Confidence-Calibrated Adversarial Training

Generalizing to Unseen Attacks

David Stutz, Matthias Hein, Bernt Schiele

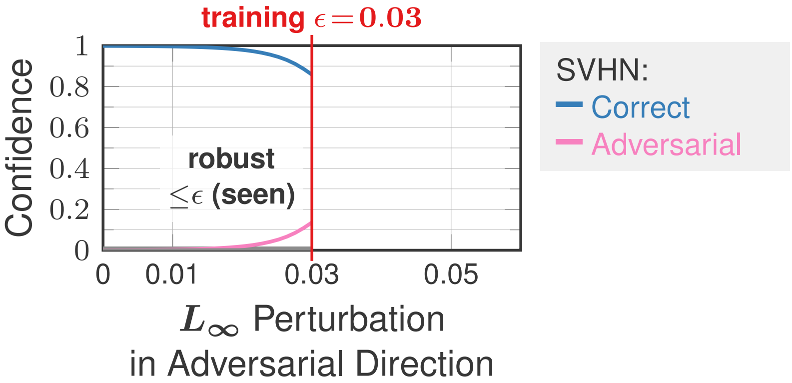
max planck institut informatik

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

2-Minute Overview

Problem: Robustness to *various* adversarial examples.

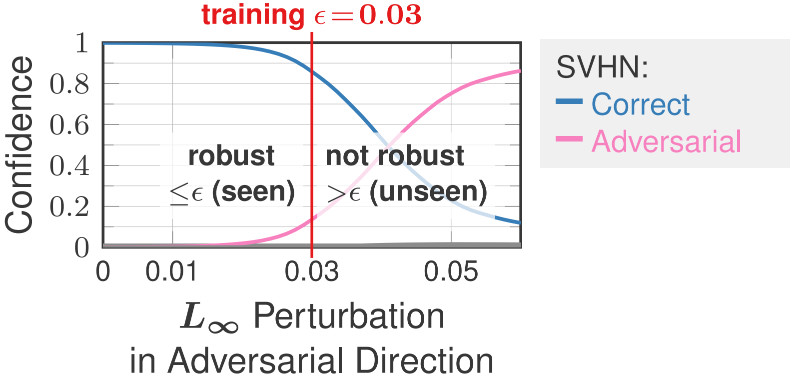
Adversarial training on L_∞ adversarial examples:



2-Minute Overview

Problem: Robustness to *various* adversarial examples.

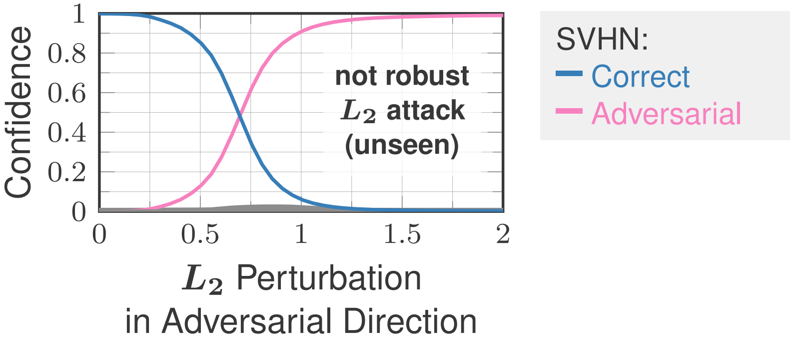
Adversarial training on L_∞ adversarial examples:



2-Minute Overview

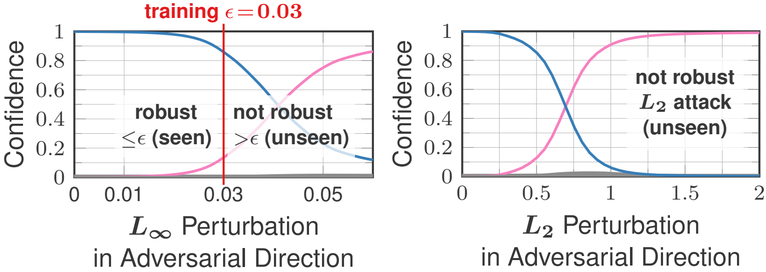
Problem: Robustness to various adversarial examples.

