Confidence-Calibrated Adversarial Training
Generalizing to Unseen Attacks
David Stutz, Matthias Hein, Bernt Schiele
**2-Minute Overview**

Problem: Robustness to *various* adversarial examples.

Adversarial training on $L_{\infty}$ adversarial examples:

![Graph showing confidence vs. perturbation]

**SVHN:**
- Correct
- Adversarial

$\text{Confidence} \leq \epsilon \text{(seen)}$

$\text{training } \epsilon = 0.03$
Problem: Robustness to *various* adversarial examples.

Adversarial training on $L_\infty$ adversarial examples:

![Graph showing confidence vs. perturbation for SVHN dataset](image)

SVHN:
- **Correct**
- **Adversarial**

Confidence-Calibrated Adversarial Training – David Stutz
2-Minute Overview

Problem: Robustness to various adversarial examples.

Adversarial training on $L_\infty$ adversarial examples:

SVHN:

- **Correct**
- **Adversarial**

![Graph showing confidence and $L_2$ perturbation in adversarial direction](graph.png)
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 $\epsilon$. 
Confidence-calibrated adversarial training ($L_\infty$ only):

2-Minute Overview

Confidence-calibrated adversarial training ($L_\infty$ only):

SVHN:
- Correct
- Adversarial
Confidence-calibrated adversarial training ($L_\infty$ only):

$$\begin{align*}
\text{confidence threshold} & \leq \epsilon \text{ seen} \\
& > \epsilon \text{ unseen}
\end{align*}$$

Robust by rejecting $\epsilon = 0.03$

SVHN:
- Correct
- Adversarial
2-Minute Overview

Confidence-calibrated adversarial training ($L_\infty$ only):

Confidence
\[
\begin{array}{c}
\text{Confidence-calibrated adversarial training ($L_\infty$ only):} \\
\text{unseen $L_2$ attack} \\
\text{confidence threshold} \\
\text{robust by rejecting} \\
\end{array}
\]

$L_2$ Perturbation
in Adversarial Direction

SVHN:
- Correct
- Adversarial
2-Minute Overview

Adversarial training:

- High-confidence on adversarial examples.
- No robustness to \textit{unseen} perturbations.

Confidence-calibrated adversarial training:

- Low-confidence on adversarial examples.
- \textbf{Robustness to \textit{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:

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

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.
1. Transition to uniform distribution on adversarial examples within the $\epsilon$-ball:

- Low-confidence extrapolated beyond $\epsilon$-ball.
1. Transition to low confidence on adversarial examples within the $\epsilon$-ball.

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

\[
\begin{align*}
&\text{Confidence on} \\
&\text{Adversarial Examples}
\end{align*}
\]

Confidence-Calibrated Adversarial Training – David Stutz
1 Transition to Low Confidence

1. Compute high-confidence adversarial examples:

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

(Confidence of class \( k \))

2. Impose target distribution via cross-entropy loss:

\[ \tilde{y} = \lambda \text{one_hot}(y) + (1 - \lambda)^{1/K} \]

\[ \lambda = (1 - \min(1, \|\delta\|_\infty / \epsilon))^{\rho} \]
1. Compute high-confidence adversarial examples:

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

2. Impose target distribution via cross-entropy loss:

\[ \tilde{y} = \lambda \text{one}_\text{hot}(y) + (1 - \lambda)^{1/K} \]

\[ \lambda = (1 - \min(1, \|\delta\|_\infty / \epsilon))^{\rho} \]

\[ L_\infty \text{ Perturbation (}\|\delta\|_\infty) \]

Transition to Low Confidence

Confidence-Calibrated Adversarial Training – David Stutz
Robustness by Confidence Thresholding

\[ L_\infty \text{ Perturbation in Adversarial Direction} \]

Confidence-Calibrated Adversarial Training – David Stutz

SVHN:
- Correct
- Adversarial

\[ \leq \epsilon \text{ seen} \]

\[ \text{training } \epsilon = 0.03 \]
Robustness by Confidence Thresholding

Confidence-Calibrated Adversarial Training – David Stutz
Robustness by Confidence Thresholding

