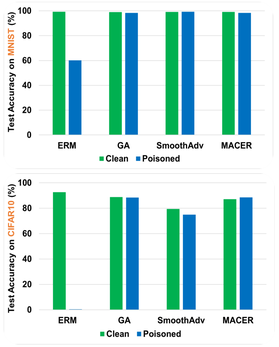
How Robust are Randomized Smoothing based Defenses to Data Poisoning?

Akshay Mehra[1], Bhavya Kailkhura[2], Pin-Yu Chen[3] and Jihun Hamm[1]
[1]Tulane University, [2]Lawrence Livermore National Laboratory,[3]IBM Research

Motivation

- Recent works have shown failure of models trained with ERM in achieving high certified robustness and have proposed new training methods.
- However, the role of the data distribution in achieving these high certified robustness guarantees has been ignored.
- Studying performance in adversarial setting such as in presence of data poisoning can be beneficial. But current poisoning attacks fail when certified defenses are used.

Can poisoning degrade certified adversarial robustness guarantees?



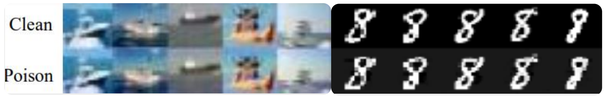
Poisons optimized against ERM fail when certified defenses are used.

Contributions

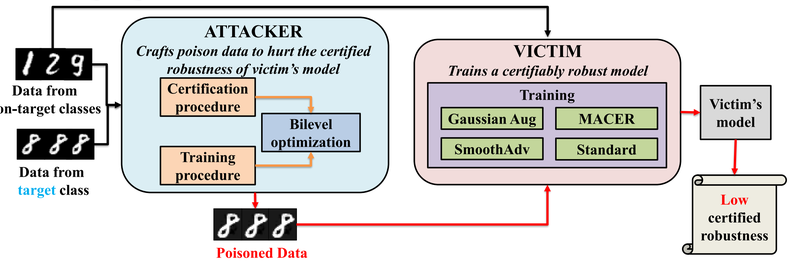
- We study the problem of using data poisoning attacks to affect the robustness guarantees of classifiers trained using certified defense methods.
- We propose a bilevel optimization based clean label poisoning attack to generate poison data against robust training and certification methods.
- We demonstrate the effectiveness of our attack at reducing certified adversarial robustness obtained using randomized smoothing on models trained with state-of-the-art certified defenses.

Poison data that hurts certified adversarial robustness

Presence of imperceptibly distorted poison data can significantly hurt certified adversarial robustness guarantees.



Poisoning Against Certified Defenses (PACD)



Generate poison data (u) such that when the victim trains a model (with parameters θ) on the poisoned data ($\mathcal{D}^{\text{clean}} \cup \mathcal{D}^{\text{poison}}$) the certified robustness guarantees of the target class are significantly diminished on a validation set $\mathcal{D}^{val}$.

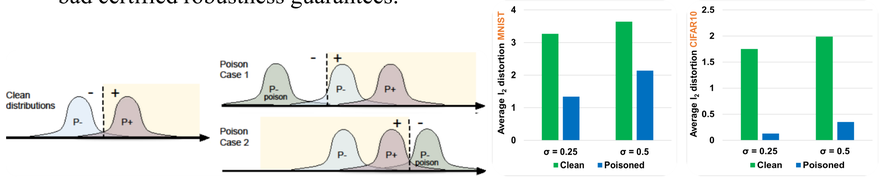
$$\min_{u \in \mathcal{U}} \mathcal{R}(\mathcal{D}^{val}; \theta^*)$$
$$s.t.\ \ \theta^* = \arg\min_{\theta} \mathcal{L}_{robust}\left(\mathcal{D}^{clean} \cup \mathcal{D}^{poison}; \theta\right)$$

Lower-level problem trains a model using training procedures that lead to models with high certified robustness such as MACER, SmoothAdv etc.

Upper-level problem lowers average certified radius obtained from a certification procedure such as Randomized Smoothing

Effect of poisoning on the decision boundary

Average margin of the smoothed classifiers on the target data is reduced leading to bad certified robustness guarantees.

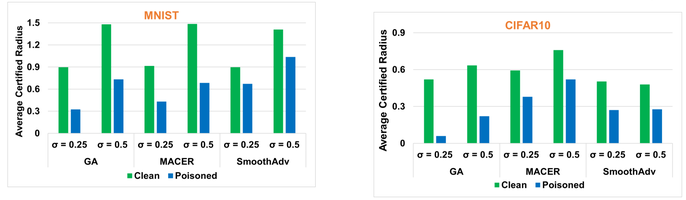


Linear Classifiers

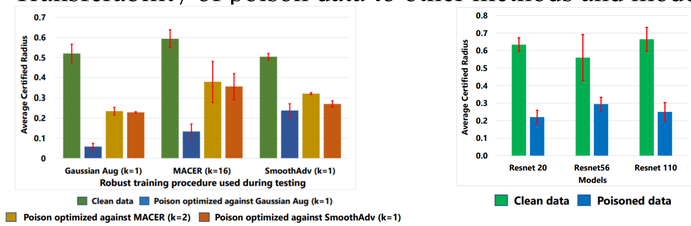
Non-Linear Classifiers

Key Results

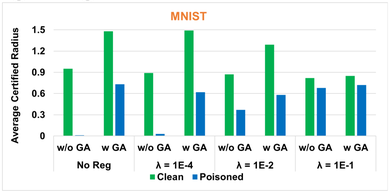
- Reduction in Average Certified Radius with poisoning



- Transferability of poison data to other methods and models



- Failure of weight regularization-based defense



References

- Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. "Certified adversarial robustness via randomized smoothing." *International Conference on Machine Learning*. PMLR, 2019.
- Zhai, Runtian, et al. "Macer: Attack-free and scalable robust training via maximizing certified radius." *arXiv preprint arXiv:2001.02378* (2020).
- Salman, Hadi, et al. "Provably robust deep learning via adversarially trained smoothed classifiers." *arXiv preprint arXiv:1906.04584* (2019).
- Pedregosa, Fabian. "Hyperparameter optimization with approximate gradient." *International conference on machine learning*. PMLR, 2016.
- Mei, Shike, and Xiaojin Zhu. "Using machine teaching to identify optimal training-set attacks on machine learners." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. No. 1. 2015.

This work was supported by the NSF EPSCoR-Louisiana Materials Design Alliance (LAMDA) program #OIA1946231 and by LLNL Laboratory Directed Research and Development project 20-ER-014 (LLNL-CONF-817233). This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC