ICIC Express Letters
Volume **18**, Number **5**, May **2024**

 ISSN 1881-803X
pp. **485–492**

IMPROVING CERTIFIED ROBUSTNESS OF SENSING-REASONING MODELS VIA LIPSCHITZ NEURAL NETWORKS

Yicong Li, Kuanjiu Zhou*, Usman Arshad, Shangzhao Zhai and Mingyu Fan

School of Software
Dalian University of Technology
No. 321, Tuqiang Street, Jinzhou District, Dalian 116620, P. R. China
{ 17640284112; fmy }@mail.dlut.edu.cn; Usman@dlut.edu.cn; 2997152047@qq.com
*Corresponding author: zhoukj@dlut.edu.cn

Received August 2023; accepted October 2023

Abstract. *Sensing-reasoning models are a promising way to relax the effective perturbation radius for robustness certification. However, such models require the participation of power-hungry Gaussian noise training to achieve random smoothing of each sensing-CNN in exchange for certifiable robustness of the components. Moreover, the robustness of this certification approach to l_∞ perturbations is deficient. To avoid the sacrifice of this high certification cost and the adaptation defect of l_∞ perturbations, this paper proposes an optimization model based on Lipschitz property, which bypasses the perturbation radius by introducing Lipschitz neural networks as sensing components and solving the perturbation radius directly with norm-bounded affine transformations and order statistics property. The random smoothing training of CNN is used to trade off the certification overhead and classification performance. Extensive experiments illustrate that our model substantially improves the validation efficiency compared to state-of-the-art models. Our model also obtains excellent l_∞ perturbations certified accuracy and exhibits stable defense against adversarial attacks.*

Keywords: Certified robustness, Lipschitz neural networks, Sensing-reasoning models, Adversarial attacks, Perturbation radius

1. **Introduction.** Certified robustness of deep neural networks has received significant attention in recent years, especially in highly safety-sensitive areas such as autonomous driving, medical diagnosis, and aviation decision making. It replaces simply improving the empirical robustness of the model, and can effectively defend against adaptive adversarial attacks that can learn robustness training methods [1,16,20]. Certified robustness involves verifying that the model can still produce the correct output when the attacking perturbations are limited within a parametric range [2,17,18].

Unfortunately, the perturbation bounds required for such robustness certification are pretty narrow, while actual adversarial attacks usually impose perturbations on model inputs larger than this bound, making it difficult to vouch for robustness certification for a large range of perturbed input samples [19]. In recent years, researchers have proposed sensing-reasoning models to relax the boundaries of perturbations. These models embed exogenous knowledge behind a data-driven perception model to facilitate logical inference. By combining domain knowledge to constrain the output of the perception model, sensing-reasoning models generalize the impact of attack perturbations on the model output [3-6,20,23]. This approach somewhat increases the bound of the perturbation paradigm for which robustness certification is capable of guaranteeing, but at the expense of certification efficiency [22]. This logical inference requires the output of many sensing components with convolutional models as the base stem as the input to the knowledge

DOI: 10.24507/icicel.18.05.485

inference model, and each sensing component requires several times more perturbation training samples than normal training by random smoothing to obtain verifiable robustness. For example, assuming that each input image samples 5 scrambled inputs, random smoothing training for a perceptual model consisting of 6 convolutional components requires 30 training samples [21]. In addition, this random smoothing-based training of the sensing components can only provide 2-parameter robustness guarantees and cannot handle infinite paradigms, which are mostly used for practical adversarial attacks.

Inspired by the Lipschitz property, the adversarial robustness of neural networks is closely related to their Lipschitz continuity. In this paper, we propose a novel sensing-reasoning model based on Lipschitz neural networks in order to improve the certification efficiency of sensing components and to provide certified robustness with infinite parametrization. We utilize the Lipschitz neural network to replace the existing perceptual component, and use the output of the Lipschitz neural network as the input of the inference component. In addition, we improve the overall robustness validation for the Lipschitz neural network-based sensing-reasoning model. Adequate experiments demonstrate that our framework achieves competitive certification accuracy while reducing the verification cost of the perceptual component and obtaining robust certification under infinite parametric perturbation bounds.

