AI 活用で挑む学問の革新と創成　　2022 年度
2021 年度採択研究代表者　　年次報告書

Zhang Jingfeng

理化学研究所 革新知能統合研究センター
博士研究員

Discouraging adversarial attacks through improving the adversarial training

研究成果の概要

Deep neural networks (DNNs) in Artificial Intelligence are susceptible to human-imperceptibly adversarial noise, which raises security concerns for high-stake applications. Since many real-world applications are equipped with DNNs, it is critical to achieve adversarial robustness against these vulnerabilities. I improve the adversarial robustness of AI-powered methods by utilizing adversarial attacks to evaluate machine learning based methods such as deep image denoising and non-parametric two sample tests. Then, I utilize adversarial training methods to enhance their robustness while maintaining their superior performance. My research on this topic was accepted at top machine learning conferences IJCAI2022 and ICML2022.

【代表的な原著論文情報】

1) Towards Adversarially Robust Image Denoising.
H. Yan, **J. Zhang**, J. Feng, M. Sugiyama, and V. Y. F. Tan.
The 31st International Joint Conference on Artificial Intelligence (IJCAI 2022)
2) Adversarial Attacks and Defense For Non-parametric Two Sample Tests.
X. Xu*, **J. Zhang***, F. Liu, M. Sugiyama, and M. Kankanhalli.
The 39th International Conference on Machine Learning (ICML 2022)