Stealthy Poisoning Attack on Certified Robustness

Akshay Mehra*, Bhavya Kailkhura†, Pin-Yu Chen‡ and Jihun Hamm*

*Tulane University, †Lawrence Livermore National Laboratory, ‡IBM Research

Motivation

- Certified robustness has emerged as the gold standard to gauge with certainty the susceptibility of machine learning models to test-time attacks.
- Robust training methods are required to train models with high certified robustness since models trained with standard training are not robust.
- Poisoning attacks that optimize the poison data to hurt the accuracy of models trained with standard training fail against robust training methods.

Contributions

- We propose a bilevel optimization based **data poisoning attack** that **degrades** the **robustness guarantees** of **certifiably robust models**.
- Significant **reduction** in the certifiable robustness of models trained with **robust training methods**, highlights the importance of training-data quality in achieving high robustness guarantees.
- **Imperceptibly** distorted poisoned data with **clean-labels** make our attack difficult to detect.

Poisoning attack on certified robustness

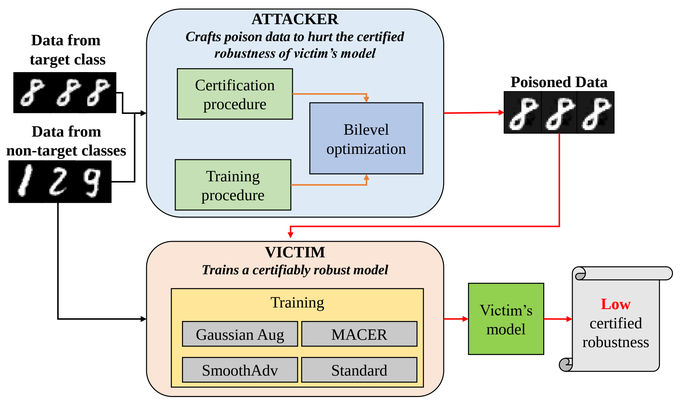
Attacker's Goal: Find poisoning points (u) such that when the victim trains a model (with parameters θ) on the poisoned data ($\mathcal{D}^{\text{clean}} \bigcup \mathcal{D}^{\text{poison}}$), the certified robustness guarantees of the target class are significantly diminished.

$$\min_{u \in \mathcal{U}} \ \mathcal{R}(\mathcal{D}^{\text{val}}; \theta^*)$$
$$\text{s.t.} \ \ \theta^* = \arg\min_{\theta} \ \mathcal{L}_{\text{robust}}(\mathcal{D}^{\text{clean}} \bigcup \mathcal{D}^{\text{poison}}; \theta),$$

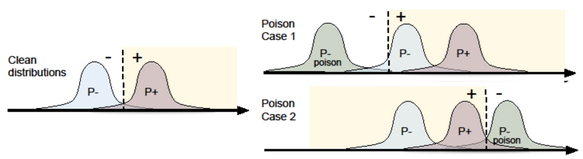
where $\mathcal{D}^{\text{clean}}$ is the clean data, $\mathcal{D}^{\text{val}}$is the validation data of the target class, ϵ is maximum permissible perturbation to the features of the base data $\{(x_i^{\text{base}}, y_i^{\text{base}})\}_{i=1}^{N_{\text{poison}}}$ used to initialize poison data, $u = \{u_1, ..., u_n\}$ are the features of the poison data $\mathcal{D}^{\text{poison}} = \{(u_i, y_i^{\text{base}})\}_{i=1}^{N_{\text{poison}}}$, $\mathcal{R}$ is the measure of certified robustness and $\mathcal{L}_{\text{robust}}$ is the loss function of a robust training method.

Imperceptibly distorted poison data: Attack points generated against certifiably robust model trained with Gaussian data augmentation [1] to reduce the average certified radius obtained using randomized smoothing for digit 8 of MNIST and class "ship" of CIFAR10 .