The contributions of this paper are as follows.

- We propose a sensing-reasoning model based on Lipschitz neural networks that bypasses the random smoothing training of the convolutional model. To the best of our knowledge, this is the first time that a Lipschitz neural network is introduced into a sensing-reasoning model.
- We provide incomplete robustness certification on our sensing-reasoning model as a whole, where we first compute the output bound of the Lipschitz neural network and use the output bound as the input perturbation range of the inference model.

The rest of the paper is organized as follows. We will first outline related work on sensing-reasoning models and Lipschitz neural networks in Section 2. Then, the details of our model will be elaborated in Section 3.1, i.e., the use of Lipschitz neural networks instead of the sensing component of the sensing-reasoning model. Next, the certified robustness of sensing and reasoning are obtained in Section 3.2. Our work is comprehensively evaluated in terms of both attack defense and verification accuracy in Section 4. Finally, conclusion and future research are summarized in Section 5.

2. **Related Work.**

2.1. **Sensing-reasoning models.** Extensive previous work [3-5] has demonstrated that sensing-reasoning models can constrain the input of neural network models by embedding external knowledge and output robust results after being subject to perturbations. And [6] provides the first verification of the robustness of sensing-reasoning models. Unfortunately, this certified robustness is obtained by random smoothing training of multiple perceptual components, which severely increases the model cost.

2.2. **Lipschitz neural networks.** Previous work [7-10] has illustrated that Lipschitz networks essentially imply robustness of authentication and lead to simpler authentication processes based on output bounds. Much of the work has dealt with the ‘2-norm Lipschitz case by using specific mathematical properties such as spectral criteria or weight matrices. In contrast, for the ∞ criterion, the standard Lipschitz network is not well certified. Huster et al. [11] found that the standard Lipschitz ReLU network cannot represent simple functions, such as absolute value functions, which inspired the first expressive GroupSort network [12]. Recently, Zhang et al. [13] first proposed a practical 1-Lipschitz architecture based on a special ∞-norm of neurons, called ∞-distance neurons, which can be extended to TinyImageNet with state-of-the-art certified robustness over relaxation-based methods.

[14] rethought Lipschitz at infinity-norm Lipschitz according to GroupSort and verified its superiority. And to the best of our knowledge, there are no studies applying Lipschitz properties to sensing-reasoning models.

3. Method.

3.1. The sensing-reasoning model based on Lipschitz neural network.

Our model consists of a perception module, which contains n binary classifiers L_i constructed by a Lipschitz neural network, and an inference component, which is constructed by a probabilistic graphical model, and an inference module, which we apply the Markov logic network (MLN) from the work of [15].

Figure 1 shows the overall architecture of our sensing-reasoning model based on Lipschitz's neural network. For each Lipschitz neural network, a unified framework is employed from [14]. This framework is essentially a fully connected neural network with H layers. The framework consists of three components: (i) parametric bounded affine transformation; (ii) Lipschitz unitary activation function; and (iii) sequential statistics. Specifically, each Lipschitz sensing component is shown in Equation (1):

$$x_j^h = \left(W_j^h\right)^T \cdot sort\left(laf\left(x^{h-1} + s_j^h\right)\right), \quad \left\|W_j^h\right\|_1 \leq 1 \tag{1}$$

Here, x_j^h is the value of the j-th neuron of the h-th layer of the Lipschitz neural network, subject to the Lipschitz property that the 1-parameter of the network weight W must be less than 1. $sort(\cdot)$ calculates all the order statistics of the $h-1$ layer $x \in R_J$, $laf(\cdot)$ denotes the Lipschitz activation function, and s denotes the bias parameter of the j-th neuron of the h-th layer. The network so propagates the H layers until the binary classification probability is calculated.

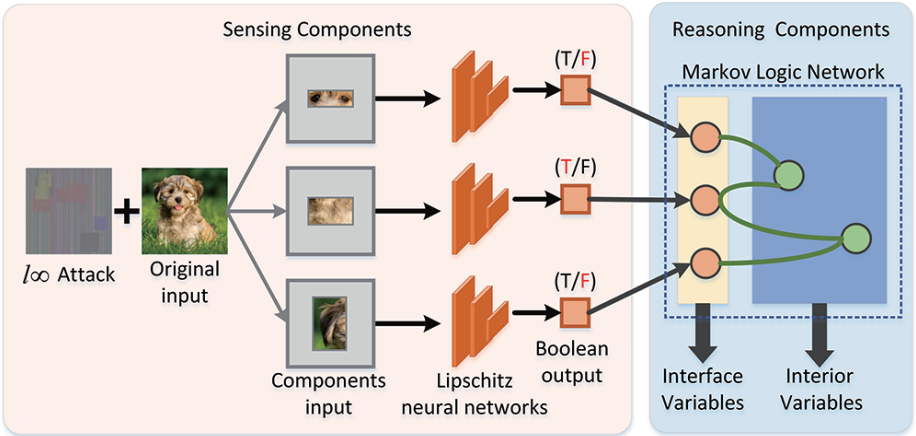

FIGURE 1. The sensing-reasoning model based on Lipschitz neural network

In the sensing component illustrated in Figure 1, given an input image X, it is manually segmented into n semantic sub-images respectively n as inputs to L_i. L_i to recognize the n semantic features e^i of the input image X respectively, each Lipschitz neural network L_i outputs a probability $P_{e^i}(X)$ corresponding to the semantics, and we set a threshold w such that when $P_{e^i}(X) > w$, a Boolean variable r_i is output that is true, otherwise it is false. In the reasoning component, MLN obtains n output Boolean values from the Lipschitz neural network as interface variables. The exogenous knowledge summarized by the experts is embedded in the internal variables and logical inference is performed based on the Boolean values of the interface variables. The joint probability P_{MLN} of the MLN is calculated based on the potential function, and if P_{MLN} is less than a threshold d, the input X is attacked, resulting in logically mutually exclusive output results from the n perceptual modules, and the MLN masks the interface variable that affects the joint

probability most significantly by backward calculation through gradient propagation. Due to space constraints, the process of calculating P_{MLN} is not elaborated in this paper, which is beyond the contribution.

3.2. **Robustness certification of the model as a whole.** This section describes the robustness certification of our optimized sensing-reasoning model. Suppose an infinite or two-parametric perturbation λ_i, $\|\lambda_i\|_\infty \vee \|\lambda_i\|_2 \leq \varepsilon$ is applied to the input X, the robustness certification of the sensing-reasoning model can be simply formalized as Equation (2):

$$\begin{aligned} & P_{MLN}\left(\{r_i \Leftarrow P_{e^i}(X)\}_{i\in n}\right) - P_{MLN}\left(\{r_i \Leftarrow P_{e^i}(X+\lambda_i)\}_{i\in n}\right) \\ \leq\ & \rho \wedge P_{s-r}(k_1) - P_{s-r}(k_2) \\ >\ & \rho \end{aligned} \tag{2}$$

where k_1 denotes the ground truth classification label of input X and k_2 denotes the maximum candidate label. Under perturbation λ_i, P_{MLN} fluctuates less than ρ, and the probability P_{s-r} of classifying X as k_1 for the sensing-reasoning model is greater than that of classifying it as k_2 for ρ. Then ε can be proved to be the robust radius of the model in Equation (3).

$$\begin{aligned} & \|\lambda_i\|_\infty \leq \varepsilon \\ \Rightarrow\ & |P_{e^i}(X) - P_{e^i}(X+\lambda_i)| \leq \upsilon \\ \Rightarrow\ & P_{MLN}\left(\{r_i \Leftarrow P_{e^i}(X)\}_{i\in n}\right) - P_{MLN}\left(\{r_i \Leftarrow P_{e^i}(X+\lambda_i)\}_{i\in n}\right) \leq \rho \end{aligned} \tag{3}$$

We propagate the robustness bounds as in Equation (3) according to the division of sensing-reasoning components to obtain bounds υ on the output of the Lipschitz network in the sensing phase.

We illustrate how Lipschitz networks to improve the efficiency of robustness certification and support for l_∞ parametric perturbations.

