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

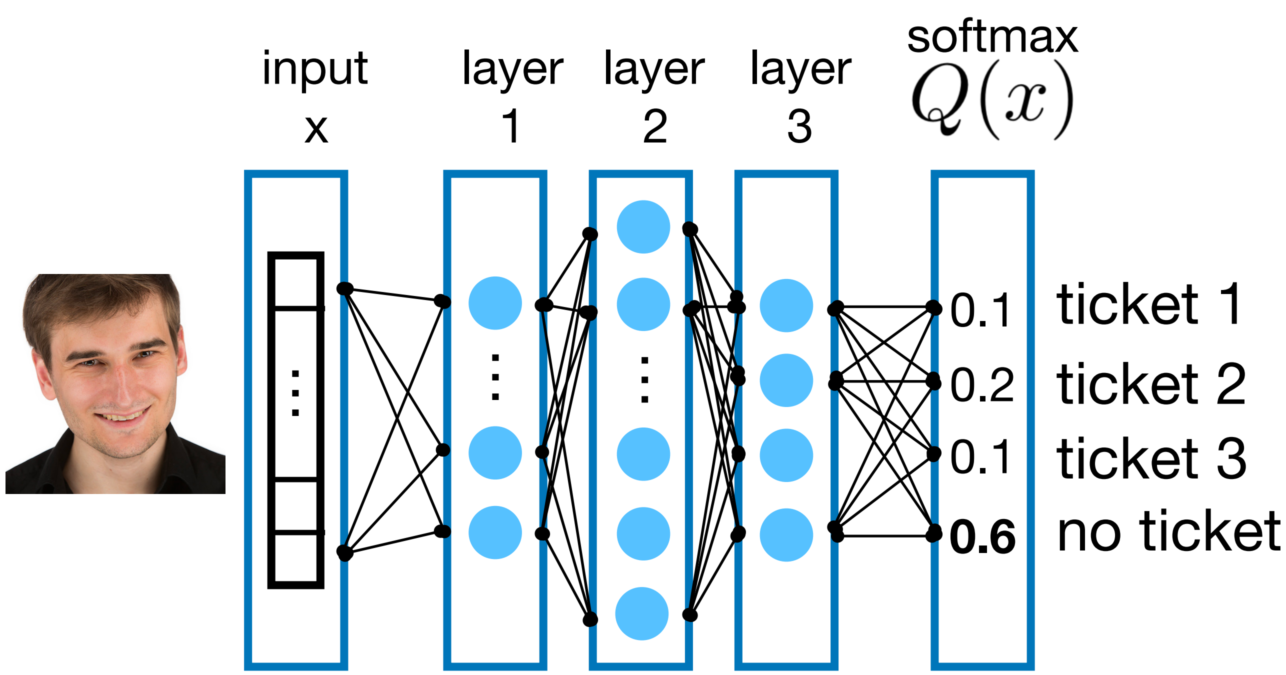
- Deep Neural Networks (DNNs) deliver remarkable performance on many tasks.
- DNNs are increasingly deployed, including in attack-prone contexts:

The New York Times

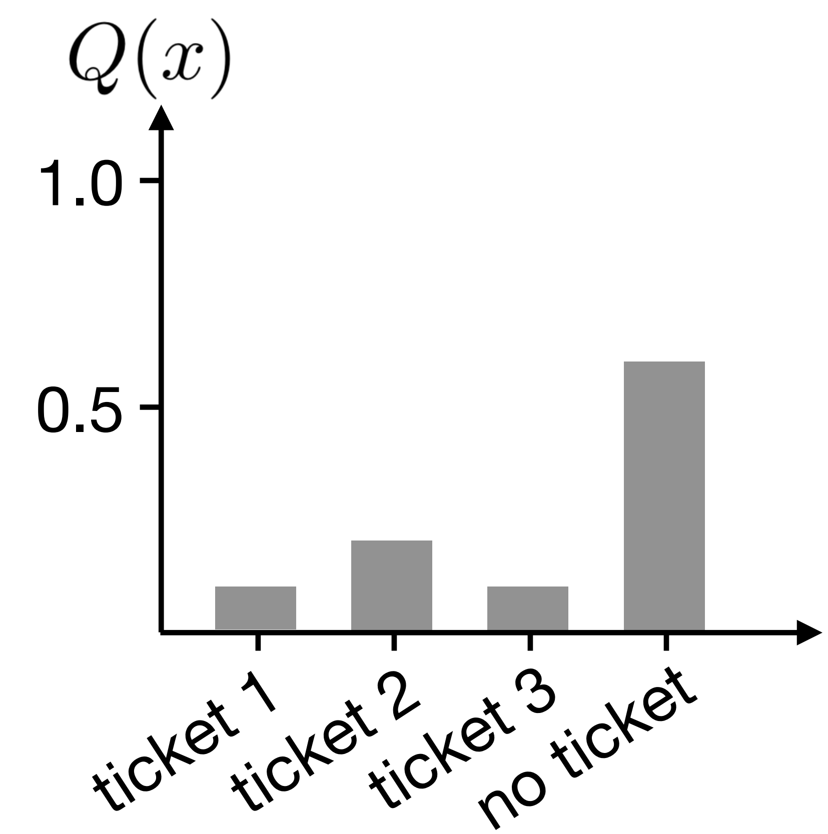
Taylor Swift Said to Use Facial Recognition to Identify Stalkers

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

Example



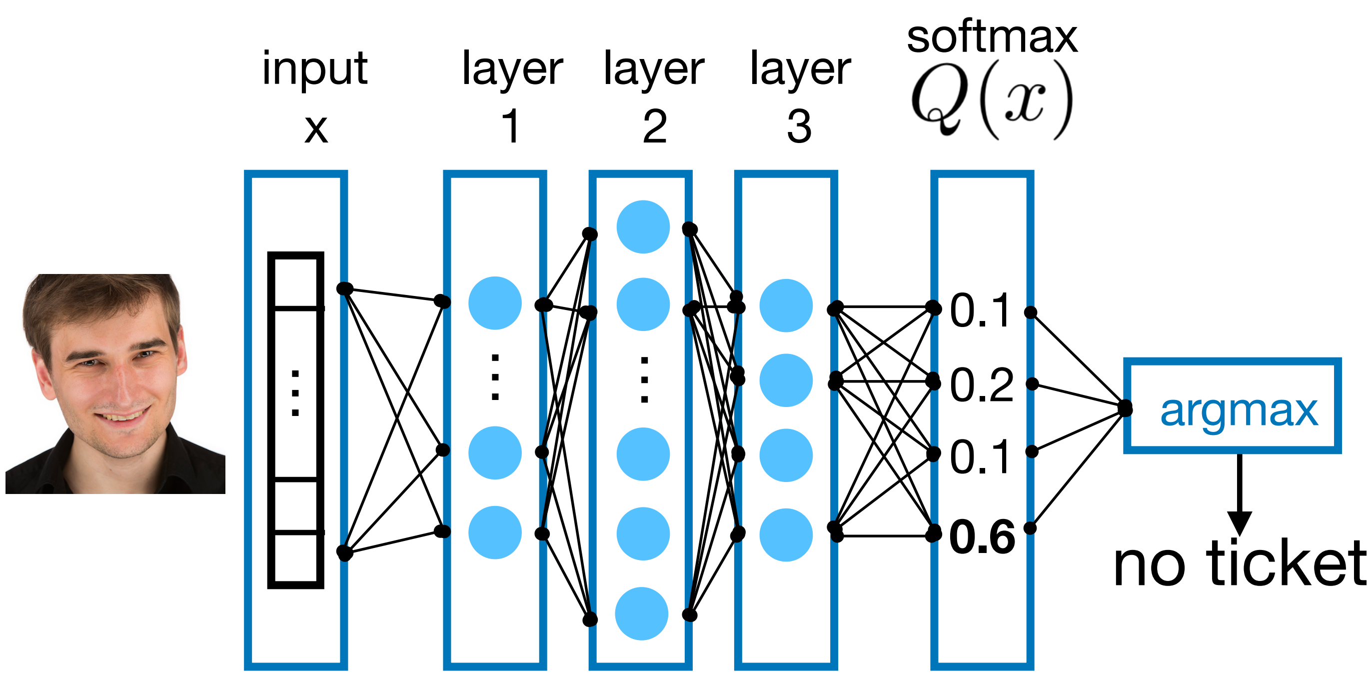
input
x
layer
1
layer
2
layer
3
softmax
$Q(x)$
0.1
ticket 1
0.2
ticket 2
0.1
ticket 3
0.6
no ticket

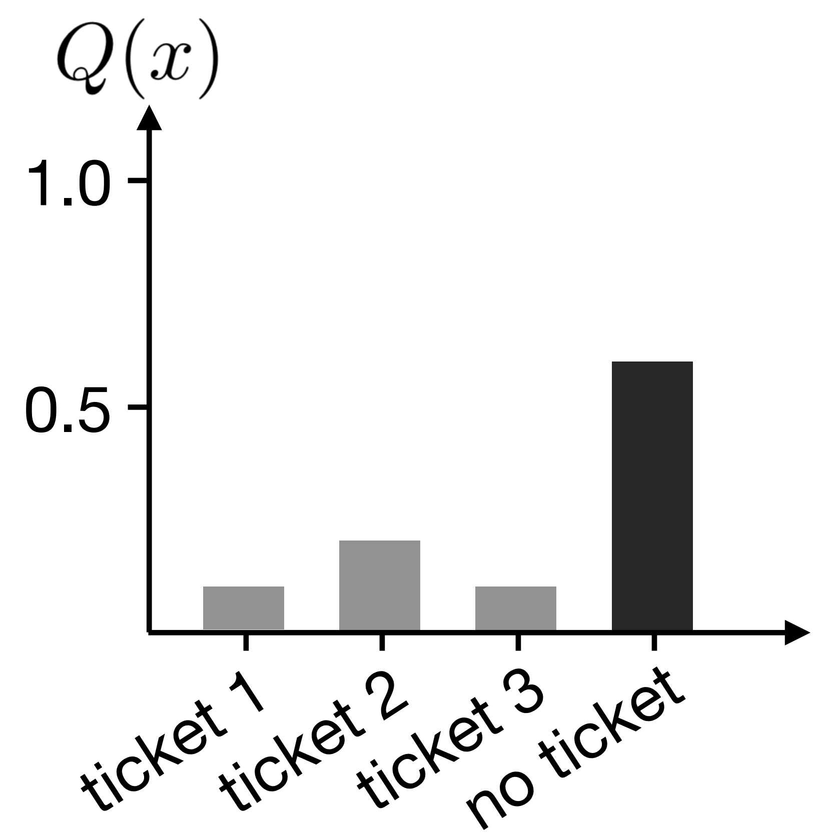


$Q(x)$
1.0
0.5
ticket 1
ticket 2
ticket 3
no ticket

Example

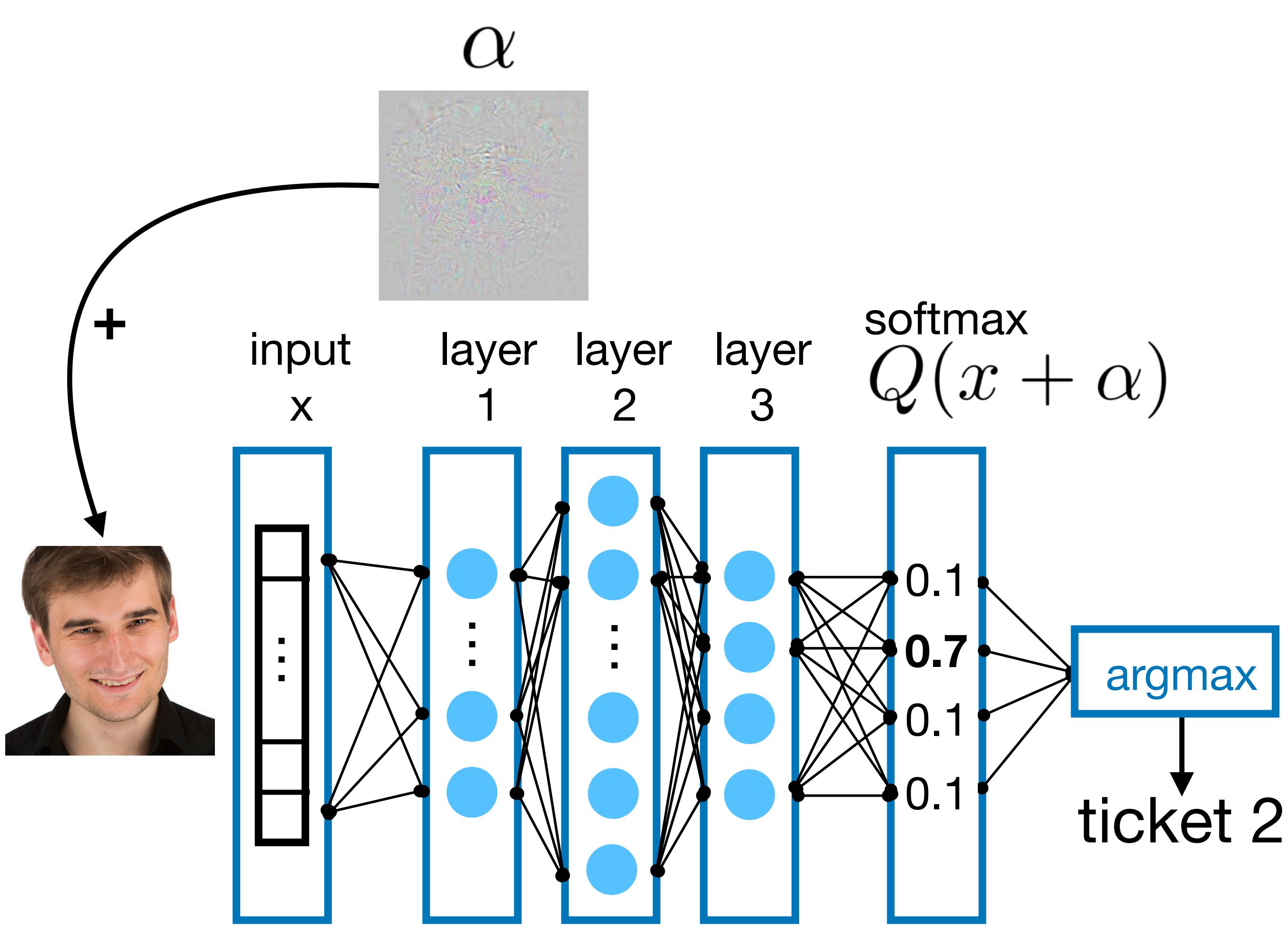
But DNNs are vulnerable to adversarial example attacks.

