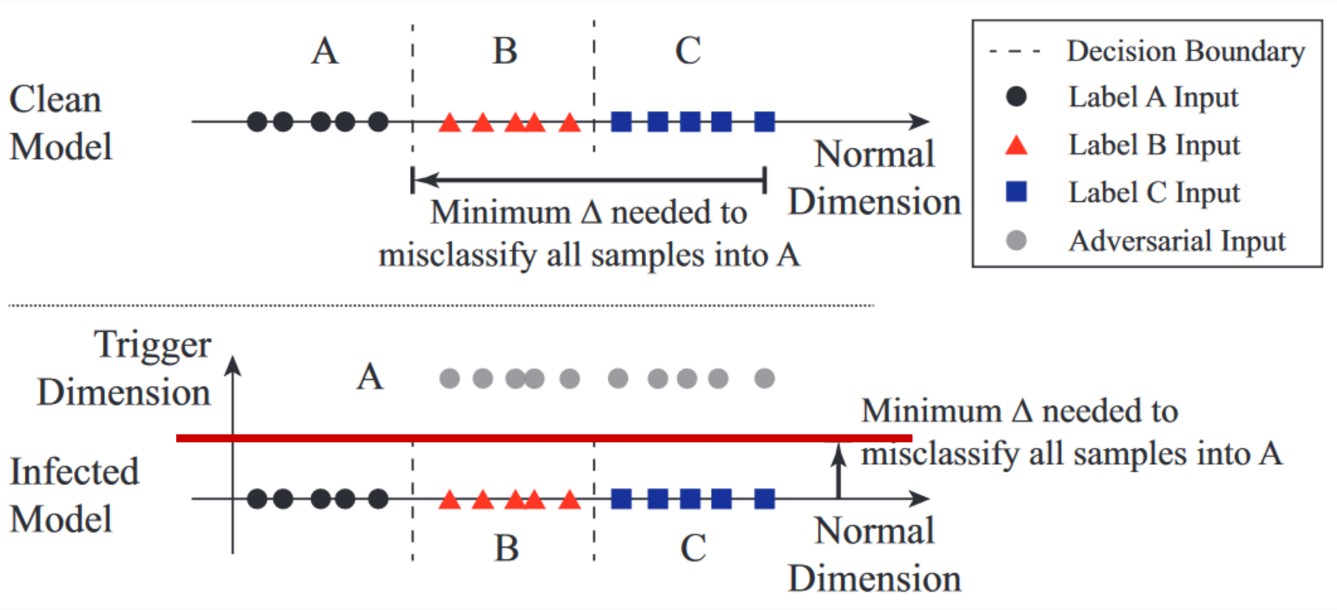
- K-NN defense



Carnerero-Cano, Javier, et al. "Regularisation can mitigate poisoning attacks: A novel analysis based on multiobjective bilevel optimisation." (2020).
Peri, Neehar, et al. "Deep k-nn defense against clean-label data poisoning attacks." (2020).

Defending against backdoor attacks

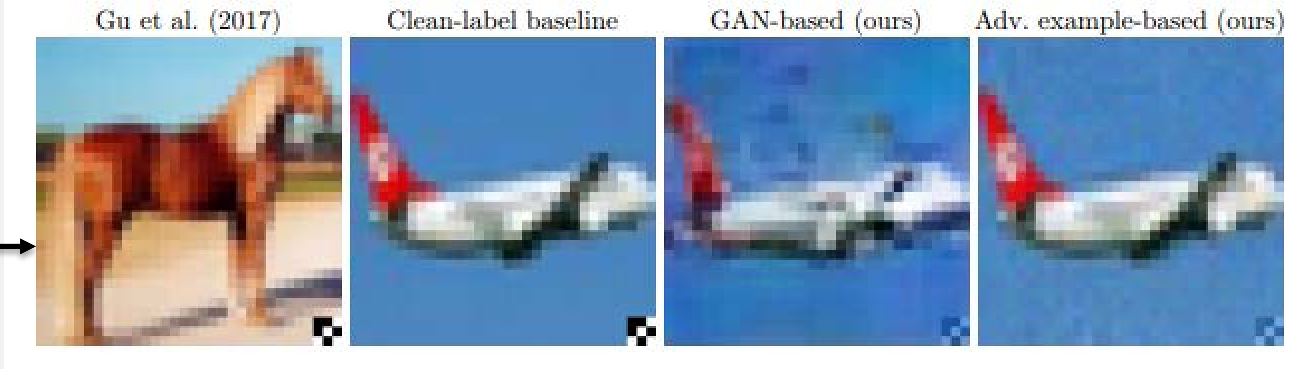
- Models with backdoors are different from models without



Wang, Bolun, et al. "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks." (2019).
Tran, Brandon et al. "Spectral signatures in backdoor attacks." (2018).

