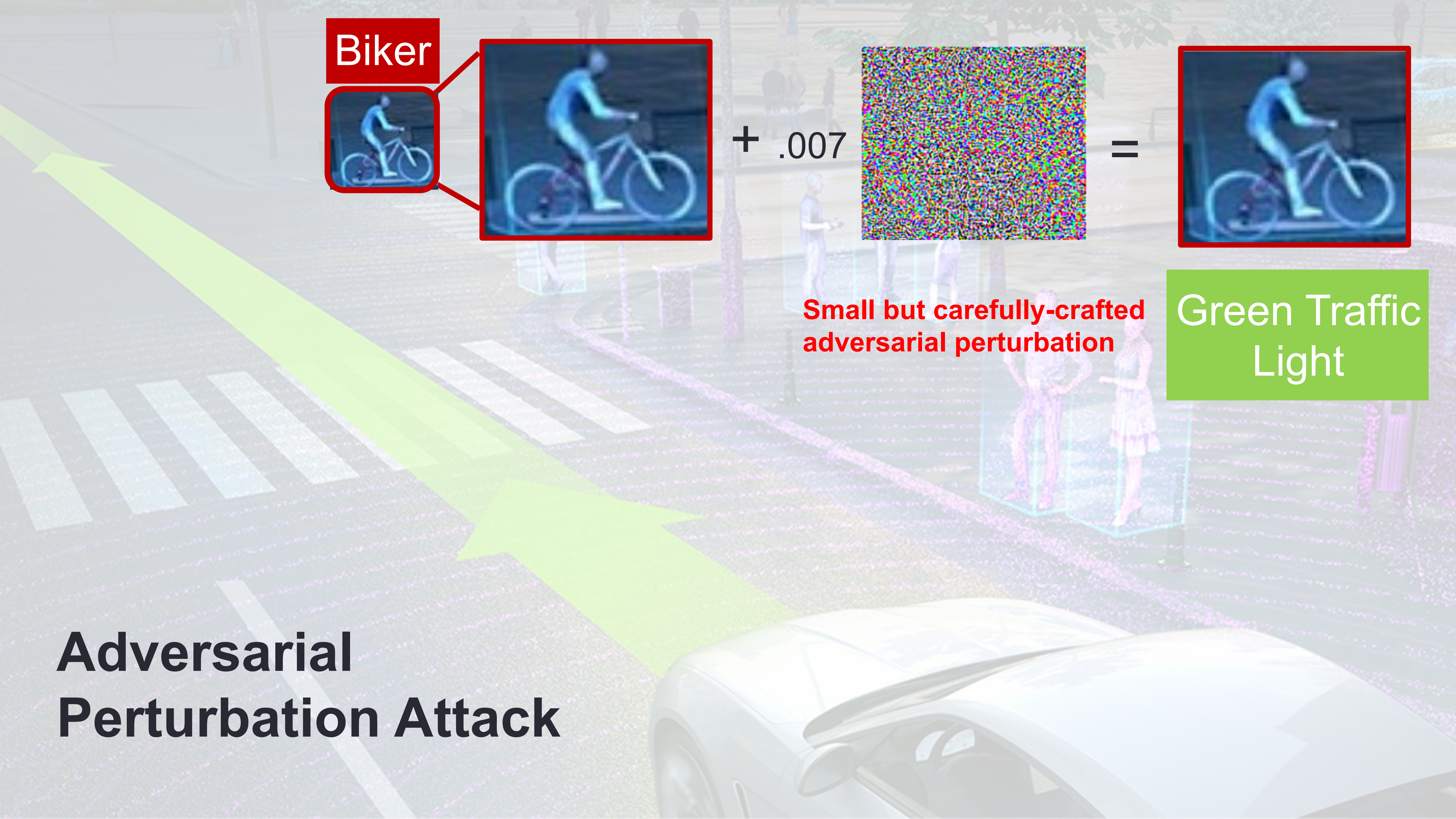
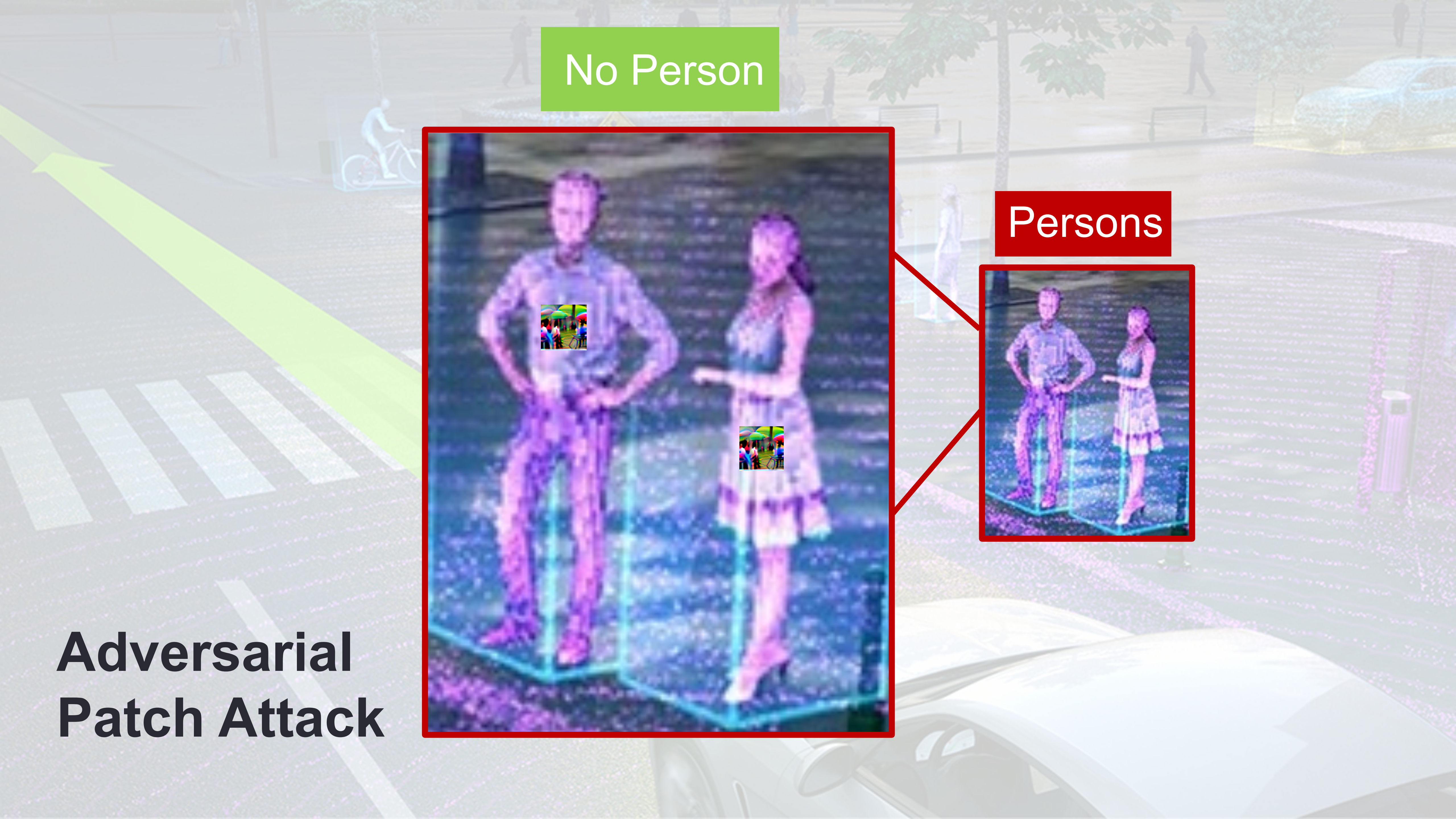
New Advances in (Adversarially) Robust and Secure Machine Learning

Hongyang Zhang
Toyota Technological Institute at Chicago
Carnegie Mellon University


Biker
Pedestrian Sign
Persons


Biker
+ .007
=
Small but carefully-crafted
adversarial perturbation
Green Traffic
Light
Adversarial
Perturbation Attack


Pedestrian Sign
(Minimal) Speed Limit Sign
Adversarial Rotation Attack


No Person
Persons
Adversarial
Patch Attack

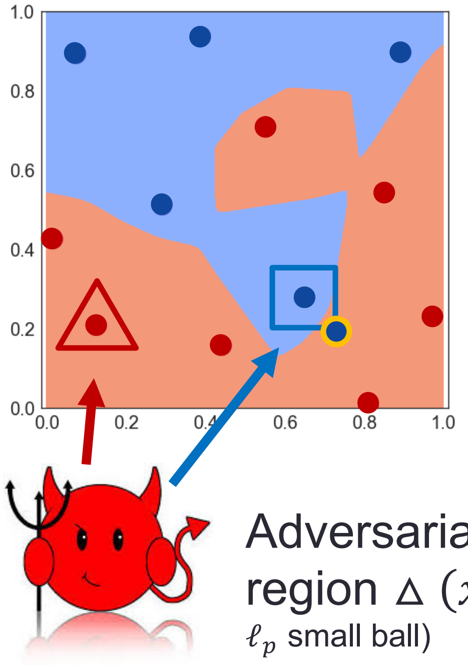
Training Scenario (night)

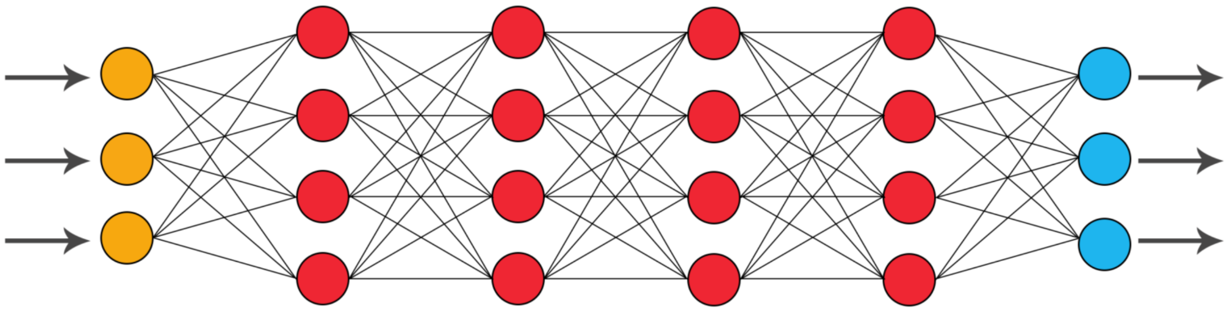
Test Scenario (daytime)

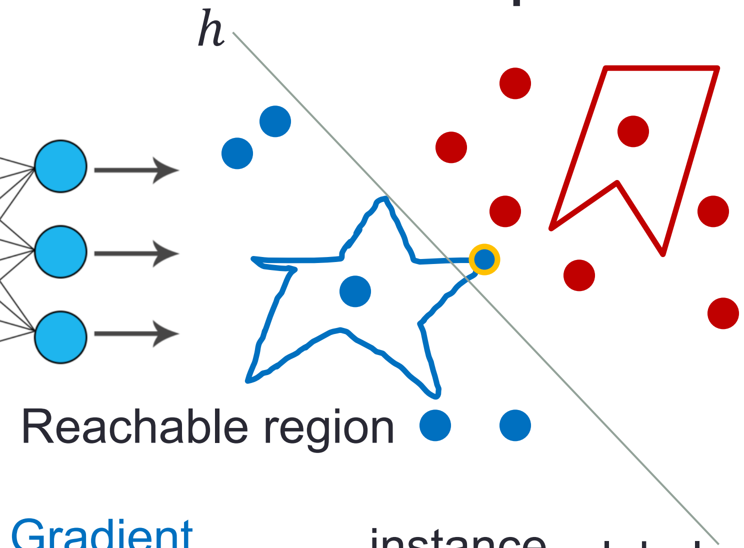
Robust, Secure and Trustworthy functioning of machine learning is the foundation of autopilot systems and AI-landing problems.

What causes adversarial examples?

Input Space **Deep Neural Network f** **Feature Space**







Adversarial reachable region $\Delta\,(x)$ (not necessarily ℓ_p small ball)

Gradient Descent

Gradient Ascent

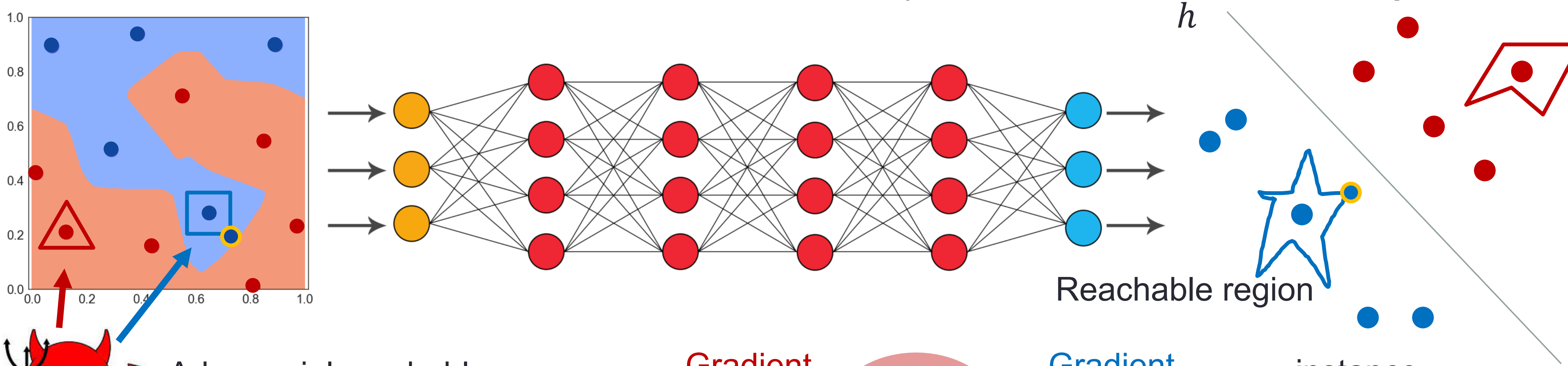
instance

label

Adversarial Training: $\min_{h,f} \mathbb{E}_{x,y\sim\mathcal{D}^n} \max_{x'\in\Delta(x)} \mathrm{Loss}(h(f(x')), y)$

What causes adversarial examples?

