Harnessing Robust Statistics for Trustworthy AI

Xiaorui Liu

Department of Computer Science
North Carolina State University
xliu96@ncsu.edu

Machine learning (ML) techniques are notably vulnerable to natural or adversarial perturbations, which can cause significant economic, ethical, and societal risks. In this New Faculty Highlight Talk, I will showcase my research on harnessing robust statistics to build robust and trustworthy AI systems. In particular, I will highlight my research breakthroughs in the areas of *graph learning* (GNNs), *large language models* (LLMs), *deep equilibrium models* (DEQs), and *deep representation learning*. These breakthroughs stem from a **unified and principled robust statistics framework** that incorporates robustness as the core inductive bias in deep learning architecture. This approach has enabled significant improvements in intrinsic robustness and generalization, even in challenging environments.

Graph Learning. We demonstrated that robust GNNs can be designed from robust non-parametric regression and signal denoising perspectives. These models significantly improve robustness, as shown in my works such as ElasticGNN (Liu et al. 2021b), AirGNN (Liu et al. 2021a), GTN (Fan et al. 2022), and RUNG (Hou et al. 2024a). This line of research has inspired numerous research aimed at strengthening GNN robustness and safety in adversarial, noisy, or missing data environments.

Large Language Models. Our work ProTransformer (Hou et al. 2024b), established a novel connection between attention mechanisms and weighted least square estimators. Leveraging this insight, we developed a set of novel robust attention mechanisms that enhances the resilience based on robust statistics, which significantly enhances the robustness of Transformers across a variety of prediction tasks, attack mechanisms, backbone architectures, and data domains.

Deep Equilibrium Models. Our paper (Gao et al. 2024) introduced the first randomized smoothing certified defense for DEQs to address efficiency limitations. Our proposed Serialized Randomized Smoothing approach accelerates the certification by up to 7 times. It not only provides an efficient robustness certification for DEQs but also opens a promising direction for certifying the robustness of any models developed within our robust statistics framework.

Representation Learning. My recent work (Hou et al. 2024c) tackles an intriguing and fundamental open challenge in representation learning: Given a well-trained deep learning model, can it be reprogrammed to enhance its robustness against adversarial or noisy input perturbations without altering its parameters? We introduce three model reprogramming paradigms to offer flexible control of robustness under different efficiency requirements.

These research breakthroughs demonstrated the transformative potential of harnessing robust statistics in enhancing the robustness and trustworthiness of deep learning. We will continue advancing this area by advocating the design of **robustness-informed neural networks** across various areas.

References

Fan, W.; Liu, X.; Jin, W.; Zhao, X.; Tang, J.; and Li, Q. 2022. Graph trend filtering networks for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 112–121.

Gao, W.; Hou, Z.; Xu, H.; and Liu, X. 2024. Certified Robustness for Deep Equilibrium Models via Serialized Random Smoothing. *Advances in Neural Information Processing Systems*.

Hou, Z.; Feng, R.; Derr, T.; and Liu, X. 2024a. Robust Graph Neural Networks via Unbiased Aggregation. *Advances in Neural Information Processing Systems*.

Hou, Z.; Gao, W.; Shen, Y.; Wang, F.; and Liu, X. 2024b. ProTransformer: Robustify Transformers via Plug-and-Play Paradigm. *Advances in Neural Information Processing Systems*.

Hou, Z.; Torkamani, M.; Krim, H.; and Liu, X. 2024c. Robustness Reprogramming for Representation Learning. arXiv:2410.04577.

Liu, X.; Ding, J.; Jin, W.; Xu, H.; Ma, Y.; Liu, Z.; and Tang, J. 2021a. Graph neural networks with adaptive residual. *Advances in Neural Information Processing Systems*, 34: 9720–9733.

Liu, X.; Jin, W.; Ma, Y.; Li, Y.; Liu, H.; Wang, Y.; Yan, M.; and Tang, J. 2021b. Elastic graph neural networks. In *International Conference on Machine Learning*, 6837–6849. PMLR.