Axiom 3.1. For a classifier $L(X)$ defined by a Lipschitz neural network, it is shown to be provably robust under perturbation $\|\lambda_i\|_\infty < \frac{\psi}{Q} \cdot |P_{e^i}(k_1) - P_{e^i}(k_2)|$. This Lipschitz axiom can be formalized as Equation (4):

$$L_i(X) = L_i\left(X+\lambda_i\right) \text{ for all } \lambda_i \text{ with } \|\lambda_i\|_\infty < \frac{\psi}{Q} \cdot |P_{e^i}(k_1) - P_{e^i}(k_2)| \tag{4}$$

where ψ is $\sqrt[2]{2}/2$ in perturbation in l_2 parameter space and is $1/2$ in l_∞, and Q denotes the Lipschitz constraint. $P_{e^i}(k_1) - P_{e^i}(k_2)$ is the margin between the ground truth label probability output by the perceptual component and the candidate label with the second largest probability. The unknowns to be computed in the robust radius are margin of $P_{e^i}(\cdot)$, compared to random smoothing for training of perturbed samples, Lipschitz networks only need to train the original dataset, and the robust radius can be easily derived instead of the tedious Gaussian CDF. In addition, previous work [14] has shown that for l_∞, Lipschitz networks are internally computed equivalent to bounded l_∞ of the constraint power matrix, add with the Lipschitz activation function.

We take $\frac{\psi}{Q} \cdot |P_{e^i}(k_1) - P_{e^i}(k_2)|$ as ε and compute output bounds for P_{MLN} using Lagrangian method, as Equations (5) and (6).

$$\begin{aligned} & \underset{\{\|\lambda_i\|_\infty<\varepsilon\}}{Max} \ln P_{MLN}\left(\{r_i \Leftarrow P_{e^i}(X+\lambda_i)\}_{i\in n}\right) \\ \leq\ & \underset{\{\|\lambda_i\|_\infty<\varepsilon\}}{Max} \ln P_1\left(\{\lambda_i\}_{i\in n}\right) + \sum_i \eta_i\lambda_i - \underset{\{\|\lambda_i\|_\infty<\varepsilon\}}{Min} \ln P_2\left(\{\lambda_i'\}_{i\in n}\right) - \sum_i \eta_i'\lambda_i' \end{aligned} \tag{5}$$

$$\begin{aligned} &\underset{\{\|\lambda_i\|_\infty<\varepsilon\}}{Min} \ln P_{MLN}\left(\{r_i \Leftarrow P_{e^i}(X+\lambda_i)\}_{i\in n}\right) \\ \geq &\underset{\{\|\lambda_i\|_\infty<\varepsilon\}}{Min} \ln P_1\left(\{\lambda_i\}_{i\in n}\right) + \sum_i \eta_i\lambda_i - \underset{\{\|\lambda_i\|_\infty<\varepsilon\}}{Max} \ln P_2\left(\{\lambda_i'\}_{i\in n}\right) - \sum_i \eta_i'\lambda_i' \end{aligned} \quad (6)$$

where P_1 and P_2 are the joint probability and marginal probability of the MLN interface variable, respectively. The maximum P_{MLN} is constantly smaller than the ratio of the minimum joint probability to the maximum marginal probability, and the opposite is true for the smallest value of P_{MLN}. η_i is the Lagrangian coefficient. We can solve it using the method from [15].

4. **Experiments.**

4.1. **Setup.**

We validated our model on MNIST and CIFAR-10 two standard image datasets to perform classification tasks using 2 RTX4000 GPUs. Note that our experiments do not involve large-scale datasets like ImageNet, since our contribution does not lie in classification performance and scalability (previous work [14] has excellently validated performance on multiple large-scale datasets), but rather in the optimization of the efficiency of robustness validation. Therefore, large-scale datasets are not necessary.