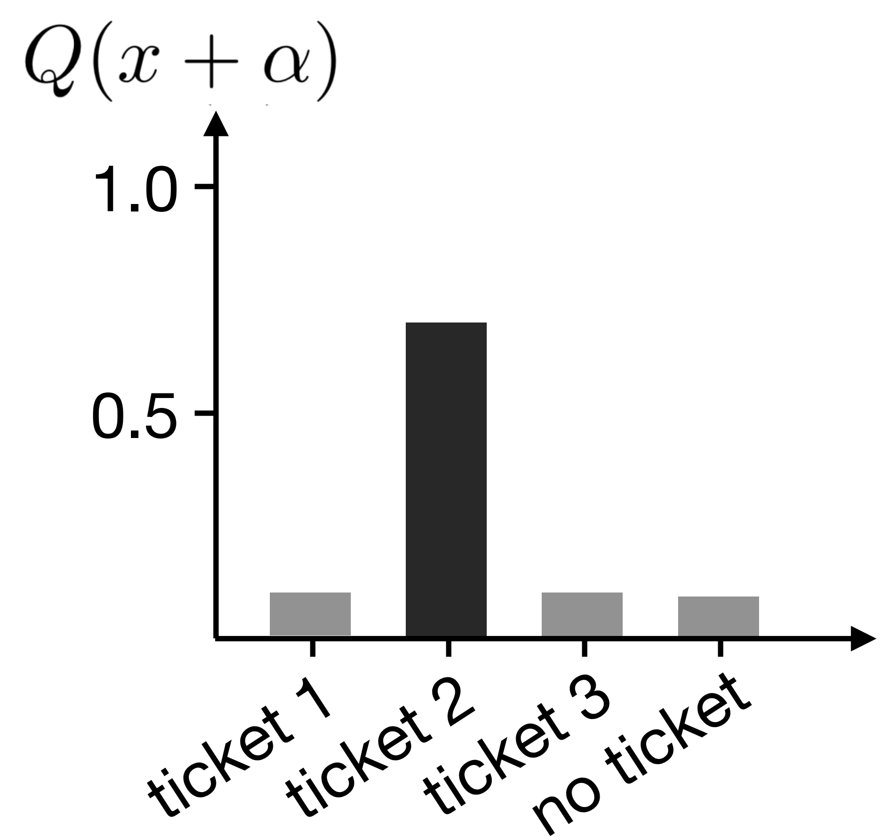


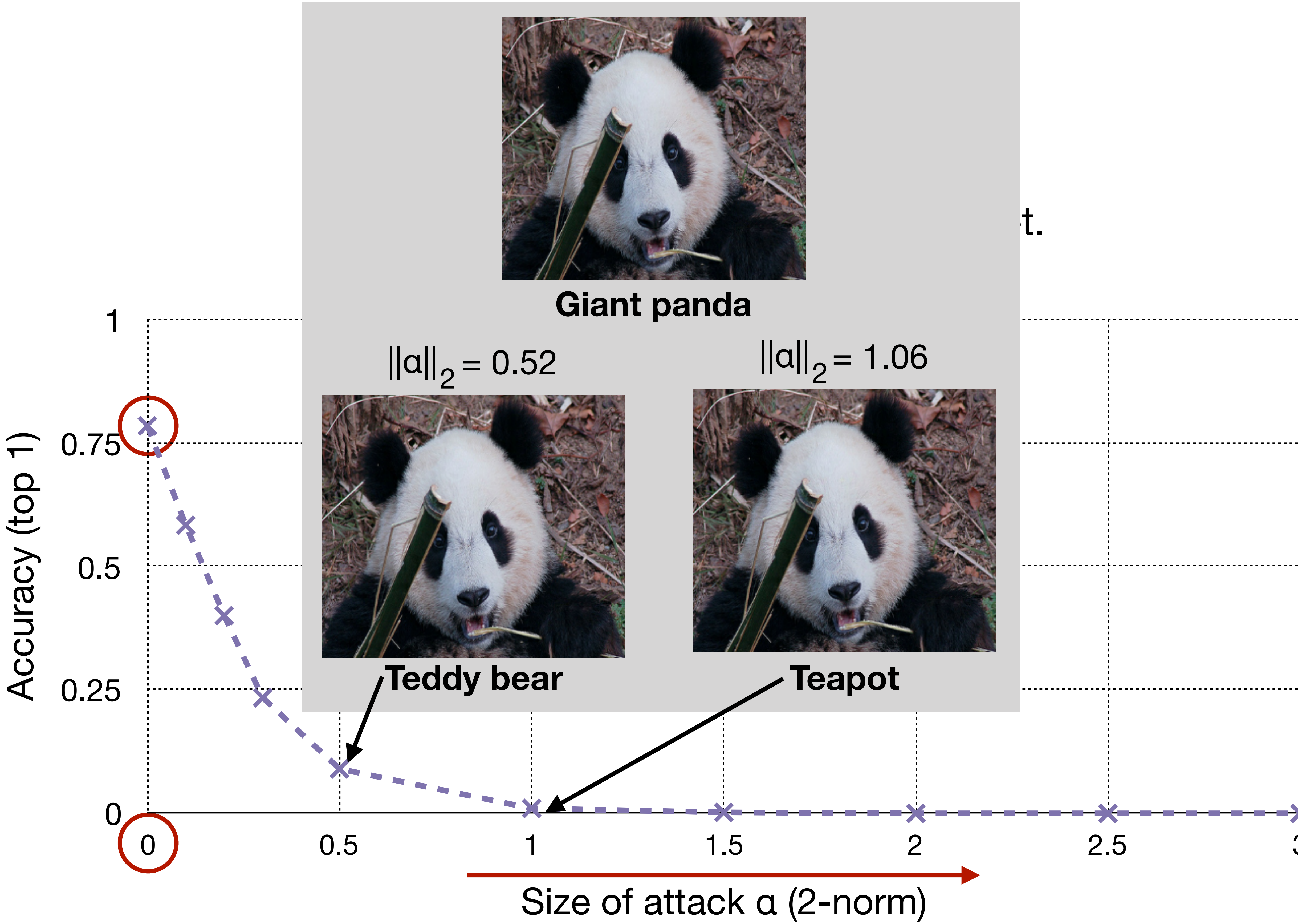


Example

But DNNs are vulnerable to adversarial example attacks.







t.
Giant panda
||α||2 = 0.52
||α||2 = 1.06
Teddy bear
Teapot
Accuracy (top 1)
1
0.75
0.5
0.25
0
0
0.5
1
1.5
2
2.5
3
Size of attack α (2-norm)

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

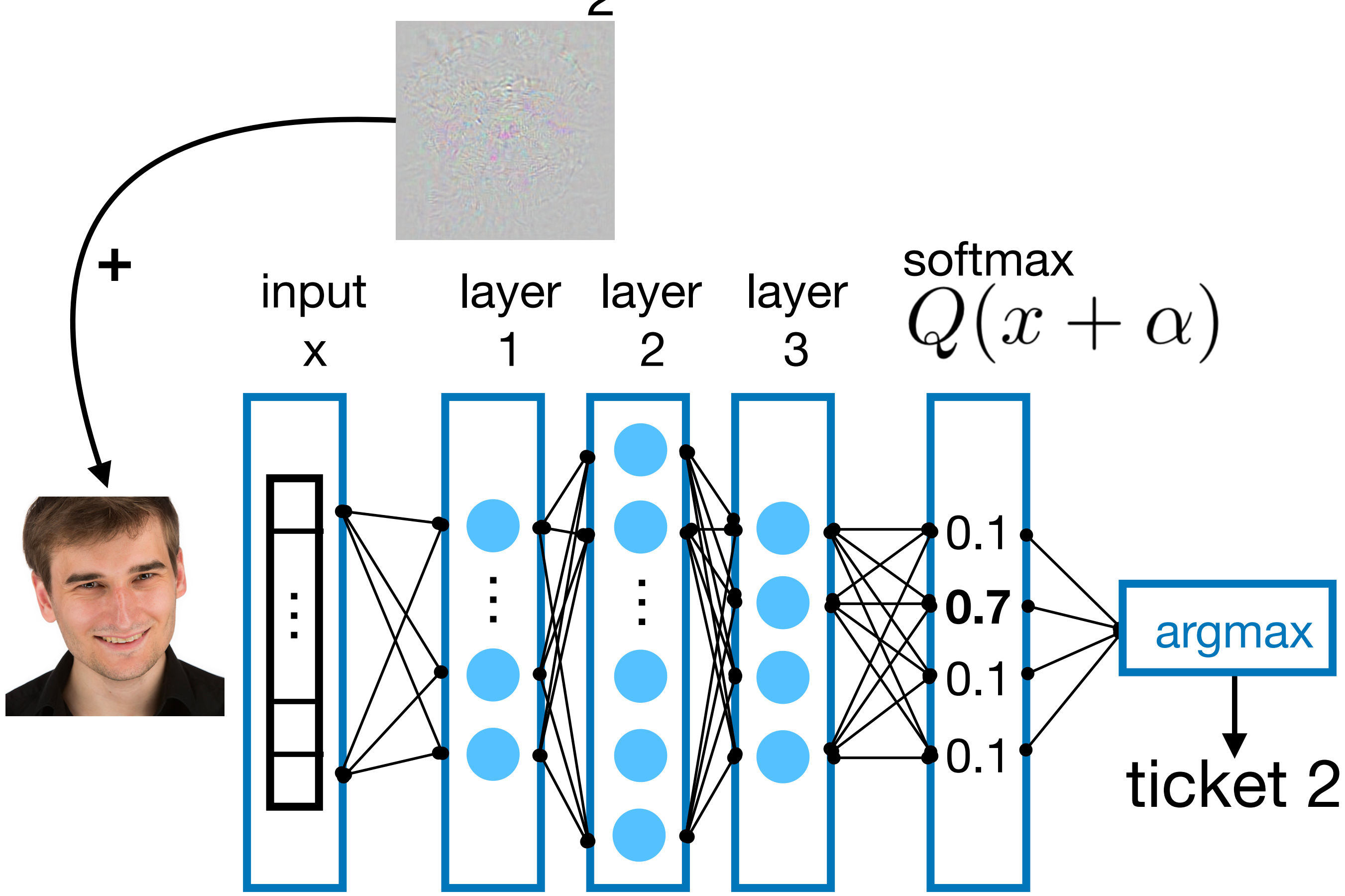
Design

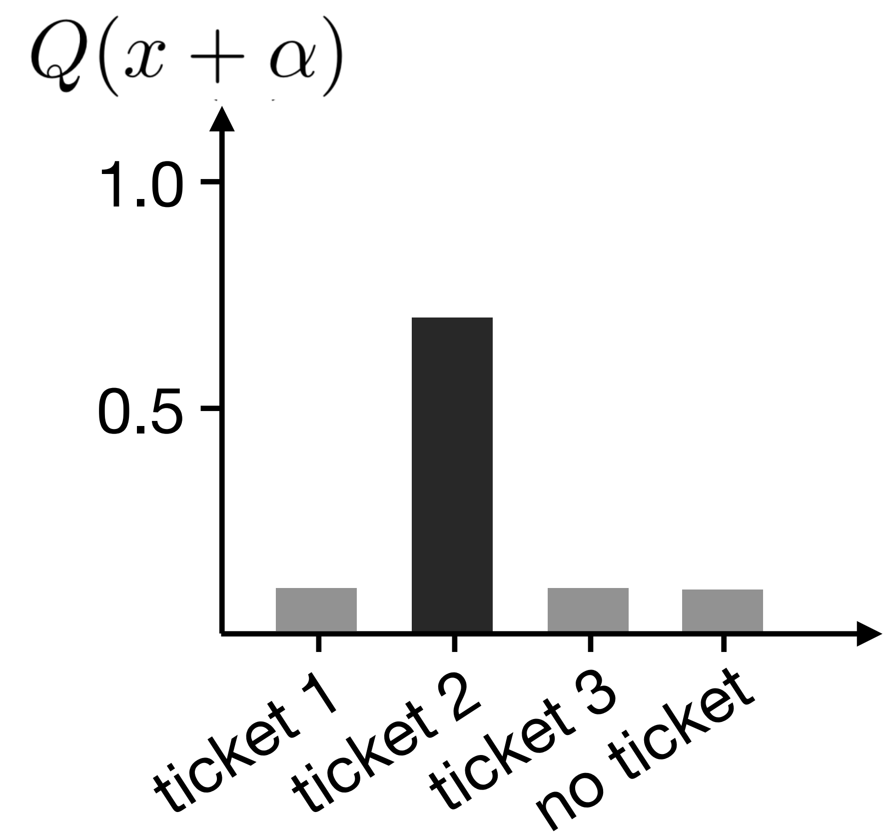
Evaluation

Key idea

- Problem: small input perturbations create large score changes.

