Project Proposal
Defense Against Gradient-Based Attacks via Randomized Smoothing

Instructor: Dr. M. H. Yassaee
Mentor: Dr. M. R. Rahmani
MohammadParsa Dini - std id: 400101204
Erfan Moeini - std id: 402212293

December 20, 2024

Abstract

Adversarial training has emerged as a crucial technique in enhancing the robustness of machine learning models against adversarial attacks. This proposal outlines a project aimed at exploring two primary ideas: (1) Proof of Certified Robustness for alternative distributions such as the Poisson Distribution, and (2) Finding a Better Bound with Conditions on the Classifier. This study builds on the foundation laid by recent literature, including works such as [5] and [6]. Our goal is to develop novel methods for improving model robustness and to establish theoretical guarantees for their effectiveness.

Introduction

Adversarial attacks pose significant challenges to the robustness of machine learning models. These attacks exploit vulnerabilities in the model by adding imperceptible perturbations to the input data, thereby causing the model to make incorrect predictions. Adversarial training is a defensive strategy designed to improve the robustness of models against such attacks. The objective of this project is to investigate new directions in adversarial training by focusing on certified robustness for alternative data distributions and improving robustness bounds under specific classifier conditions.

1 Background, Related Work, and Motivation

Several key studies have contributed to the understanding and development of adversarial training techniques. Mao et al. [1] explored the connection between certified and adversarial training, highlighting methods for enhancing model robustness. Tramèr et al. [2] examined the transferability of adversarial examples, while Shafahi et al. [3] discussed the inevitability of adversarial examples. FGSM (Fast Gradient Sign Method) and PGD (Projected Gradient Descent) are prominent techniques used in adversarial attacks, where FGSM

is a white-box method that assumes access to the gradient of the loss function. Another approach involves adding noise to the gradients, thereby increasing the likelihood of model errors.

In these methods, the goal is to solve the optimization problem:

$$\max_{\delta} \mathcal{L}(f_\theta(x+\delta), y)$$

where $\mathcal{L}$ is the loss function, f_θ represents the model with parameters θ, x is the input data, y is the true label, and δ is the adversarial perturbation subject to a constraint, such as $\|\delta\| \leq \epsilon$.

Our project is motivated by the need to establish stronger theoretical guarantees for model robustness and to explore the efficacy of adversarial training under different distributional assumptions. Specifically, we aim to extend the concept of certified robustness to the Poisson distribution and to find better robustness bounds with additional classifier conditions.

2 Project Ideas

2.1 Proof of Certified Robustness for Poisson Distribution

The first idea is to provide a theoretical proof of certified robustness for models trained with data following a Poisson distribution. This involves extending the framework used for Gaussian distributions to accommodate the Poisson distribution, which is common in various real-world applications. We will derive certified robustness, within which adversarial perturbations do not affect model predictions, ensuring robustness.

Mathematically, let X be the input data following a Poisson distribution with parameter λ. The certified radius r is defined as the radius within which:

$$\forall \delta \text{ with } \|\delta\| \leq r, \quad f_\theta(X+\delta) = f_\theta(X)$$

2.2 Finding a Better Bound with Conditions on the Classifier

The goal is to derive tighter bounds on the robustness radius r_p^* by taking into account specific properties of the classifier being smoothed. Current bounds depend only on the smoothing distribution and the dimensionality d, without leveraging the behavior of the classifier [6].

$$r_p^* \leq \frac{\sigma}{2\sqrt{2} d^{\frac{1}{2}-\frac{1}{p}}} \left(\frac{1}{\sqrt{1-p_1(x)}} + \frac{1}{\sqrt{p_2(x)}} \right)$$

where $p_1(x)$ and $p_2(x)$ are the probabilities of the first and second most probable labels.

Expected Outcomes

Through this project, we expect to achieve the following outcomes:

- A theoretical framework for certified robustness applicable to Poisson-distributed data.
- Classifier-dependent robustness bounds that improve certifiable robustness for randomized smoothing techniques.

References

[1] Yuhao Mao, Mark Niklas Müller, Marc Fischer, and Martin Vechev. *Connecting Certified and Adversarial Training*. Available at: arXiv:2305.04574.

[2] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. *The Space of Transferable Adversarial Examples*. Available at: arXiv:1704.03453.

[3] Ali Shafahi, W. Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. *Are adversarial examples inevitable?*. In Proceedings of the 7th International Conference on Learning Representations (ICLR 2019). Available at: arXiv:1809.02104.

[4] S. M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard. *Universal adversarial perturbations*. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). Available at: arXiv:1704.08945.

[5] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. *Certified Adversarial Robustness via Randomized Smoothing*. Available at: arXiv:2305.04574.

[6] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. *The Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness*. Available at: arXiv:2002.03239.