4.2. **Efficiency of robustness certification.**

We test the training time and certification time of the model on MNIST dataset and CIFAR-10 dataset using l_2 and l_∞ perturbation, respectively. Table 1 and Table 2 report the experimental results of CNN and Lipschitz neural network as sensing components after applying l_2 and l_∞ perturbations on the MNIST dataset, respectively. Table 3 and Table 4 report the experimental results after applying l_2 and l_∞ perturbations on the CIFAR-10 dataset, respectively.

Under l_2 perturbation, our model achieves comparable classification performance and certification accuracy to CNN. On the CIFAR-10 dataset, our model also did not lag behind other methods. However, our model significantly outperforms the CNN-based one in terms of training and validation time, thanks to the fact that Lipschitz neural network validation does not require propagation of output bounds as in IBP, and does not require adding perturbation samples to the training set, which saves 28.9% of training time, and 36.2 of validation time in CIFAR-10, and 28.9% of training time, and 78.2% of validation

Table 1. Comparison of results on the l_2 perturbed MNIST dataset

Model	Method	Train times		$\lambda_i = 0.1$			$\lambda_i = 0.25$		
		Train	Certify	Clean	PGD	Certify	Clean	PGD	Certify
CNN	IBP	21.5	4.1	94.28	84.2	91.3	93.2	94.3	91.4
	CAP	18.7	–	91.67	83.1	97.4	89.0	94.2	93.1
	CROWN-IBP	27.3	8.2	91.75	90.2	97.9	89.2	96.5	94.2
	Random smoothing	47.9	23.7	93.66	91.3	97.4	–	96.2	90.0
Lipschitz	Radius solution	**10.4**	**3.1**	93.2	**93.2**	94.1	92.5	95.1	**98.2**

Table 2. Comparison of results on the l_∞ perturbed MNIST dataset

Model	Method	Train times		$\lambda_i = 0.1$			$\lambda_i = 0.25$		
		Train	Certify	Clean	PGD	Certify	Clean	PGD	Certify
CNN	IBP	20.5	6.1	80.20	86.2	89.1	83.2	84.3	81.0
	CAP	20.1	–	85.9	83.9	91.4	89.0	84.2	83.8
	CROWN-IBP	32.5	7.9	88.4	85.9	88.9	89.2	86.5	84.9
	Random smoothing	41.0	21.4	80.7	79.0	87.4	–	86.2	88.2
Lipschitz	Radius solution	**9.2**	**3.1**	**93.2**	**93.2**	**94.1**	**92.4**	**92.1**	**93.9**

TABLE 3. Comparison of results on the l_2 perturbed CIFAR-10 dataset

Model	Method	Train times		$\lambda_i = 2/255$			$\lambda_i = 8/255$		
		Train	Certify	Clean	PGD	Certify	Clean	PGD	Certify
CNN	IBP	91.5	101.1	62.7	54.2	41.3	53.7	34.3	21.5
	CAP	510.7	7021.9	64.6	63.8	57.7	43.8	34.9	33.7
	CROWN-IBP	97.3	48.2	61.5	65.2	42.3	42.9	36.8	26.2
	Random smoothing	107.9	23.7	63.66	52.3	55.4	–	33.2	30.2
Lipschitz	Radius solution	**76.4**	**15.1**	63.2	53.2	**60.2**	52.9	**45.8**	**37.8**

TABLE 4. Comparison of results on the l_∞ perturbed CIFAR-10 dataset

Model	Method	Train times		$\lambda_i = 2/255$			$\lambda_i = 8/255$		
		Train	Certify	Clean	PGD	Certify	Clean	PGD	Certify
CNN	IBP	97.5	101.6	24.68	24.2	21.3	23.6	23.6	21.4
	CAP	525.1	8122.5	26.32	33.1	26.5	29.2	24.3	16.2
	CROWN-IBP	95.4	43.2	23.95	30.2	27.2	26.1	25.1	17.2
	Random smoothing	101.6	27.8	23.76	21.3	35.6	–	26.7	20.4
Lipschitz	Radius solution	**76.7**	**15.1**	**53.7**	**53.2**	**54.1**	**42.5**	**45.1**	**43.2**

time in MNIST. Under l_∞ perturbation, with perturbation radius of 0.1 and 2/255, our model achieves a validation accuracy of 94.1 and 54.1 on the two datasets, respectively. Perturbation radius of 0.25 and 8/255 achieved 93.9 and 43.2 certification accuracies.