Input Space **Deep Neural Network f** **Feature Space**



Adversarial reachable region $\triangle(x)$ (not necessarily ℓ_p small ball)

Gradient Descent

Gradient Ascent

instance

label

Adversarial Training: (Empirical Defense)

$$\min_{h,f} \mathbb{E}_{x,y\sim\mathcal{D}^n} \max_{x'\in\Delta(x)} \mathrm{Loss}(h(f(x')), y)$$

How to test robustness on x_{test} as an adversary:

$$\max_{x'\in\Delta(x_{\text{test}})} \mathrm{Loss}(h(f(x')), y_{\text{test}})$$

No guarantee to get its global optimality

Certified Defense: provide certification about its optimality via smoothing

Overview of This Talk

Paradigms

Robustness

Adversarial Example

Random Noise

Mixed Random/Adversarial Corruption

Part I: Empirical Defense --- TRADES

Empirical Defense

Certified Defense

Adversarial Defenses

Norm-Bounded Adversarial Example

Unrestricted Adversarial Example

Positive Result

Hardness Result

Applications

Adversarial Vision Challenge Model Track

Adversarial Vision Challenge Untargeted Attack

Google Unrestricted Adversarial Examples Challenge

RobustBench A standardized benchmark for adversarial robustness

GLUE

Trade-off between Robustness and Accuracy

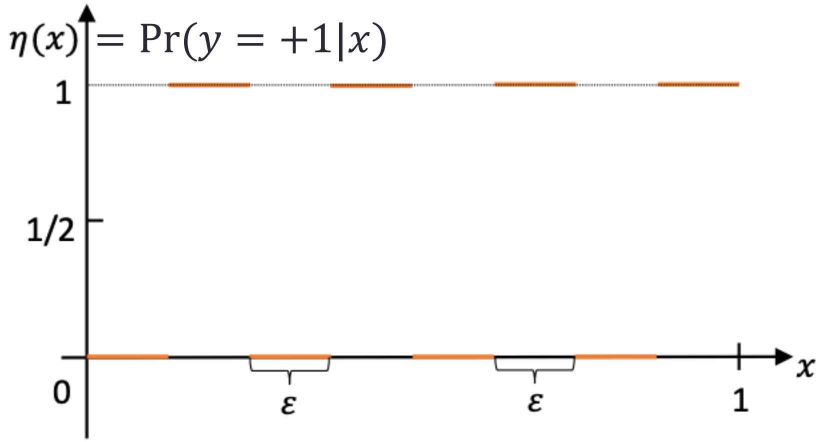
$$R_{rob}(f) := \mathbb{E}_{x,y\sim\mathcal{D}} 1\{\exists x' \in \triangle(x) \text{ s.t. } f(x')y \leq 0\}$$

$y \in \{+1, -1\}$, classifier $f: \mathcal{X} \to \mathbb{R}$

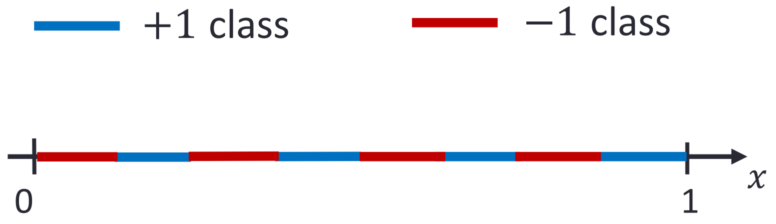
Indicator function

$$R_{nat}(f) := \mathbb{E}_{x,y\sim\mathcal{D}} 1\{f(x)y \leq 0\}$$

- An example of trade-off (for norm-bounded threat model when $\triangle(x) = \mathbb{B}_p(x, \varepsilon)$):



$x \sim \text{Uniform}[0,1]$



	Bayes Optimal Classifier
$R_{nat}(f)$	0 (minimal R_{nat})
$R_{rob}(f)$	1

Trade-off between Robustness and Accuracy

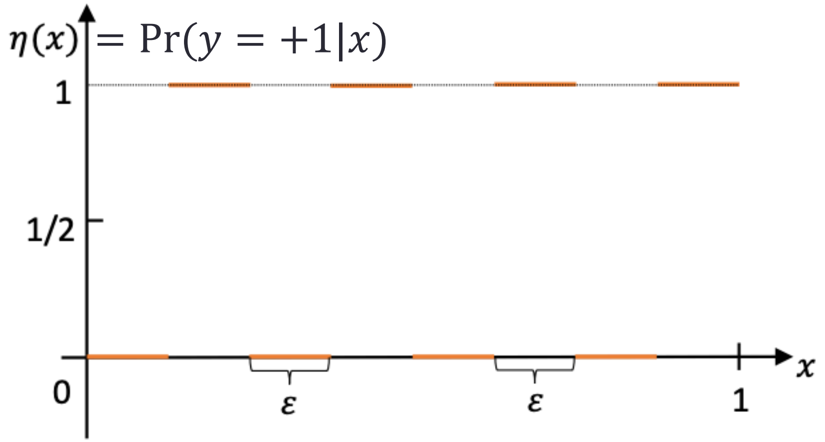
$$R_{rob}(f) := \mathbb{E}_{x,y\sim\mathcal{D}} 1\{\exists x' \in \triangle(x) \text{ s.t. } f(x')y \le 0\}$$

$y \in \{+1, -1\}$, classifier $f: \mathcal{X} \to \mathbb{R}$

Indicator function

$$R_{nat}(f) := \mathbb{E}_{x,y\sim\mathcal{D}} 1\{f(x)y \le 0\}$$

- An example of trade-off (for norm-bounded threat model when $\triangle(x) = \mathbb{B}_p(x, \varepsilon)$):



$x \sim \text{Uniform}[0,1]$

+1 class   −1 class

	Bayes Optimal Classifier	All +1 Classifier
$R_{nat}(f)$	0 (minimal R_{nat})	1/2
$R_{rob}(f)$	1	1/2 (minimal R_{rob})

Solution: minimize $\min_f R_{nat}(f) + R_{rob}(f)/\lambda$

Computationally, weighted average $R_{nat}(f) + R_{rob}(f)/\lambda$ is non-differentiable.

Classification-Calibrated Surrogate Loss

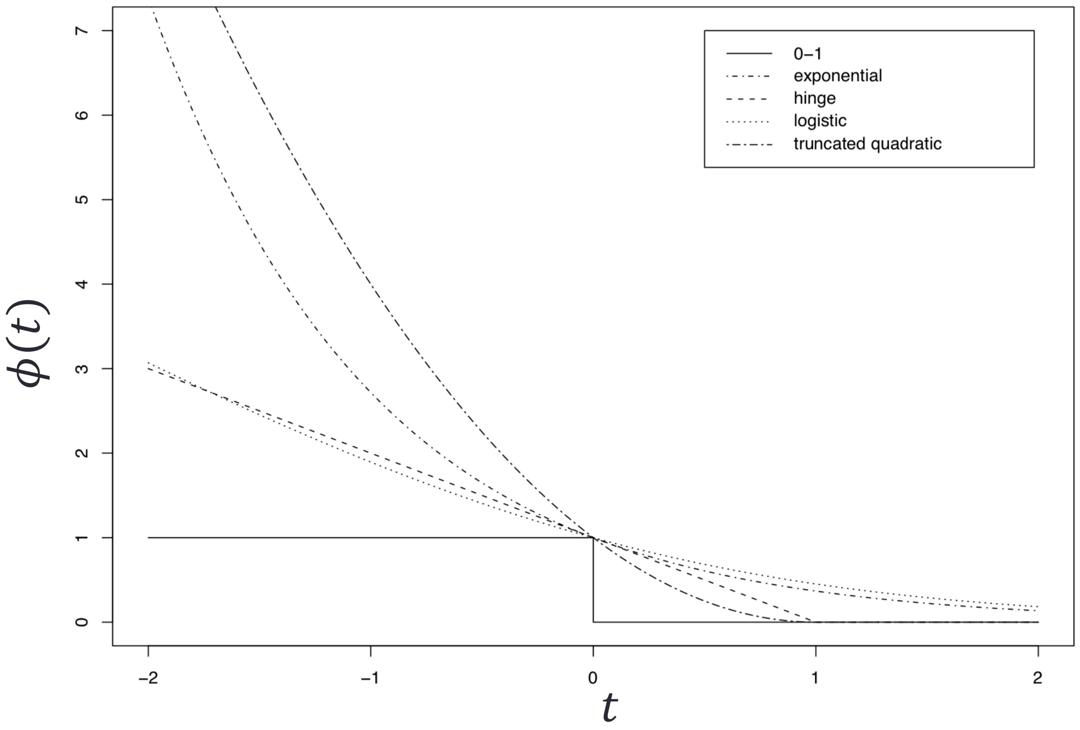
$$R_{rob}(f) := \mathbb{E}_{x,y\sim\mathcal{D}} 1\{\exists x' \in \Delta(x) \text{ s.t. } f(x')y \le 0\}$$

Can we design a differentiable surrogate loss for the trade-off?

$$R_{nat}(f) := \mathbb{E}_{x,y\sim\mathcal{D}} 1\{f(x)y \le 0\}$$

[Bartlett et al.'06]
approximate

$$R_{\phi}(f) := \mathbb{E}_{x,y\sim\mathcal{D}} \phi(f(x)y)$$



[Bartlett et al.'06] Convexity, Classification, and Risk Bounds, Journal of the American Statistical Association, 2006

Our Methodology --- TRADES

$$\min_f [\overbrace{\mathbb{E}_{x,y\sim\mathcal{D}}\, \phi(f(x)y)}^{\text{minimize difference between } f(x) \text{ and } y \text{ for accuracy}} + \mathbb{E}_{x,y\sim\mathcal{D}} \max_{x'\in\Delta(x)} \phi(f(x)f(x')/\lambda)]$$

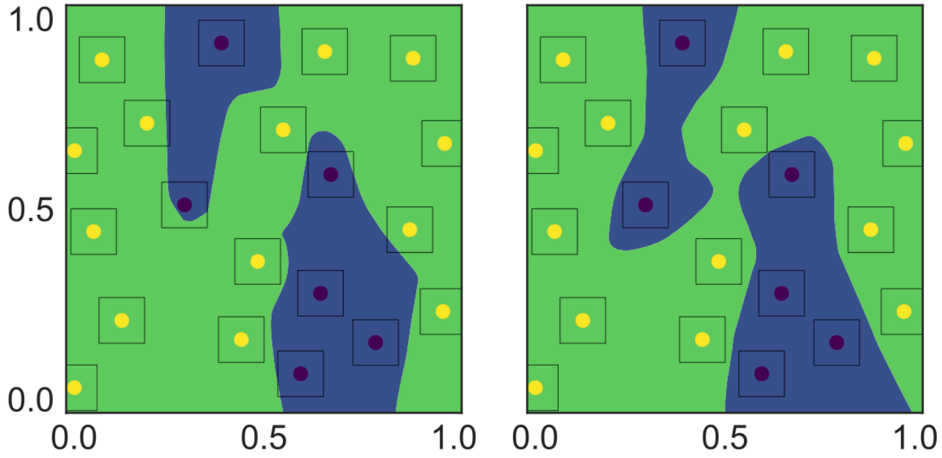
[Zhang, Yu, Jiao, Xing, Ghaoui, Jordan'19] Theoretically Principled Trade-off between Robustness and Accuracy, ICML 2019

Our Methodology --- TRADES

$$\min_f [\mathbb{E}_{x,y\sim\mathcal{D}}\, \phi(f(x)y) + \underbrace{\mathbb{E}_{x,y\sim\mathcal{D}} \max_{x'\in\Delta(x)} \phi(f(x)f(x')/\lambda)}_{\text{minimize difference between } f(x) \text{ and } f(x') \text{ for robustness}}]$$

Gradient Ascent

Replace $\mathcal{D}$ with $\mathcal{D}^n$ and do ERM



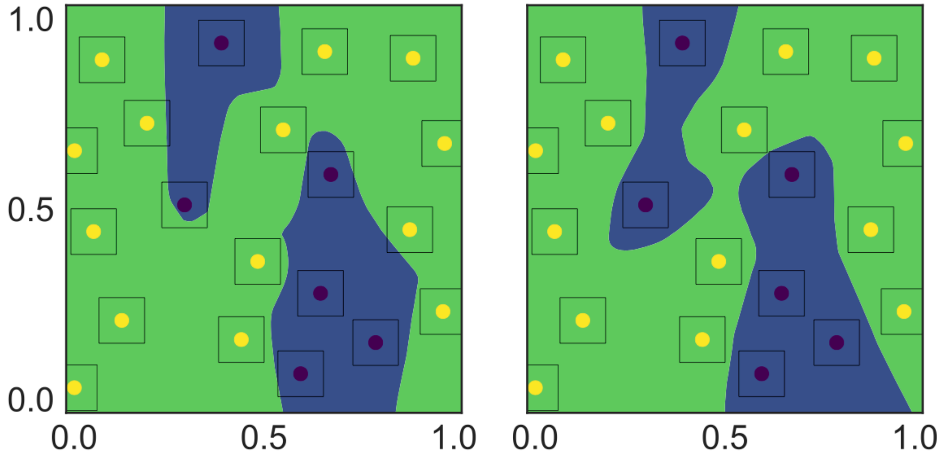
$\min_f \mathbb{E}\, \phi(f(x)y)$ $\min_f [\mathbb{E}\, \phi(f(x)y) + \mathbb{E} \max_{x'\in\Delta(x)} \phi(f(x)f(x')/\lambda)]$

[Zhang, Yu, Jiao, Xing, Ghaoui, Jordan'19] Theoretically Principled Trade-off between Robustness and Accuracy, ICML 2019

Our Methodology --- TRADES

$$\min_f [\overbrace{\mathbb{E}_{x,y\sim\mathcal{D}}\, \phi(f(x)y) + \mathbb{E}_{x,y\sim\mathcal{D}} \boxed{\max_{x'\in\Delta(x)}}\, \phi(f(x)f(x')/\lambda)}^{\text{TRADES Loss}(f)\text{:}}]$$

Gradient Ascent



$\min_f \mathbb{E}\, \phi(f(x)y)$ $\quad$ $\min_f [\mathbb{E}\, \phi(f(x)y) + \mathbb{E} \max_{x'\in\Delta(x)} \phi(f(x)f(x')/\lambda)]$

[Zhang, Yu, Jiao, Xing, Ghaoui, Jordan'19] Theoretically Principled Trade-off between Robustness and Accuracy, ICML 2019

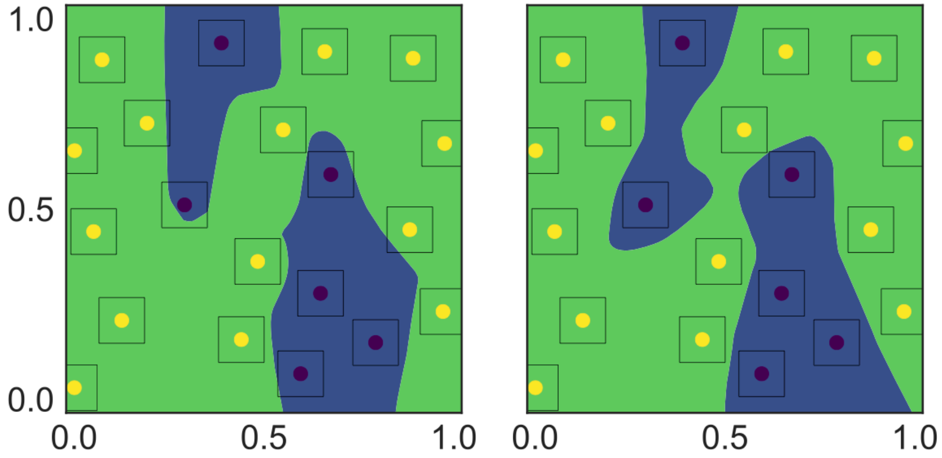
Theoretical Results

Theorem 1 (Informal, upper bound, Zhang et al.'19):

For any distribution $\mathcal{D}$, f, $\Delta(x)$ and $\lambda > 0$, we have $R_{rob}(f) - R_{nat}^* \le$ TRADES Loss$(f) - R_\phi^*$.