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

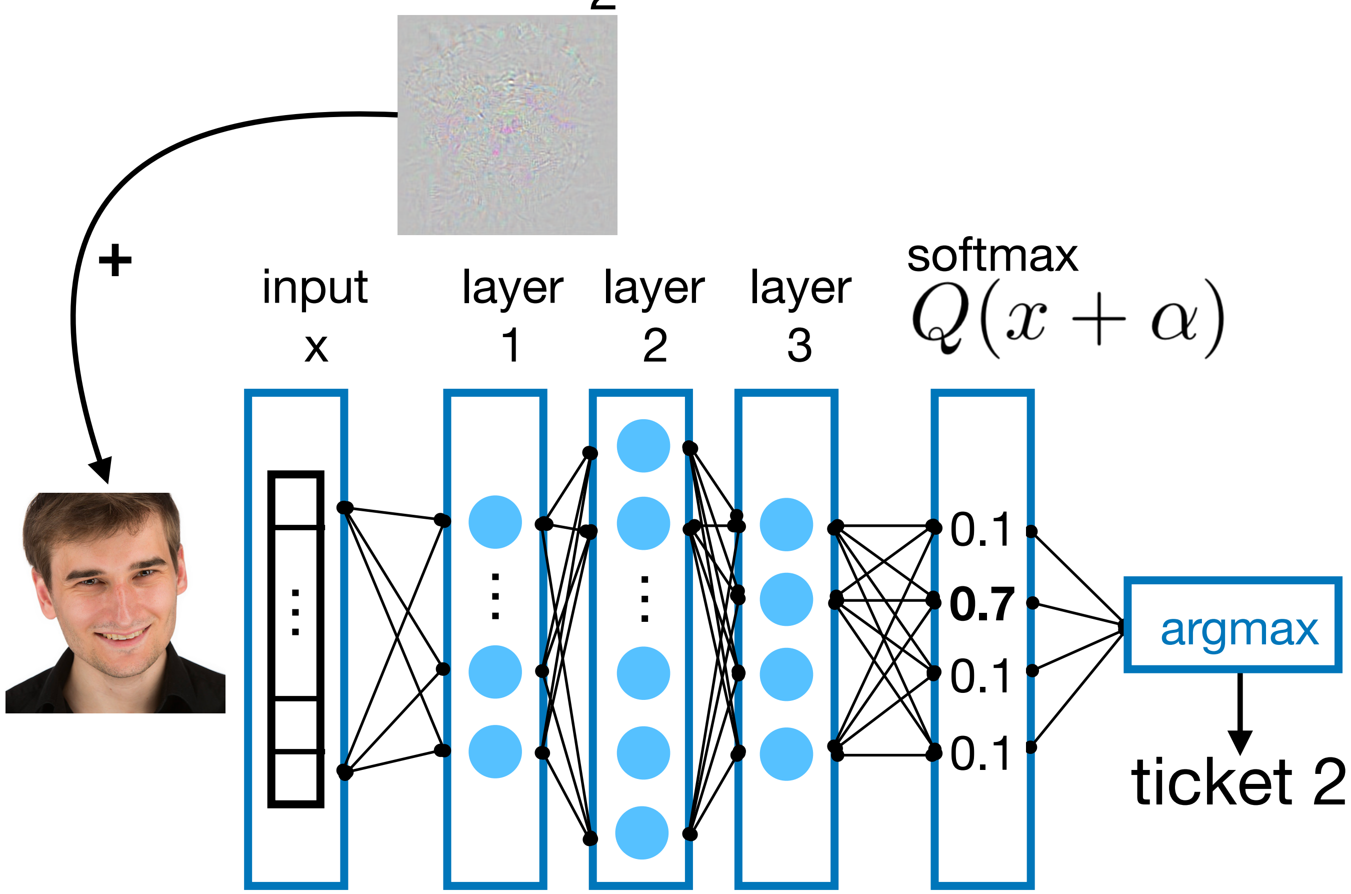


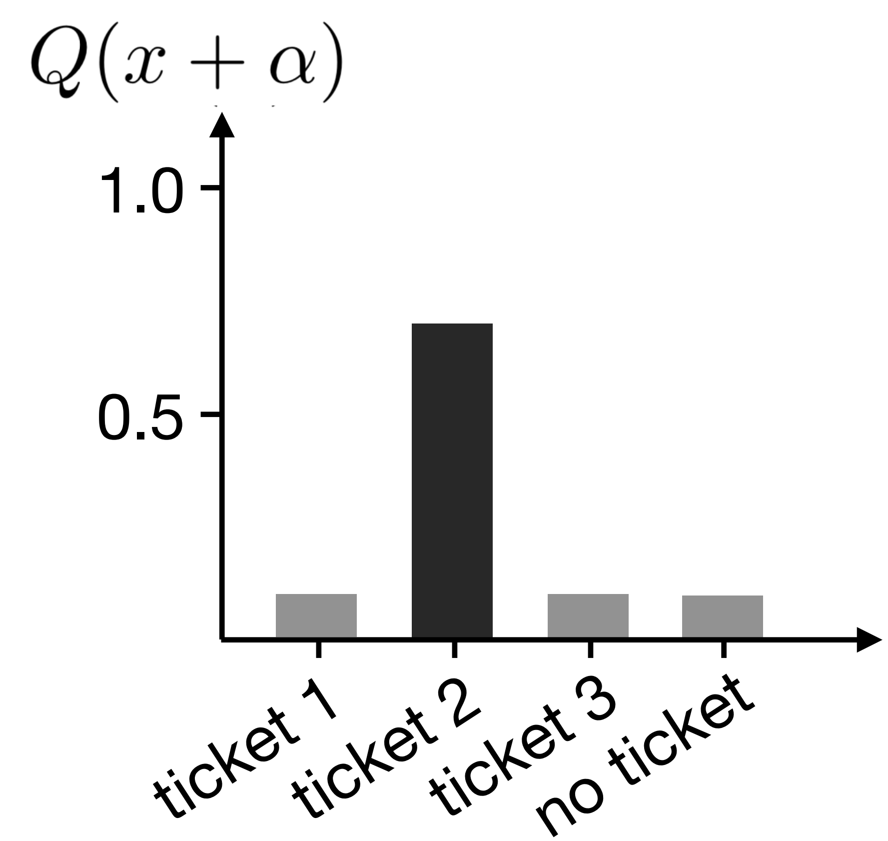


Key idea

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

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





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 (ε, δ)-DP randomized computation A_f:

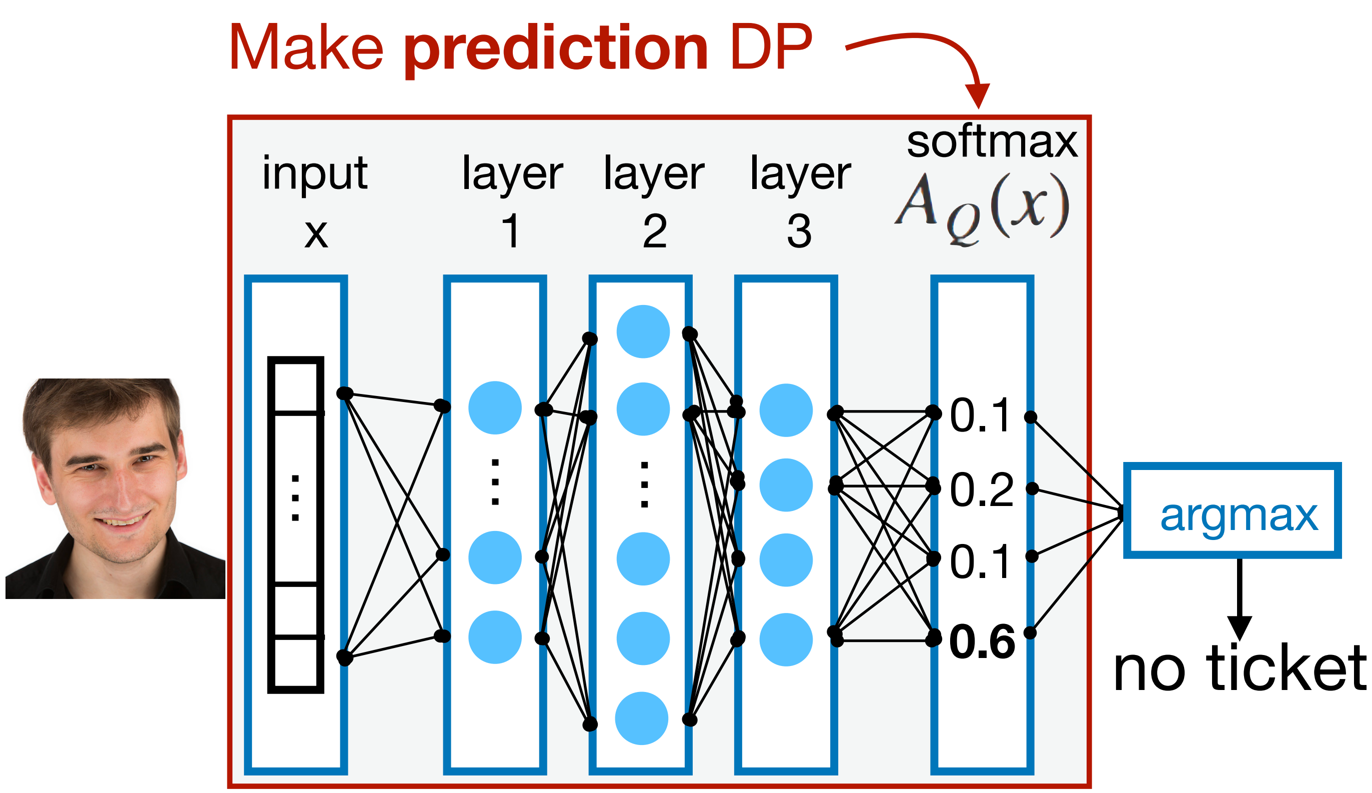
$$P(A_f(d) \in S) \leq e^{\epsilon} P(A_f(d') \in S) + \delta$$

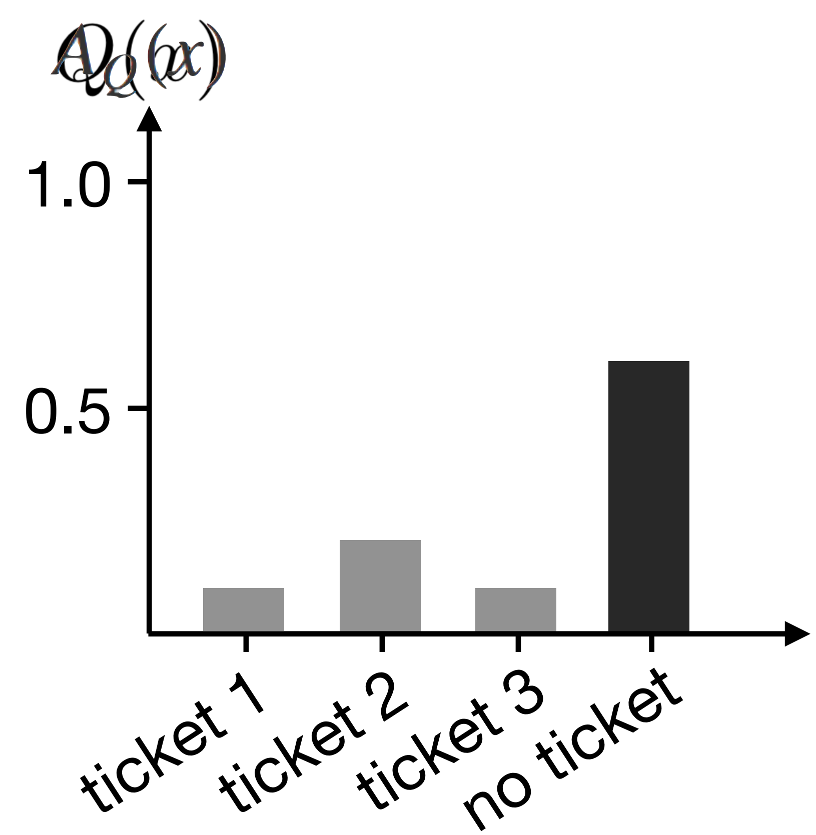
- We prove the Expected Output Stability Bound. For any DP mechanism with bounded outputs in [0, 1] we have:

$$\boxed{\mathbb{E}(A_f(d)) \leq e^{\epsilon} \mathbb{E}(A_f(d')) + \delta}$$

Key idea

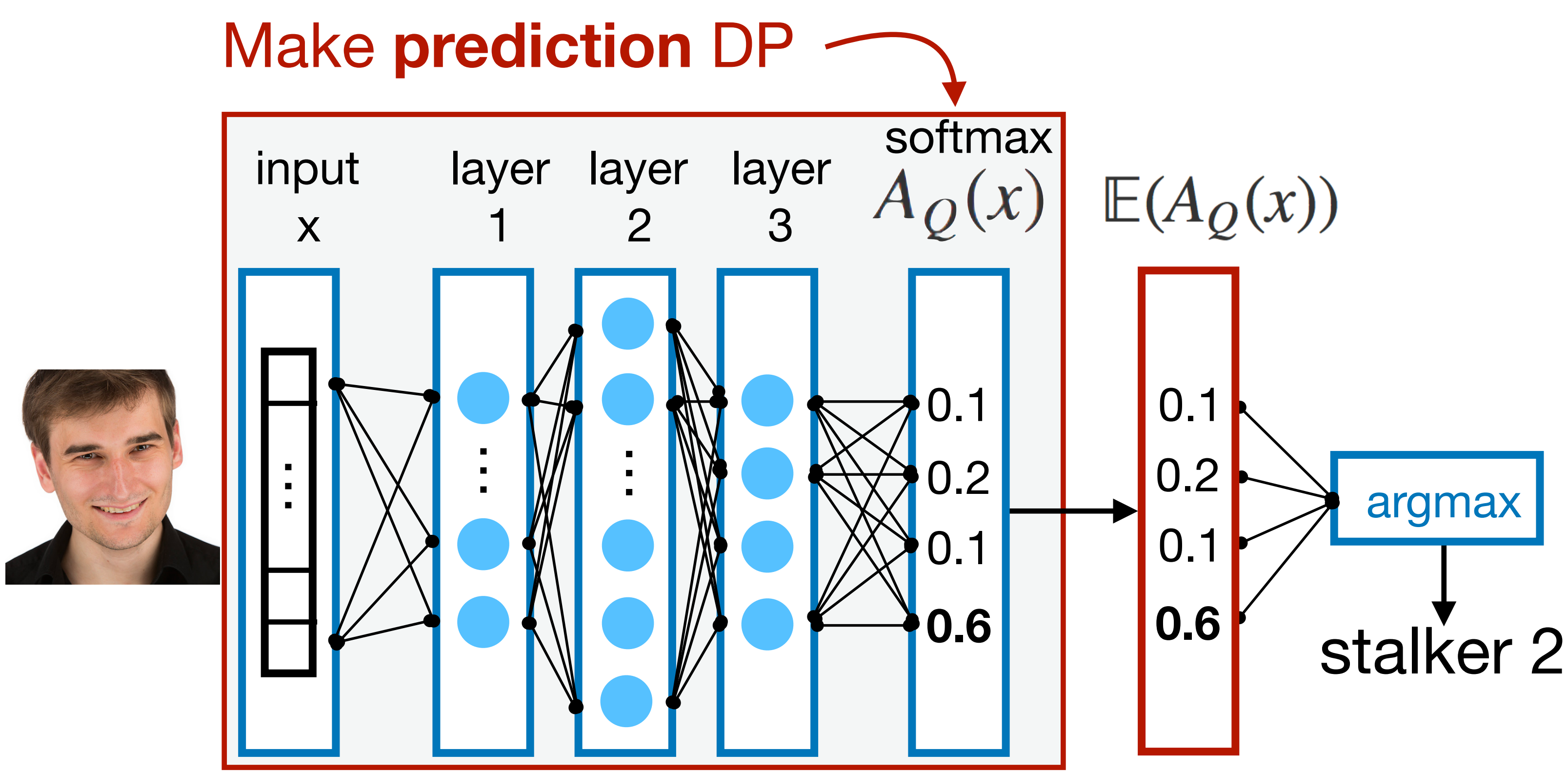
- Problem: small input perturbations create large score changes.
- Idea: design a DNN with bounded maximum score changes (leveraging Differential Privacy theory).