4.3. **Defensive capability of Lipschitz network-based under adversarial attacks.** We trained 8 Lipschitz neural networks as perceptual components on the CIFAR-10 dataset, extracting 8 features e_i on image categories, respectively, and poisoning a certain size of training samples against attacks on 1 to 4 sensing components. We use the difference between the predicted probability minimum of the ground truth label and the probability maximum of the second candidate label as a metric to measure the defense capability of the model. To make the metric constant > 0, we add 1 to the difference.

Figure 2 reports the difference in the probability of our model with increasing number of attacked sensing components under l_2 perturbation, which is comparable to that of CNN, and even more stable for Lipschitz prediction when the number of attacked components

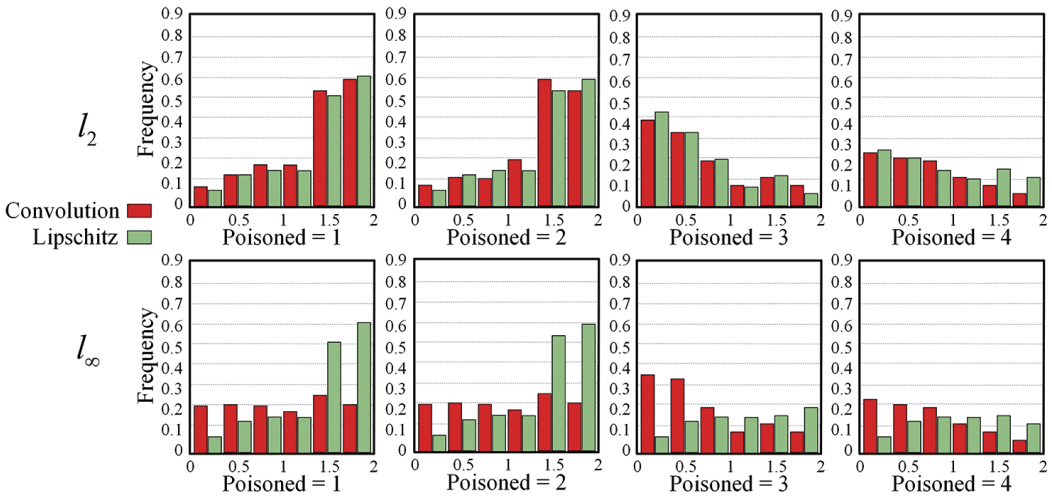

FIGURE 2. The defensive capabilities of CNN and Lipschitz at different poisoned levels

is 4. Although all models suffer from a decrease in stability due to an increase in the number of attacks, our model still achieves superior probability differences, while the CNN is affected by the perturbation, with the second candidate label over the true label becoming incorrectly predicted as often as 0.35 when the number of attacks is greater than 2.

5. **Conclusions.** In this paper, we propose a novel sensing-reasoning model based on the Lipschitz property, which builds on the original framework by (1) introducing a Lipschitz neural network as the perceptual component, bypassing the random smoothing-based CNN Gaussian noise training, and optimizing the computational complexity of robustness verification of the perceptual pipeline with better perturbation robustness at l_∞ perturbation, and (2) we provide robustness verification of the pipeline as a whole for the sensing-reasoning model. In the future, we will further investigate the optimization of the robustness certification cost for the relaxed utilization of Lipschitz neural network.

Acknowledgment. This work is partially supported by Open Fund of State Key Laboratory of Power Grid Safety and Energy Conservation (China Electric Power Research Institute) (No. DZB51202101256). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] N. Carlini and D. Wagner, Towards evaluating the robustness of neural networks, *2017 IEEE Symposium on Security and Privacy (SP)*, pp.39-57, 2017.

[2] K. Eykholt, I. Evtimov, E. Fernandes et al., Robust physical-world attacks on deep learning visual classification, *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1625-1634, 2018.

[3] N. M. Gürel, X. Qi, L. Rimanic, C. Zhang and B. Li, Knowledge enhanced machine learning pipeline against diverse adversarial attacks, *International Conference on Machine Learning*, pp.3976-3987, 2021.