- R_{nat}^*: minimal value of $R_{nat}(f)$ over all classifiers f
- R_ϕ^*: minimal value of $R_\phi(f) := \mathbb{E}_{x,y\sim\mathcal{D}}\phi(f(x)y)$ over all classifiers f
- ϕ: classification-calibrated surrogate loss

$$\min_f [\mathbb{E}_{x,y\sim\mathcal{D}}\, \phi(f(x)y) + \underbrace{\mathbb{E}_{x,y\sim\mathcal{D}} \max_{x'\in\Delta(x)} \phi(f(x)f(x')/\lambda)}_{\text{TRADES Loss}(f):}]$$



$\min_f \mathbb{E}\,\phi(f(x)y)$ $\qquad$ $\min_f[\mathbb{E}\,\phi(f(x)y) + \mathbb{E}\max_{x'\in\Delta(x)} \phi(f(x)f(x')/\lambda)]$

[Zhang, Yu, Jiao, Xing, Ghaoui, Jordan'19] Theoretically Principled Trade-off between Robustness and Accuracy, ICML 2019

Theoretical Results

Theorem 1 (Informal, upper bound, Zhang et al.'19):

For any distribution $\mathcal{D}$, f, $\Delta(x)$ and $\lambda > 0$, we have $R_{rob}(f) - R_{nat}^* \le \text{TRADES Loss}(f) - R_\phi^*$.

- ❑ R_{nat}^*: minimal value of $R_{nat}(f)$ over all classifiers f
- ❑ R_ϕ^*: minimal value of $R_\phi(f) := \mathbb{E}_{x,y\sim\mathcal{D}}\phi(f(x)y)$ over all classifiers f
- ❑ ϕ: classification-calibrated surrogate loss

Theorem 2 (Informal, lower bound, Zhang et al.'19):

For any $\Delta(x)$, there exist a data distribution $\mathcal{D}$, a classifier f, and an $\lambda > 0$ such that

$$R_{rob}(f) - R_{nat}^* \ge \text{TRADES Loss}(f) - R_\phi^*.$$

[Zhang, Yu, Jiao, Xing, Ghaoui, Jordan'19] Theoretically Principled Trade-off between Robustness and Accuracy, ICML 2019

Experiments --- CIFAR10 with 8-intensity level attacks

Defense	Defense type	Under which attack	Dataset	Distance	Natural Accuracy $\mathcal{A}_{\mathrm{nat}}(f)$	Robust Accuracy $\mathcal{A}_{\mathrm{rob}}(f)$
Buckman et al. (2018)	gradient mask	Athalye et al. (2018)	CIFAR10	0.031 (ℓ_∞)	-	0%
Ma et al. (2018)	gradient mask	Athalye et al. (2018)	CIFAR10	0.031 (ℓ_∞)	-	5%
Dhillon et al. (2018)	gradient mask	Athalye et al. (2018)	CIFAR10	0.031 (ℓ_∞)	-	0%
Song et al. (2018)	gradient mask	Athalye et al. (2018)	CIFAR10	0.031 (ℓ_∞)	-	9%
Na et al. (2017)	gradient mask	Athalye et al. (2018)	CIFAR10	0.015 (ℓ_∞)	-	15%
Wong et al. (2018)	robust opt.	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	27.07%	23.54%
Madry et al. (2018)	robust opt.	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	87.30%	**47.04%**

$$\min_f \mathbb{E} \max_{x' \in \mathbb{B}(x,\varepsilon)} \phi(f(x')y)$$ (by Madry et al.)

Defense	Defense type	Under which attack	Dataset	Distance	$\mathcal{A}_{\mathrm{nat}}(f)$	$\mathcal{A}_{\mathrm{rob}}(f)$
TRADES ($1/\lambda = 1.0$)	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	88.64%	49.14%
TRADES ($1/\lambda = 6.0$)	regularization	FGSM^{20} (PGD)	CIFAR10	0.031 (ℓ_∞)	84.92%	**56.61%**

$$\min_f [\mathbb{E}\, \phi(f(x)y) + \mathbb{E} \max_{x' \in \mathbb{B}(x,\varepsilon)} \phi(f(x)f(x')/\lambda)]$$ (ours)

Defense	Defense type	Under which attack	Dataset	Distance	$\mathcal{A}_{\mathrm{nat}}(f)$	$\mathcal{A}_{\mathrm{rob}}(f)$
TRADES ($1/\lambda = 6.0$)	regularization	LBFGSAttack	CIFAR10	0.031 (ℓ_∞)	84.92%	81.58%
TRADES ($1/\lambda = 1.0$)	regularization	MI-FGSM	CIFAR10	0.031 (ℓ_∞)	88.64%	51.26%
TRADES ($1/\lambda = 6.0$)	regularization	MI-FGSM	CIFAR10	0.031 (ℓ_∞)	84.92%	57.95%
TRADES ($1/\lambda = 1.0$)	regularization	C&W	CIFAR10	0.031 (ℓ_∞)	88.64%	84.03%
TRADES ($1/\lambda = 6.0$)	regularization	C&W	CIFAR10	0.031 (ℓ_∞)	84.92%	81.24%
Samangouei et al. (2018)	gradient mask	Athalye et al. (2018)	MNIST	0.005 (ℓ_2)	-	55%
Madry et al. (2018)	robust opt.	FGSM^{40} (PGD)	MNIST	0.3 (ℓ_∞)	99.36%	96.01%
TRADES ($1/\lambda = 6.0$)	regularization	FGSM^{40} (PGD)	MNIST	0.3 (ℓ_∞)	99.48%	96.07%
TRADES ($1/\lambda = 6.0$)	regularization	C&W	MNIST	0.005 (ℓ_2)	99.48%	99.46%

Overview of This Talk

Paradigms

Robustness

Adversarial Examples

Random Noises

Mixed Random/Adversarial Corruptions

Empirical Defenses

Certified Defenses

Significant Experimental Results via Case Study

Norm-Bounded Adversarial Example

Unrestricted Adversarial Example

Positive Results

Hardness Results

Applications

Adversarial Vision Challenge Model Track

Adversarial Vision Challenge Untargeted Attack

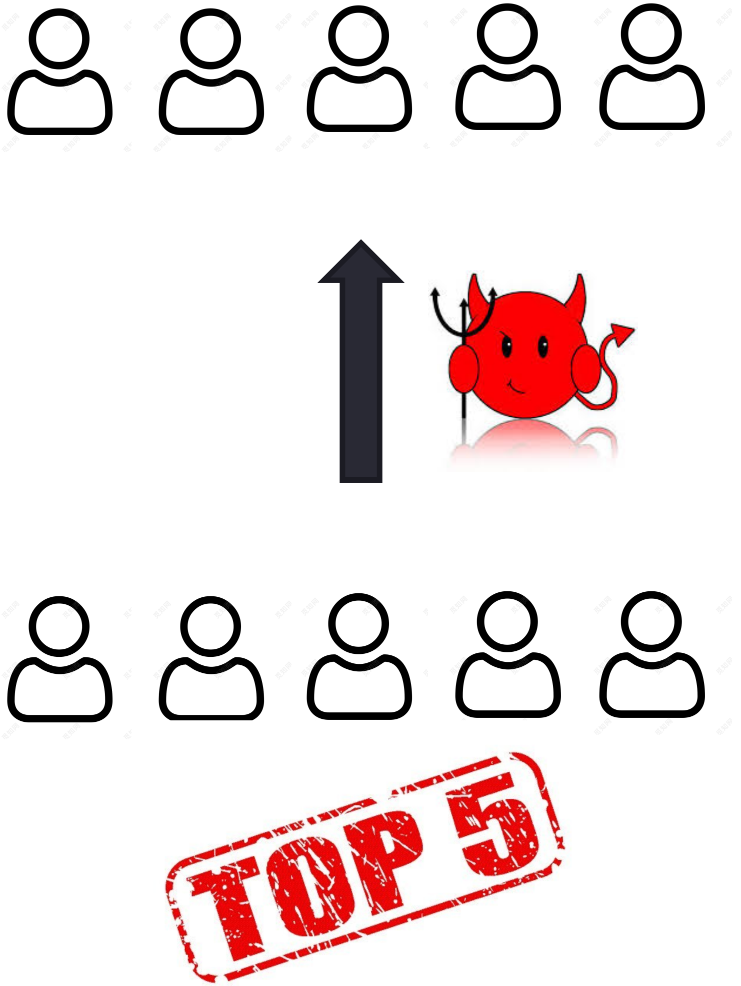
Google Unrestricted Adversarial Examples Challenge

RobustBench A standardized benchmark for adversarial robustness

GLUE

Case Study I: NeurIPS'18 Adversarial Vision Challenge

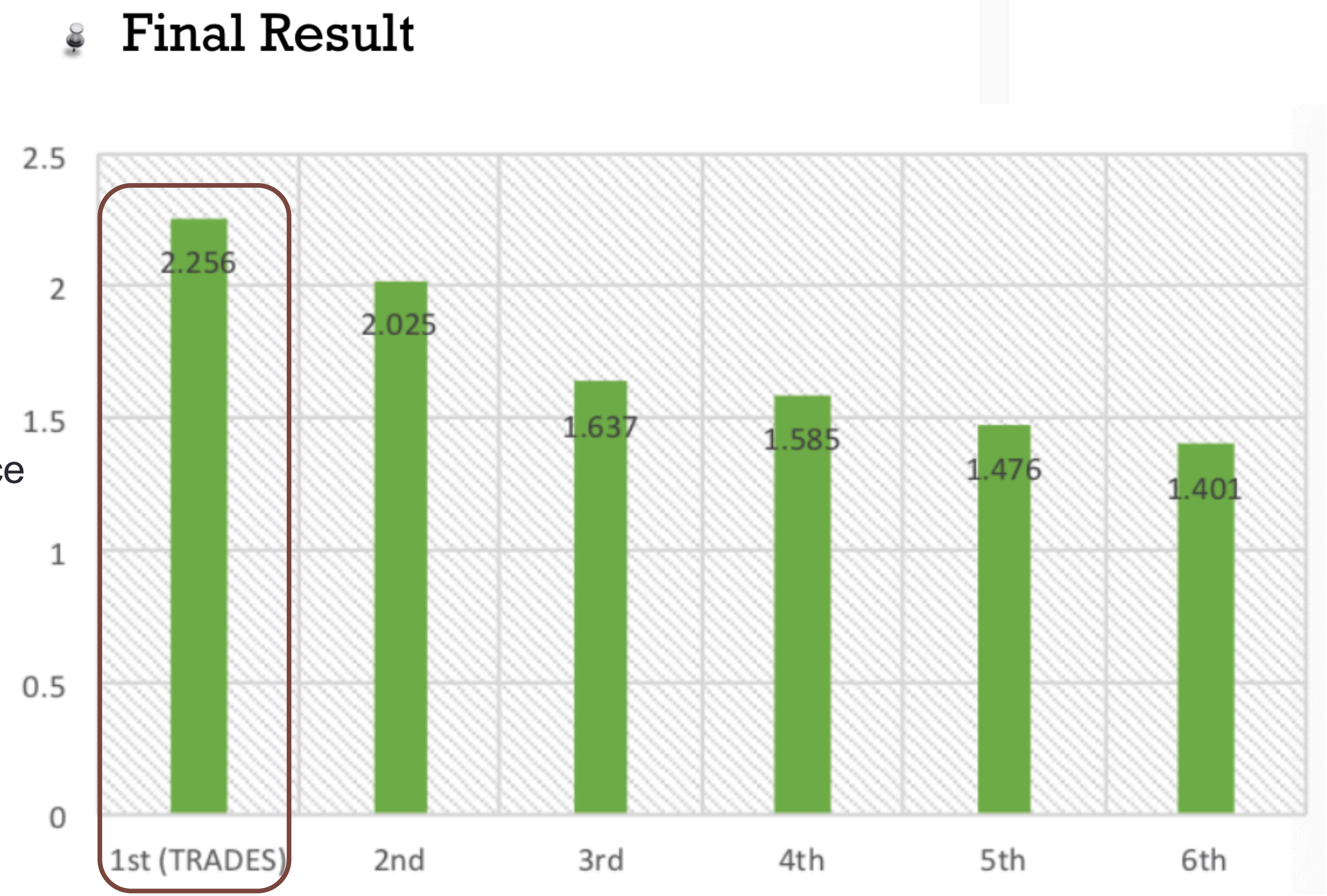
Ranking



- Evaluation criterion
  - 400+ teams, ~3,000 submissions
  - ImageNet dataset
  - Model Track and Attack Track
  - Participants in the two tracks play against each other

Case Study I: NeurIPS'18 Adversarial Vision Challenge

y-axis: mean ℓ_2 perturbation distance to let a classifier make a mistake



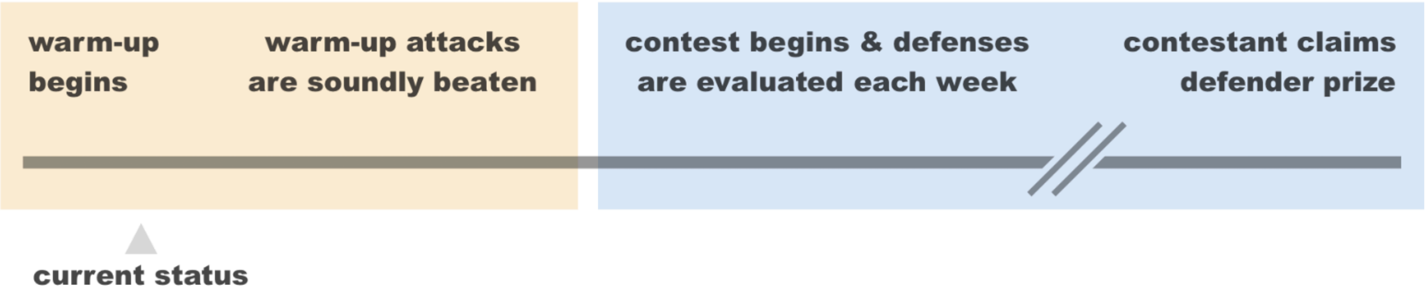
Case Study II: Unrestricted Adversarial Examples Challenge

Unrestricted Adversarial Examples Challenge build passing

In the Unrestricted Adversarial Examples Challenge, attackers submit arbitrary adversarial inputs, and defenders are expected to assign low confidence to difficult inputs while retaining high confidence and accuracy on a clean, unambiguous test set. You can learn more about the motivation and structure of the contest in our recent paper

This repository contains code for the warm-up to the challenge, as well as the public proposal for the contest. We are currently accepting defenses for the warm-up.

