A Short Survey on Adversarial Robustness

Yibin Wang
yibinwang@hust.edu.cn

Abstract

Neural network models have gained great success on clean samples, but they are known to be vulnerable to adversarial perturbation which is almost indistinguishable. To address the problem, adversarial training has been proposed to enhance model's robustness to adversarial perturbation. Nowadays, great achievements have been reached by using different methods to enhance the adversarial training. Many approaches such as data augmentation, modification to the model structure and loss function, convex relaxation and random smoothing have achieved significant effects on empirical robustness and certified robustness. However, existing method still cannot achieve robustness towards all undetectable adversarial perturbations. In this paper, we summarize the previous works on adversarial robustness based on different methods and discuss the shortcomings and future works.

1 Brief on Adversarial Training

For the classification task, a model predicts label $y \in \mathcal{Y}$ given a input $x \in \mathbb{R}^d$ and the output space $\mathcal{Y} = [k]$ contains k classes. Given a suitable loss function $L(\theta, x, y)$, we train the model parameters θ by minimizing the risk $\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(\theta, x, y)]$. For each data point x, we introduce a set of allowed perturbation $\mathcal{S} \subseteq \mathbb{R}^d$ that restrict the perturbation to be indistinguishable. The ℓ_p-ball around x is studied as a natural notion for adversarial perturbation in previous works.

Adversarial attacks aim to find the allowed perturbation to maximize the risk $\mathbb{E}_{(x,y)\sim\mathcal{D}}[L(\theta, x, y)]$ using the optimization below:

$$\underset{\delta \in \mathcal{S}}{\operatorname{argmax}}\, L(\theta, x + \delta, y). \tag{1}$$

To achieve the robustness against adversarial attack, adversarial training (Madry et al., 2017) perturb the input using adversarial attack first and then use the input to minimize the loss function:

$$\begin{aligned} &\min_\theta \rho(\theta), \quad \text{where} \\ &\rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max_{\delta \in \mathcal{S}} L(\theta, x + \delta, y)]. \end{aligned} \tag{2}$$

2 Evaluation of Robustness

Empirical Robustness The empirical robustness is evaluated by adversarial attack based on Eq. (1). Madry et al. (2017) used experiments to show PGD attack is a reliable first-order adversary, i.e., attack that rely only on first-order information. Auto-attack (Croce and Hein, 2020) is a universal and parameter-free evaluation through ensemble attacks based on PGD attack, which has been taken as the baseline of robustness evaluation.

Certified Robustness Empirical robustness cannot provide the guarantee that the worst situation which can be achieved through exhaustive search. The technique of adversarial attack does not always find the worst-case attack (a phenomenon also observed by Tjeng et al. (2017)). In contrast, certified robustness guarantees that a model is robust to all adversarial perturbations of a given input, regardless of the attacks for evaluation. Certified robustness provides a lower bound on the robust accuracy of a model in the face of various adversarial attacks.

3 Empirical Robustness Based on Different Methods

3.1 Modification of Loss Function

The trade off between robust accuracy and clean accuracy is the crucial problem of adversarial robustness which is firstly observed by Tsipras et al. (2018). TRADES (Zhang et al., 2019) decomposes the prediction error for adversarial examples (robust error) as the sum of the natural (classification) error and boundary error, and provides a differentiable upper bound using the theory of classification-calibrated loss. The classification-calibrated loss trades adversarial robustness off against accuracy to achieve better performance on robustness than the vanilla adversarial training.

3.2 Data augmentation

Adversarial training meets a serious over-fitting which is observed by a series of works (Rice et al. 2020; Rebuffi et al. 2021;). Data augmentation is naturally taken to mitigate over-fitting. (Alayrac et al. 2019; Carmon et al. 2019) use semi-supervised learning with unlabeled data to bridge sample complexity gap between standard and robust classification. Data generation with DDPM (Ho et al., 2020) is used to augment data and get the SOTA result (Rebuffi et al. 2021; Gowal et al. 2021; Wang et al. 2023).