[4] Z. Xu, I. Gavran, Y. Ahmad, R. Majumdar, D. Neider, U. Topcu and B. Wu, Joint inference of reward machines and policies for reinforcement learning, *Proc. of the International Conference on Automated Planning and Scheduling*, vol.30, pp.590-598, 2020.

[5] M. Richardson and P. Domingos, Markov logic networks, *Machine Learning*, vol.62, pp.107-136, 2006.

[6] Z. Yang, Z. Zhao, B. Wang et al., Improving certified robustness via statistical learning with logical reasoning, *Advances in Neural Information Processing Systems*, vol.35, pp.34859-34873, 2022.

[7] F. Farnia, J. M. Zhang and D. Tse, Generalizable adversarial training via spectral normalization, *arXiv Preprint*, arXiv: 1811.07457, 2018.

[8] H. Qian and M. N. Wegman, L2-nonexpansive neural networks, *arXiv Preprint*, arXiv: 1802.07896, 2018.

[9] C. Anil, J. Lucas and R. Grosse, Sorting out Lipschitz function approximation, *International Conference on Machine Learning*, pp.291-301, 2019.

[10] K. Leino, Z. Wang and M. Fredrikson, Globally-robust neural networks, *International Conference on Machine Learning*, pp.6212-6222, 2021.

[11] T. Huster, C.-Y. J. Chiang and R. Chadha, Limitations of the Lipschitz constant as a defense against adversarial examples, *Proc. of ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018*, Dublin, Ireland, pp.16-29, 2019.

[12] J. E. Cohen, T. Huster and R. Cohen, Universal Lipschitz approximation in bounded depth neural networks, *arXiv Preprint*, arXiv: 1904.04861, 2019.

[13] B. Zhang, T. Cai, Z. Lu, D. He and L. Wang, Towards certifying l_∞ robustness using neural networks with l_∞-dist neurons, *International Conference on Machine Learning*, pp.12368-12379, 2021.

[14] B. Zhang, D. Jiang, D. He and L. Wang, Rethinking Lipschitz neural networks and certified robustness: A Boolean function perspective, *Advances in Neural Information Processing Systems*, 2022.

[15] Z. Yang, Z. Zhao, B. Wang et al., Improving certified robustness via statistical learning with logical reasoning, *Advances in Neural Information Processing Systems*, vol.35, pp.34859-34873, 2022.

[16] Y. Zeng, Z. Shi, M. Jin, F. Kang, L. Lyu, C. J. Hsieh and R. Jia, Towards robustness certification against universal perturbations, *The 11th International Conference on Learning Representations*, 2023.

[17] H. Wu, T. Tagomori, A. Robey et al., Toward certified robustness against real-world distribution shifts, *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp.537-553, 2023.

[18] G. K. Nayak, R. Rawal and A. Chakraborty, DE-CROP: Data-efficient certified robustness for pre-trained classifiers, *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.4622-4631, 2023.

[19] K. Kakizaki, K. Fukuchi and J. Sakuma, Certified defense for content based image retrieval, *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.4561-4570, 2023.

[20] Z. Hammoudeh and D. Lowd, Reducing certified regression to certified classification for general poisoning attacks, *The 1st IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

[21] T. Rumezhak, F. G. Eiras, P. H. Torr and A. Bibi, RANCER: Non-axis aligned anisotropic certification with randomized smoothing, *Proc. of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.4672-4680, 2023.

[22] N. Levy, R. Yerushalmi and G. Katz, gRoMA: A tool for measuring deep neural networks global robustness, *arXiv Preprint*, arXiv: 2301.02288, 2023.

[23] M. Lechner, A. Amini, D. Rus and T. A. Henzinger, Revisiting the adversarial robustness-accuracy tradeoff in robot learning, *IEEE Robotics and Automation Letters*, vol.8, no.3, pp.1595-1602, 2023.

[24] L. Hsiung, Y. Y. Tsai, P. Y. Chen and T. Y. Ho, Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.24658-24667, 2023.