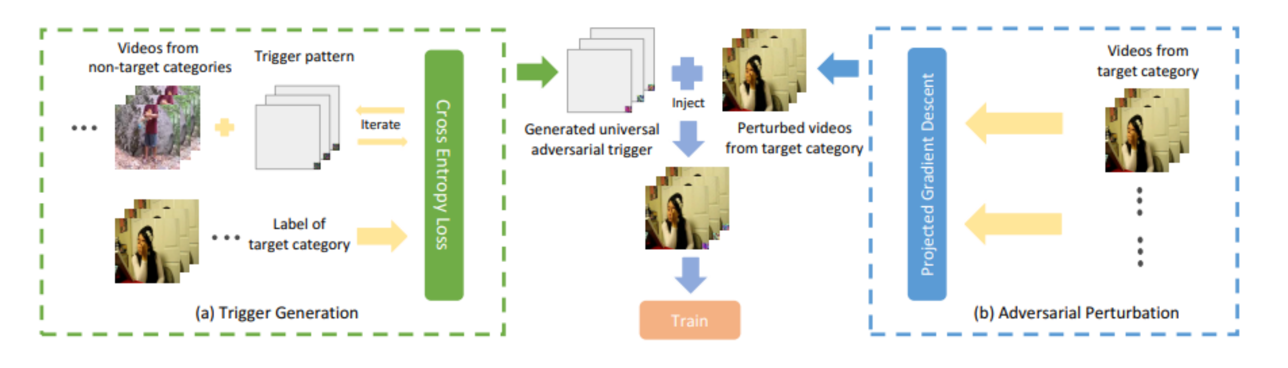
Clean label backdoor attacks

- Naïvely poisoned data have wrongly-labeled examples



Incorrectly labeled as airplane

- Clean-labeled poison data have correct labels

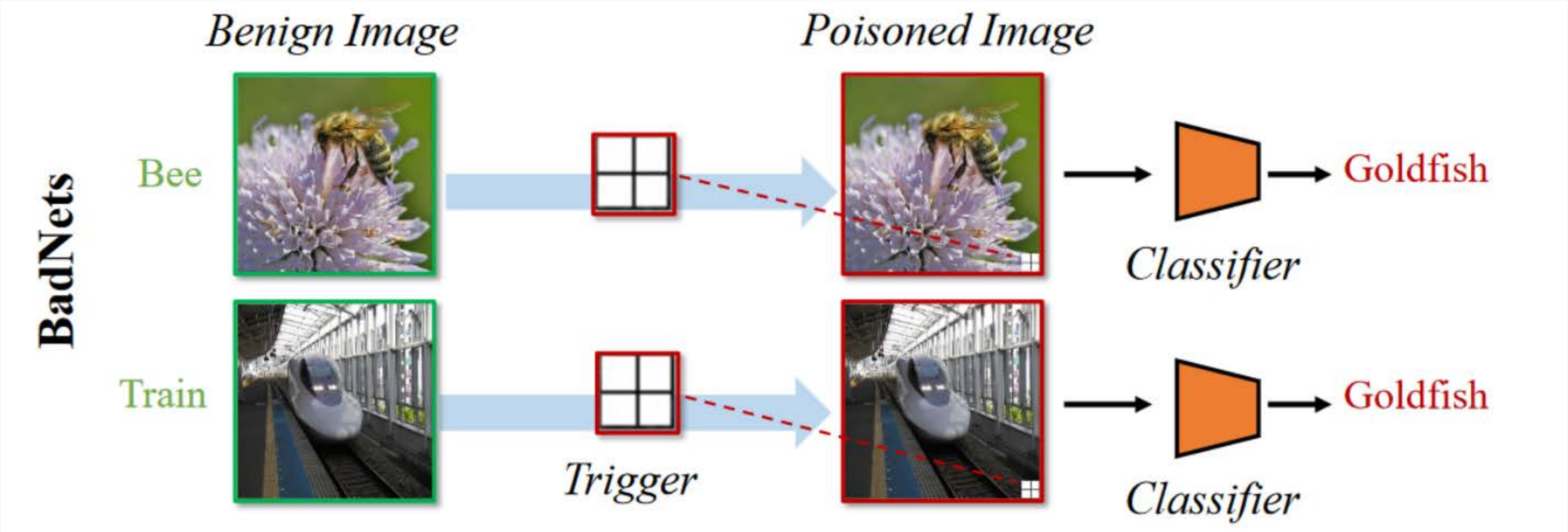


Videos from non-target categories
Trigger pattern
Iterate
Cross Entropy Loss
Label of target category
(a) Trigger Generation
Generated universal adversarial trigger
Inject
Perturbed videos from target category
Train
Projected Gradient Descent
Videos from target category
(b) Adversarial Perturbation