Adversarial training on L_∞ adversarial examples:



2-Minute Overview

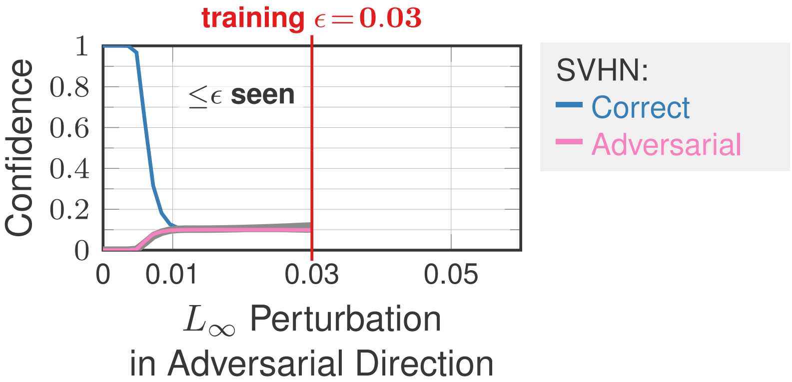
Summary of adversarial training:



- High-confidence on adversarial examples ($\leq\epsilon$).
- *No* generalization to larger/other L_p perturbations.
- Behavior not meaningful for arbitrarily large ϵ.

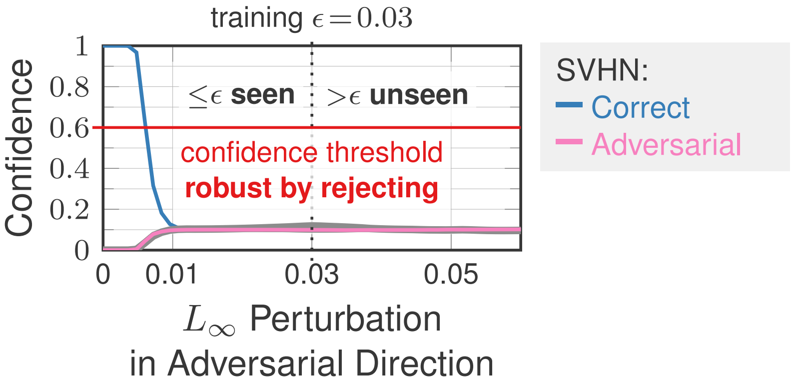
2-Minute Overview

Confidence-calibrated adversarial training (L_∞ *only*):



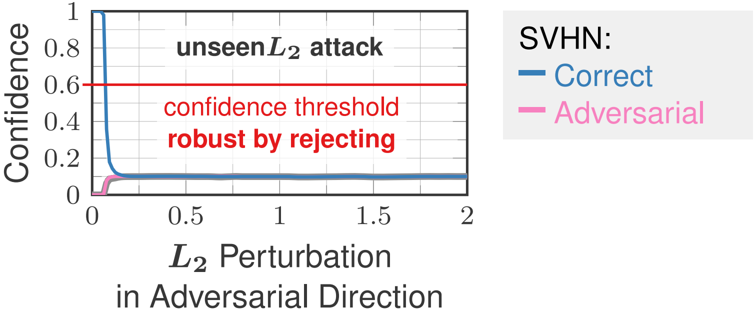
2-Minute Overview

Confidence-calibrated adversarial training (L_∞ *only*):



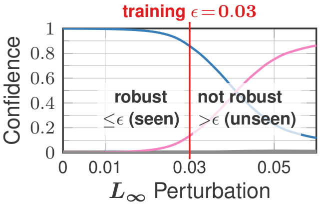
2-Minute Overview

Confidence-calibrated adversarial training (L_∞ *only*):



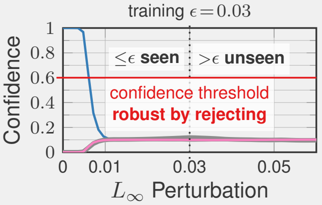
2-Minute Overview

Adversarial training:



- High-confidence on adversarial examples.
- No robustness to *unseen* perturbations.