4 Certified Robustness Based on Different Methods

4.1 Certified Robustness through convex relaxation

For the classification task, a model $f : \mathcal{X} \to \mathcal{Y}$ predicts label $y \in \mathcal{Y}$ given a input $x \in \mathcal{X}$, where the output space $\mathcal{Y} = \{y_1, y_2, \cdots, y_M\}$ contains M classes.

Formally, we use $\mathcal{B}_{adv}(x)$ to denote the adversarial set of the input x. We formulate the set consisting of all the adversarial examples with allowed perturbations of x:

$$\mathcal{B}_{adv}(x) = x + \delta, \quad \delta \in \mathcal{S} \tag{3}$$

Our goal is to defend against the adversarial word substitutions and train models with certified robustness, *i.e.*,

$$\forall\, \mathbf{x}' \in \mathcal{B}_{adv}(x), \quad f(x') = f(x) = y. \tag{4}$$

If Eq. (4) holds and the model classifies the instance correctly, that is, $y = y_{true}$, then we call the model prediction on input x is certified.

We can easily decompose certified robust accuracy into *robustness* and *standard accuracy*. Robustness cares about whether the model prediction is consistent under perturbations. Clearly, achieving robustness is a necessary condition for obtaining models with high certified robust accuracy. We next illustrate the conditions to be satisfied for robustness in terms of interval bound. Assuming we can calculate the interval bound of the output logits $\mathbf{z}^K$: $\underline{\mathbf{z}}^K \leq \mathbf{z}^K \leq \overline{\mathbf{z}}^K$ of all the perturbed inputs $x' \in \mathcal{B}_{adv}(x)$, a model with robustness satisfies that the lower bound of the model's largest logit $\mathbf{z}^K_{y_{max}}$ is greater than the upper bound of other logits, *i.e.*,

$$\underline{\mathbf{z}}^K_{y_{max}} \geq \overline{\mathbf{z}}^K_y, \quad \forall y \in \mathcal{Y}, y \neq y_{max}. \tag{5}$$

To evaluate the model's certified robust accuracy, we just need to replace the model's largest logit $\mathbf{z}^K_{y_{max}}$ with the logit of the true class $\mathbf{z}^K_{y_{true}}$ in Eq. (5).

Interval Bound Propagation IBP provides the solution to estimate the interval bound layer by layer. We could represent a K-layer neural network model as a series of transformations f_k:

$$\mathbf{z}^k = f_k(\mathbf{z}^{k-1}), \quad k = 1, \cdots, K, \tag{6}$$

where $\mathbf{z}^k$ is the vector of activations in the k^{th} layer. To calculate the interval bound of the output logits, we need to construct the interval bound of the input vector and propagate it through the network. Let $\varphi(x_i) \in \mathbb{R}^D$ denote the embedding word vector of word x_i with D dimensions. The word vector input is $\mathbf{z}^0 = \langle \varphi(x_0), \varphi(x_1), \cdots, \varphi(x_N) \rangle$. We obtain the interval bounds of the word vector input $\mathbf{z}^0$ by constructing the convex hull of $\mathcal{S}(x_i)$ in the embedding space:

$$\begin{aligned} \underline{z}^0_{ij} &= \min_{x_i \in \mathcal{S}(x_i) \cup \{x_i\}} \varphi(x_i)_j, \\ \overline{z}^0_{ij} &= \max_{x_i \in \mathcal{S}(x_i) \cup \{x_i\}} \varphi(x_i)_j, \end{aligned} \tag{7}$$

where $\varphi(x_i)_j$ is the j^{th} element of the word vector of word x_i. $\underline{\mathbf{z}}^0$ and $\overline{\mathbf{z}}^0$ are the lower and upper bounds of $\mathbf{z}^0$, respectively.