Warm-up & Contest Timeline



Case Study II: Unrestricted Adversarial Examples Challenge

The class of bicycle

The class of bird

Case Study II: Unrestricted Adversarial Examples Challenge

Our methodology:

$$\min_f [\mathbb{E}\,\phi(f(x)y) + \mathbb{E} \max_{x'\in\Delta(x)} \phi(f(x)f(x')/\lambda)]$$

Choose the adversarial reachable region as the union of these threat models

Defense	Submitted by	Clean data	Common corruptions	Spatial grid attack	SPSA attack	Boundary attack	Submission Date
Pytorch ResNet50 (trained on bird-or-bicycle extras)	TRADES	100.0%	100.0%	99.5%	100.0%	95.0%	Jan 17th, 2019 (EST)
Keras ResNet (trained on ImageNet)	Google Brain	100.0%	99.2%	92.2%	1.6%	4.0%	Sept 29th, 2018
Pytorch ResNet (trained on bird-or-bicycle extras)	Google Brain	98.8%	74.6%	49.5%	2.5%	8.0%	Oct 1st, 2018

Case Study II: Unrestricted Adversarial Examples Challenge

Clean image:

Defense	Submitted by	Clean data	Common corruptions	Spatial grid attack	SPSA attack	Boundary attack	Submission Date
Pytorch ResNet50 (trained on bird-or-bicycle extras)	TRADES	100.0%	100.0%	99.5%	100.0%	95.0%	Jan 17th, 2019 (EST)
Keras ResNet (trained on ImageNet)	Google Brain	100.0%	99.2%	92.2%	1.6%	4.0%	Sept 29th, 2018
Pytorch ResNet (trained on bird-or-bicycle extras)	Google Brain	98.8%	74.6%	49.5%	2.5%	8.0%	Oct 1st, 2018

Case Study II: Unrestricted Adversarial Examples Challenge

Clean image:

Corrupted image:

Defense	Submitted by	Clean data	Common corruptions	Spatial grid attack	SPSA attack	Boundary attack	Submission Date
Pytorch ResNet50 (trained on bird-or-bicycle extras)	TRADES	100.0%	100.0%	99.5%	100.0%	95.0%	Jan 17th, 2019 (EST)
Keras ResNet (trained on ImageNet)	Google Brain	100.0%	99.2%	92.2%	1.6%	4.0%	Sept 29th, 2018
Pytorch ResNet (trained on bird-or-bicycle extras)	Google Brain	98.8%	74.6%	49.5%	2.5%	8.0%	Oct 1st, 2018

Case Study II: Unrestricted Adversarial Examples Challenge

Clean image:

Corrupted image:

Defense	Submitted by	Clean data	Common corruptions	Spatial grid attack	SPSA attack	Boundary attack	Submission Date
Pytorch ResNet50 (trained on bird-or-bicycle extras)	TRADES	100.0%	100.0%	99.5%	100.0%	95.0%	Jan 17th, 2019 (EST)
Keras ResNet (trained on ImageNet)	Google Brain	100.0%	99.2%	92.2%	1.6%	4.0%	Sept 29th, 2018
Pytorch ResNet (trained on bird-or-bicycle extras)	Google Brain	98.8%	74.6%	49.5%	2.5%	8.0%	Oct 1st, 2018

Case Study II: Unrestricted Adversarial Examples Challenge

Clean image:

Corrupted image:

Defense	Submitted by	Clean data	Common corruptions	Spatial grid attack	SPSA attack	Boundary attack	Submission Date
Pytorch ResNet50 (trained on bird-or-bicycle extras)	TRADES	100.0%	100.0%	99.5%	100.0%	95.0%	Jan 17th, 2019 (EST)
Keras ResNet (trained on ImageNet)	Google Brain	100.0%	99.2%	92.2%	1.6%	4.0%	Sept 29th, 2018
Pytorch ResNet (trained on bird-or-bicycle extras)	Google Brain	98.8%	74.6%	49.5%	2.5%	8.0%	Oct 1st, 2018

Case Study II: Unrestricted Adversarial Examples Challenge

Clean image:

Corrupted image:

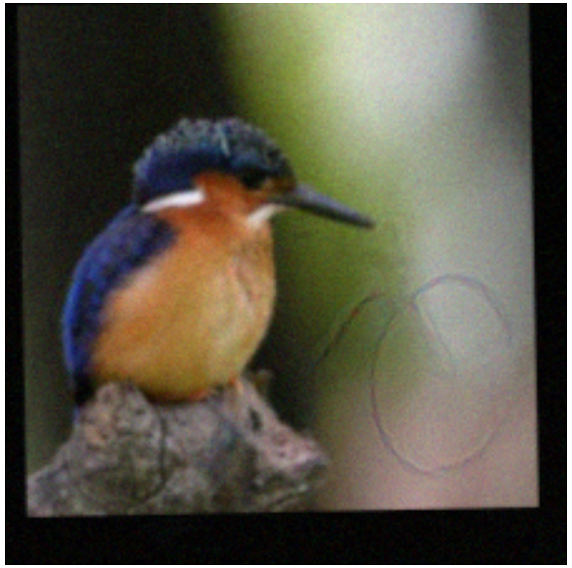
Adversarial example around the decision boundary

Defense	Submitted by	Clean data	Common corruptions	Spatial grid attack	SPSA attack	Boundary attack	Submission Date
Pytorch ResNet50 (trained on bird-or-bicycle extras)	TRADES	100.0%	100.0%	99.5%	100.0%	95.0%	Jan 17th, 2019 (EST)
Keras ResNet (trained on ImageNet)	Google Brain	100.0%	99.2%	92.2%	1.6%	4.0%	Sept 29th, 2018
Pytorch ResNet (trained on bird-or-bicycle extras)	Google Brain	98.8%	74.6%	49.5%	2.5%	8.0%	Oct 1st, 2018

Interpretability of TRADES --- Adversarial Examples by Boundary Attack

The class of bicycle

The class of bird



Overview of This Talk

Paradigms

Robustness

Adversarial Examples

Random Noises

Mixed Random/Adversarial Corruptions

Empirical Defenses

Certified Defenses

Adversarial Defenses

Norm-Bounded Adversarial Example

Unrestricted Adversarial Example

Positive Results

Hardness Results

Applications

Adversarial Vision Challenge Model Track

Adversarial Vision Challenge Untargeted Attack

Unrestricted Adversarial Examples Challenge

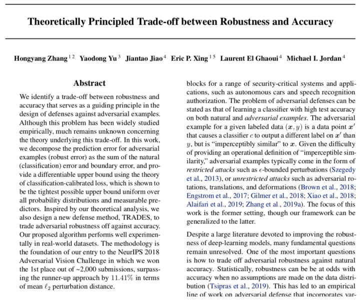
RobustBench

A standardized benchmark for adversarial robustness

GLUE

Significant Impact of TRADES

Significant Impact of TRADES



Theoretically Principled Trade-off between Robustness and Accuracy

Hongyang Zhang [1 2] Yaodong Yu [3] Jiantao Jiao [4] Eric P. Xing [1 5] Laurent El Ghaoui [4] Michael I. Jordan [4]

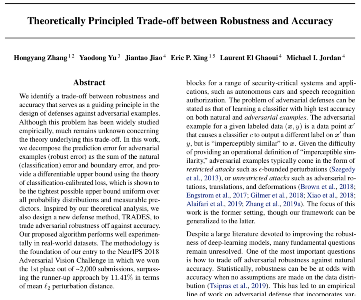
Abstract

We identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples. Although this problem has been widely studied empirically, much remains unknown concerning the theory underlying this trade-off. In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors. Inspired by our theoretical analysis, we also design a new defense method, TRADES, to trade adversarial robustness off against accuracy. Our proposed algorithm performs well experimentally in real-world datasets. The methodology is the foundation of our entry to the NeurIPS 2018 Adversarial Vision Challenge in which we won the 1st place out of ~2,000 submissions, surpassing the runner-up approach by 11.41% in terms of mean ℓ_2 perturbation distance.

blocks for a range of security-critical systems and applications, such as autonomous cars and speech recognition authorization. The problem of adversarial defenses can be stated as that of learning a classifier with high test accuracy on both natural and *adversarial examples*. The adversarial example for a given labeled data $(\boldsymbol{x}, y)$ is a data point $\boldsymbol{x}'$ that causes a classifier c to output a different label on $\boldsymbol{x}'$ than y, but is "imperceptibly similar" to $\boldsymbol{x}$. Given the difficulty of providing an operational definition of "imperceptible similarity," adversarial examples typically come in the form of *restricted attacks* such as ϵ-bounded perturbations (Szegedy et al., 2013), or *unrestricted attacks* such as adversarial rotations, translations, and deformations (Brown et al., 2018; Engstrom et al., 2017; Gilmer et al., 2018; Xiao et al., 2018; Alaifari et al., 2019; Zhang et al., 2019a). The focus of this work is the former setting, though our framework can be generalized to the latter.

Despite a large literature devoted to improving the robustness of deep-learning models, many fundamental questions remain unresolved. One of the most important questions is how to trade off adversarial robustness against natural accuracy. Statistically, robustness can be be at odds with accuracy when no assumptions are made on the data distribution (Tsipras et al., 2019). This has led to an empirical line of work on adversarial defense that incorporates var-

Significant Impact of TRADES



Theoretically Principled Trade-off between Robustness and Accuracy

Hongyang Zhang[1,2] **Yaodong Yu**[3] **Jiantao Jiao**[4] **Eric P. Xing**[1,5] **Laurent El Ghaoui**[4] **Michael I. Jordan**[4]

Abstract

We identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples. Although this problem has been widely studied empirically, much remains unknown concerning the theory underlying this trade-off. In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors. Inspired by our theoretical analysis, we also design a new defense method, TRADES, to trade adversarial robustness off against accuracy. Our proposed algorithm performs well experimentally in real-world datasets. The methodology is the foundation of our entry to the NeurIPS 2018 Adversarial Vision Challenge in which we won the 1st place out of ~2,000 submissions, surpassing the runner-up approach by 11.41% in terms of mean ℓ_2 perturbation distance.

blocks for a range of security-critical systems and applications, such as autonomous cars and speech recognition authorization. The problem of adversarial defenses can be stated as that of learning a classifier with high test accuracy on both natural and *adversarial examples*. The adversarial example for a given labeled data $(\boldsymbol{x}, y)$ is a data point $\boldsymbol{x}'$ that causes a classifier c to output a different label on $\boldsymbol{x}'$ than y, but is "imperceptibly similar" to $\boldsymbol{x}$. Given the difficulty of providing an operational definition of "imperceptible similarity," adversarial examples typically come in the form of *restricted attacks* such as ϵ-bounded perturbations (Szegedy et al., 2013), or *unrestricted attacks* such as adversarial rotations, translations, and deformations (Brown et al., 2018; Engstrom et al., 2017; Gilmer et al., 2018; Xiao et al., 2018; Alaifari et al., 2019; Zhang et al., 2019a). The focus of this work is the former setting, though our framework can be generalized to the latter.

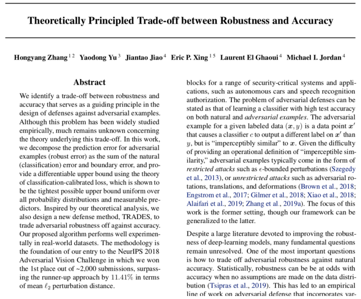
Despite a large literature devoted to improving the robustness of deep-learning models, many fundamental questions remain unresolved. One of the most important questions is how to trade off adversarial robustness against natural accuracy. Statistically, robustness can be be at odds with accuracy when no assumptions are made on the data distribution (Tsipras et al., 2019). This has led to an empirical line of work on adversarial defense that incorporates var-

TRADES — Jan. 2019

Acceleration — April 2019
Univ. of Maryland
Peking University
(2 concurrent works)

- Achieved 30x speed-up on ImageNet, almost as fast as natural training

Significant Impact of TRADES



Theoretically Principled Trade-off between Robustness and Accuracy

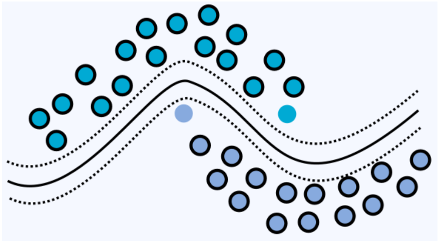
Hongyang Zhang [1,2] Yaodong Yu [3] Jiantao Jiao [4] Eric P. Xing [1,5] Laurent El Ghaoui [4] Michael I. Jordan [4]

Abstract

We identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples. Although this problem has been widely studied empirically, much remains unknown concerning the theory underlying this trade-off. In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors. Inspired by our theoretical analysis, we also design a new defense method, TRADES, to trade adversarial robustness off against accuracy. Our proposed algorithm performs well experimentally in real-world datasets. The methodology is the foundation of our entry to the NeurIPS 2018 Adversarial Vision Challenge in which we won the 1st place out of ~2,000 submissions, surpassing the runner-up approach by 11.41% in terms of mean ℓ_2 perturbation distance.

blocks for a range of security-critical systems and applications, such as autonomous cars and speech recognition authorization. The problem of adversarial defenses can be stated as that of learning a classifier with high test accuracy on both natural and *adversarial examples*. The adversarial example for a given labeled data $(\boldsymbol{x}, y)$ is a data point $\boldsymbol{x}'$ that causes a classifier c to output a different label on $\boldsymbol{x}'$ than y, but is "imperceptibly similar" to $\boldsymbol{x}$. Given the difficulty of providing an operational definition of "imperceptible similarity," adversarial examples typically come in the form of *restricted attacks* such as ϵ-bounded perturbations (Szegedy et al., 2013), or *unrestricted attacks* such as adversarial rotations, translations, and deformations (Brown et al., 2018; Engstrom et al., 2017; Gilmer et al., 2018; Xiao et al., 2018; Alaifari et al., 2019; Zhang et al., 2019a). The focus of this work is the former setting, though our framework can be generalized to the latter.