Confidence-calibrated adversarial training:



- Low-confidence on adversarial examples.
- **Robustness to *unseen* perturbations** by confidence thresholding.

Interested?

More details:

Paper & code: `davidstutz.de/ccat`

Contact: `david.stutz@mpi-inf.mpg.de`

Interested?

More details:

Paper & code: `davidstutz.de/ccat`

Contact: `david.stutz@mpi-inf.mpg.de`

Outline:

1. Problems of adversarial training
2. *Confidence-calibrated adversarial training*
3. Confidence-thresholded robust test error
4. Results on SVHN and CIFAR10

Problems of Adversarial Training

Min-max formulation:

classifier

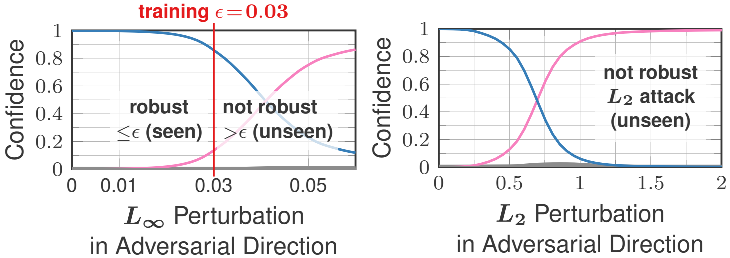
$$\min_w \mathbb{E}_{p(x,y)} \left[\max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}(f(x + \delta; w), y) \right].$$

minimizing cross-entropy yields high-confidence

Problems of Adversarial Training

Min-max formulation:

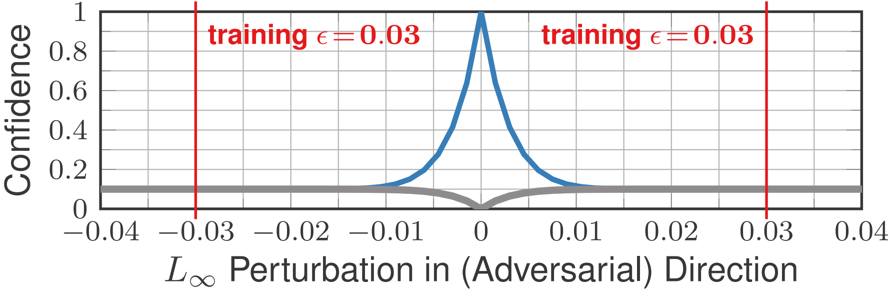
$$\min_w \mathbb{E}_{p(x,y)} \left[\max_{\|\delta\|_\infty \leq \epsilon} \mathcal{L}(f(x + \delta; w), y) \right].$$



- Robustness does *not* generalize to unseen attacks.

Confidence-Calibrated Adversarial Training

1 Transition to uniform distribution on adversarial examples within the ϵ-ball:

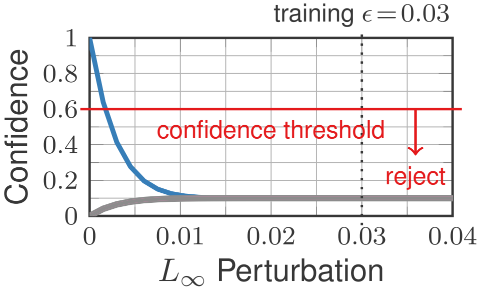


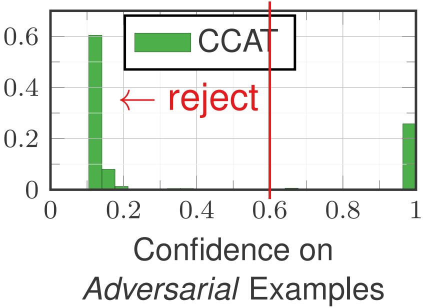
- Low-confidence extrapolated beyond ϵ-ball.