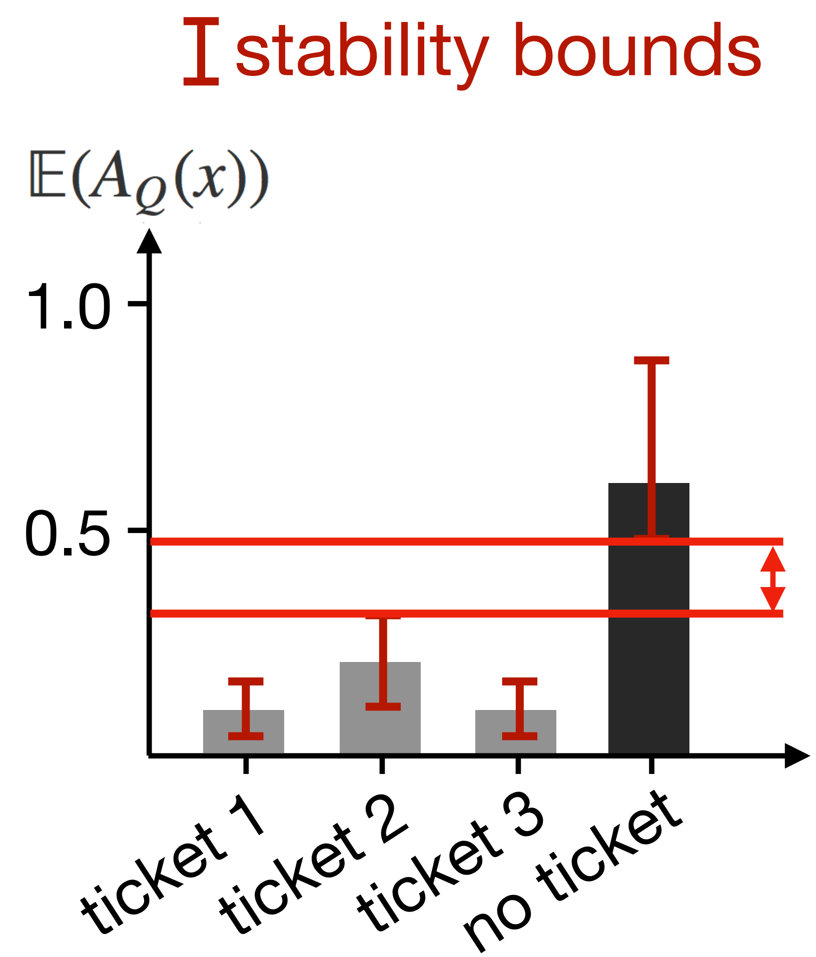




Key idea

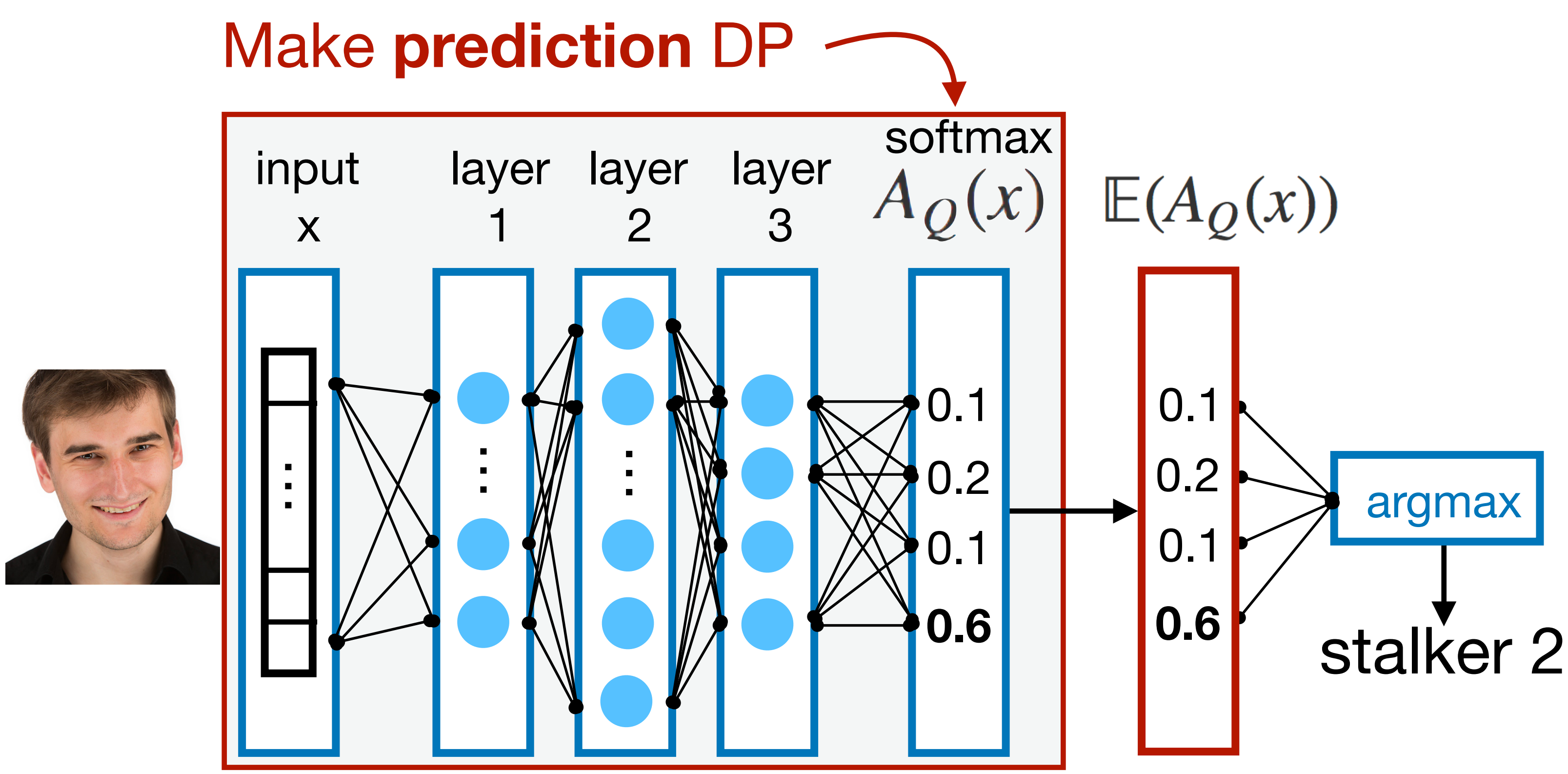
- Problem: small input perturbations create large score changes.
- Idea: design a DNN with bounded maximum score changes (leveraging Differential Privacy theory).

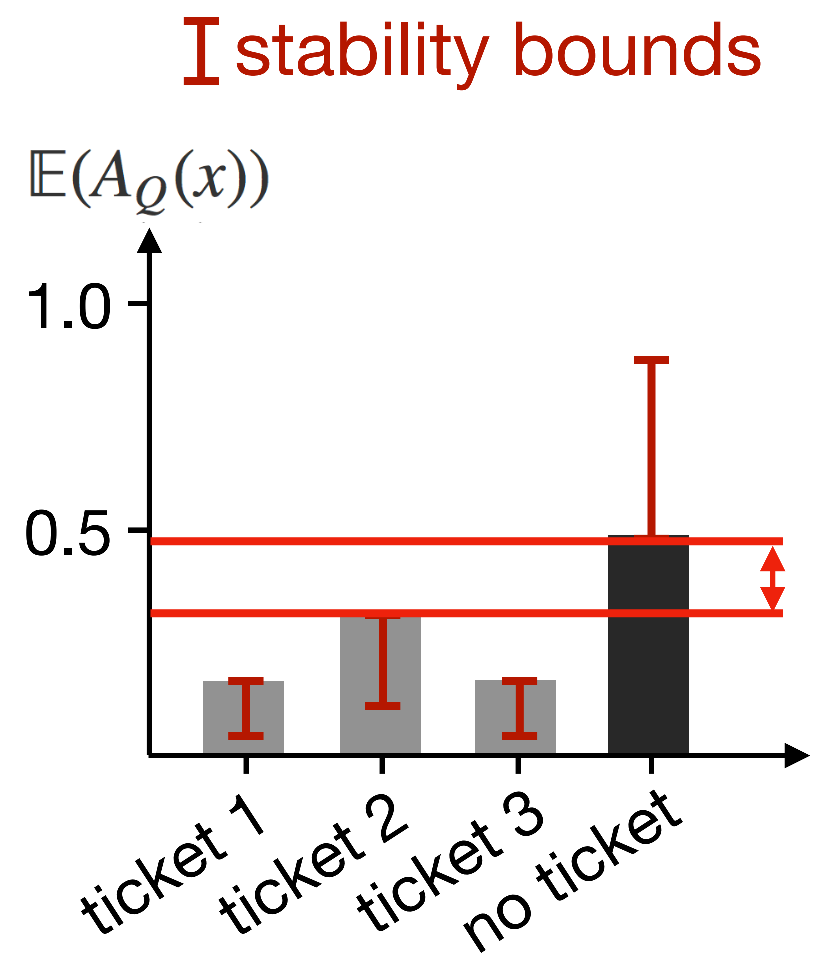




Key idea

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

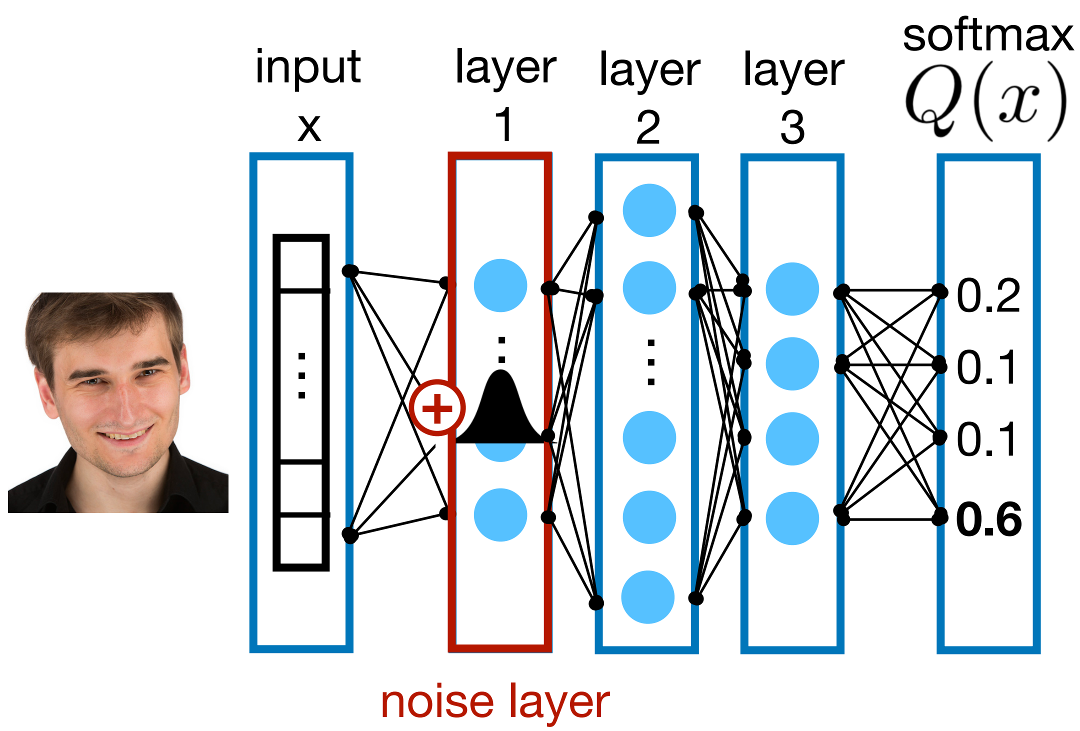




PixelDP architecture

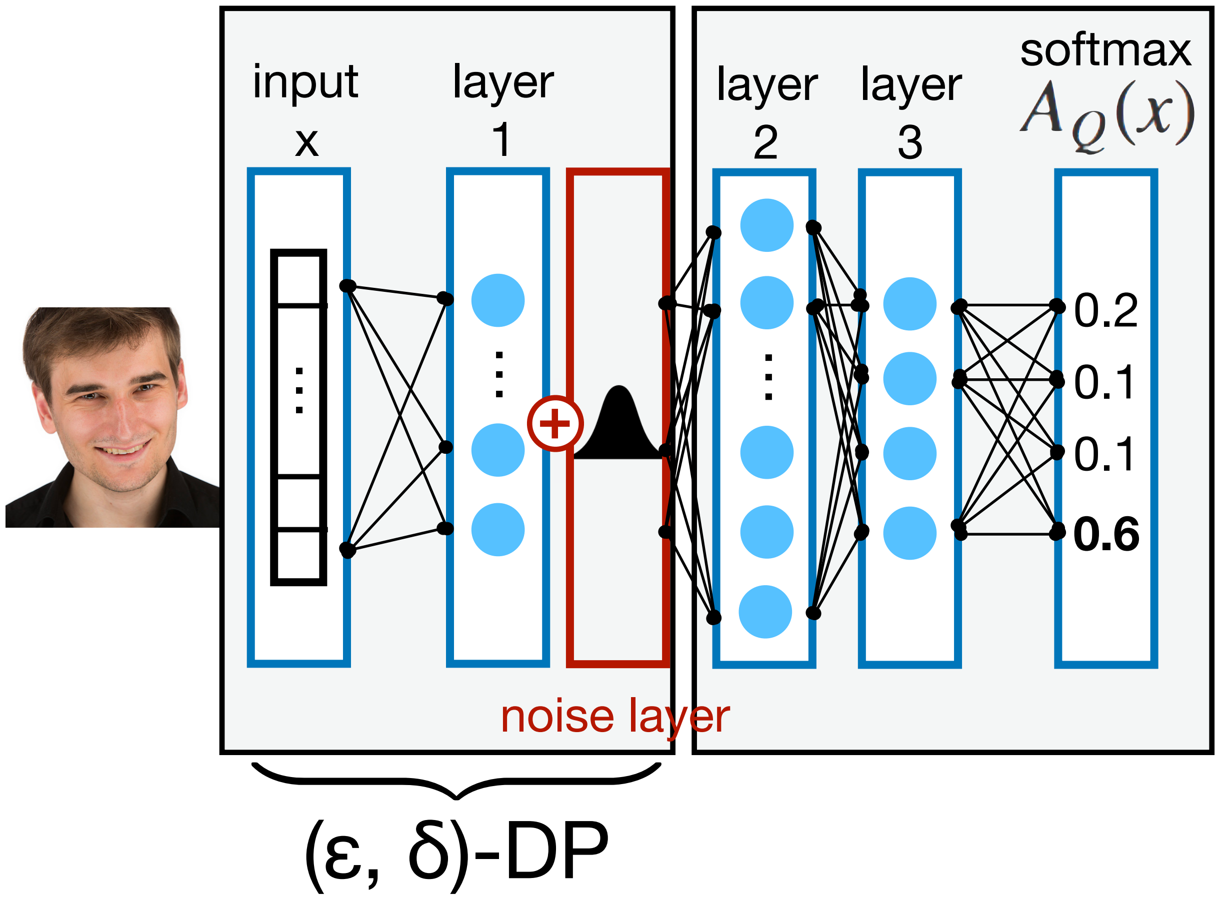
1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.

PixelDP architecture



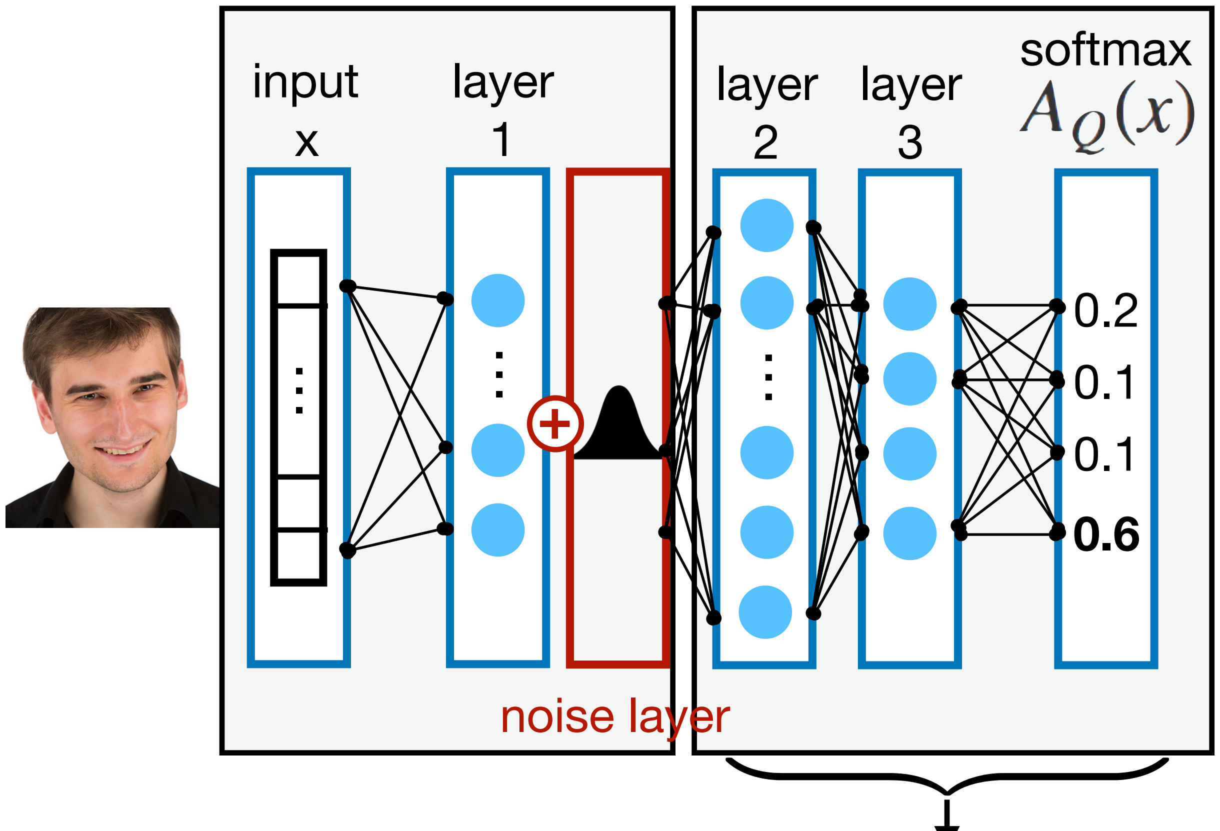
1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.