Despite a large literature devoted to improving the robustness of deep-learning models, many fundamental questions remain unresolved. One of the most important questions is how to trade off adversarial robustness against natural accuracy. Statistically, robustness can be be at odds with accuracy when no assumptions are made on the data distribution (Tsipras et al., 2019). This has led to an empirical line of work on adversarial defense that incorporates var-



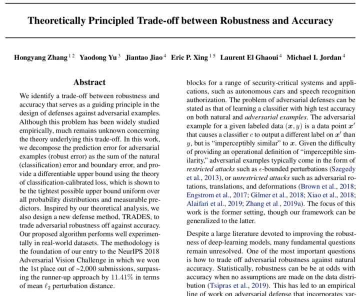
TRADES — Jan. 2019

Acceleration — April 2019
Univ. of Maryland
Peking University
(2 concurrent works)

Semi-Supervision — June 2019
Stanford
DeepMind
Peking University
(3 concurrent works)

- TRADES + 500K extra unlabeled data can improve robust accuracy by +5% on CIFAR10

Significant Impact of TRADES



Theoretically Principled Trade-off between Robustness and Accuracy

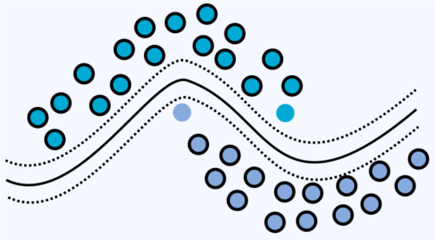
Hongyang Zhang [1,2] Yaodong Yu [3] Jiantao Jiao [4] Eric P. Xing [1,5] Laurent El Ghaoui [4] Michael I. Jordan [4]

Abstract

We identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples. Although this problem has been widely studied empirically, much remains unknown concerning the theory underlying this trade-off. In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors. Inspired by our theoretical analysis, we also design a new defense method, TRADES, to trade adversarial robustness off against accuracy. Our proposed algorithm performs well experimentally in real-world datasets. The methodology is the foundation of our entry to the NeurIPS 2018 Adversarial Vision Challenge in which we won the 1st place out of ~2,000 submissions, surpassing the runner-up approach by 11.41% in terms of mean ℓ_2 perturbation distance.

blocks for a range of security-critical systems and applications, such as autonomous cars and speech recognition authorization. The problem of adversarial defenses can be stated as that of learning a classifier with high test accuracy on both natural and *adversarial examples*. The adversarial example for a given labeled data $(\boldsymbol{x}, y)$ is a data point $\boldsymbol{x}'$ that causes a classifier c to output a different label on $\boldsymbol{x}'$ than y, but is "imperceptibly similar" to $\boldsymbol{x}$. Given the difficulty of providing an operational definition of "imperceptible similarity," adversarial examples typically come in the form of *restricted attacks* such as ϵ-bounded perturbations (Szegedy et al., 2013), or *unrestricted attacks* such as adversarial rotations, translations, and deformations (Brown et al., 2018; Engstrom et al., 2017; Gilmer et al., 2018; Xiao et al., 2018; Alaifari et al., 2019; Zhang et al., 2019a). The focus of this work is the former setting, though our framework can be generalized to the latter.

Despite a large literature devoted to improving the robustness of deep-learning models, many fundamental questions remain unresolved. One of the most important questions is how to trade off adversarial robustness against natural accuracy. Statistically, robustness can be be at odds with accuracy when no assumptions are made on the data distribution (Tsipras et al., 2019). This has led to an empirical line of work on adversarial defense that incorporates var-



GLUE

TRADES — Jan. 2019

Acceleration — April 2019
Univ. of Maryland
Peking University
(2 concurrent works)

Semi-Supervision — June 2019
Stanford
DeepMind
Peking University
(3 concurrent works)

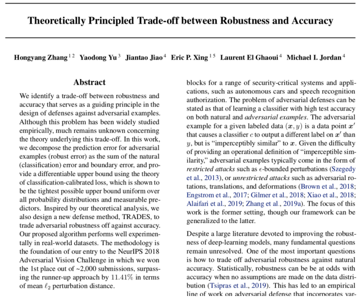
NLP Champion — Dec. 2019
GaTech, Microsoft

- Won 1st place in GLUE (on Dec. 9th, 2019), beating largest NLP T5 model of 11 billion parameters

	Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
+	1	Microsoft D365 AI & MSR AI	MT-DNN-SMART		89.9	69.5	97.5	93.7/91.6	92.9/92.5	73.9/90.2	91.0	90.8	99.2	89.7	94.5	50.2
	2	T5 Team - Google	T5		89.7	70.8	97.1	91.9/89.2	92.5/92.1	74.6/90.4	92.0	91.7	96.7	92.5	93.2	53.1
	3	ALBERT-Team Google Language	ALBERT (Ensemble)		89.4	69.1	97.1	93.4/91.2	92.5/92.0	74.2/90.5	91.3	91.0	99.2	89.2	91.8	50.2

[Jiang et al.'20] SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization, ACL 2020

Significant Impact of TRADES



Theoretically Principled Trade-off between Robustness and Accuracy

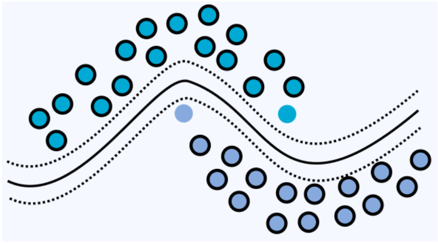
Hongyang Zhang [1,2] Yaodong Yu [3] Jiantao Jiao [4] Eric P. Xing [1,5] Laurent El Ghaoui [4] Michael I. Jordan [4]

Abstract

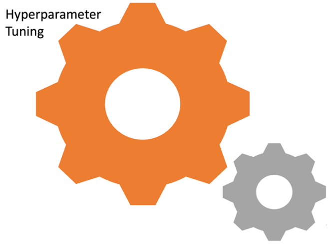
We identify a trade-off between robustness and accuracy that serves as a guiding principle in the design of defenses against adversarial examples. Although this problem has been widely studied empirically, much remains unknown concerning the theory underlying this trade-off. In this work, we decompose the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provide a differentiable upper bound using the theory of classification-calibrated loss, which is shown to be the tightest possible upper bound uniform over all probability distributions and measurable predictors. Inspired by our theoretical analysis, we also design a new defense method, TRADES, to trade adversarial robustness off against accuracy. Our proposed algorithm performs well experimentally in real-world datasets. The methodology is the foundation of our entry to the NeurIPS 2018 Adversarial Vision Challenge in which we won the 1st place out of ~2,000 submissions, surpassing the runner-up approach by 11.41% in terms of mean ℓ_2 perturbation distance.

blocks for a range of security-critical systems and applications, such as autonomous cars and speech recognition authorization. The problem of adversarial defenses can be stated as that of learning a classifier with high test accuracy on both natural and *adversarial examples*. The adversarial example for a given labeled data $(\boldsymbol{x}, y)$ is a data point $\boldsymbol{x}'$ that causes a classifier c to output a different label on $\boldsymbol{x}'$ than y, but is "imperceptibly similar" to $\boldsymbol{x}$. Given the difficulty of providing an operational definition of "imperceptible similarity," adversarial examples typically come in the form of *restricted attacks* such as ϵ-bounded perturbations (Szegedy et al., 2013), or *unrestricted attacks* such as adversarial rotations, translations, and deformations (Brown et al., 2018; Engstrom et al., 2017; Gilmer et al., 2018; Xiao et al., 2018; Alaifari et al., 2019; Zhang et al., 2019a). The focus of this work is the former setting, though our framework can be generalized to the latter.

Despite a large literature devoted to improving the robustness of deep-learning models, many fundamental questions remain unresolved. One of the most important questions is how to trade off adversarial robustness against natural accuracy. Statistically, robustness can be be at odds with accuracy when no assumptions are made on the data distribution (Tsipras et al., 2019). This has led to an empirical line of work on adversarial defense that incorporates var-



GLUE



TRADES — Acceleration — Semi-Supervision — NLP Champion — Hyper-Parameters

Jan. 2019

April 2019
Univ. of Maryland
Peking University
(2 concurrent works)

June 2019
Stanford
DeepMind
Peking University
(3 concurrent works)

Dec. 2019
GaTech, Microsoft

Sep. 2020
DeepMind
Tsinghua University
(2 concurrent works)

- Hyper-parameter tuning of TRADES can further improve robust accuracy by 5% on CIFAR-10

RobustBench

A standardized benchmark for adversarial robustness

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

EPFL

PRINCETON UNIVERSITY

5 out of top 5 and **9 out of top 10** methods use TRADES as their training algorithms.

Rank	Method	Standard accuracy	Robust accuracy	Extra data	Architecture	Venue
1	Uncovering the Limits of Adversarial Training against Norm-Bounded Adversarial Examples *We show the robust accuracy reported in the paper since AutoAttack performs slightly worse (65.88%).*	91.10%	65.87%	✓	WideResNet-70-16	arXiv, Oct 2020
2	[illegible] Adversarial Training against Norm-Bounded Adversarial Examples *[illegible] AutoAttack performs slightly worse (62.[illegible]0%).*	89.48%	62.76%	✓	WideResNet-28-10	arXiv, Oct 2020
3	Adversarial Weight Perturbation Helps Robust Generalization	88.25%	60.04%	✓	WideResNet-28-10	NeurIPS 2020
4	Does Network Width Really Help Adversarial Robustness?	85.60%	59.78%	✓	WideResNet-34-15	arXiv, Oct 2020
5	Unlabeled Data Improves Adversarial Robustness	89.69%	59.53%	✓	WideResNet-28-10	NeurIPS 2019

Significant Impact of TRADES (based on Our CIFAR Challenge)

- TRADES motivates new attacks:

*Powered by TRADES CIFAR-10 Challenge on GitHub

Attack	Submitted by	Attack Model	Robust Acc	Time
PGD-20	(initial entry)	ℓ_∞, 8 intensity	56.61%	Jan 24, 2019
PGD-1,000	(initial entry)	ℓ_∞, 8 intensity	56.43%	Jan 24, 2019

[Croce et al.'20] Minimally Distorted Adversarial Examples with A Fast Adaptive Boundary Attack, ICML 2020.
[Gowal et al.'19] An Alternative Surrogate Loss for PGD-based Adversarial Testing, arXiv 2019.
[Tashiro et al.'20] Diversity Can Be Transferred, NeurIPS 2020.

Significant Impact of TRADES (based on Our CIFAR Challenge)

- TRADES motivates new attacks:

*Powered by TRADES CIFAR-10 Challenge on GitHub

Attack	Submitted by	Attack Model	Robust Acc	Time
PGD-20	(initial entry)	ℓ_∞, 8 intensity	56.61%	Jan 24, 2019
PGD-1,000	(initial entry)	ℓ_∞, 8 intensity	56.43%	Jan 24, 2019
fab-attack	U. of Tubingen	ℓ_∞, 8 intensity	53.44%	Jun 7, 2019

[Croce et al.'20] Minimally Distorted Adversarial Examples with A Fast Adaptive Boundary Attack, ICML 2020.
[Gowal et al.'19] An Alternative Surrogate Loss for PGD-based Adversarial Testing, arXiv 2019.
[Tashiro et al.'20] Diversity Can Be Transferred, NeurIPS 2020.

Significant Impact of TRADES (based on Our CIFAR Challenge)

- TRADES motivates new attacks:

*Powered by TRADES CIFAR-10 Challenge on GitHub

Attack	Submitted by	Attack Model	Robust Acc	Time
PGD-20	(initial entry)	ℓ_∞, 8 intensity	56.61%	Jan 24, 2019
PGD-1,000	(initial entry)	ℓ_∞, 8 intensity	56.43%	Jan 24, 2019
fab-attack	U. of Tubingen	ℓ_∞, 8 intensity	53.44%	Jun 7, 2019
MultiTargeted	DeepMind	ℓ_∞, 8 intensity	53.07%	Oct 31, 2019

[Croce et al.'20] Minimally Distorted Adversarial Examples with A Fast Adaptive Boundary Attack, ICML 2020.
[Gowal et al.'19] An Alternative Surrogate Loss for PGD-based Adversarial Testing, arXiv 2019.
[Tashiro et al.'20] Diversity Can Be Transferred, NeurIPS 2020.

Significant Impact of TRADES (based on Our CIFAR Challenge)

- TRADES motivates new attacks:

*Powered by TRADES CIFAR-10 Challenge on GitHub

Attack	Submitted by	Attack Model	Robust Acc	Time
PGD-20	(initial entry)	ℓ_∞, 8 intensity	56.61%	Jan 24, 2019
PGD-1,000	(initial entry)	ℓ_∞, 8 intensity	56.43%	Jan 24, 2019
fab-attack	U. of Tubingen	ℓ_∞, 8 intensity	53.44%	Jun 7, 2019
MultiTargeted	DeepMind	ℓ_∞, 8 intensity	53.07%	Oct 31, 2019
ODI-PGD	Stanford	ℓ_∞, 8 intensity	53.01%	Feb 16, 2020

[Croce et al.'20] Minimally Distorted Adversarial Examples with A Fast Adaptive Boundary Attack, ICML 2020.
[Gowal et al.'19] An Alternative Surrogate Loss for PGD-based Adversarial Testing, arXiv 2019.
[Tashiro et al.'20] Diversity Can Be Transferred, NeurIPS 2020.

Significant Impact of TRADES (based on Our CIFAR Challenge)

- TRADES motivates new attacks:

*Powered by TRADES CIFAR-10 Challenge on GitHub

Attack	Submitted by	Attack Model	Robust Acc	Time
PGD-20	(initial entry)	ℓ_∞, 8 intensity	56.61%	Jan 24, 2019
PGD-1,000	(initial entry)	ℓ_∞, 8 intensity	56.43%	Jan 24, 2019
fab-attack	U. of Tubingen	ℓ_∞, 8 intensity	53.44%	Jun 7, 2019
MultiTargeted	DeepMind	ℓ_∞, 8 intensity	53.07%	Oct 31, 2019
ODI-PGD	Stanford	ℓ_∞, 8 intensity	53.01%	Feb 16, 2020
CAA	Xiaofeng Mao	ℓ_∞, 8 intensity	52.94%	Dec 14, 2020
EWR-PGD	Ye Liu	ℓ_∞, 8 intensity	52.92%	Dec 20, 2020

… … Can we give a certified lower bound for the robust acc.?

[Croce et al.'20] Minimally Distorted Adversarial Examples with A Fast Adaptive Boundary Attack, ICML 2020.
[Gowal et al.'19] An Alternative Surrogate Loss for PGD-based Adversarial Testing, arXiv 2019.
[Tashiro et al.'20] Diversity Can Be Transferred, NeurIPS 2020.