Confidence-Calibrated Adversarial Training

1 Transition to **low confidence on adversarial examples** within the ϵ-ball.

2 **Reject low-confidence (adversarial) examples** via confidence-thresholding:





1 Transition to Low Confidence

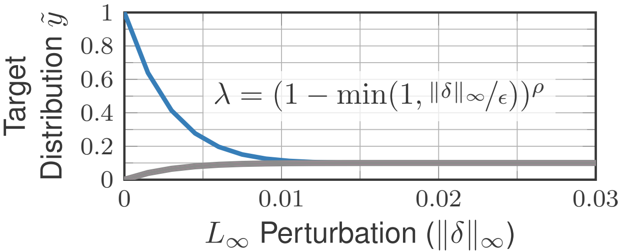
1. Compute high-confidence adversarial examples:

$$\tilde{\delta} = \max_{\|\delta\|_\infty \leq \epsilon} \max_{k \neq y} f_k(x + \delta; w)$$

confidence of class k

2. Impose target distribution via cross-entropy loss:

$$\tilde{y} = \lambda \text{ one\_hot}(y) + (1 - \lambda)^1/K$$



1 Transition to Low Confidence

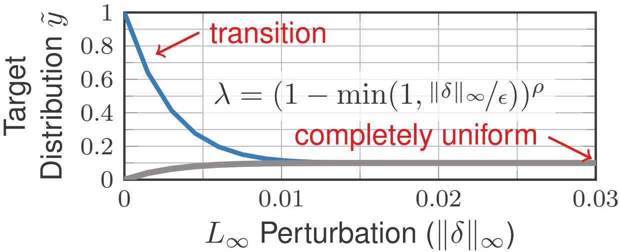
1. Compute high-confidence adversarial examples:

$$\tilde{\delta} = \max_{\|\delta\|_\infty \leq \epsilon} \max_{k \neq y} f_k(x + \delta; w)$$

