AI 活用で挑む学問の革新と創成
2021 年度採択研究者

2021 年度
年次報告書

Zhang Jingfeng

Imperfect information learning team RIKEN AIP
Postdoctoral researcher

Discouraging adversarial attacks through improving the adversarial training

§１. 研究成果の概要

Deep neural networks (DNNs) are vulnerable to human-imperceptibly adversarial noise, bringing security concerns to high-stake applications. It is urgent and critical to obtain the adversarial robustness against the adversarial noises. To enhance DNNs' adversarial robustness, I leverage the adversarial distillation that deals with the interactions between student and teacher models [1] and the collaboration scheme that deal with multiple sub-models [3]. Besides, I also understand the interactions between noisy labels and adversarial robustness [4] and explore how to leverage noisy labels to enhance adversarial robustness further [2].

【代表的な原著論文情報】

1 Zhu, Jianing, Jiangchao Yao, Bo Han, **Jingfeng Zhang**, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, and Hongxia Yang. Reliable Adversarial Distillation with Unreliable Teachers", International Conference of Learning Representation (ICLR), April, 2022

2 **Jingfeng Zhang***, Xilie Xu*, Bo Han, Tongliang Liu, Gang Niu, Lizhen Cui, Masashi Sugiyama. NoiLin: Improving adversarial training and correcting stereotype of noisy labels, Transactions on Machine Learning Research (TMLR), June, 2022

3 Sen Cui*, **Jingfeng Zhang***, Jian Liang, Bo Han, Masashi Sugiyama, Changshui Zhang. Synergy-of-Experts: Collaborate to Improve Adversarial Robustness, 2022

4 Jianing Zhu*, **Jingfeng Zhang***, Bo Han, Tongliang Liu, Gang Niu, Hongxia Yang, Mohan Kankanhalli, Masashi Sugiyama. Understanding the Interaction of Adversarial Training with Noisy Labels, 2022