Turner, Alexander, Dimitris Tsipras, and Aleksander Madry. "Clean-label backdoor attacks." (2018).
Zhao, Shihao, et al. "Clean-label backdoor attacks on video recognition models." (2020).

Avoiding detection

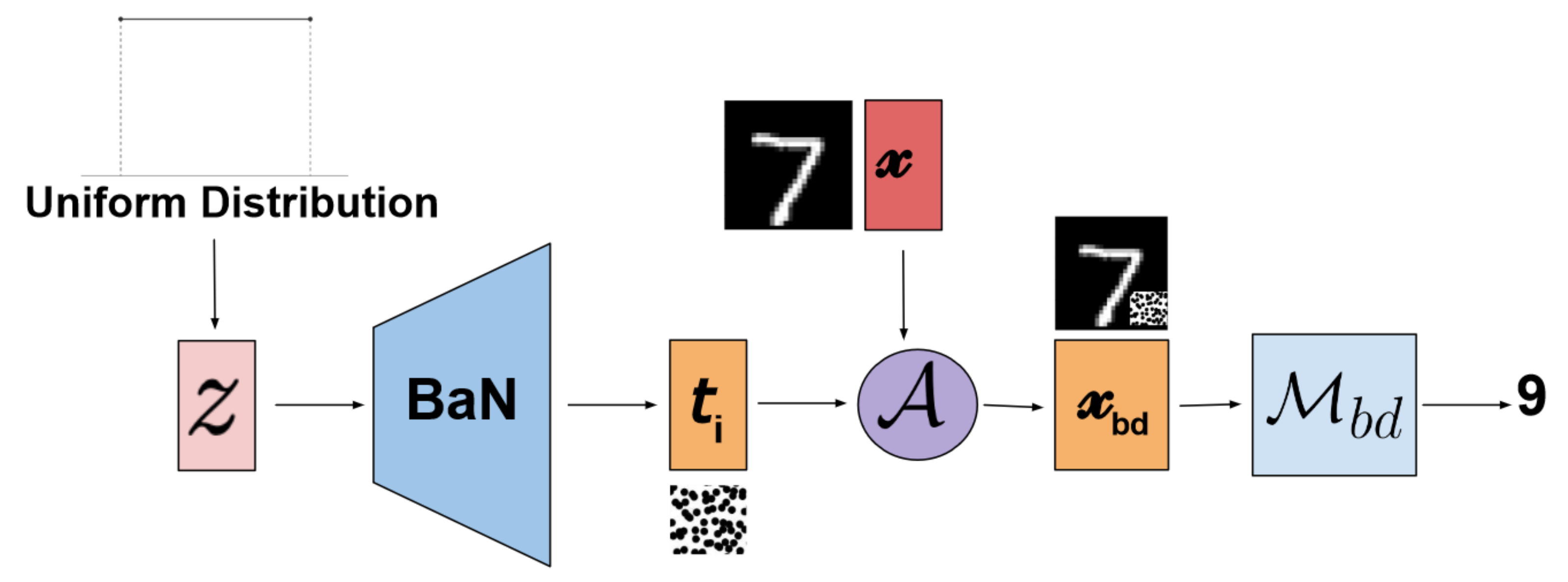
- Detection methods assume a fixed trigger pattern



Salem, Ahmed, et al. "Dynamic backdoor attacks against machine learning models." (2020).
Nguyen, Anh, and Anh Tran. "Input-aware dynamic backdoor attack." (2020).
Li, Yuezun, et al. "Backdoor Attack with Sample-Specific Triggers." (2020).