Similarly, for subsequent layers $k > 0$, we denote the lower and upper bounds of activations in the k^{th} layer as $\underline{\mathbf{z}}^k$ and $\overline{\mathbf{z}}^k$, respectively. The bounds on the $\mathbf{z}^k$ can be obtained from the bounds of previous layer $\mathbf{z}^{k-1}$:

$$\begin{aligned} \underline{z}^k_i &= \min_{\underline{\mathbf{z}}^{k-1} \leq \mathbf{z}^{k-1} \leq \overline{\mathbf{z}}^{k-1}} \mathbf{e}_i^\top f_k(\mathbf{z}^{k-1}), \\ \overline{z}^k_i &= \max_{\underline{\mathbf{z}}^{k-1} \leq \mathbf{z}^{k-1} \leq \overline{\mathbf{z}}^{k-1}} \mathbf{e}_i^\top f_k(\mathbf{z}^{k-1}), \end{aligned} \tag{8}$$

where $\mathbf{e}_i$ is the one-hot vector with 1 in the i^{th} position.

Interval Bound Propagation (IBP) (Gowal et al., 2018) gives a simple way to solve the above problems for affine layers and monotonic activation functions as described in Appendix A. Shi et al.

(2021) proposed a series of improvements on IBP training mitigating exploded bounds at initialization, and the imbalance in ReLU activation states. SABR (Müller et al., 2022) builds a smaller perturbation simplex and achieves tighter verifiable bound with selecting the subset of the adversarial input. L-infinity network (Zhang et al., 2022, Zhang et al., 2021) is a novel structure inherently resists ℓ_∞ perturbations and achieves the SOTA result on certified robustness of ℓ_∞ perturbations. But the property of L-infinity network that only restricts on ℓ_∞ perturbations leads to its lack of resistance to large ℓ_2 perturbations.

4.2 Certified Robustness through random smoothing

Randomized smoothing (Cohen et al., 2019, Lecuyer et al., 2019, Zhang et al., 2020) is a model structure-free method towards certified robustness. However, Yang et al. (2020) and Blum et al. (2020) showed that randomized smoothing cannot achieve meaningful results for a relatively large ℓ_∞ perturbation due to the curse of dimensionality.

5 Discussion

In this section, we discuss the shortcomings of the existing methods and future works.

Firstly, the trade-off between robust accuracy and clean accuracy is the crucial problem of adversarial robustness. The existing methods have shown improvement in empirical robustness, but they often come at the cost of clean accuracy. Further exploration of loss functions that can provide a better trade-off between the two accuracy metrics is needed.

Secondly, data augmentation has been shown to be effective in mitigating over-fitting in adversarial training. However, the current data augmentation methods are limited in their ability to generate diverse and realistic adversarial examples. Developing more powerful data augmentation methods that can better capture the diversity of adversarial examples could further improve the robustness of models.

Thirdly, while certified robustness provides a lower bound on the robust accuracy of a model, it often provides conservative estimates of the true robust accuracy. Finding more efficient and accurate methods for certified robustness is an important direction for future research.

Finally, it is worth noting that the adversarial robustness of neural network models is a challenging and open problem. New adversarial attack methods are continuously being proposed, and existing models can be easily broken by these attacks. Therefore, developing more robust models that can defend against a wide range of attacks is an ongoing challenge in the field of machine learning.

References

Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. 2019. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32.

Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. 2020. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *The Journal of Machine Learning Research*, 21(1):8726–8746.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. 2019. Unlabeled data improves adversarial robustness. *Advances in neural information processing systems*, 32.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320. PMLR.

Francesco Croce and Matthias Hein. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. 2018. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. 2021. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34:4218–4233.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mark Niklas Müller, Franziska Eckert, Marc Fischer, and Martin Vechev. 2022. Certified training: Small boxes are all you need. *arXiv preprint arXiv:2210.04871*.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. 2021. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*.

Leslie Rice, Eric Wong, and Zico Kolter. 2020. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR.

Zhouxing Shi, Yihan Wang, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. 2021. Fast certified robust training with short warmup. *Advances in Neural Information Processing Systems*, 34:18335–18349.

Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. 2017. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. 2023. Better diffusion models further improve adversarial training. *arXiv preprint arXiv:2302.04638*.

Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. 2020. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR.