PixelDP architecture



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.

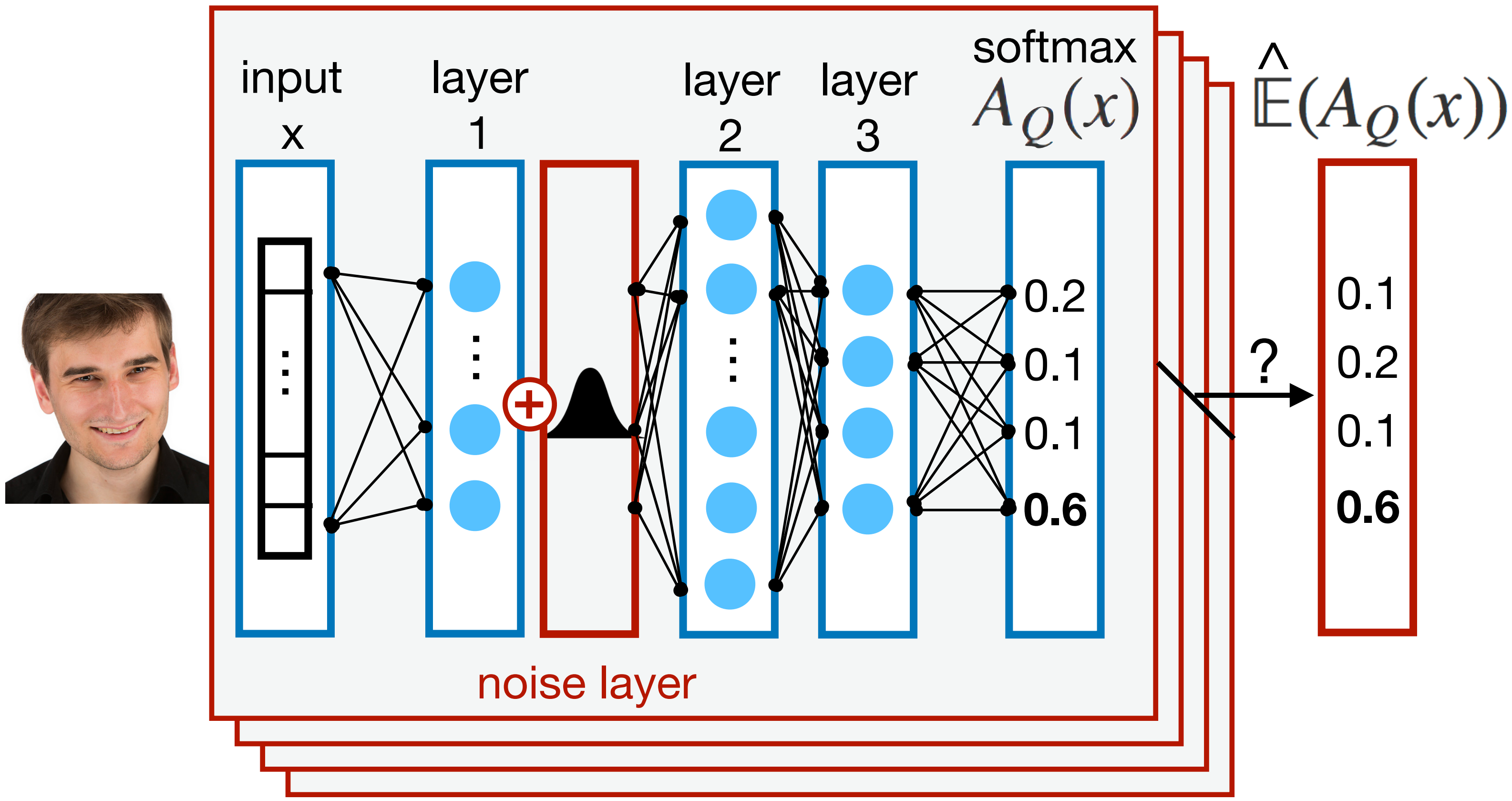
PixelDP architecture



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

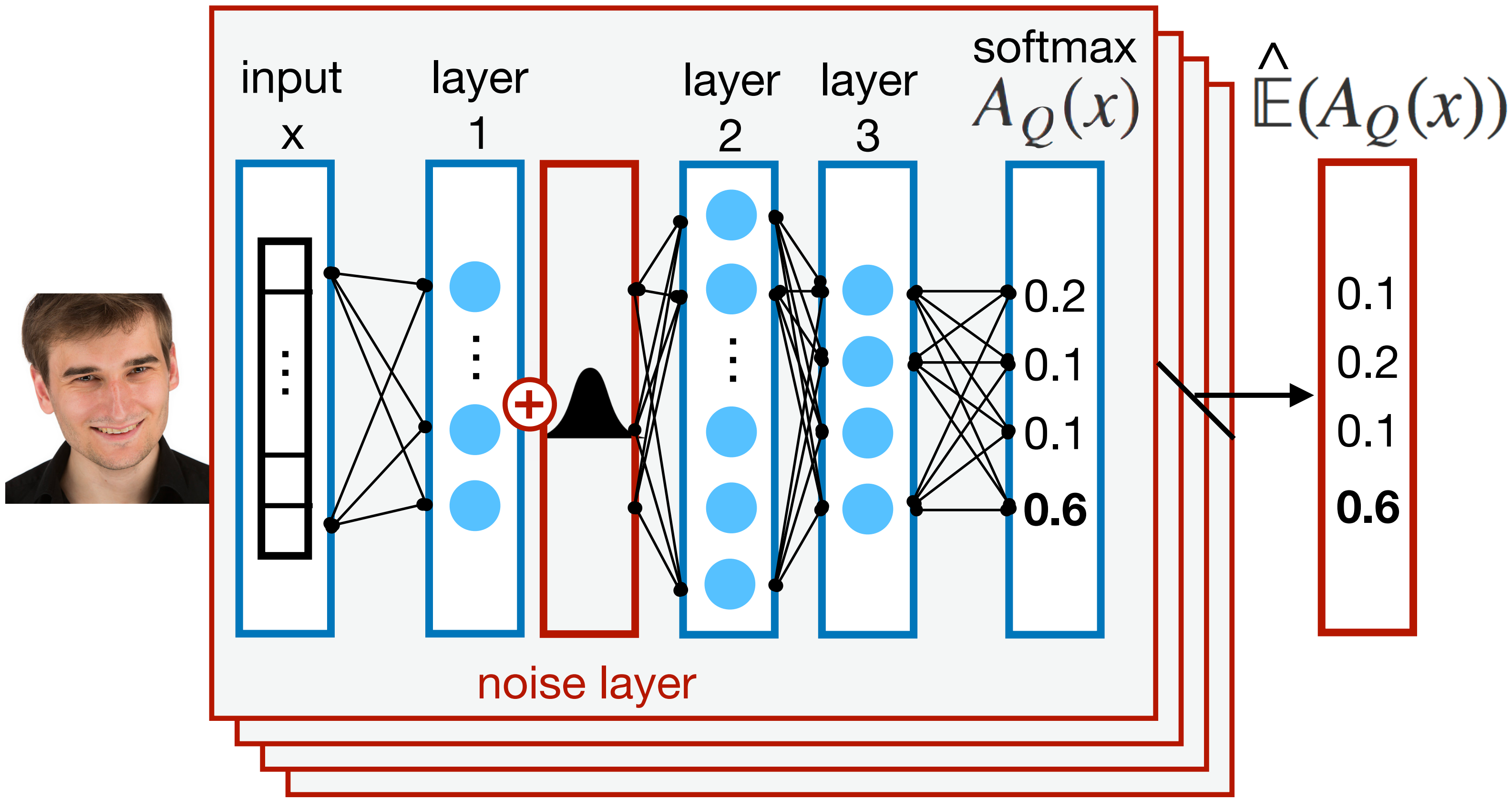
PixelDP architecture

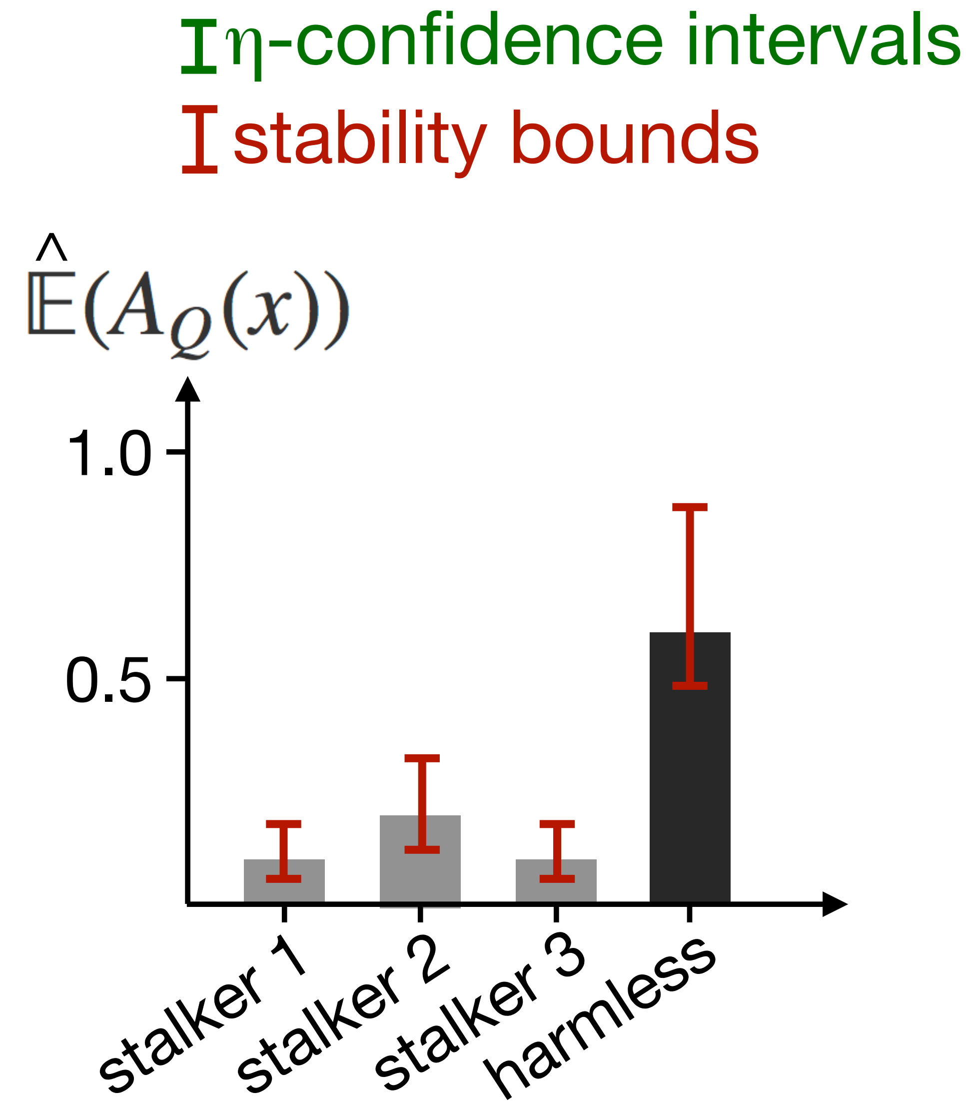


Compute empirical mean with standard Monte Carlo estimate.

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.