Overview of attack generation and evaluation



Effect of poisoning on the decision boundary

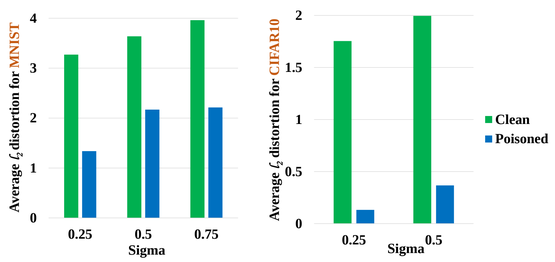


For **linear classifiers**, the certified radius of a point is it's distance to the decision boundary. Assuming P^- is the distribution of the negative class, x_i^+ and x_i^- are training points of the positive and the negative class, the analytical solution to the poisoning problem (Eq. 1) is to shift all x_i^- either towards left or right by ϵ.

$$\min_{u} \ \ \mathbb{E}_{P_-}[\max(\text{sign}(w^*)(-b^*/w^* - x), 0)]$$
$$w^*, b^* = \arg\min_{w,b} \frac{1}{2n}\Big[\sum_{i=1}^{n} l(x_i^+, 1) + \sum_{i=1}^{n} l(u_i, -1)\Big]. \tag{1}$$

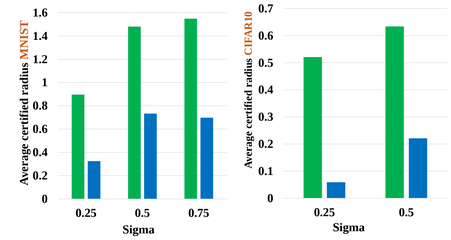
Theorem 1. *If the perturbation is large enough, i.e., $\epsilon \geq \frac{\sum_i x_i^+ - \sum_i x_i^-}{n}$ then there are two locally optimal solutions to (1) which are $u_i = x_i^- - \epsilon$ (Case 1) and $u_i = x_i^- + \epsilon$ (Case 2) for $i = 1, ..., n$. Otherwise, there is a unique globally optimal solution which is $u_i = x_i^- - \epsilon$ (Case 1) for $i = 1, ..., n$.*

For **non-linear classifiers** the average ℓ_2 distortion to generate adversarial examples for the smooth classifier is reduced. This shows that the decision boundary is closer to the original test distribution similar to the case of linear classifiers.

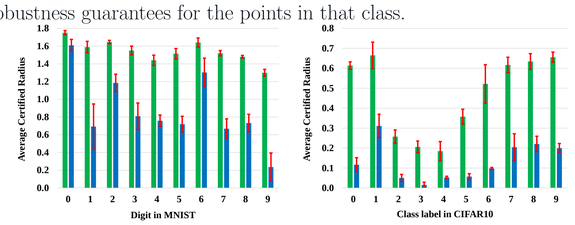


Our key results

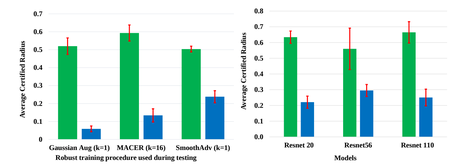
- Our attack **reduces** the **average certified radius** and **approximate certified accuracy** obtained using randomized smoothing, of the target class.



- The attacker can **target any class** in the dataset and reduce the certified robustness guarantees for the points in that class.



- Attack points are **transferable** to models trained with **state-of-the-art robust training methods** and to models with **different architectures**.



Clean data **Poisoned data**

Main references

[1] Cohen, Rosenfeld, and Kolter. "Certified adversarial robustness via randomized smoothing". In: *arXiv preprint arXiv:1902.02918* (2019).

[2] Mehra and Hamm. "Penalty Method for Inversion-Free Deep Bilevel Optimization". In: *arXiv preprint arXiv:1911.03432* (2019).

[3] Mehra, Kailkhura, Chen, and Hamm. "How Robust are Randomized Smoothing based Defenses to Data Poisoning?" In: *arXiv preprint arXiv:2012.01274* (2020).

[4] Mei and Zhu. "Using Machine Teaching to Identify Optimal Training-Set Attacks on Machine Learners." In: *AAAI*. 2015, pp. 2871–2877.