Towards Certifying ℓ_∞ Robustness using Neural Networks with ℓ_∞-dist Neurons

Bohang Zhang, Tianle Cai, Zhou Lu, Di He, Liwei Wang

Peking University, Princeton University, Microsoft Research

June 20, 2021

Index

1. Introduction
2. Existing Approaches
3. Certified ℓ_∞ Robustness via ℓ_∞-dist Net
   - An 1-Lipschitz Network: ℓ_∞-dist Net
   - Theoretical Properties of ℓ_∞-dist Nets
4. Training ℓ_∞-dist Nets
5. Experiments & Results
6. Conclusion

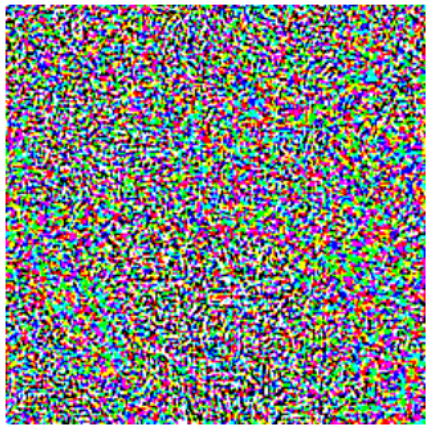
Introduction

- Modern neural networks are usually sensitive to small, adversarially chosen perturbations to the inputs.
- Given an image $\boldsymbol{x}$, an indistinguishable small adversarial perturbation $\boldsymbol{\delta}$ is able to fool the classifier f to produce a wrong class using $f(\boldsymbol{x}+\delta)$.
  - We focus on ℓ_∞-norm bounded perturbations, i.e. $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$.

$\boldsymbol{x}$

"panda"
57.7% confidence

$+\,.007\times$



"nematode"
8.2% confidence

$=$

$\boldsymbol{x}+\boldsymbol{\delta}$

"gibbon"
99.3 % confidence

Existing Approaches

- To improve models' robustness to adversarial examples, many attempts have been made.
- Adversarial training methods:
  - First generate on-the-fly adversarial examples, then train model parameters using these perturbed samples together with the original labels.
  - Can achieve decent empirical robustness against some particular attack methods (e.g. PGD), but cannot give formal (certified) guarantees.
- Training provably robust models:
  - Calculate certified radius provided by robust certification methods (typically based on convex relaxation), then train models to maximize such certified radius.
  - Drawbacks: sophisticated to implement and computationally expensive.

Existing Approaches

- Randomized Smoothing:
  - If a Gaussian random noise is added to the input, a certified guarantee on small ℓ_2 perturbation can be computed for Gaussian smoothed classifiers.
  - Can not achieve good results for relatively large ℓ_∞ perturbations.

- Fundamentally different from these approaches, **we propose a novel network that is inherently 1-Lipschitz with respect to ℓ_∞-norm, and that can provide ℓ_∞ robustness guarantee by its nature**.

ℓ_∞-dist Neuron

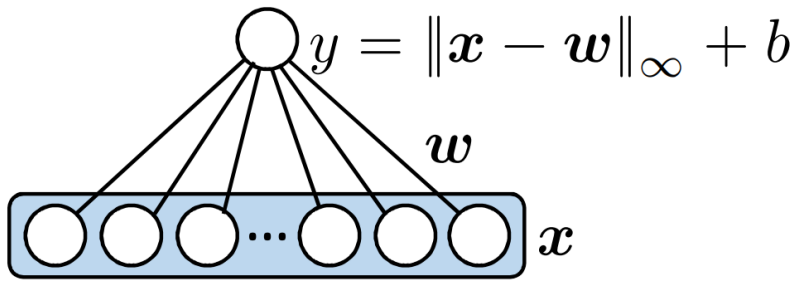
- We introduce a new type of neuron called ℓ_∞-dist neuron, using ℓ_∞ distance as the basic operation:

$$u(\boldsymbol{z}, \theta) = \|\boldsymbol{z} - \boldsymbol{w}\|_\infty + b$$

  - $\theta = \{\boldsymbol{w}, b\}$ is the parameter set.
  - There is no need to further apply a non-linear activation function since the neuron itself is non-linear.

- Similar to dot-product, ℓ_∞-distance is also a similarity measure. A smaller ℓ_∞-distance indicates a stronger similarity.

- Contrast to the conventional neuron, if the perturbation $\|\boldsymbol{\delta}\|_\infty \leq \epsilon$, y can change at most ϵ.



ℓ_∞-dist Net

- Using ℓ_∞-dist Neuron, we can construct ℓ_∞-dist net.
- For example, consider a simple MLP network as follows:
  - The network takes $\boldsymbol{x}^{(0)}$ as input.
  - The k-th unit in the l-th hidden layer $x_k^{(l)}$ is computed by $x_k^{(l)} = u(\boldsymbol{x}^{(l-1)}, \theta^{(l,k)}) = \|\boldsymbol{x}^{(l-1)} - \boldsymbol{w}^{(l,k)}\|_\infty + b^{(l,k)}$.
  - The network outputs $\boldsymbol{x}^{(L)}$.
- We can similarly consider other neural network architectures, such as convolutional networks with weight sharing.
- For classification tasks, $\boldsymbol{x}^{(L)} \in \mathbb{R}^C$, and we can apply any standard loss function on the ℓ_∞-dist net, such as the cross-entropy loss or hinge loss.