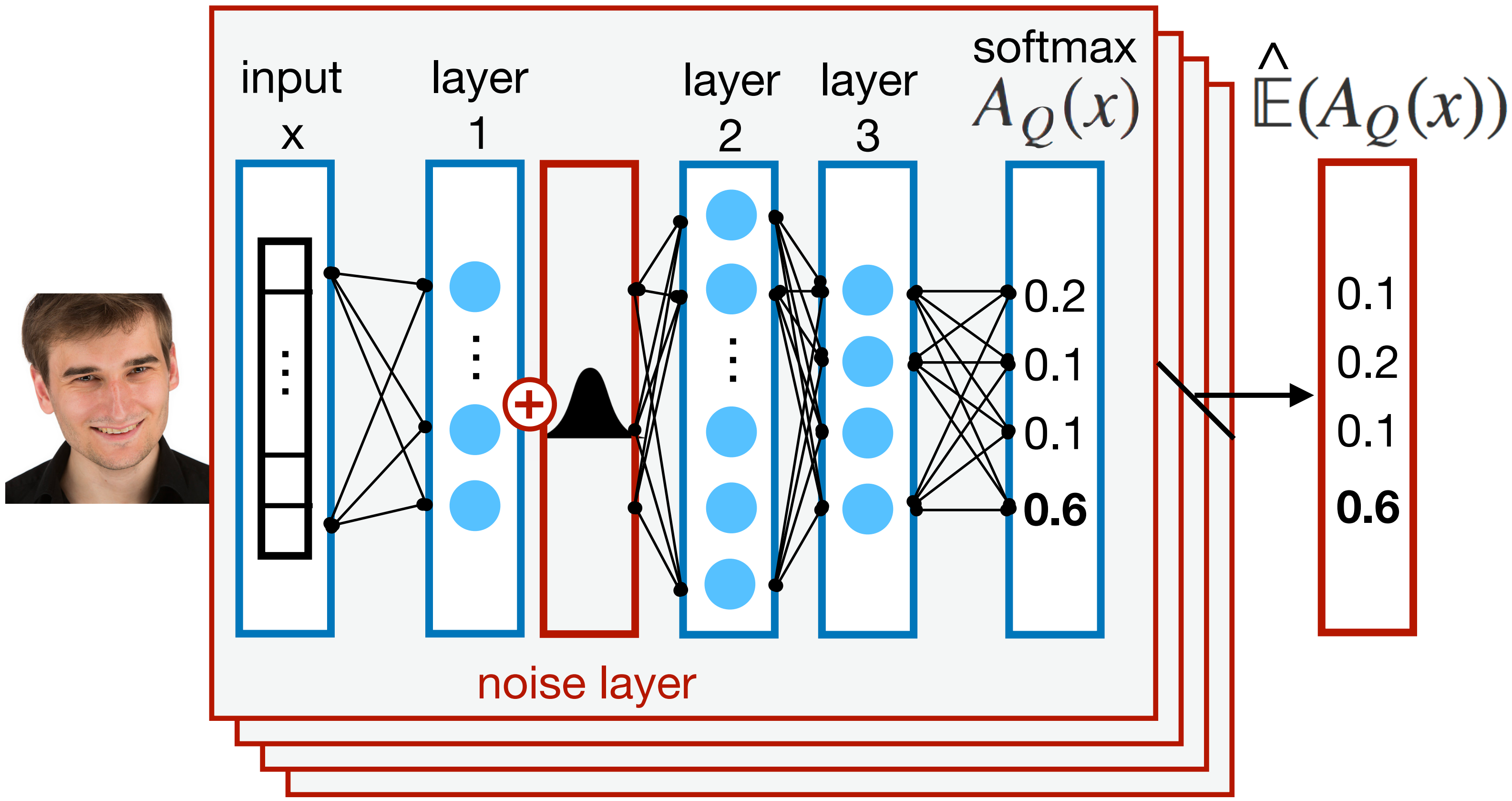
PixelDP architecture

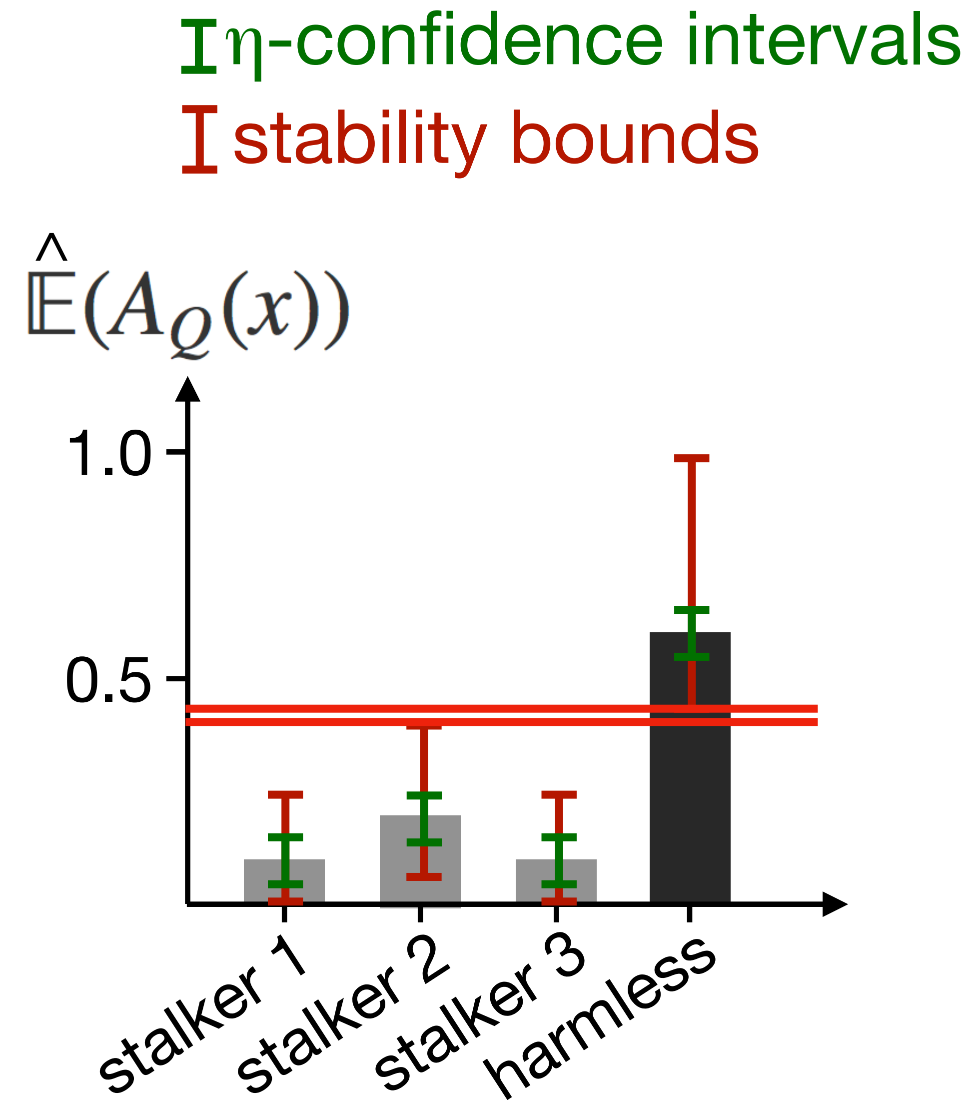




1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.

PixelDP architecture





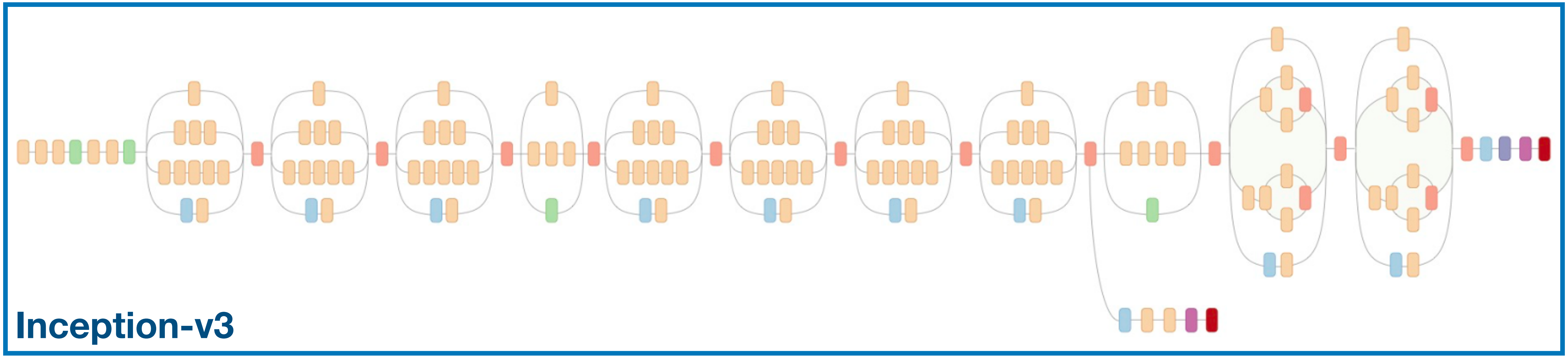
1. Add a new noise layer to make DNN DP.
2. Estimate the DP DNN's mean scores.
3. Add estimation error in the stability bounds.

Further challenges

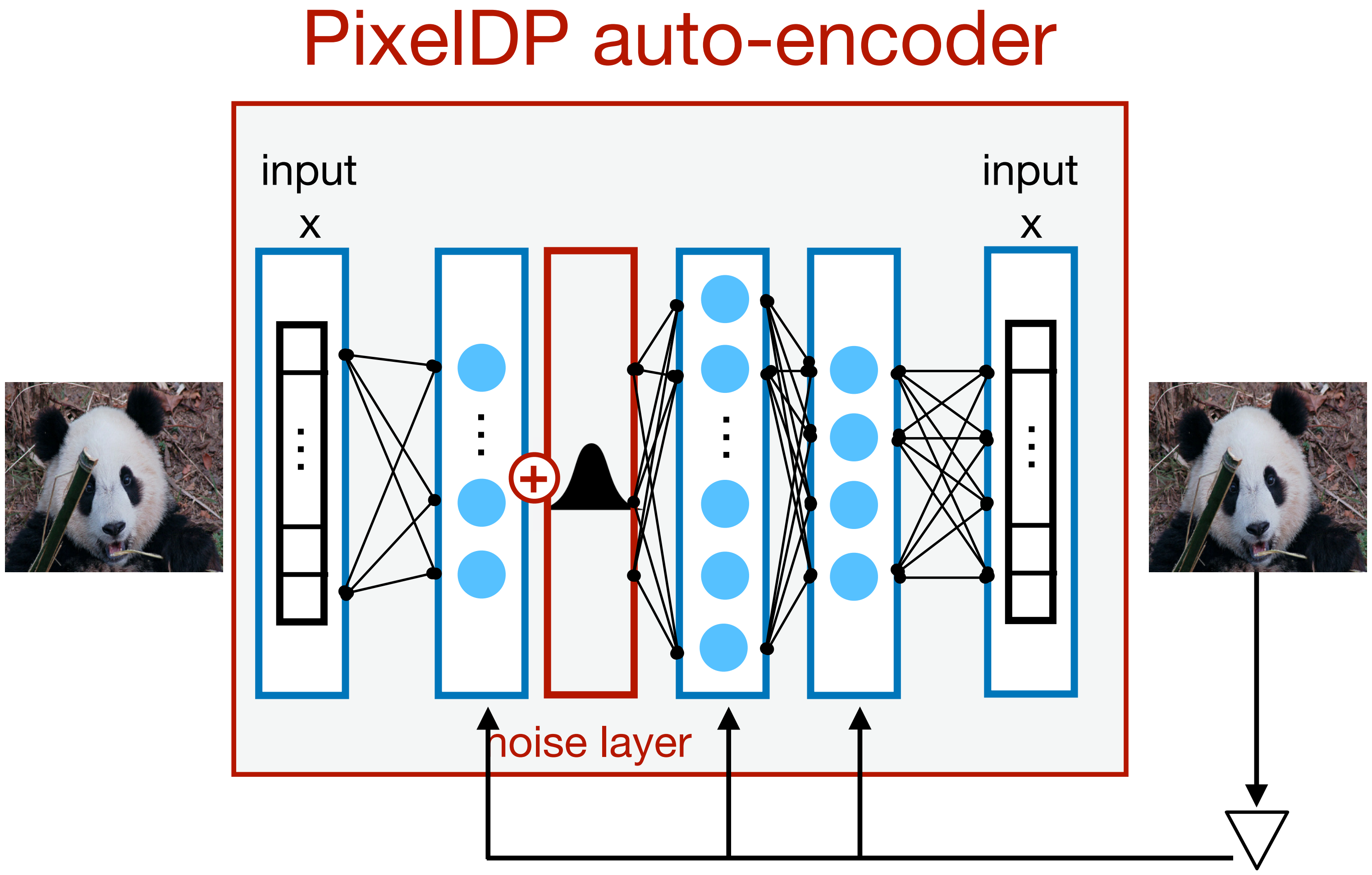
- Train DP DNN with noise.
- Control pre-noise sensitivity during training.
- Support various attack norms (L_1, L_2, L_∞).
- Scale to large DNNs and datasets.

Scaling to Inception on ImageNet

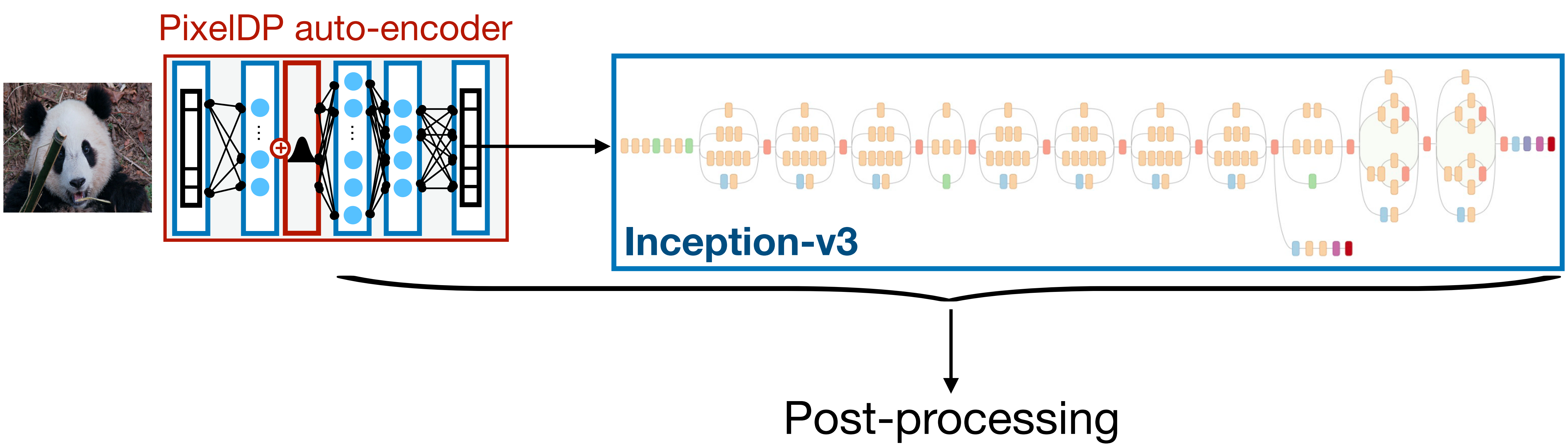
- Large dataset: image resolution is 300x300x3.
- Large model:
    - 48 layers deep.
    - 23 millions parameters.
    - Released pre-trained by Google on ImageNet.



Scaling to Inception on ImageNet



Scaling to Inception on ImageNet



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:

Dataset	Image size	Number of Classes
ImageNet	299x299x3	1000
CIFAR-100	32x32x3	100
CIFAR-10	32x32x3	10
SVHN	32x32x3	10
MNIST	28x28x1	10

Three models:

Dataset	Number of Layers	Number of Parameters
Inception-v3	48	23M
Wide ResNet	28	36M
CNN	3	3M

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.

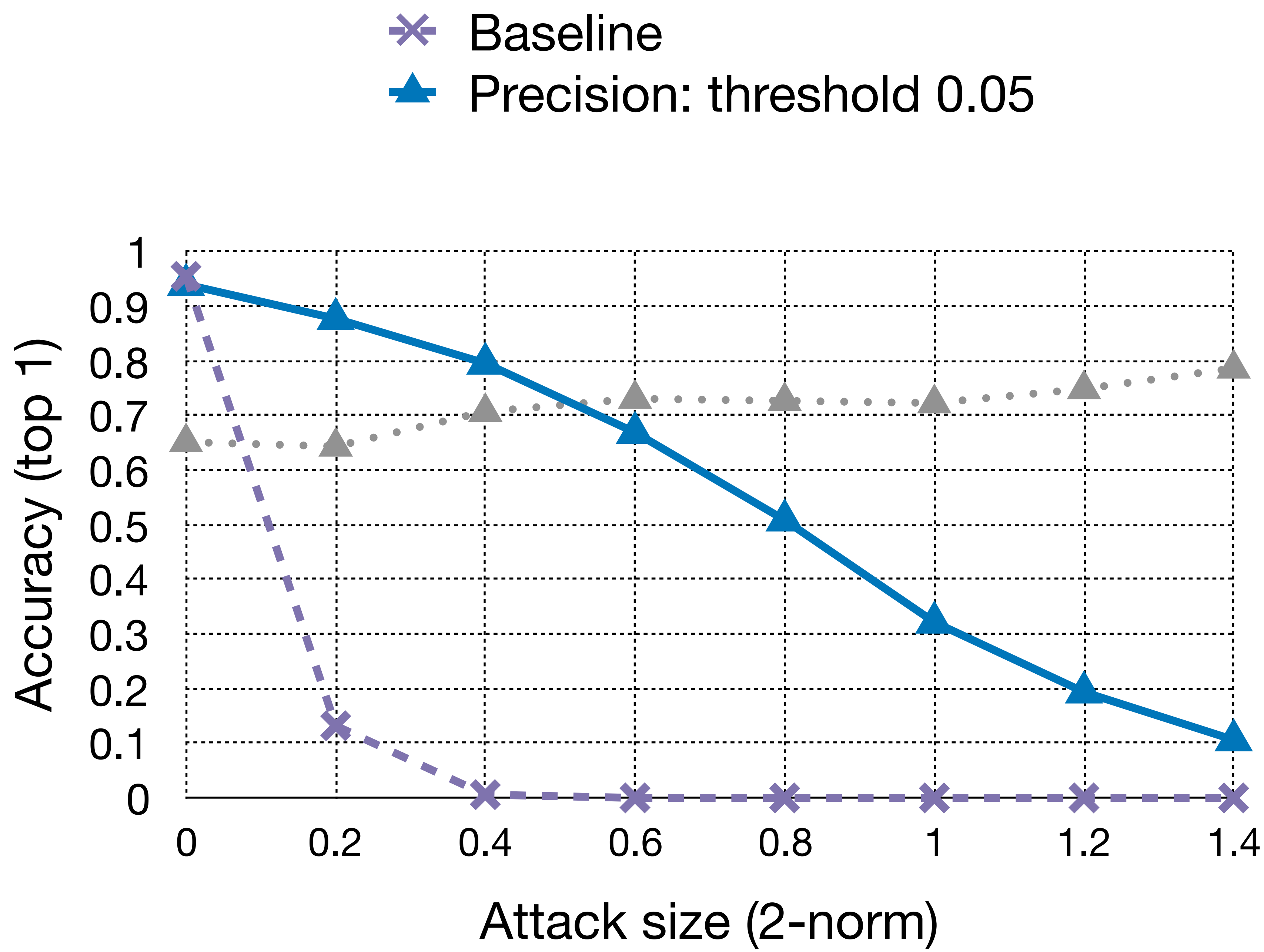
Guaranteed accuracy on ImageNet with Inception-v3

Model	Accuracy (%)	Guaranteed accuracy (%) 0.05	0.1	0.2
Baseline	78	-	-	-
PixelDP: L=0.25	68	63	0	0
PixelDP: L=0.75	58	53	49	40

More DP noise

Meaningful guaranteed accuracy for ImageNet!