Bohang Zhang, Du Jiang, Di He, and Liwei Wang. 2021. Boosting the certified robustness of l-infinity distance nets. In *International Conference on Learning Representations*.

Bohang Zhang, Du Jiang, Di He, and Liwei Wang. 2022. Rethinking lipschitz neural networks for certified l-infinity robustness. *arXiv preprint arXiv:2210.01787*.

Dinghuai Zhang, Mao Ye, Chengyue Gong, Zhanxing Zhu, and Qiang Liu. 2020. Black-box certification with randomized smoothing: A functional optimization based framework. *Advances in Neural Information Processing Systems*, 33:2316–2326.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR.

A Interval Bound Propagation

Here we give a brief description of Interval Bound Propagation (IBP) (Gowal et al., 2018) on its calculation of bound propagation and training loss.

Bound Propagation For Eq. (8), IBP provides corresponding calculation methods for affine layers and monotonic activation functions:

- For the affine transformation, denoted by $\mathbf{z}^{k+1} = \mathbf{W}\mathbf{z}^k + \mathbf{b}$, we have:

$$\begin{aligned} \mathbf{u}^{k+1} &= \frac{1}{2}\mathbf{W}(\overline{\mathbf{z}}^k + \underline{\mathbf{z}}^k), \\ \mathbf{r}^{k+1} &= \frac{1}{2}\left|\mathbf{W}\right|(\overline{\mathbf{z}}^k - \underline{\mathbf{z}}^k), \\ \overline{\mathbf{z}}^{k+1} &= \mathbf{u}^{k+1} + \mathbf{r}^{k+1}, \\ \underline{\mathbf{z}}^{k+1} &= \mathbf{u}^{k+1} - \mathbf{r}^{k+1}, \end{aligned} \tag{9}$$

  where $|\cdot|$ is the element-wise absolute value operator.

- For the element-wise monotonic activation function (*e.g.* ReLU, tanh, sigmoid), denoted by $\mathbf{z}^{k+1} = h(\mathbf{z}^k)$, we have:

$$\begin{aligned} \overline{\mathbf{z}}^{k+1} &= h(\overline{\mathbf{z}}^k), \\ \underline{\mathbf{z}}^{k+1} &= h(\underline{\mathbf{z}}^k). \end{aligned} \tag{10}$$

IBP Loss For the interval bounds calculated by Eq. (7), the IBP method scales them with scalar ϵ:

$$\begin{aligned} \underline{z}^0_{ij}(\epsilon) &= z^0_{ij} - \epsilon(z^0_{ij} - \underline{z}^0_{ij}), \\ \overline{z}^0_{ij}(\epsilon) &= z^0_{ij} + \epsilon(\overline{z}^0_{ij} - z^0_{ij}). \end{aligned} \tag{11}$$

Using bound propagation, we can get the lower bound and upper bound of logits with the scalar ϵ: $\underline{\mathbf{z}}^K(\epsilon)$ and $\overline{\mathbf{z}}^K(\epsilon)$, respectively. Similar to Eq. (5), we can get the worst-case logits and use them to construct the IBP loss:

$$\mathcal{L}_{IBP}(\epsilon) = \mathcal{L}_{CE}(\mathbf{z}^K_{worst}(\epsilon), y_{true}), \tag{12}$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss and $\mathbf{z}^K_{worst}(\epsilon)$ is the worst-case logits:

$$\mathbf{z}^K_{worst}(\epsilon) = \begin{cases} \underline{\mathbf{z}}^K_{y_{true}}(\epsilon) & \text{if } y = y_{true}, \\ \overline{\mathbf{z}}^K_y(\epsilon) & \text{otherwise.} \end{cases} \tag{13}$$

Then, IBP loss can be combined with normal cross-entropy loss to train the model and boost the certified robust accuracy:

$$\mathcal{L}_{model} = (1-\beta)\mathcal{L}_{CE}(\mathbf{z}^K, y_{true}) + \beta\mathcal{L}_{IBP}(\epsilon). \tag{14}$$