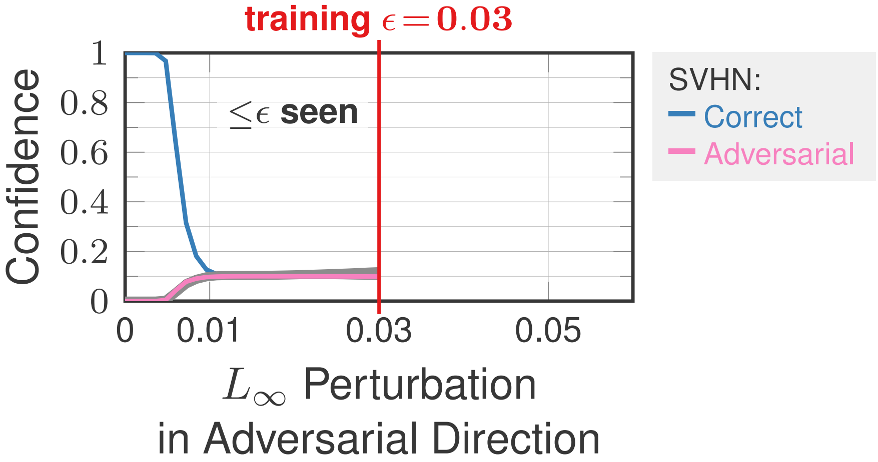
confidence of class k

2. Impose target distribution via cross-entropy loss:

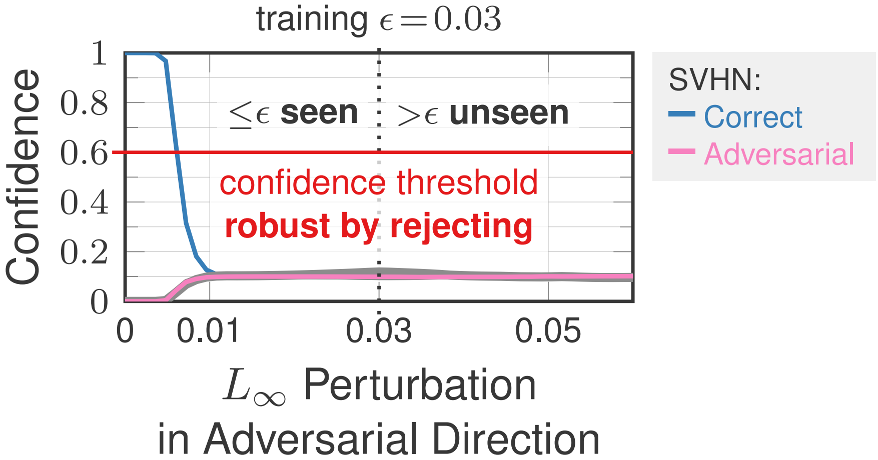
$$\tilde{y} = \lambda \text{ one\_hot}(y) + (1 - \lambda)1/K$$



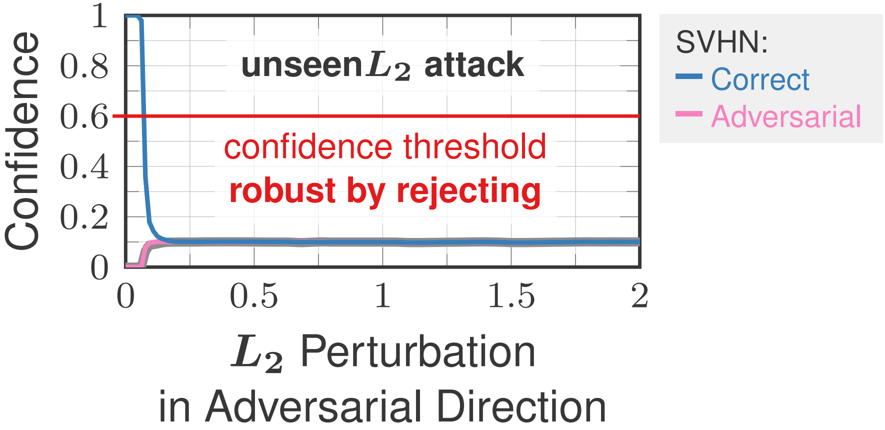
2 Robustness by Confidence Thresholding



2 Robustness by Confidence Thresholding

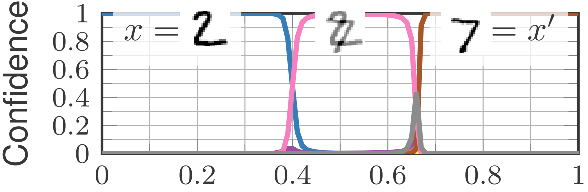


2 Robustness by Confidence Thresholding

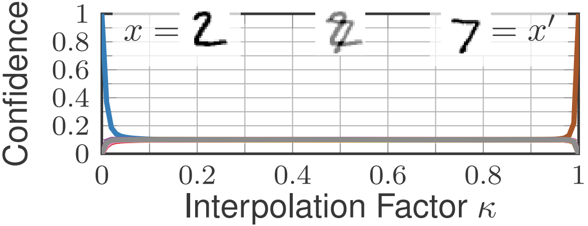


2 Meaningful Extrapolation of Confidence

Adversarial training:



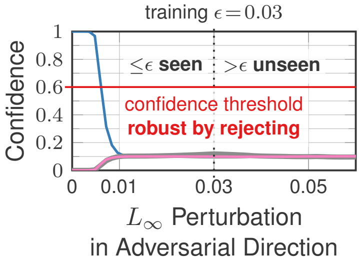
Confidence-calibrated adversarial training:

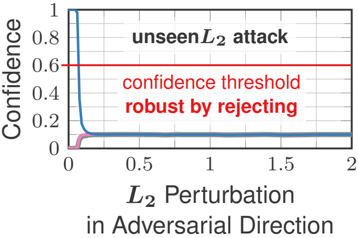


Summary: Generalizable Robustness

Confidence-calibrated adversarial training:

1. Transition: low confidence on adversarial examples.
2. Reject low-confidence (adversarial) examples.





- **Robustness to previously *unseen* perturbations.**

“Standard” Robust Test Error RErr

= error on test examples that are “attacked”.

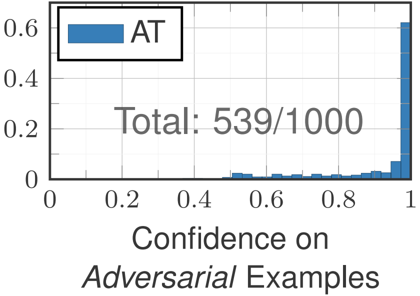
Adversarial Training (AT):
57.3% RErr

Ours (CCAT):
97.8% RErr

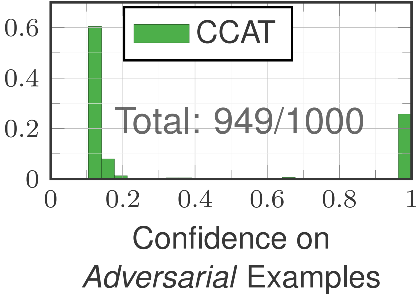
“Standard” Robust Test Error RErr

= error on test examples that are “attacked”.

Adversarial Training (AT):
57.3% RErr



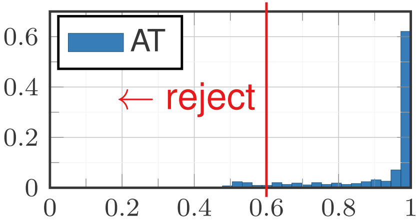
Ours (CCAT):
97.8% RErr



Confidence-Thresholded RErr

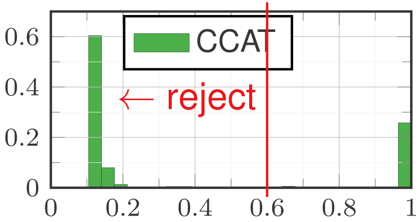
= error on test examples that are "attacked"
and ***pass*** *confidence thresholding*.

Adversarial Training (AT):
56% (−1.3%)



Confidence on
Adversarial Examples

Ours (CCAT):
39.1% (−58.7%)

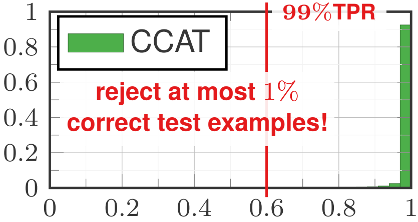


Confidence on
Adversarial Examples

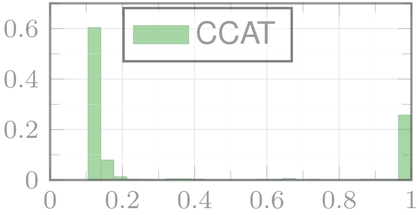
Determine Confidence Threshold

- Independent of adversarial examples.
- Avoid incorrectly rejecting (clean) test examples.

Confidence threshold at 99% true positive rate TPR:



Confidence on
Test Examples



Confidence on
Adversarial Examples

Results

Datasets: SVHN, CIFAR10, 1000 test examples.

Per-example, worst-case (thresholded) RErr across:

Attack	Iterations	Restarts
PGD	200-1000	10-50
Query-Limited†	1000	11
Simple†	1000	10
Square†	5000	1
Geometry†	1000	1
Random†	–	5000

(† Black-box attacks.)

- Attacks adapted to maximize confidence.

SVHN: Generalization to Unseen Attacks

SVHN: RErr $\downarrow$ in % at 99%TPR					
	L_∞ $\epsilon = 0.03$				
	seen				
AT	56.0				
CCAT	**39.1**				

(Lower RErr $\downarrow$ means “better” robustness.)

SVHN: Generalization to Unseen Attacks

SVHN: RErr ↓ in % at 99%TPR					
	L_∞ $\epsilon = 0.03$	L_∞ $\epsilon = 0.06$	L_2 $\epsilon = 2$	L_1 $\epsilon = 24$	L_0 $\epsilon = 10$
	seen	**unseen**	**unseen**	**unseen**	**unseen**
AT	56.0				
CCAT	**39.1**				

(Lower RErr ↓ means “better” robustness.)

SVHN: Generalization to Unseen Attacks

SVHN: RErr $\downarrow$ in % at 99%TPR					
	L_∞ $\epsilon = 0.03$	L_∞ $\epsilon = 0.06$	L_2 $\epsilon = 2$	L_1 $\epsilon = 24$	L_0 $\epsilon = 10$
	seen	**unseen**	**unseen**	**unseen**	**unseen**
AT	56.0	88.4	99.4	99.5	73.6
CCAT	**39.1**	**53.1**	**29.0**	**31.7**	**3.5**

(Lower RErr $\downarrow$ means “better” robustness.)

Cifar10: Generalization to Unseen Attacks

CIFAR10: RErr $\downarrow$ in % at 99%TPR					
	L_∞ $\epsilon = 0.03$	L_∞ $\epsilon = 0.06$	L_2 $\epsilon = 2$	L_1 $\epsilon = 24$	L_0 $\epsilon = 10$
	seen	**unseen**	**unseen**	**unseen**	**unseen**
AT	**62.7**	93.7	98.4	98.4	72.4
CCAT	67.9	92.0	**51.8**	**58.5**	**20.3**

(Lower RErr $\downarrow$ means "better" robustness.)

"Unconventional" Attacks

CIFAR10: RErr, FPR and CErr at 99%TPR			
	adv. frames	distal	corrupted
	unseen	**unseen**	**unseen**
	RErr $\downarrow$	FPR$\downarrow$	CErr$\downarrow$
Normal	96.6	83.3	12.3
AT	78.7	75.0	16.2
CCAT	**65.1**	**0**	**8.5**

(FPR: false positive rate, fraction of non-rejected adv. examples.)

(CErr: test error on corrupted examples after thresholding.)

Improved Accuracy

	SVHN: Err ↓ in %		CIFAR10: Err ↓ in %	
	no reject	99% TPR	no reject	99% TPR
Normal	3.6	2.6	**8.3**	7.4
AT	3.4	2.5	16.6	15.5
CCAT	**2.9**	**2.1**	10.1	**6.7**

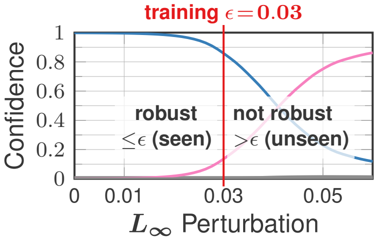
(Err: test error before and after thresholding.)

Confidence-Calibrated Adversarial Training

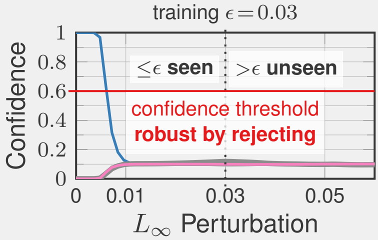
Low-confidence on adversarial examples and *beyond*.

- Robustness generalizes to unseen attacks.
- Accuracy improves.

Adversarial training:



Ours:



Paper & code: `davidstutz.de/ccat`