Part II: Hardness of Certified Defense against Adversarial Examples

Random Smoothing --- A wrapper to robustify base classifier

Certified robust radius by [Cohen et al.'19]:

Confidence of majority vote

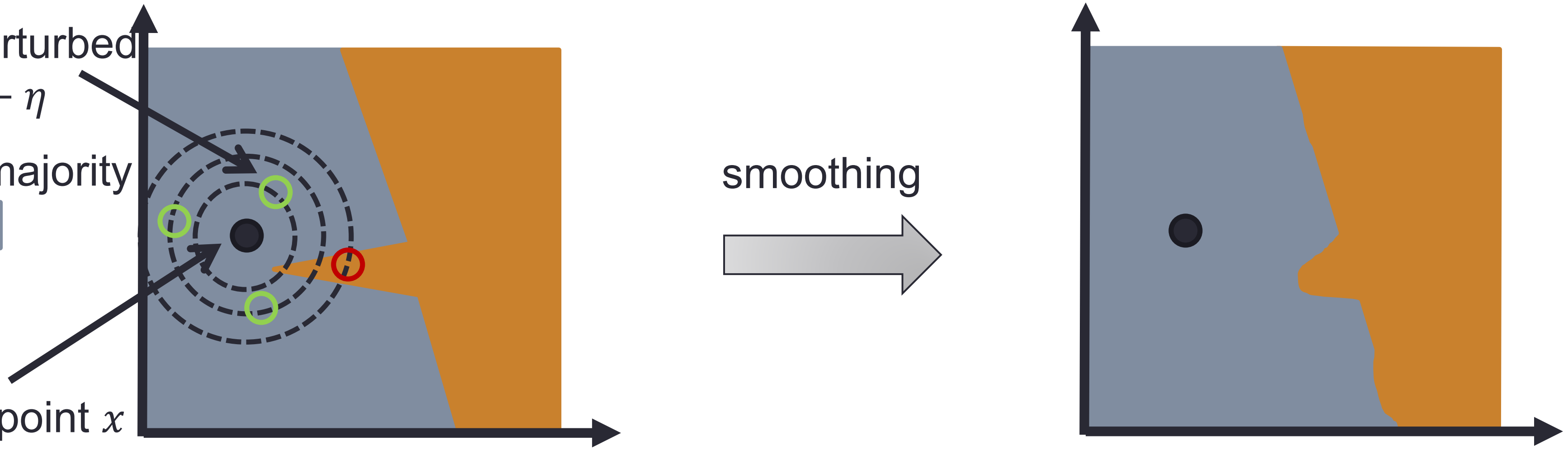
Given any input $x \in \mathbb{R}^d$, let η be Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ and $p = \max_y \mathbb{P}_\eta[f(x+\eta) = y]$. Then $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \leq \Phi^{-1}(p)\sigma/\sqrt{d}$, where Φ is CDF of standard Gaussian.

Randomly perturbed data point $x + \eta$

Output majority vote:

Proposed by [Cohen et al.'19], [Lecuyer et al.'19], [Li et al.'19]

Data point x



Base Classifier $f(x)$

Smoothed Classifier $g(x) = \operatorname{argmax}_y \mathbb{P}_\eta\ [f(x+\eta) = y]$

[Cohen et al.'19] Certified Adversarial Robustness via Randomized Smoothing, ICML 2019.
[Lecuyer et al.'19] Certified Robustness to Adversarial Examples with Differential Privacy, S&P 2019.
[Li et al.'19] Certified Adversarial Robustness with Additive Noise, NeurIPS 2019.

Our Experiments on Random Smoothing

Certified robust radius by [Cohen et al.'19]:

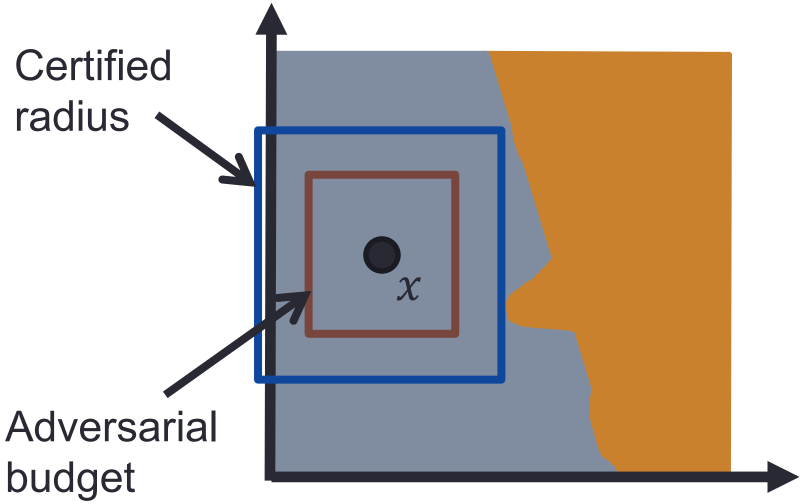
Confidence of majority vote

Given any input $x \in \mathbb{R}^d$, let η be Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ and $p = \max_y \mathbb{P}_\eta[f(x+\eta) = y]$. Then $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \leq \Phi^{-1}(p)\sigma/\sqrt{d}$, where Φ is CDF of standard Gaussian.

Computable certified radius for x

Method	2/255 Certified Robust Acc.
Random Smoothing (TRADES)	62.6%
Random Smoothing (Adv. Training)	60.8%
Random Smoothing (Nat. Training)	50.0%
Zhang et al. (2020)	54.0%
Wong et al. (2018)	53.9%
Mirman et al. (2018)	52.2%
Gowal et al. (2018)	50.0%
Xiao et al. (2019)	45.9%

Table 1: Certified ℓ_∞ robustness at a radius of 2/255 on the CIFAR-10 dataset.



Smoothed Classifier

$$g(x) = \operatorname*{argmax}_y \mathbb{P}_\eta\ [f(x+\eta) = y]$$

Our Experiments on Random Smoothing

Certified robust radius by [Cohen et al.'19]:

Confidence of majority vote

Given any input $x \in \mathbb{R}^d$, let η be Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ and $p = \max_y \mathbb{P}_\eta[f(x+\eta) = y]$. Then $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \leq \Phi^{-1}(p)\sigma/\sqrt{d}$, where Φ is CDF of standard Gaussian.

Computable certified radius for x

Method	2/255 Certified Robust Acc.
Random Smoothing (TRADES)	62.6%
Random Smoothing (Adv. Training)	60.8%
Random Smoothing (Nat. Training)	50.0%
Zhang et al. (2020)	54.0%
Wong et al. (2018)	53.9%
Mirman et al. (2018)	52.2%
Gowal et al. (2018)	50.0%
Xiao et al. (2019)	45.9%

Table 1: Certified ℓ_∞ robustness at a radius of 2/255 on the CIFAR-10 dataset.

8/255 Certified Robust Acc.

~10% (for random smoothing on all base classifiers)

Our Experiments on Random Smoothing

Certified robust radius by [Cohen et al.'19]:

Confidence of majority vote

Given any input $x \in \mathbb{R}^d$, let η be Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ and $p = \max_y \mathbb{P}_\eta[f(x+\eta) = y]$. Then $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \leq \Phi^{-1}(p)\sigma/\sqrt{d}$, where Φ is CDF of standard Gaussian.

Computable certified radius for x

Method	2/255 Certified Robust Acc.
Random Smoothing (TRADES)	62.6%
Random Smoothing (Adv. Training)	60.8%
Random Smoothing (Nat. Training)	50.0%
Zhang et al. (2020)	54.0%
Wong et al. (2018)	53.9%
Mirman et al. (2018)	52.2%
Gowal et al. (2018)	50.0%
Xiao et al. (2019)	45.9%

Table 1: Certified ℓ_∞ robustness at a radius of 2/255 on the CIFAR-10 dataset.

8/255 Certified Robust Acc.

~10% (for random smoothing on all base classifiers)

Random Smoothing with dimension-independent ℓ_∞ radius?

Certified robust radius by [Cohen et al.'19]:

Confidence of majority vote

Given any input $x \in \mathbb{R}^d$, let η be Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ and $p = \max_y \mathbb{P}_\eta[f(x+\eta) = y]$. Then $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \le \Phi^{-1}(p)\sigma/\sqrt{d}$, where Φ is CDF of standard Gaussian.

Computable certified radius for x

Can we improve the $\sigma/\sqrt{d}$ dependence by looking at other noise distributions or is it inevitable? Why?

Our Hardness Result concerning Random Smoothing

Certified robust radius by [Cohen et al.'19]:

Confidence of majority vote

Given any input $x \in \mathbb{R}^d$, let η be Gaussian noise $\mathcal{N}(0, \sigma^2 I)$ and $p = \max_y \mathbb{P}_\eta[f(x+\eta) = y]$. Then $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \le \Phi^{-1}(p)\sigma/\sqrt{d}$, where Φ is CDF of standard Gaussian.

Theorem 1 (Our hardness result, JMLR'20):

Given any input x, let η be noise from any distribution with variance of η_i being σ_i^2. If $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \le \varepsilon$, then $\varepsilon < c_p \sigma_i/\sqrt{d}$ for 99% entries i.

[Blum, Dick, Manoj, Zhang (α-β)'20] Random Smoothing Might be Unable to Certify ℓ_∞ Robustness, JMLR 2020

Intuition behind the Hardness Result

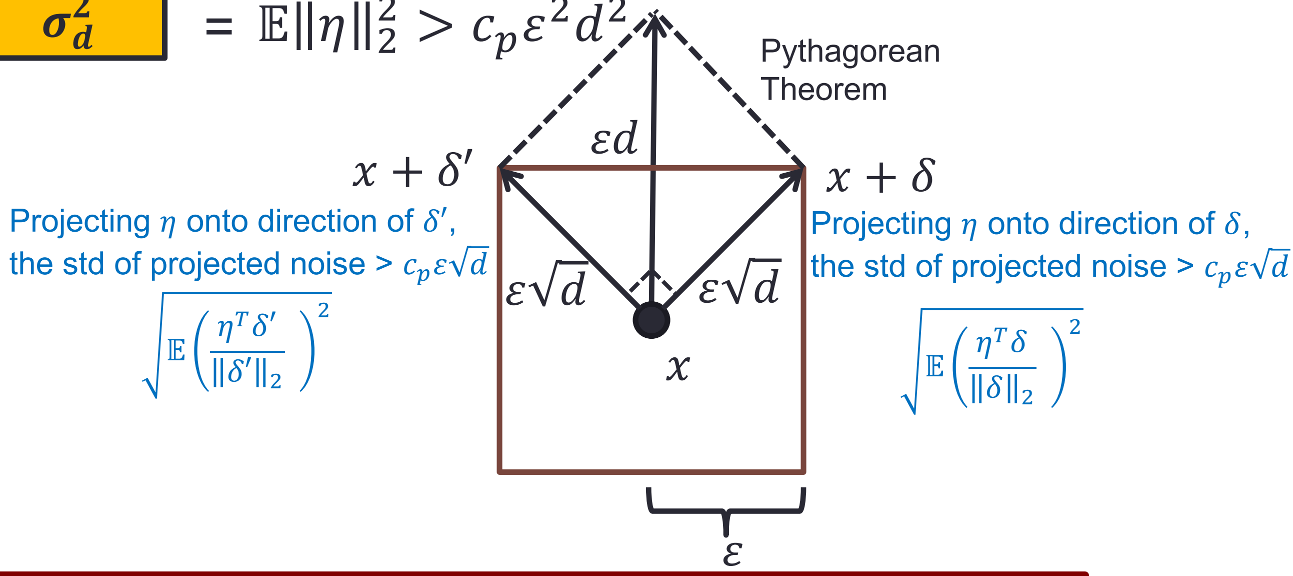
Reasonable question: why $\varepsilon < c_p \sigma_i / \sqrt{d}$ is inevitable?

Step 1: d-dimensional case

$$\overbrace{\boxed{\sigma_1^2} + \boxed{\sigma_2^2} + \boxed{\sigma_3^2} + \cdots\cdots + \boxed{\sigma_d^2}}^{d \text{ such entries}} \overset{\text{by def.}}{=} \mathbb{E}\|\eta\|_2^2 > c_p \varepsilon^2 d^2$$

$$\exists i, \ \boxed{\sigma_i^2} > c_p \varepsilon^2 d$$

Key intuition: The magnitude (std) of random noise in the direction should overwhelm that of adv. perturbation to cancel out its effect



Theorem 1 (Our hardness result, JMLR'20):

Given any input x, let η be noise from any distribution with variance of η_i being σ_i^2. If $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \leq \varepsilon$, then $\varepsilon < c_p \sigma_i / \sqrt{d}$ for 99% entries i.

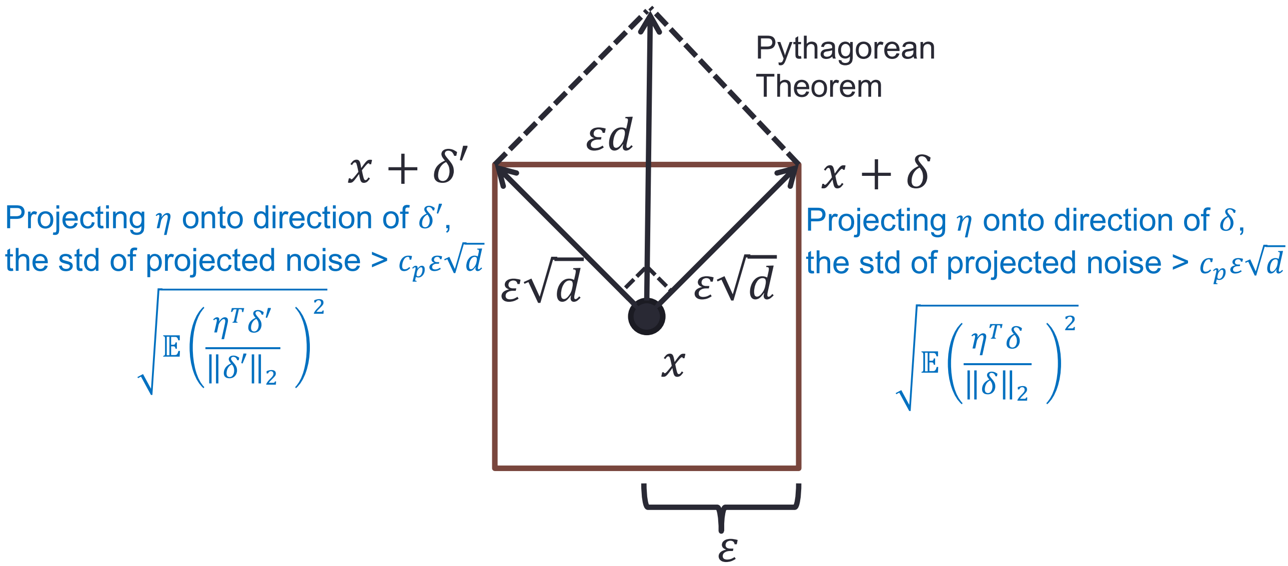
[Blum, Dick, Manoj, Zhang (α-β)'20] Random Smoothing Might be Unable to Certify ℓ_∞ Robustness, JMLR 2020

Intuition behind the Hardness Result

Reasonable question: why $\varepsilon < c_p \sigma_i / \sqrt{d}$ is inevitable?

Step 1: d-dimensional case

Step 2: repeat Step 1 for $(d-1)$-dimensional case by projecting out dimension i



Theorem 1 (Our hardness result, JMLR'20):

Given any input x, let η be noise from any distribution with variance of η_i being σ_i^2. If $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \le \varepsilon$, then $\varepsilon < c_p \sigma_i / \sqrt{d}$ for 99% entries i.