Lipschitzness of ℓ_∞-dist Net

- Recall a function $\boldsymbol{g}(\boldsymbol{z}) : \mathbb{R}^m \to \mathbb{R}^n$ is called λ-Lipschitz with respect to ℓ_p-norm $\|\cdot\|_p$, if for any $\boldsymbol{z}_1, \boldsymbol{z}_2$, the following holds:

$$\|\boldsymbol{g}(\boldsymbol{z}_1) - \boldsymbol{g}(\boldsymbol{z}_2)\|_p \leq \lambda \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_p.$$

- **ℓ_∞-dist is 1-Lipschitz with respect to ℓ_∞-norm**
  - First, ℓ_∞-dist neuron is 1-Lipschitz;
  - Then the mapping from one layer to the next layer $\boldsymbol{x}^{(\ell)} \to \boldsymbol{x}^{(\ell+1)}$ is 1-Lipschitz;
  - Finally, the whole network is 1-Lipschitz by composition.

How to Compute Certified Robustness

- The Lipschitz property can be used to certify robustness as follows:
  - Let $f(\boldsymbol{x}) = \arg\max_i \boldsymbol{g}$ and $\boldsymbol{x}$ is correctly classified. Define $\mathrm{margin}(\boldsymbol{x}; \boldsymbol{g})$ as the difference between the largest and second-largest elements of $\boldsymbol{g}(\boldsymbol{x})$. Then the certified robust radius is at least $\mathrm{margin}(\boldsymbol{x}; \boldsymbol{g})/2$.
- Note that **the certification process is very efficient (*only a forward pass required*)**.

The expressive power of ℓ_∞-dist Nets

- Since ℓ_∞-dist nets is Lipschitz with respect to ℓ_∞-norm, it is natural to ask whether ℓ_∞-dist nets can approximate *any* 1-Lipschitz function.

Theorem (Lipschitz-Universal Approximation Theorem for ℓ_∞-dist Nets)

For any 1-Lipschitz function $\tilde{g}(\boldsymbol{x})$ (with respect to ℓ_∞-norm) on a bounded domain $\mathbb{K} \in \mathbb{R}^{d_{\text{input}}}$ and any $\epsilon > 0$, there exists an ℓ_∞-dist net $g(\boldsymbol{x})$ with width no more than $d_{input} + 2$, such that for all $\boldsymbol{x} \in \mathbb{K}$, we have $\|g(\boldsymbol{x}) - \tilde{g}(\boldsymbol{x})\|_\infty \leq \epsilon$.

- This theorem implies that **an ℓ_∞-dist net can approximate any 1-Lipschitz function** with respect to ℓ_∞-norm on a compact set, using width barely larger than the input dimension.

The robust generalization ability of ℓ_∞-dist Nets

- The remaining question is whether ℓ_∞-dist nets can generalize well on unseen test data.
- Consider the following two-class classification problem: let $(\boldsymbol{x}, y) \sim \mathcal{D}$ be an instance-label couple where $y \in \{1, -1\}$. For a function $g(\boldsymbol{x}) : \mathbb{R}^{d_{\text{input}}} \to \mathbb{R}$, we use $\text{sign}(g(\boldsymbol{x}))$ as the classifier. The r-robust test error γ_r of a classifier g is defined as

$$\gamma_r = \mathbb{E}_{\mathcal{D}} \left[\sup_{\|\boldsymbol{x}' - \boldsymbol{x}\|_\infty \leq r} \mathbb{I}_{yg(\boldsymbol{x}') \leq 0} \right]$$

The robust generalization ability of ℓ_∞-dist Nets

Theorem (Robust Generalization Error of ℓ_∞-dist Nets)

Let $\mathbb{F}$ denote the set of all g represented by an ℓ_∞-dist net with width W and depth L. For every $t > 0$, with probability at least $1 - 2e^{-2t^2}$ over the random drawing of n samples, for all $r > 0$ and $g \in \mathbb{F}$ we have that

$$\gamma_r \leq \inf_{\delta \in (0,1]} \left[\frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{I}_{y_i g(\boldsymbol{x}_i) \leq \delta + r}}_{\text{large training margin}} + \underbrace{\tilde{O}\left(\frac{LW^2}{\delta \sqrt{n}} \right)}_{\text{network size}} + \left(\frac{\log \log_2 (\frac{2}{\delta})}{n} \right)^{\frac{1}{2}} \right] + \frac{t}{\sqrt{n}}. \tag{1}$$

- This theorem demonstrates that when a large margin classifier is found on training data, and the size of the ℓ_∞-dist net is not too large, then with high probability, the model can generalize well in terms of adversarial robustness.

