Certified Robustness to Adversarial Examples with Differential Privacy

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Suman Jana
Columbia University

Code: https://github.com/columbia/pixeldp
Contact: mathias@cs.columbia.edu
Deep Learning

• Deep Neural Networks (DNNs) deliver remarkable performance on many tasks.

• DNNs are increasingly deployed, including in attack-prone contexts:

The New York Times

Taylor Swift Said to Use Facial Recognition to Identify Stalkers

By Sopan Deb, Natasha Singer - Dec. 13, 2018
Example

- **Input x**
  - Layer 1
  - Layer 2
  - Layer 3
  - Softmax

- Q(x):
  - ticket 1: 0.1
  - ticket 2: 0.2
  - ticket 3: 0.1
  - no ticket: 0.6

- Graph:
  - Q(x) vs. ticket
  - ticket 1: 0.1
  - ticket 2: 0.2
  - ticket 3: 0.1
  - no ticket: 0.6
Example

But DNNs are vulnerable to adversarial example attacks.
Example

But DNNs are vulnerable to adversarial example attacks.
Accuracy under attack

Inception-v3 DNN on ImageNet dataset.

<table>
<thead>
<tr>
<th>Object</th>
<th>Accuracy under attack</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giant panda</td>
<td>( | \alpha |_2 = 0.52 )</td>
</tr>
<tr>
<td>Teddy bear</td>
<td>( | \alpha |_2 = 1.06 )</td>
</tr>
<tr>
<td>Teapot</td>
<td>( | \alpha |_2 )</td>
</tr>
</tbody>
</table>

Size of attack \( \alpha \) (2-norm)

Accuracy (top 1)
Best-effort approaches

1. Evaluate accuracy under attack:
   • Launch an attack on examples in a test set.
   • Compute accuracy on the attacked examples.
2. Improve accuracy under attack:
   • Many approaches: e.g. train on adversarial examples.
     (e.g Goodfellow+ '15; Papernot+ '16; Buckman+ '18; Guo+ '18)

Problem: both steps are attack specific, leading to an arms race that attackers are winning.
  (e.g Carlini-Wagner '17; Athalye+ '18)
Key questions

- **Guaranteed accuracy**: what is my minimum accuracy under any attack?
- **Prediction robustness**: given a prediction can any attack change it?
Key questions

- **Guaranteed accuracy**: what is my minimum accuracy under any attack?
- **Prediction robustness**: given a prediction can any attack change it?

- A few recent approaches with provable guarantees. (e.g. Wong-Kolter '18; Raghunathan+ '18; Wang+ '18)
- **Poor scalability in terms of**:
  - Input dimension (e.g. number of pixels).
  - DNN size.
  - Size of training data.
Key questions

- **Guaranteed accuracy**: what is my minimum accuracy under any attack?
- **Prediction robustness**: given a prediction can any attack change it?

- My defense **PixelDP** gives answers for norm bounded attacks.
- Key idea: novel use of **differential privacy** theory at prediction time.
- The **most scalable** approach: first provable guarantees for large models on ImageNet!
PixelDP outline

Motivation

Design

Evaluation
Key idea

- Problem: small input perturbations create large score changes.
Key idea

- Problem: small input perturbations create large score changes.
- Idea: **design a DNN with bounded maximum score changes** (leveraging Differential Privacy theory).

\[ \sqrt{\sum \alpha_i^2} = \| \alpha \|_2 \leq L \]

![Diagram of a DNN with layers and softmax output](image)

Graph showing the score changes for different tickets.
Differential Privacy

• Differential Privacy (DP): technique to randomize a computation over a database, such that changing one data point can only lead to bounded changes in the distribution over possible outputs.

• For $(\varepsilon, \delta)$-DP randomized computation $A_f$:

$$P(A_f(d) \in S) \leq e^\varepsilon P(A_f(d') \in S) + \delta$$

• We prove the Expected Output Stability Bound. For any DP mechanism with bounded outputs in $[0, 1]$ we have:

$$\mathbb{E}(A_f(d')) \leq e^\varepsilon \mathbb{E}(A_f(d')) + \delta$$
Key idea

- Problem: small input perturbations create large score changes.
- Idea: design a DNN with bounded maximum score changes (leveraging Differential Privacy theory).
Key idea

• Problem: small input perturbations create large score changes.
• Idea: design a DNN with bounded maximum score changes (leveraging Differential Privacy theory).
Key idea

• Problem: small input perturbations create large score changes.

• Idea: **design a DNN with bounded maximum score changes** (leveraging Differential Privacy theory).
1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.
1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.
1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.
PixelDP architecture

Resilience to *post-processing*: any computation on the output of an \((\varepsilon, \delta)\)-DP mechanism is still \((\varepsilon, \delta)\)-DP.

1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.
1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.
1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.
1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.
Further challenges

• Train DP DNN with noise.
• Control pre-noise sensitivity during training.
• Support various attack norms ($L_1$, $L_2$, $L_\infty$).
• Scale to large DNNs and datasets.
Scaling to Inception on ImageNet

- Large dataset: image resolution is 300x300x3.
- Large model:
  - 48 layers deep.
  - 23 millions parameters.
  - Released pre-trained by Google on ImageNet.

Inception-v3
Scaling to Inception on ImageNet

PixelDP auto-encoder

Draws for prediction
Given an image to classify, we want to detect the highest probability label.

1. Fixed bounds
We first ask how many draws we need to distinguish the highest probability with probability at least 1. We start from Hoeffding's inequality applied to a Bernouilli variable that...

\[
P(\bar{X} \leq p) \leq 2e^{-2n\varepsilon^2}
\]

We now note the difference between the highest and second highest label probability, and rewrite the variable so that the bounds do not overlap... Finally, we apply a union bound over the k possible labels, and end up with:

\[
P(\mid \bar{X} - p \mid \geq \varepsilon) \leq 2ke^{-2n\varepsilon^2} \leq \frac{1}{n}
\]

For instance in datasets with k = 10, distinguishing the top label with probability at least 0.99 when it is bigger than the second one by 0.1 requires n \approx 1500 draws.
Scaling to Inception on ImageNet

Mathias Lecuyer
June 5, 2018

5

1 Draws for prediction

Given an image to classify, we want to detect the highest probability label.

1.1 Fixed bounds

We first ask how many draws we need to distinguish the highest probability with probability at least $1 - \epsilon$. We start from Hoeffding’s inequality applied to a Bernoulli variable that...

$\bar{X}$ the empirical mean, $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

$P(|\bar{X} - p| > \epsilon) \leq \frac{2e^{-2\epsilon^2 n}}{k^2}$

We now note $\epsilon$ the difference between the highest and second highest label probability, and rewrite the variable so that the bounds do not overlap $\epsilon$. Finally, we apply a union bound over the $k$ possible labels, and end up with:

$P(|\bar{X} - p| > \epsilon) \leq \frac{2ke^{-2\epsilon^2 n}}{k^2} \leq \frac{1}{n}$

For instance in datasets with $k = 10$, distinguishing the top label with probability at least 0.99 when it is bigger than the second one by 0.1 requires $n \approx 1500$ draws.

PixelDP auto-encoder

Inception-v3

Post-processing
PixelDP Outline

Motivation

Design

Evaluation
Evaluation:

1. Guaranteed accuracy on large DNNs/datasets
2. Are robust predictions harder to attack in practice?
3. Comparison with other defenses against state-of-the-art attacks.
Methodology

Five datasets:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Image size</th>
<th>Number of Classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>299x299x3</td>
<td>1000</td>
</tr>
<tr>
<td>CIFAR-100</td>
<td>32x32x3</td>
<td>100</td>
</tr>
<tr>
<td>CIFAR-10</td>
<td>32x32x3</td>
<td>10</td>
</tr>
<tr>
<td>SVHN</td>
<td>32x32x3</td>
<td>10</td>
</tr>
<tr>
<td>MNIST</td>
<td>28x28x1</td>
<td>10</td>
</tr>
</tbody>
</table>

Three models:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of Layers</th>
<th>Number of Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inception-v3</td>
<td>48</td>
<td>23M</td>
</tr>
<tr>
<td>Wide ResNet</td>
<td>28</td>
<td>36M</td>
</tr>
<tr>
<td>CNN</td>
<td>3</td>
<td>3M</td>
</tr>
</tbody>
</table>

Metrics:

- Guaranteed accuracy.
- Accuracy under attack.

Attack methodology:

- State of the art attack [Carlini and Wagner S&P'17].
- Strengthened against our defense by averaging gradients over multiple noise draws.
Guaranteed accuracy on ImageNet with Inception-v3

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
<th>Guaranteed accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>78</td>
<td>-</td>
</tr>
<tr>
<td>PixelDP: L=0.25</td>
<td>68</td>
<td>63</td>
</tr>
<tr>
<td>PixelDP: L=0.75</td>
<td>58</td>
<td>53</td>
</tr>
</tbody>
</table>

More DP noise

Meaningful **guaranteed accuracy** for ImageNet!
What if we only act on robust predictions?
(e.g. if not robust, check ticket)

**Dataset:** CIFAR-10
Accuracy on robust predictions

- Baseline
- Precision: threshold 0.1
- Recall: threshold 0.1

Dataset: CIFAR-10

Comparison: Madry+ '17

If we increase the robustness threshold: better accuracy, less predictions.
Comparison with other provable defenses

PixelDP scales to larger models, yielding better accuracy and robustness.
PixelDP summary

• PixelDP is the first defense that:
  • Gives **attack-independent guarantees** against norm-bounded adversarial attacks.
  • And **scales** to the largest models and datasets.

• Already extensions by others!
  • Improve the bounds at a given noise level (Li+ '18; Cohen+ '19).
  • Use other noise distributions (Pinot+ '19).
  • Adapt optimization (Rakin+ '18).
Appendix
Comparison with best-effort techniques

Dataset: CIFAR-10

Comparison: Best effort defense by Madry+ '17

PixelDP is empirically competitive with the state-of-the-art best-effort defense.
## Related work

### Best effort

<table>
<thead>
<tr>
<th>+ Scale:</th>
<th>- No robustness guarantees:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Run a best effort attack per gradient step [Goodfellow+ '15, Madry+ '17].</td>
<td>- Often broken soon after release [Athalye+ '18].</td>
</tr>
<tr>
<td>- Preprocess inputs [Buckman+ '18, Guo+ '18].</td>
<td></td>
</tr>
<tr>
<td>- Train a second model based on the first one [Papernot+ '16].</td>
<td></td>
</tr>
<tr>
<td>+ Flexible:</td>
<td></td>
</tr>
<tr>
<td>- Support most architectures.</td>
<td></td>
</tr>
</tbody>
</table>

### Certified

<table>
<thead>
<tr>
<th>+ Provable guarantees:</th>
<th>- Hard to scale:</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Per prediction [Wong-Kolter+ '18, Wong+ '18, Raghunathan+ '18, Wang+ '18].</td>
<td>- Requires orders of magnitude more computation [Wong-Kolter+ '18, Wong+ '18, Wang+ '18].</td>
</tr>
<tr>
<td>- In expectation [Sinha+ '17].</td>
<td>- Support only 1 hidden layer [Raghunathan+ '18].</td>
</tr>
<tr>
<td>- Often not flexible:</td>
<td></td>
</tr>
<tr>
<td>- No ReLU, MaxPool, or accuracy guarantees [Sinha+ '17].</td>
<td></td>
</tr>
<tr>
<td>- Only ReLU, no BatchNorm [Wong-Kolter '18].</td>
<td></td>
</tr>
</tbody>
</table>

PixelDP is the first certified defense that both achieves provable guarantees of robustness, scales and is broadly applicable to arbitrary networks.
Results - CIFAR-10

$L_\infty$ attack:

![Graph showing the performance of different attacks on CIFAR-10 dataset. The x-axis represents the empirical attack bound ($L_{\text{attack}}$) and the y-axis represents the conventional accuracy. The graph compares Baseline, Madry with $L = 0.03$ (\(\infty\)-norm), and PixelDP with $L = 0.1$ (2-norm).]
Results - SVHN

$L_{\infty}$ attack:

\[ A(d) = f(d) + N(0, \sigma^2) \]

- DP

- RobustOpt
- Baseline
- PixelDP, L = 0.1 (2-norm)

Conventional accuracy vs. Empirical attack bound, $L_{attack}$ (2-norm)
Certification on ImageNet/Inception-v3

![Graph showing certified accuracy against attack size for different attack methods. The graph includes lines for Baseline, PixelDP (L=0.1), PixelDP (L=0.3), and PixelDP (L=1.0).]
Certification on CIFAR-10

![Certification on CIFAR-10](image)

- Baseline
- PixelDP (L=0.1)
- PixelDP (L=0.3)
Comparison with Best Effort Techniques

Undefended: $\|\alpha\|_2 = 0.52$

Giant panda

Undefended: $\|\alpha\|_2 = 3.41$

Teddy bear

Teddy bear


Full references