[Blum, Dick, Manoj, Zhang (α-β)'20] Random Smoothing Might be Unable to Certify ℓ_∞ Robustness, JMLR 2020

Intuition behind the Hardness Result

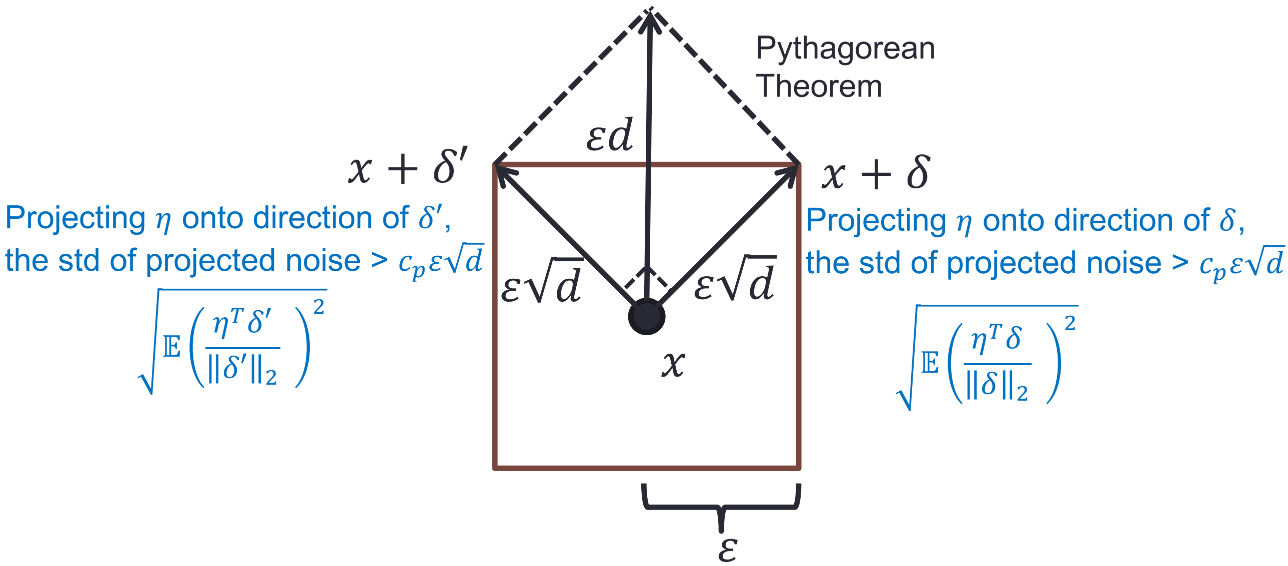
Reasonable question: why $\varepsilon < c_p\sigma_i/\sqrt{d}$ is inevitable?

Step 1: d-dimensional case

Step 2: repeat Step 1 for $(d-1)$-dimensional case by projecting out dimension i

Step 3: repeat Step 2 for $(d-2)$-dimensional case

… (repeat by $99\%d$ times)



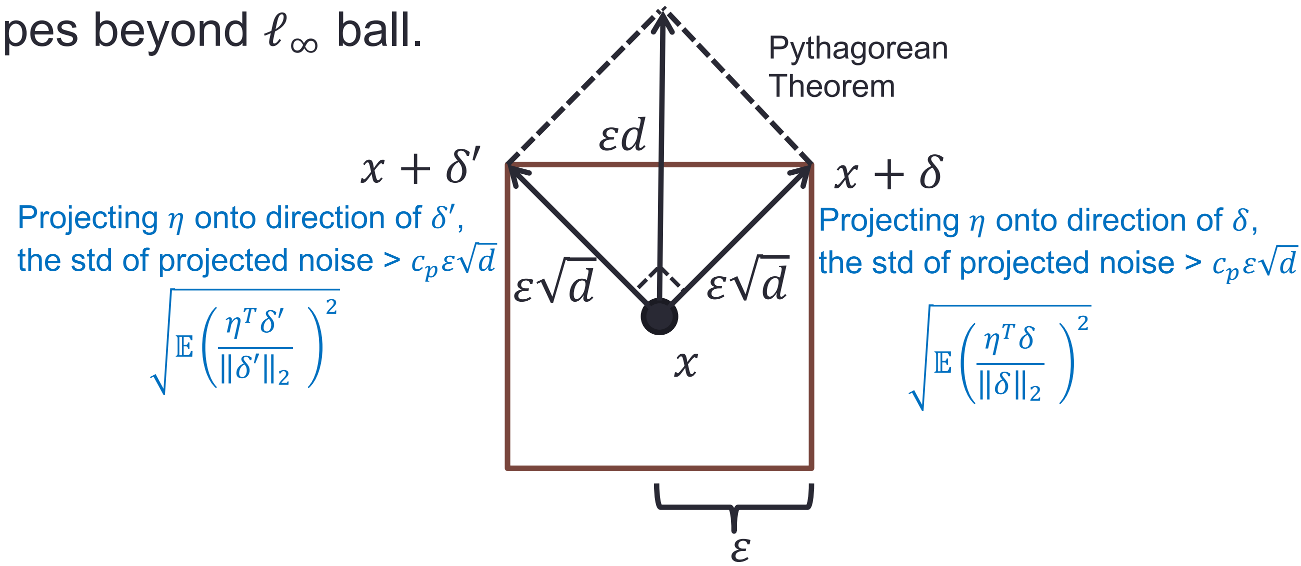
Theorem 1 (Our hardness result, JMLR'20):

Given any input x, let η be noise from any distribution with variance of η_i being σ_i^2. If $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \le \varepsilon$, then $\varepsilon < c_p\sigma_i/\sqrt{d}$ for 99% entries i.

[Blum, Dick, Manoj, Zhang (α-β)'20] Random Smoothing Might be Unable to Certify ℓ_∞ Robustness, JMLR 2020

Take-Home Message from the Hardness Result

- The $\sigma_i/\sqrt{d}$ dependence in the certified radius stems from the fact that the length of adversarial perturbation can be as large as $\varepsilon\sqrt{d}$ in the ℓ_∞ ball.
- (Current version of) random smoothing might be unable to certify ℓ_∞ robustness.
- The hardness can be extended to other shapes beyond ℓ_∞ ball.



Theorem 1 (Our hardness result, JMLR'20):

Given any input x, let η be noise from any distribution with variance of η_i being σ_i^2. If $g(x) = g(x+\delta)$ for any δ such that $\|\delta\|_\infty \le \varepsilon$, then $\varepsilon < c_p\sigma_i/\sqrt{d}$ for 99% entries i.

[Blum, Dick, Manoj, Zhang (α-β)'20] Random Smoothing Might be Unable to Certify ℓ_∞ Robustness, JMLR 2020

Overview of This Talk

Paradigms

Robustness

Adversarial Example

Random Noise

Mixed Random/Adversarial Corruption

What’s Next?

Adversarial Defenses

Empirical Defense

Norm-Bounded Adversarial Example

Unrestricted Adversarial Example

Certified Defense

Positive Result

Hardness Result

Applications

Unrestricted Adversarial Examples Challenge

ROBUSTBENCH

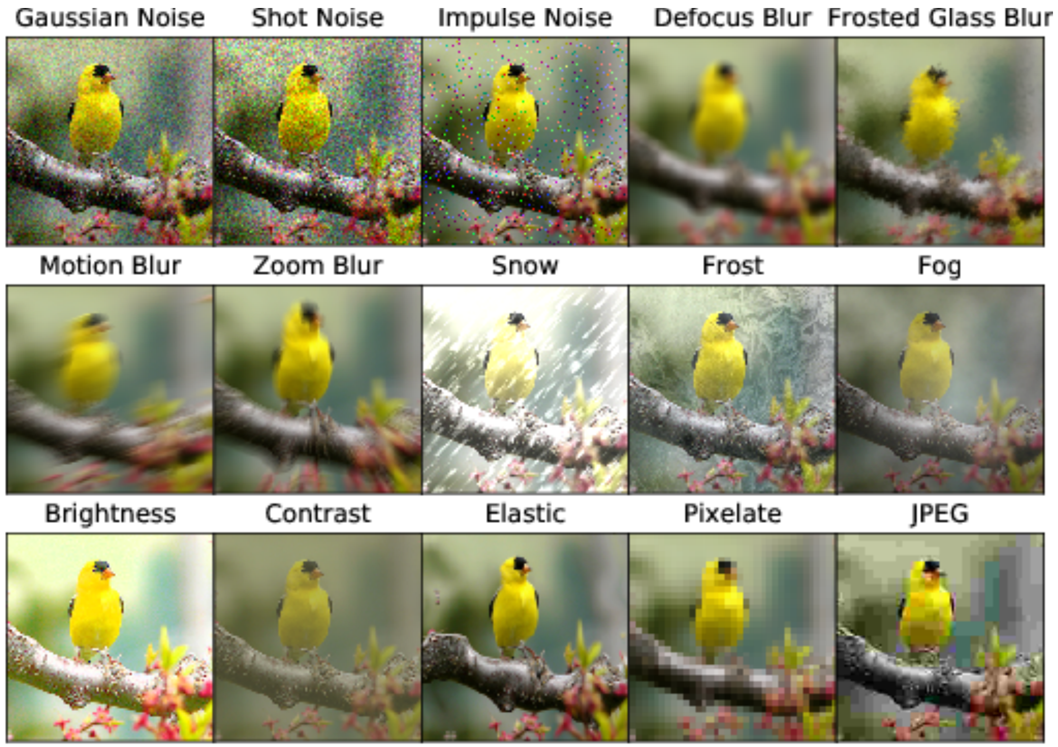
A standardized benchmark for adversarial robustness

GLUE

What's next for robustness?

- ❑ Certified robustness requires thinking beyond random smoothing
- ❑ Major issue with curve fitting: training phase should "mimic" the test phase

Out-of-distribution generalization (sample complexity) problem (ImageNet-C):



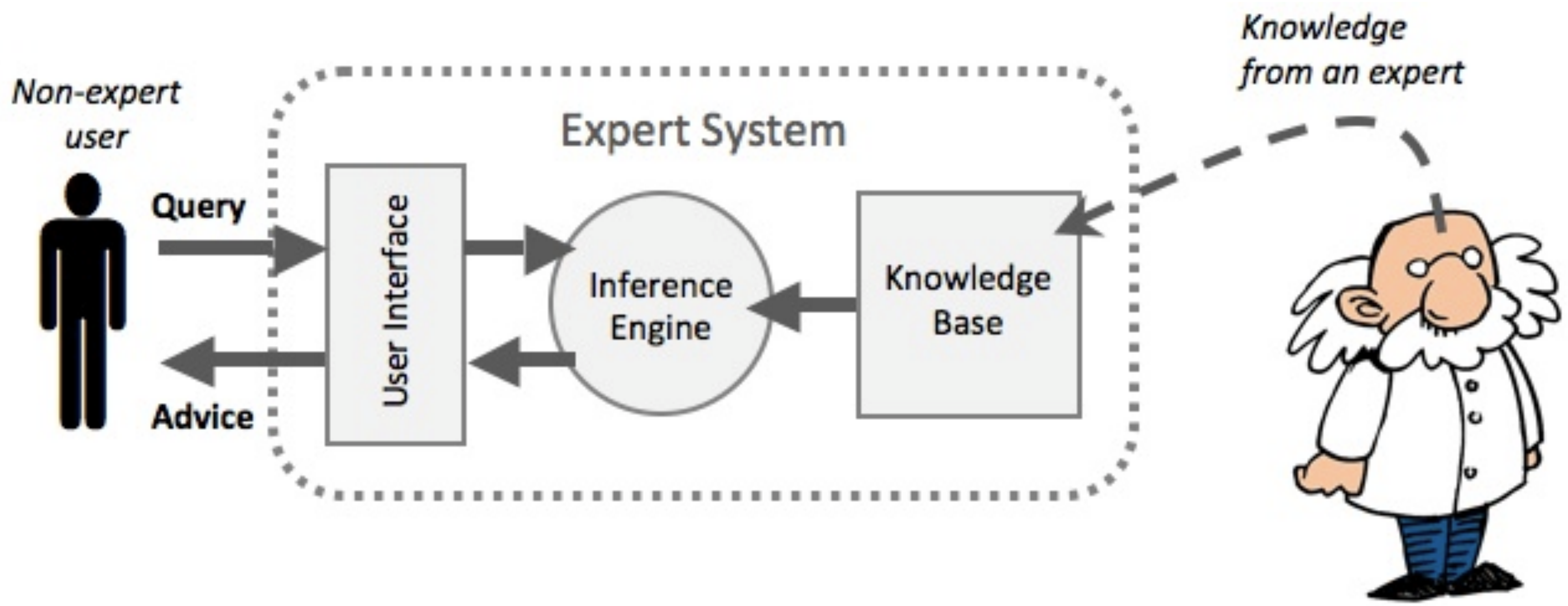
Training Phase:

$$\min_f [\mathbb{E}\, \phi(f(x)y) + \mathbb{E} \max_{X' \in \Delta(x)} \phi(f(x)f(x')/\lambda)]$$

Impossible to mimic ALL corruptions

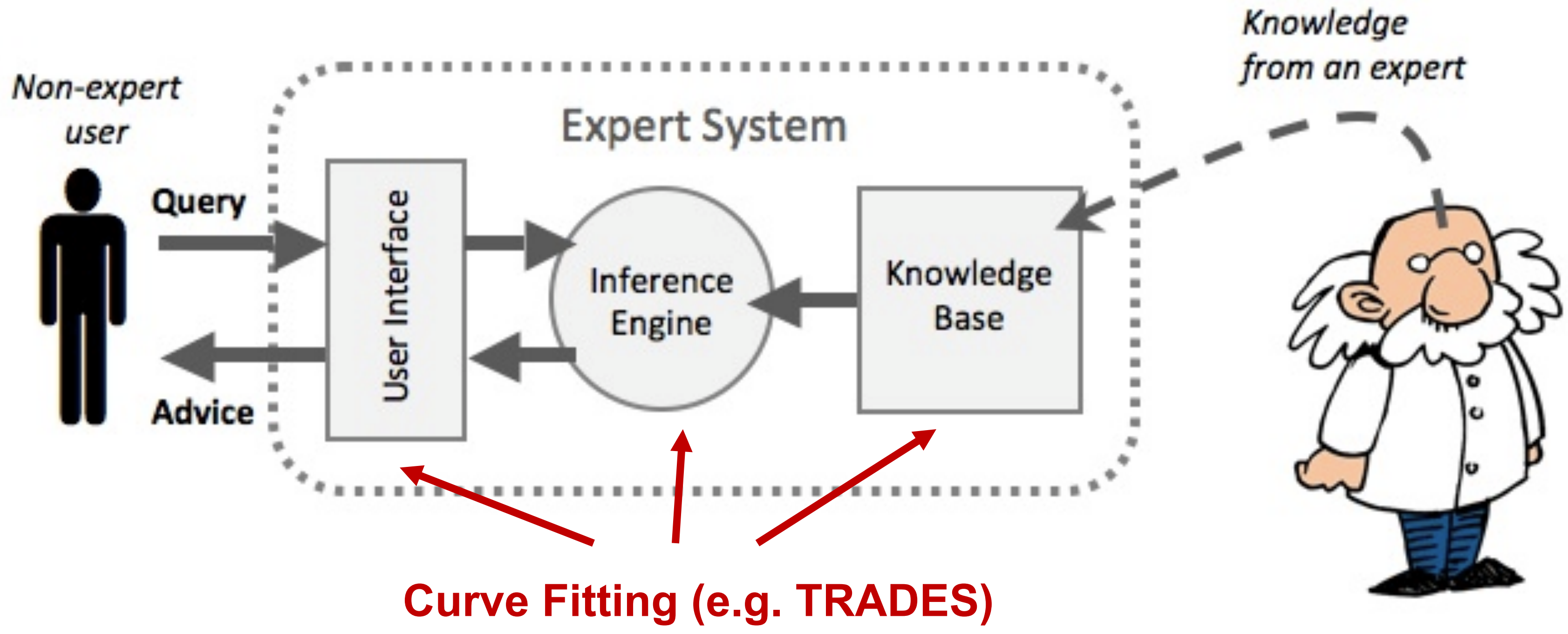
What's next for robustness?

