Research Statement

Xuanqing Liu
Computer Science Department, University of California Los Angeles

My primary research interests are the security of popular machine learning (especially deep learning) algorithms, adversarial neural networks, and optimization algorithms. The security concern arises when models are deployed for sensitive tasks, such as image or video recognition/captioning, natural language understanding, and voice synthesis. We noticed that previous researches mostly concentrate on how to improve the performance of different kinds of models but less on noticing the hidden threats -- the models themselves turn out to be very fragile to external changes [1], even for the state-of-the-art models. Therefore, it is becoming more and more critical to analyze the robustness of modern machine learning models, and further to seek some effective remedies inspired by theoretical analysis.

My past research works are a combination of two directions: in the first direction, we would like to see the stability of the classic models, either to the change of testing input (adversarial attack) or to the training input (data poisoning attack). Our preliminary findings are both are not as stable as people imagined. In particular, we found that graph-based semi-supervised learning is not trustable in the sense that an attacker can smash our models with a small price [2]. On the other hand, we found that the k-nearest model exhibits stronger stability in testing input compared with deep neural networks [3]. Another direction is to improve the robustness of existing models. Among many methods, I focused on the randomization-based approach; previous works include random input [4], random weights [5], and Neural SDE [6]. All of them are shown to be useful to strong adversarial attacks to some degree.

As to the ongoing and future works, I plan to keep seeking efficient methods to fortify current models, especially the ones that are production-ready in various areas. And beyond that, I would like to explore the connections between model robustness and model interpretability.

References

[1] Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
[2] Liu, Xuanqing, et al. “A Unified Framework for Data Poisoning Attack to Graph-based Semi-supervised Learning”, To appear at NeurIPS 2019.
[3] Wang, Lu, et al. "Evaluating the Robustness of Nearest Neighbor Classifiers: A Primal-Dual Perspective." arXiv preprint arXiv:1906.03972 (2019).

[4] Liu, Xuanqing, et al. "Towards robust neural networks via random self-ensemble." Proceedings of the European Conference on Computer Vision (ECCV). 2018.
[5] Liu, Xuanqing, et al. "Adv-bnn: Improved adversarial defense through robust bayesian neural network." arXiv preprint arXiv:1810.01279 (2018).
[6] Liu, Xuanqing, et al. "Neural SDE: Stabilizing Neural ODE Networks with Stochastic Noise." arXiv preprint arXiv:1906.02355 (2019).