SVHN:
- Correct
- Adversarial

unseen $L_2$ attack
confidence threshold
robust by rejecting

$L_2$ Perturbation
in Adversarial Direction
2 Meaningful Extrapolation of Confidence

Adversarial training:

Confidence-calibrated adversarial training:

Confidence-Calibrated Adversarial Training – David Stutz
Summary: Generalizable Robustness

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

- Robustness to previously unseen perturbations.
“Standard” Robust Test Error RErr

= error on test examples that are “attacked”.

<table>
<thead>
<tr>
<th>Adversarial Training (AT):</th>
<th>Ours (CCAT):</th>
</tr>
</thead>
<tbody>
<tr>
<td>57.3% REErr</td>
<td>97.8% REErr</td>
</tr>
</tbody>
</table>
“Standard” Robust Test Error RErr

= error on test examples that are “attacked”.

Adversarial Training (AT):
57.3% RErr

Total: 539/1000

Confidence on Adversarial Examples

Ours (CCAT):
97.8% RErr

Total: 949/1000

Confidence on Adversarial Examples
**Confidence-Thresholded RErr**

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

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

Ours (CCAT):  
39.1% (−58.7%)

Confidence on Adversarial Examples

Confidence-Calibrated Adversarial Training – David Stutz
Determine Confidence Threshold

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

Confidence threshold at 99% true positive rate TPR:

![Graph showing confidence on test and adversarial examples with CCAT, indicating rejection at most 1% correct test examples at 99% TPR.](image)
### Results

Datasets: SVHN, CIFAR10, 1000 test examples.

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

<table>
<thead>
<tr>
<th>Attack</th>
<th>Iterations</th>
<th>Restarts</th>
</tr>
</thead>
<tbody>
<tr>
<td>PGD</td>
<td>200-1000</td>
<td>10-50</td>
</tr>
<tr>
<td>Query-Limited†</td>
<td>1000</td>
<td>11</td>
</tr>
<tr>
<td>Simple†</td>
<td>1000</td>
<td>10</td>
</tr>
<tr>
<td>Square†</td>
<td>5000</td>
<td>1</td>
</tr>
<tr>
<td>Geometry†</td>
<td>1000</td>
<td>1</td>
</tr>
<tr>
<td>Random†</td>
<td>–</td>
<td>5000</td>
</tr>
</tbody>
</table>

† Black-box attacks.

- Attacks adapted to maximize confidence.
SVHN: Generalization to Unseen Attacks

<table>
<thead>
<tr>
<th>SVHN: RErr ↓ in % at 99% TPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_\infty$</td>
</tr>
<tr>
<td>$\epsilon = 0.03$</td>
</tr>
<tr>
<td>seen</td>
</tr>
<tr>
<td>AT</td>
</tr>
<tr>
<td>56.0</td>
</tr>
<tr>
<td>CCAT</td>
</tr>
<tr>
<td>39.1</td>
</tr>
</tbody>
</table>

(Lower RErr ↓ means “better” robustness.)
SVHN: Generalization to Unseen Attacks

<table>
<thead>
<tr>
<th>SVHN: RErr ↓ in % at 99% TPR</th>
<th>$L_\infty$ $\epsilon = 0.03$</th>
<th>$L_\infty$ $\epsilon = 0.06$</th>
<th>$L_2$ $\epsilon = 2$</th>
<th>$L_1$ $\epsilon = 24$</th>
<th>$L_0$ $\epsilon = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>seen</td>
<td><strong>seen</strong></td>
<td><strong>unseen</strong></td>
<td><strong>unseen</strong></td>
<td><strong>unseen</strong></td>
<td><strong>unseen</strong></td>
</tr>
</tbody>
</table>

| AT                          | 56.0                  |                     |                     |                     |                     |
| CCAT                        | **39.1**              |                     |                     |                     |                     |