- Certified robustness requires thinking beyond random smoothing
- Major issue with curve fitting: training phase should "mimic" the test phase
- Expert system: inference engine, knowledge base, human interface

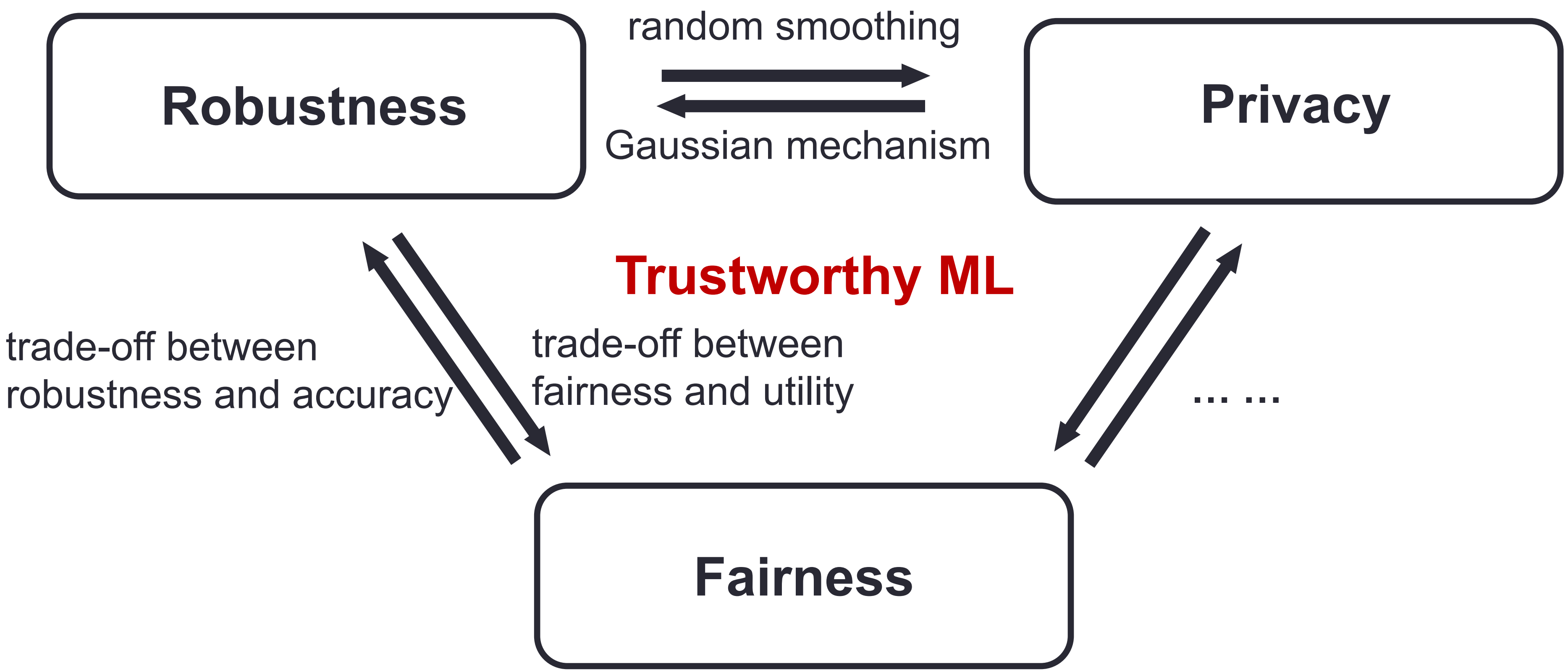


What's next for robustness?

- ❑ Knowledge Base: a huge organized set of knowledge about a particular subject
- ❑ Inference Engine: a set of rules on which to make decisions
- ❑ User Interface: human in the loop and human-computer interaction



Towards Trustworthy Machine Learning



Overview of This Talk

Robustness

Paradigms

Adversarial Example

Random Noise

Mixed Random/Adversarial Corruption

My Other Works on Machine Learning

Empirical Defense

Certified Defense

Adversarial Defenses

Norm-Bounded Adversarial Example

Unrestricted Adversarial Example

Positive Result

Hardness Result

Applications

Adversarial Vision Challenge

Model Track

Adversarial Vision Challenge

Untargeted Attack

Google

Unrestricted Adversarial Examples Challenge

RobustBench

A standardized benchmark for adversarial robustness

GLUE

Overview of My Works

Paradigms	Robustness		
	Adversarial Example	**Random Noise**	**Mixed Random/Adversarial Corruption**
Works	• **[ICML'19]** (TRADES) • **[JMLR'20]** (hardness of random smoothing) • **[NeurIPS'18]** (Adversarial Vision Challenge) • **[NeurIPS'20]** (trade-off between robustness and accuracy) • **[ECCV'20]** (adversarial patch attack)	• **[COLT'16]** (learning with Massart noise) • **[NeurIPS'17]** (s-concave dist.) • **[NeurIPS'17]** (power of comparison) • **[NeurIPS'20]** (new DL training method)	• **[JMLR'19], [ITCS'18]** (strong duality of robust PCA) • **[SODA'19]** (testing problem) • **[IEEE Trans. Info Theory'16]** (exact recoverability of robust PCA) • **[NeurIPS'16]** (online Robust PCA) • **[Proceeding of IEEE'18], [Book'17]** (applications in CV)
Other Works	[NeurIPS'19], [NeurIPS'19], [AISTATS'19], [ICALP'18], [ICML'17], [AAAI'17], [AAAI'15], [Neural Computation'15], [ECML'15], [BMVC'15], [Neurcomputing'14] [ICML'20], [ICML'20], [ICML'19], [ICML'16], [ICML'16]		

Overview of My Works

Paradigms	Robustness: Adversarial Example	Robustness: Random Noise	Robustness: Mixed Random/Adversarial Corruption
Works	• **[ICML'19]** (TRADES) • **[JMLR'20]** (hardness of random smoothing) • **[NeurIPS'18]** (Adversarial Vision Challenge) • **[NeurIPS'20]** (trade-off between robustness and accuracy) • **[ECCV'20]** (adversarial patch attack)	• **[COLT'16]** (learning with Massart noise) • **[NeurIPS'17]** (s-concave dist.) • **[NeurIPS'17]** (power of comparison) • **[NeurIPS'20]** (new DL training method)	• **[JMLR'19], [ITCS'18]** (strong duality of robust PCA) • **[SODA'19]** (testing problem) • **[IEEE Trans. Info Theory'16]** (exact recoverability of robust PCA) • **[NeurIPS'16]** (online Robust PCA) • **[Proceeding of IEEE'18], [Book'17]** (applications in CV)
Other Works	[NeurIPS'19], [NeurIPS'19], [AISTATS'19], [ICALP'18], [ICML'17], [AAAI'17], [AAAI'15], [Neural Computation'15], [ECML'15], [BMVC'15], [Neurcomputing'14] [ICML'20], [ICML'20], [ICML'19], [ICML'16], [ICML'16]			

Overview of My Works

Paradigms — **Robustness**

Adversarial Example

Works

- **[ICML’19]** (TRADES)
- **[JMLR’20]** (hardness of random smoothing)
- **[NeurIPS’18]** (Adversarial Vision Challenge)
- **[NeurIPS’20]** (trade-off between robustness and accuracy)
- **[ECCV’20]** (adversarial patch attack)

Random Noise

- **[COLT’16]** (learning with Massart noise)
- **[NeurIPS’17]** (s-concave dist.)
- **[NeurIPS’17]** (power of comparison)
- **[NeurIPS’20]** (new DL training method)

Mixed Random/Adversarial Corruption

- **[JMLR’19], [ITCS’18]** (strong duality of robust PCA)
- **[SODA’19]** (testing problem)
- **[IEEE Trans. Info Theory’16]** (exact recoverability of robust PCA)
- **[NeurIPS’16]** (online Robust PCA)
- **[Proceeding of IEEE’18], [Book’17]** (applications in CV)

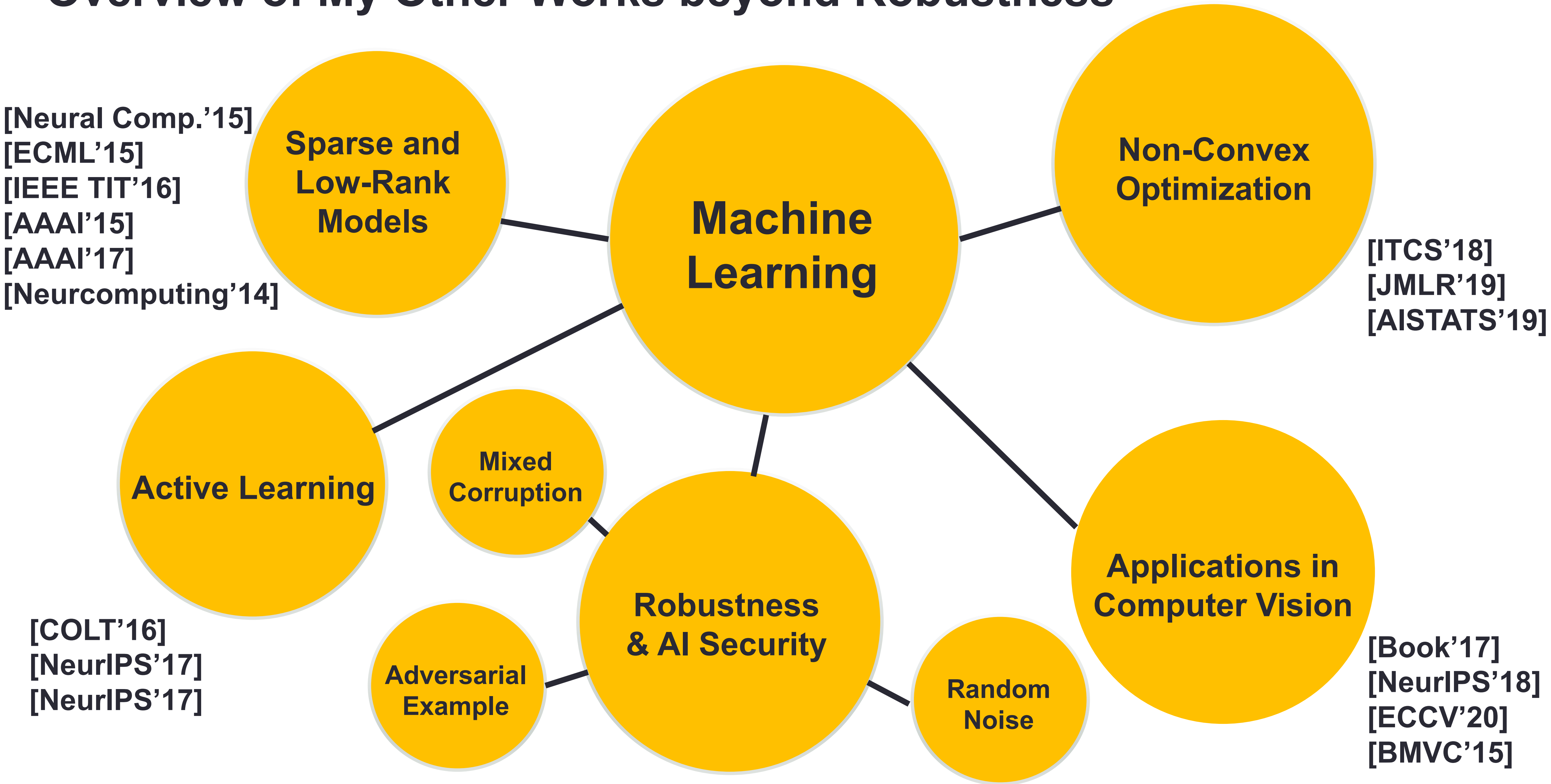
Other Works

[NeurIPS’19], [NeurIPS’19], [AISTATS’19], [ICALP’18], [ICML’17], [AAAI’17], [AAAI’15], [Neural Computation’15], [ECML’15], [BMVC’15], [Neurcomputing’14] [ICML’20], [ICML’20], [ICML’19], [ICML’16], [ICML’16]

Overview of My Works

Paradigms	Robustness		
	Adversarial Example	**Random Noise**	**Mixed Random/Adversarial Corruption**
Works	• **[ICML’19]** (TRADES) • **[JMLR’20]** (hardness of random smoothing) • **[NeurIPS’18]** (Adversarial Vision Challenge) • **[NeurIPS’20]** (trade-off between robustness and accuracy) • **[ECCV’20]** (adversarial patch attack)	• **[COLT’16]** (learning with Massart noise) • **[NeurIPS’17]** (s-concave dist.) • **[NeurIPS’17]** (power of comparison) • **[NeurIPS’20]** (new DL training method)	• **[JMLR’19], [ITCS’18]** (strong duality of robust PCA) • **[SODA’19]** (testing problem) • **[IEEE Trans. Info Theory’16]** (exact recoverability of robust PCA) • **[NeurIPS’16]** (online Robust PCA) • **[Proceeding of IEEE’18], [Book’17]** (applications in CV)
Other Works	**[NeurIPS’19], [NeurIPS’19], [AISTATS’19], [ICALP’18], [ICML’17], [AAAI’17], [AAAI’15], [Neural Computation’15], [ECML’15], [BMVC’15], [Neurcomputing’14] [ICML’20], [ICML’20], [ICML’19], [ICML’16], [ICML’16]**		

Overview of My Other Works beyond Robustness



Acknowledgements

DARPA

NSF

Almaden Institute IBM

Petuum

TTIC TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

Carnegie Mellon University

Thank You!

hongyanz@ttic.edu