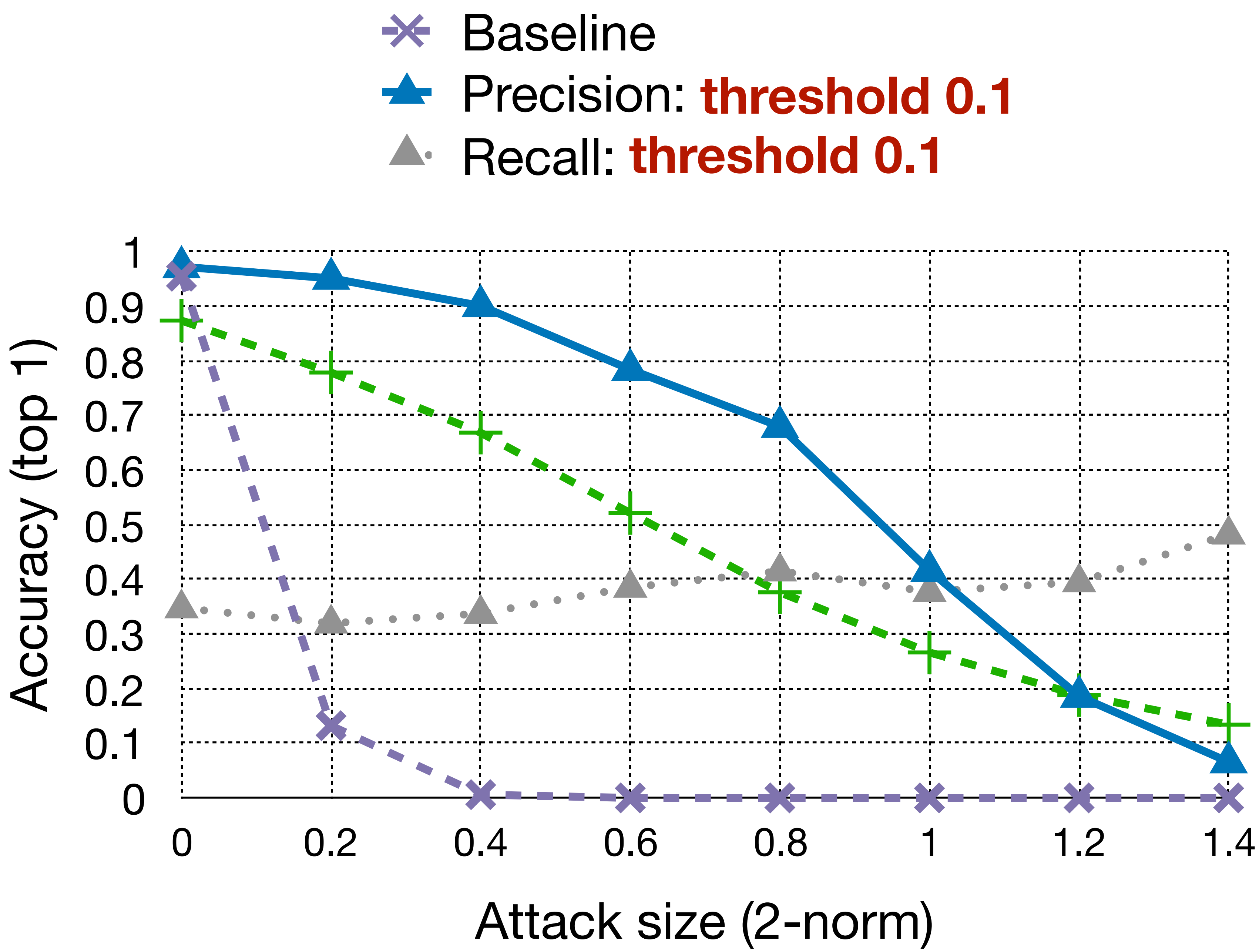
Accuracy on robust predictions



Dataset: CIFAR-10

What if we only act on robust predictions?
(e.g. if not robust, check ticket)

Accuracy on robust predictions



Dataset: CIFAR-10

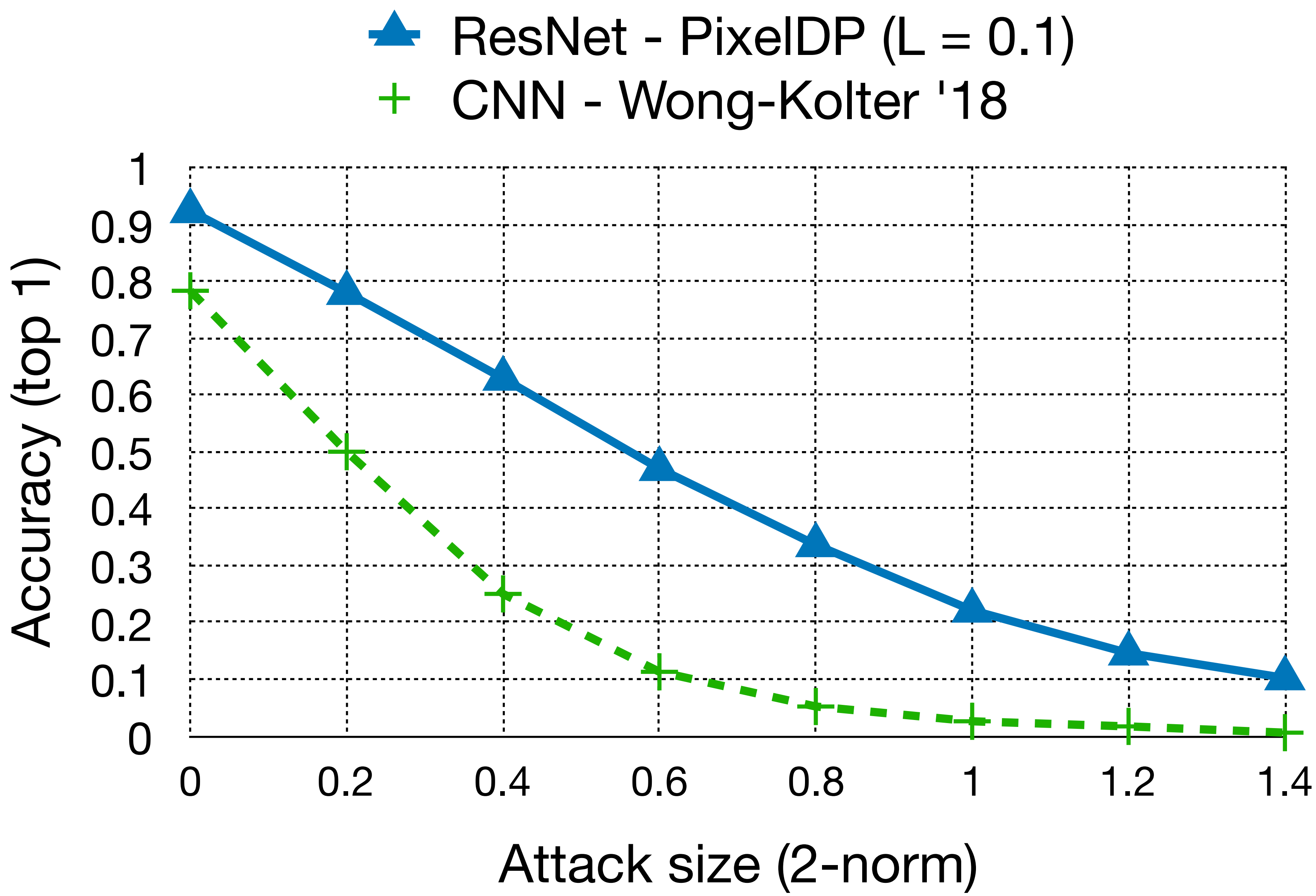
Comparison:
Madry+ '17

If we increase the robustness threshold:
better accuracy, less predictions.

Comparison with other provable defenses

Dataset: SVHN

Comparison:
Wong-Kolter '18



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.

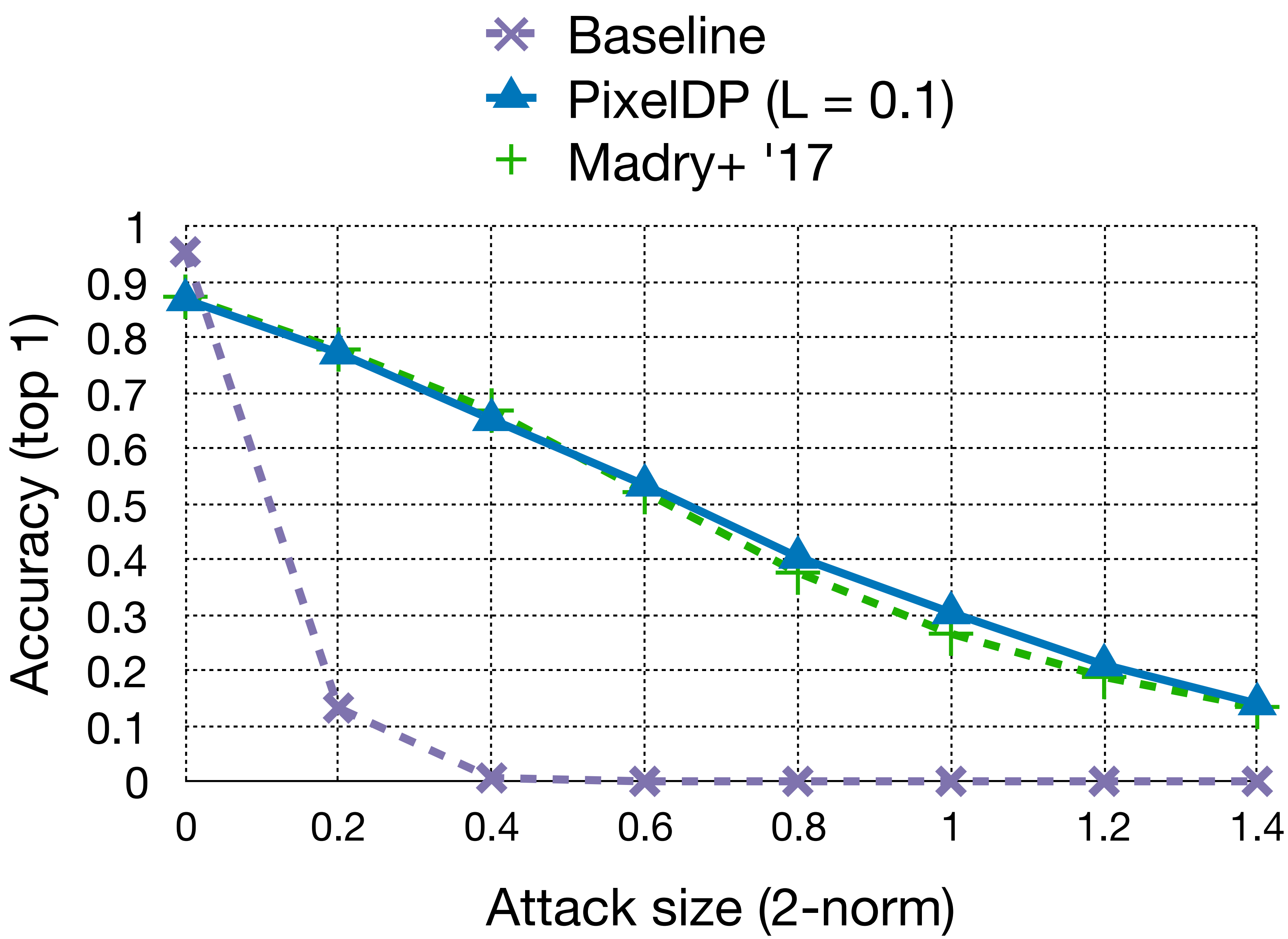
- Already extensions by others!
  - Improve the bounds at a given noise level (Li+ '18; Cohen+ '19).
  - Use other noise distributions (Pinot+ '19).
  - Adapt optimization (Rakin+ '18).