(Lower RErr ↓ means “better” robustness.)
## SVHN: Generalization to Unseen Attacks

<table>
<thead>
<tr>
<th></th>
<th>$L_\infty$ $\epsilon = 0.03$</th>
<th>$L_\infty$ $\epsilon = 0.06$</th>
<th>$L_2$ $\epsilon = 2$</th>
<th>$L_1$ $\epsilon = 24$</th>
<th>$L_0$ $\epsilon = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>seen</strong></td>
<td>56.0</td>
<td>88.4</td>
<td>99.4</td>
<td>99.5</td>
<td>73.6</td>
</tr>
<tr>
<td><strong>unseen</strong></td>
<td>39.1</td>
<td>53.1</td>
<td>29.0</td>
<td>31.7</td>
<td>3.5</td>
</tr>
</tbody>
</table>

(Lower REr↓ means “better” robustness.)
CIFAR10: RErr ↓ in % at 99% TPR

<table>
<thead>
<tr>
<th></th>
<th>$L_\infty$</th>
<th>$L_\infty$</th>
<th>$L_2$</th>
<th>$L_1$</th>
<th>$L_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon$ = 0.03</td>
<td>$\epsilon$ = 0.06</td>
<td>$\epsilon$ = 2</td>
<td>$\epsilon$ = 24</td>
<td>$\epsilon$ = 10</td>
<td></td>
</tr>
<tr>
<td>seen</td>
<td>unseen</td>
<td>unseen</td>
<td>unseen</td>
<td>unseen</td>
<td></td>
</tr>
<tr>
<td>AT</td>
<td>62.7</td>
<td>93.7</td>
<td>98.4</td>
<td>98.4</td>
<td>72.4</td>
</tr>
<tr>
<td>CCAT</td>
<td>67.9</td>
<td>92.0</td>
<td>51.8</td>
<td>58.5</td>
<td>20.3</td>
</tr>
</tbody>
</table>

(Lower RErr ↓ means “better” robustness.)
## CIFAR10: RErr, FPR and CErr at 99% TPR

<table>
<thead>
<tr>
<th></th>
<th>adv. frames</th>
<th>distal</th>
<th>corrupted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>unseen</td>
<td>unseen</td>
<td>unseen</td>
</tr>
<tr>
<td><strong>RErr ↓</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>96.6</td>
<td>83.3</td>
<td>12.3</td>
</tr>
<tr>
<td>AT</td>
<td>78.7</td>
<td>75.0</td>
<td>16.2</td>
</tr>
<tr>
<td>CCAT</td>
<td>65.1</td>
<td>0</td>
<td>8.5</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th></th>
<th><strong>SVHN:</strong></th>
<th></th>
<th><strong>CIFAR10:</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Err ↓ in %</td>
<td></td>
<td>Err ↓ in %</td>
</tr>
<tr>
<td>no</td>
<td>99%</td>
<td></td>
<td>99%</td>
</tr>
<tr>
<td>reject</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>3.6</td>
<td>2.6</td>
<td>8.3</td>
</tr>
<tr>
<td>AT</td>
<td>3.4</td>
<td>2.5</td>
<td>16.6</td>
</tr>
<tr>
<td>CCAT</td>
<td>2.9</td>
<td>2.1</td>
<td>10.1</td>
</tr>
</tbody>
</table>

(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:

```
Confidence  | L_∞ Perturbation
-----------------|------------------
0.00  | 0.01  | 0.03  | 0.05  | 0.2  | 0.4  | 0.6  | 0.8  | 1
---   |      |      |      |      |     |     |     |     |
0.20  |      |      |      |      |     |     |     |     |
0.40  |      |      |      |      |     |     |     |     |
0.60  |      |      |      |      |     |     |     |     |
0.80  |      |      |      |      |     |     |     |     |
1.00  |      |      |      |      |     |     |     |     |
```

confidence threshold

robust by rejecting

Paper & code: davidstutz.de/ccat