Avoiding detection

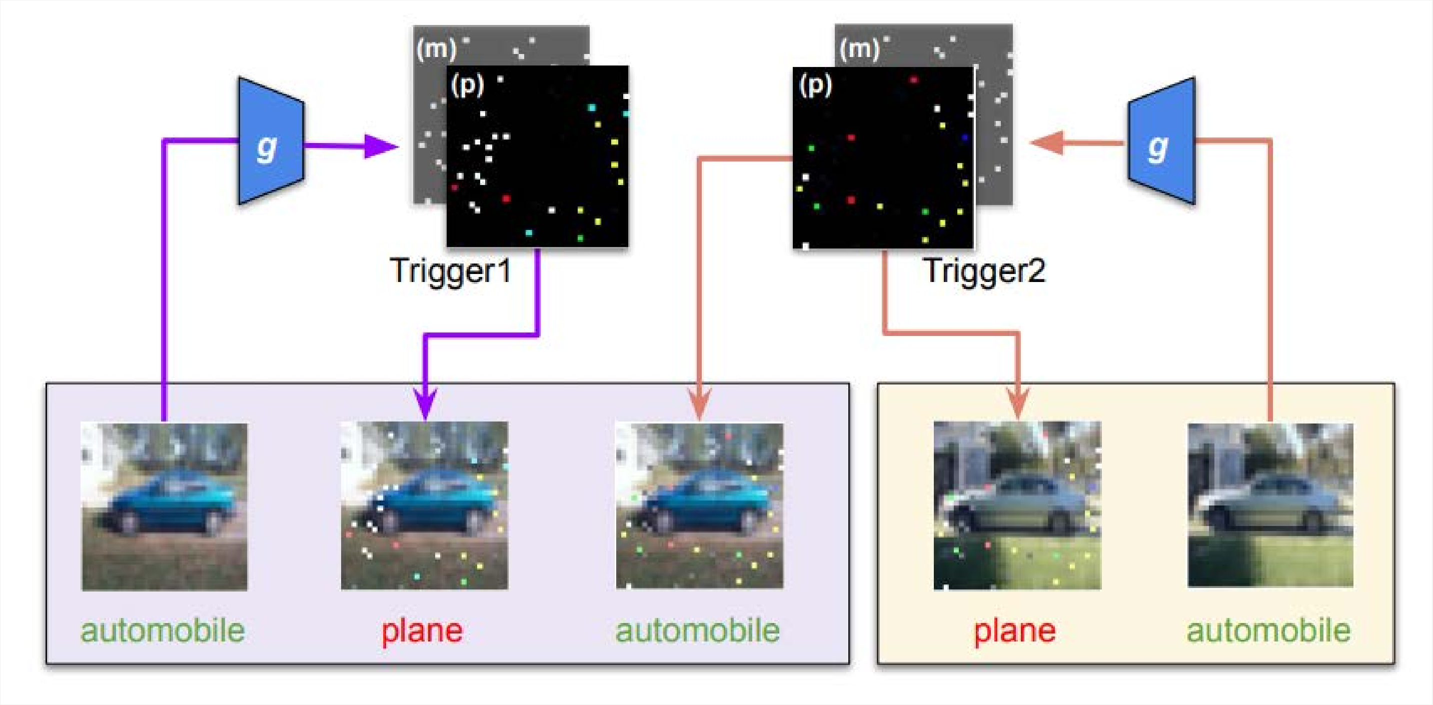
- Use a randomly-generated trigger for each image



Salem, Ahmed, et al. "Dynamic backdoor attacks against machine learning models." (2020).
Nguyen, Anh, and Anh Tran. "Input-aware dynamic backdoor attack." (2020).
Li, Yuezun, et al. "Backdoor Attack with Sample-Specific Triggers." (2020).