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

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

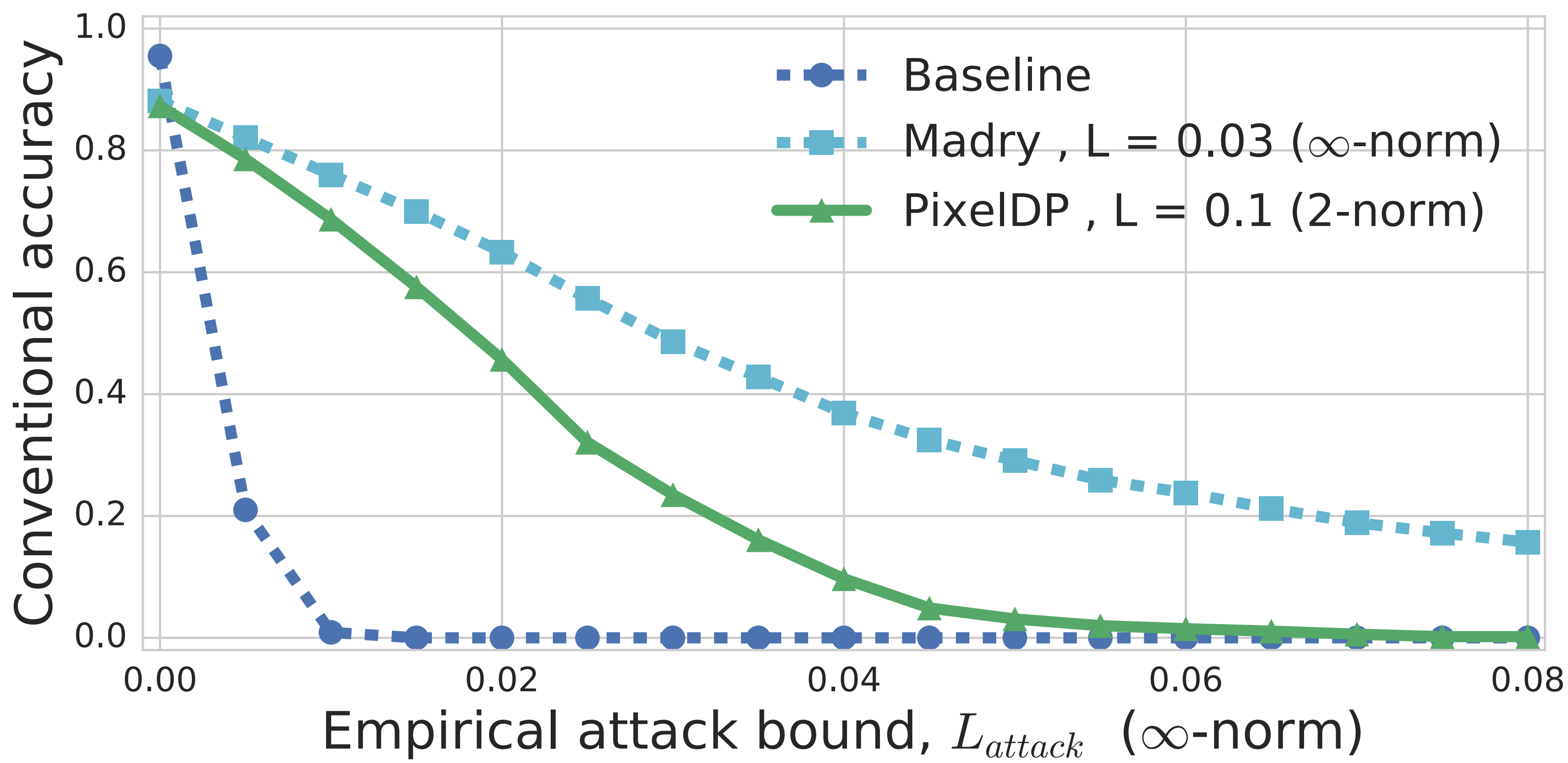
Certified

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

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

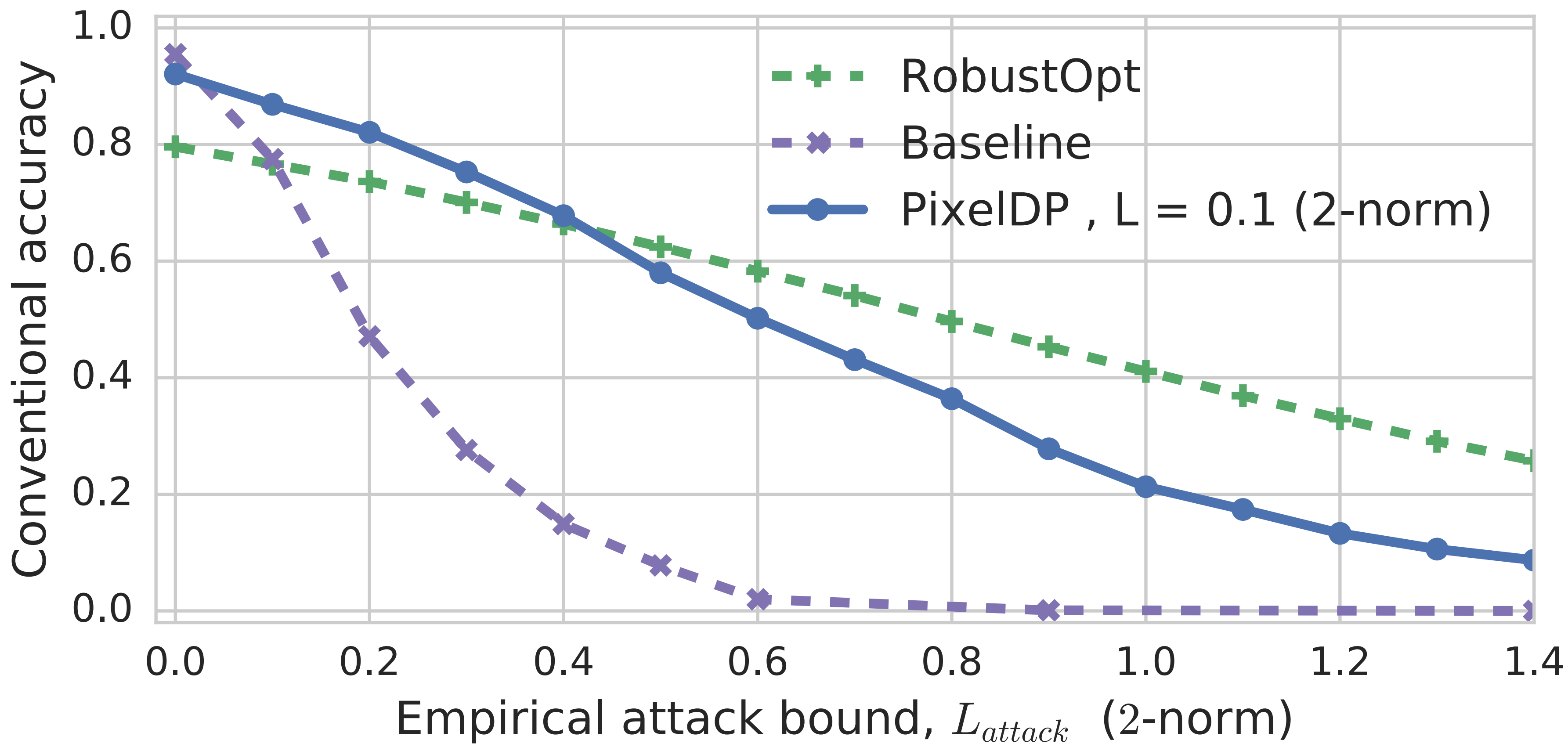
Results - CIFAR-10

L_∞ attack:

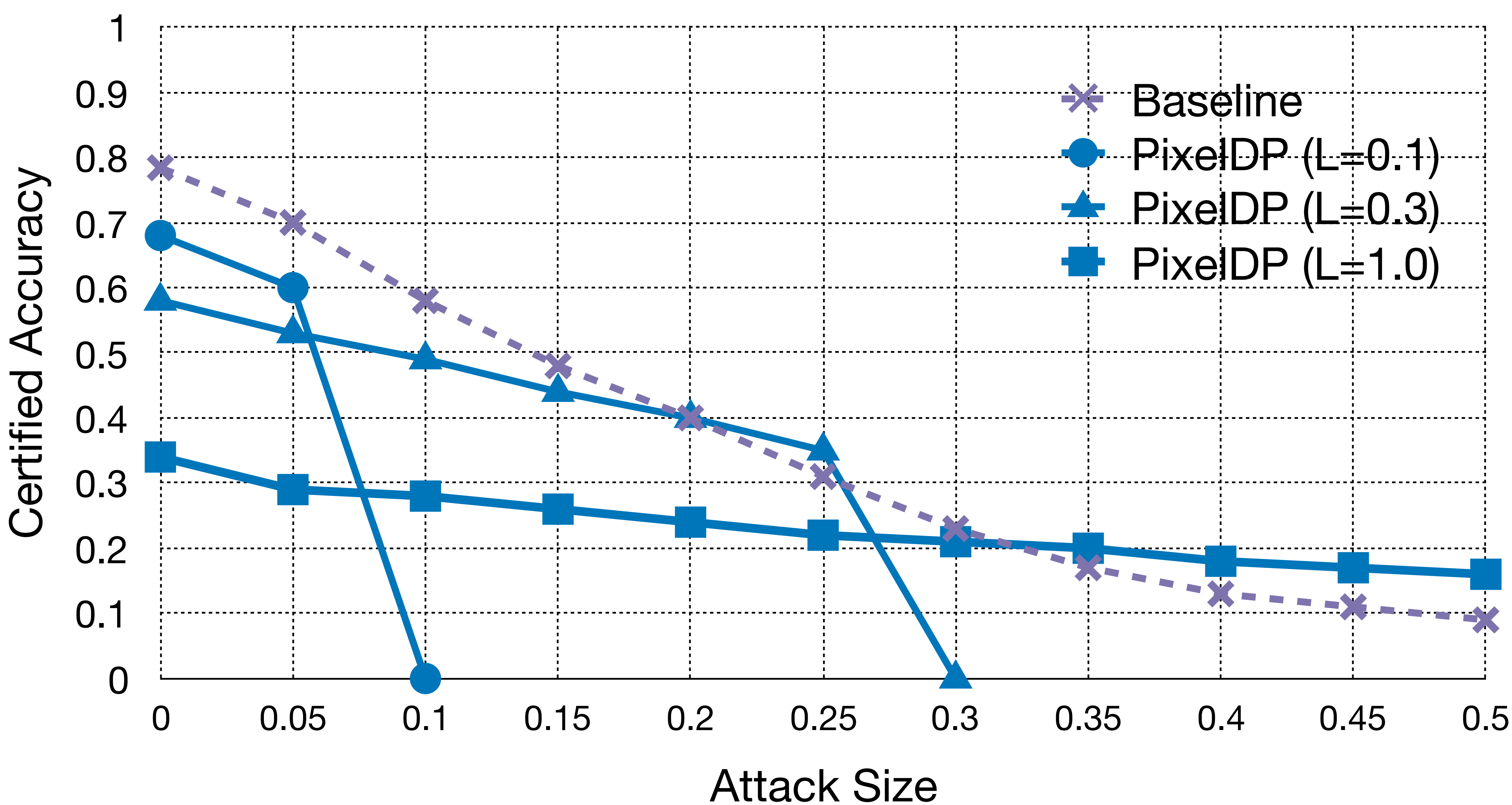


Results - SVHN

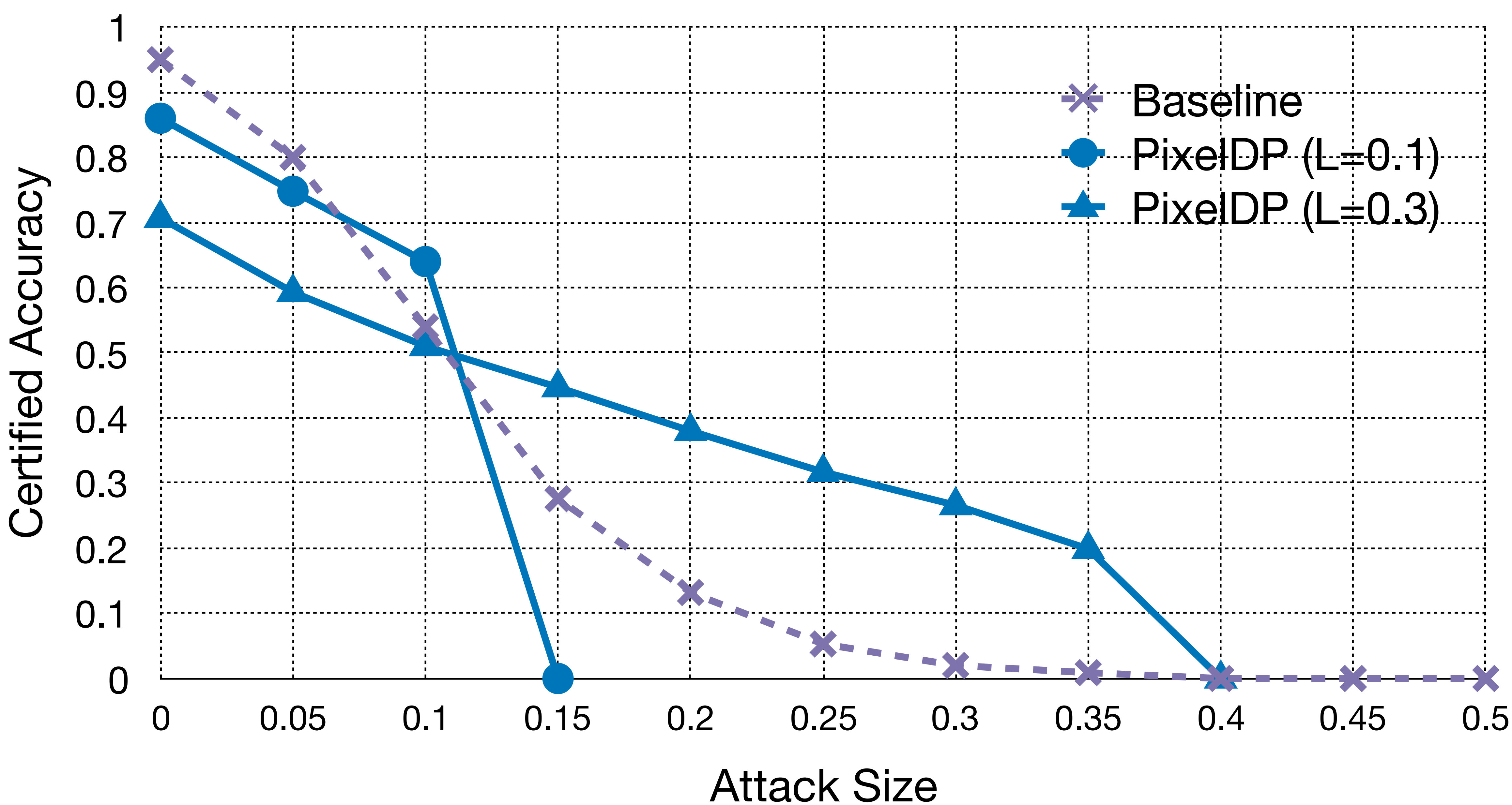
L_∞ attack:



Certification on ImageNet/Inception-v3



Certification on CIFAR-10



Comparison with Best Effort Techniques

Giant panda

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

Teddy bear

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

Teddy bear

Full references

- [Goodfellow+ '15] I. Goodfellow, J. Shlens, and C. Szegedy. *Explaining and harnessing adversarial examples*. ICLR 2015.
- [Papernot+ '16] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. *Distillation as a defense to adversarial perturbations against deep neural networks*. S&P 2016.
- [Buckman+ '18] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow. *Thermometer encoding: One hot way to resist adversarial examples*. ICLR 2018.
- [Guo+ '18] C. Guo, M. Rana, M. Cisse, and L. van der Maaten. *Countering adversarial images using input transformations*. ICLR 2018.
- [Madry+ '17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. *Towards deep learning models resistant to adversarial attacks*. arXiv 2017.

Full references

- [Carlini-Wagner '17]] N. Carlini and D. Wagner. *Towards evaluating the robustness of neural networks*. S&P 2017.
- [Athalye+ '18] A. Athalye, N. Carlini, and D. Wagner. *Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples*. ICML 2018.
- [Wong-Kolter '18] E. Wong and Z. Kolter. *Provable defenses against adversarial examples via the convex outer adversarial polytope*. ICML 2018.
- [Raghunathan+ '18] A. Raghunathan, J. Steinhardt, and P. Liang. *Certified defenses against adversarial examples*. arXiv 2018.
- [Wang+ '18] S. Wang, K. Pei, W. Justin, J. Yang, and S. Jana. *Efficient formal safety analysis of neural networks.* NeurIPS 2018.
- [Li+ '18] B. Li, C. Chen, W. Wang, and L. Carin. *Second-Order Adversarial Attack and Certifiable Robustness. arXiv 2018.*

Full references

- [Rakin+ '18] A.S. Rakin, Z. He, and D. Fan. *Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness against Adversarial Attack*. arXiv 2018.
- [Cohen+ '19] J. Cohen, E. Rosenfeld, and Z. Kolter. *Certified Adversarial Robustness via Randomized Smoothing*. arXiv 2019.
- [Pinot+ '19] R. Pinot, L. Meunier, A. Araujo, H. Kashima, F. Yger, C. Gouy-Pailler, and J. Atif. *Theoretical evidence for adversarial robustness through randomization: the case of the Exponential family*. arXiv 2019.