Training ℓ_∞-dist Nets

- We empirically find that the optimization is challenging and directly training the network usually *fails* to obtain a good performance.
- Furthermore, as this architecture is entirely new, the tricks and techniques for conventional network training may not be appropriate in our setting.

Normalization

- The output of an ℓ_∞-dist neuron is biased (always being non-negative, assuming no bias term).
- This will cause the output scale in upper layers linearly increase.
- However, we can not apply batch normalization in ℓ_∞-dist nets, since the Lipschitz constant will change due to the scaling operation, and the robustness of the model cannot be guaranteed.
- Fortunately, we find that using the shift operation alone already helps the optimization.
- Similar to BatchNorm, we use the running mean during inference, which serves as additional bias terms in ℓ_∞-dist neurons and does not affect the Lipschitz constant of the model.

Smoothed Approximated Gradients

- Another optimization difficulty: the gradients of the ℓ_∞-dist operation are very **sparse** which typically contain only one non-zero element.
- In practice, we observe that there are **less than 1%** parameters updated in an epoch if we directly train the ℓ_∞-dist net using SGD/Adam from random initialization.
- To improve the optimization, we relax the ℓ_∞-dist neuron by using the ℓ_p-dist neuron for the whole network to get an approximate and non-sparse gradient of the model parameters.
- During training, we set p to be a small value in the beginning and increase it in each iteration until it approaches infinity. For the last few epochs, we set p to infinity and train the model to the end.

Experiments & Results

Table 1. Comparison of our results with existing methods[1].

Dataset	Method	FLOPs	Test	Robust	Certified
MNIST ($\epsilon = 0.3$)	Group Sort (Anil et al., 2019)	2.9M	97.0	34.0	2.0
	COLT (Balunovic & Vechev, 2020)	4.9M	97.3	-	85.7
	IBP (Gowal et al., 2018)	114M	97.88	93.22	91.79
	CROWN-IBP (Zhang et al., 2020b)	114M	98.18	93.95	92.98
	ℓ_∞-dist Net	82.7M	98.54	94.71	92.64
	ℓ_∞-dist Net+MLP	85.3M	**98.56**	**95.28**	**93.09**
Fashion MNIST ($\epsilon = 0.1$)	CAP (Wong & Kolter, 2018)	0.41M	78.27	68.37	65.47
	IBP (Gowal et al., 2018)	114M	84.12	80.58	77.67
	CROWN-IBP (Zhang et al., 2020b)	114M	84.31	80.22	78.01
	ℓ_∞-dist Net	82.7M	**87.91**	79.64	77.48
	ℓ_∞-dist Net+MLP	85.3M	**87.91**	**80.89**	**79.23**
CIFAR-10 ($\epsilon = 8/255$)	PVT (Dvijotham et al., 2018a)	2.4M	48.64	32.72	26.67
	DiffAI (Mirman et al., 2019)	96.3M	40.2	-	23.2
	COLT (Balunovic & Vechev, 2020)	6.9M	51.7	-	27.5
	IBP (Gowal et al., 2018)	151M	50.99	31.27	29.19
	CROWN-IBP (Zhang et al., 2020b)	151M	45.98	34.58	33.06
	CROWN-IBP (loss fusion) (Xu et al., 2020a)	151M	46.29	35.69	33.38
	ℓ_∞-dist Net	121M	**56.80**	**37.46**	33.30
	ℓ_∞-dist Net+MLP	123M	50.80	37.06	**35.42**

Experiments & Results

Table 2. Comparison of our results with Xu et al. (2020a) on TinyImageNet dataset ($\epsilon = 1/255$).

Method	Model	FLOPs	Test	Robust	Certified
CROWN-IBP (loss fusion) (Xu et al., 2020a)	CNN7+BN	458M	21.58	19.04	12.69
	ResNeXt	64M	21.42	20.20	13.05
	DenseNet	575M	22.04	19.48	14.56
	WideResNet	5.22G	27.86	20.52	15.86
ℓ_∞-dist net	ℓ_∞-dist Net+MLP	156M	21.82	18.09	**16.31**

Table 4. Comparison of per-epoch training speed for different methods on CIFAR-10 dataset.

Method	Per-epoch Time (seconds)
IBP	17.4
CROWN-IBP	112.4
CROWN-IBP (loss fusion)	43.3
ℓ_∞-dist Net	19.7
ℓ_∞-dist Net+MLP	19.7

Conclusion & Future Work

- We design a novel neuron that uses ℓ_∞ distance as its basic operation.
- We show that ℓ_∞-dwist net is naturally a 1-Lipschitz function with respect to ℓ_∞ norm, which provides a theoretical guarantee of the certified robustness based on the margin of the prediction outputs.
- We further formally analyze the expressive power and the robust generalization ability of the network, and provide a holistic training strategy to handle optimization difficulties encountered in training ℓ_∞-dist nets.
- Experiments show promising results on MNIST, Fashion-MNIST, CIFAR-10 and TinyImageNet datasets.
- We hope this work can bring a new research direction in the area of certified robustness.

Thank You!