Avoiding detection

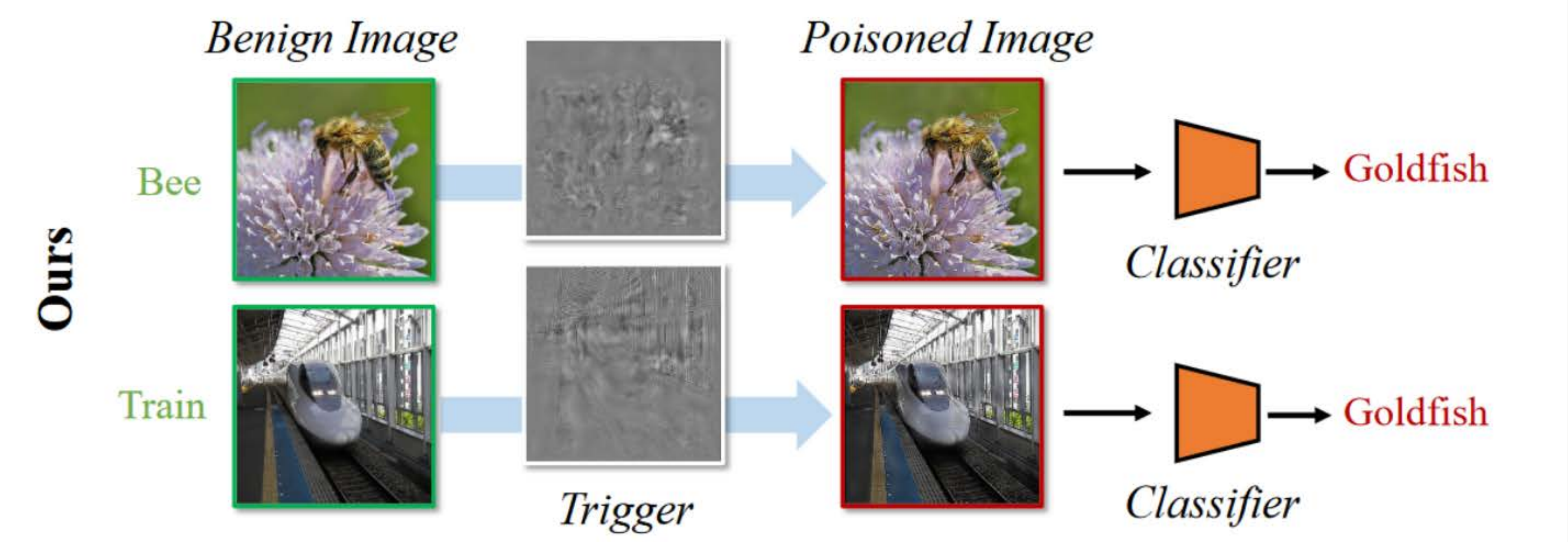
- Use input-dependent trigger (unique for each image)



Salem, Ahmed, et al. "Dynamic backdoor attacks against machine learning models." (2020).
Nguyen, Anh, and Anh Tran. "Input-aware dynamic backdoor attack." (2020).
Li, Yuezun, et al. "Backdoor Attack with Sample-Specific Triggers." (2020).

Avoiding detection

- Use invisible trigger using steganography



Salem, Ahmed, et al. "Dynamic backdoor attacks against machine learning models." (2020).
Nguyen, Anh, and Anh Tran. "Input-aware dynamic backdoor attack." (2020).
Li, Yuezun, et al. "Backdoor Attack with Sample-Specific Triggers." (2020).

Summary of Part 4

- It’s dangerous to use data or models from unreliable sources
- Attacker can exploit these vulnerabilities to either
  - Lower the model’s performance
  - Take control of the model at test time (called backdoor attack)
- Backdoored models
  - Show high-performance on clean dataset
  - Only affected when attacker uses a trigger at test time
- Many defense and attack methods proposed
- Many open problems remain

Conclusion

- Machine learning has seen tremendous success in several areas, however the key assumption about training/test distributions may not hold in practice, exposing significant vulnerabilities
- Adversarial examples demonstrate the failure of state-of-the-art models when the assumptions are broken
- Several heuristic ways were proposed to make models robust to adversarial examples but were later broken by stronger adversaries
- Certified robustness is emerging as the gold standard to measure the performance of models at test time but they are computationally demanding and practically very small
- The data hungry nature of machine learning and the difficulty of obtaining well curated labeled data makes machine learning vulnerable to poisoning
- Outlier detection can help find poisoned data; model cleansing can help find pointed model. The cat-and-mouse game is still ongoing