公開

研 究 終 了 報 告 書

「Discouraging adversarial attacks through improving the adversarial training」

研究期間： 2021 年 10 月～2023 年 6 月
研 究 者： Zhang Jingfeng

1． 研究のねらい

Although deep neural networks (DNNs) have been widely deployed, they are suspectable to adversarial attacks. Attackers can add human-imperceptible noise to the natural data, which leads the DNNs to make the wrong predictions. The adversarial attack has posed a huge threat to the reliable deployment of DNNs in many security-related fields such as medicine, finance, face recognition, autonomous driving, etc. Therefore, it is urgently needed to develop adversarially robust DNNs so that earn people's trust.

Fortunately, adversarial training (AT) is the most effective training method for obtaining the adversarial robustness of DNNs against those crafted adversarial attacks. However, the AT methods has many drawbacks such as limited improvements of adversarial robustness, issues of robust overfittings, degradation of natural accuracies, etc.

Thus, this research proposal aims to develop more effective AT methods.

2． 研究成果

（1）概要

This research has narrowed the gap between realistic security and the research for adversarial robustness. In particular, from the perspective of theory and algorithms, we have developed several effective AT strategies such as injecting dynamic label noises, leveraging collaborations. Those advanced AT techniques can effectively enhance DNNs robustness. Furthermore, from the perspective of applications, we have applied the adversarial robustness into two case studies such as deep image denoiser and non-parametric two-sample tests, then we have developed the robust counterparts. The robust image denoisers and statistical test methods can have good performance and robustness against attacks at the same time. Those results are published in the top-tier machine learning conferences and journals such as ICML, NeurIPS, IJCAI, TMLR.

（2）詳細

NoiLin: Improving Adversarial Training and Correcting Stereotype of Noisy Labels -> Transactions of Machine Learning Research 2022

Purpose: Improve adversarial training (AT) for enhancing adversarial robustness.

Details: we propose our method "NoiLIn" that automatically adjusts NL injection into AT. In each training epoch, we first randomly flip a portion of labels of the training set, and then execute an AT method using the noisy-label set. As the training progresses, we increase the flipping portion if there occurs robustness degradation (evaluated on a validation set). Our

simple yet effective strategy is easily compatible with various AT methods, which does not hurt or even boost their adversarial robustness consistently over training epochs.

Synergy-of-Experts: Collaborate to Improve Adversarial Robustness. -> NeurIPS 2022

Purpose: Leverage the collaboration scheme to benefit the adversarial robustness

Details: Adversarially robust learning methods require invariant predictions to a small neighborhood of its natural inputs, thus often encountering insufficient model capacity. Learning multiple sub-models in an ensemble can mitigate this insufficiency, further improving both generalization and robustness. However, an ensemble still wastes the limited capacity of multiple models. To optimally utilizing the limited capacity, this paper proposes to learn a collaboration among multiple sub-models. Compared with the ensemble, the collaboration} enables the possibility of correct predictions even if there exists a single correct sub-model. Besides, learning a collaboration could enable every sub-model to fit its vulnerability area and reserve the rest of the sub-models to fit other vulnerability areas. To implement the idea, we propose a collaboration framework, namely Collaborate to Defend against Adversarial Attacks, which could effectively minimize the vulnerability overlap of all sub-models and then choose a representative sub-model to make correct predictions. Empirical experiments verify that our method outperforms various ensemble methods against black-box and white-box adversarial attacks.

Towards Adversarially Robust Deep Image Denoising -> IJCAI2022

Purpose: Investigate the robustness of deep image denoising models.

Details: This work systematically investigates the adversarial robustness of deep image denoisers (DIDs), i.e, how well DIDs can recover the ground truth from noisy observations degraded by adversarial perturbations. Firstly, to evaluate DIDs' robustness, we propose a novel adversarial attack, namely Observation-based Zero-mean Attack (ObsAtk), to craft adversarial zero-mean perturbations on given noisy images. We find that existing DIDs are vulnerable to the adversarial noise generated by ObsAtk. Secondly, to robustify DIDs, we propose an adversarial training strategy, hybrid adversarial training (HAT), that jointly trains DIDs with adversarial and non-adversarial noisy data to ensure that the reconstruction quality is high and the denoisers around non-adversarial data are locally smooth. The resultant DIDs can effectively remove various types of synthetic and adversarial noise. We also uncover that the robustness of DIDs benefits their generalization capability on unseen real-world noise. Indeed, HAT-trained DIDs can recover high-quality clean images from real-world noise even without training on real noisy data. Extensive experiments on benchmark datasets, including Set68, Fluorescence, and SIDD, corroborate the effectiveness of ObsAtk and HAT.

Adversarial Attacks and Defenses for Non-Parametric Two-Sample Tests -> ICML 2022

Purpose: Investigate the robustness of statistical test model, i.e., non-parametric two sample tests.

公開

Details: Non-parametric two-sample tests (TSTs) that judge whether two sets of samples are drawn from the same distribution are widely employed in scientific research such as physics, neurophysiology, and biology. TSTs becomes basic tools for many research areas. This work leverage the adversarial attacks to investigate the failure modes of TSTs, we propose the corresponding defense solutions.

GAT: Guided Adversarial Training with Pareto-optimal Auxiliary Tasks. -> ICML2023

Purpose: Investigate the robustness of foundation models that power various downstream applications.

Details: A foundation model is trained from the large-scale data, and serves as the "one-single backbone" for many AI applications (see picture below). To use the foundation model, we simply fine-tune it at our local device (e.g., personal computer). For example, the foundation can easily support our very own dialects. Then, everyone could have a "personalized" AI model for their voice. Another example is "personalized" gaming experience. In video game, the foundation model can power the AI agents. And, those AI agents can be trained locally. Therefore, every computer could have "different" AI agents, and our users could have "personalized" gaming experience in video game. Most importantly, this foundation model opens the Application Programming Interface (APIs) to allow the programmers to develop their own AI applications.

3. 今後の展開

In the future, I would like to push forward the direction of "Adversarial Robustness of Foundation Models".

I will investigate the robustness of foundation models that power various downstream applications.

4. 自己評価

In general, I am satisfied with the research achievements that are powered by this ACT-X program.

Ripple effects of research results: My developed adversarial robust methods can not only push forward the human's knowledge boundaries but also let ordinary people to trust AI methods.

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：3件

1. Hanshu Yan, **Jingfeng Zhang**, Gang Niu, Jiashi Feng, Vincent Tan, Masashi Sugiyama., CIFS: Improving Adversarial Robustness of CNNs via Channel-wise Importance-based Feature Selection, in International Joint Conferences on Artificial Intelligence (IJCAI) 2022
Abstract: This paper used adversarial attacks methods to evaluate the vulnerabilities of deep image denoisers and developed a robust deep image denoiser that can maintain good

公開

performance and also resist attacks.
2. Xilie Xu, **Jingfeng Zhang**, Feng Liu, Masashi Sugiyama, and Mohan Kankanhalli., Adversarial Attacks and Defense for Non-parametric Two Sample Tests, in International Conference of Machine Learning (ICML) 2022
Abstract: This paper investigated the robustness of statistical test model, i.e., non-parametric two sample tests, from the adversarial point of view, and provided a robust deep kernel methods that can reliably conduct two-sample tests.
3. Salah Ghamizi, **Jingfeng Zhang**, Maxime Cordy, Mike Papadakis, Masashi Sugiyama, Yves Le Traon, GAT: Guided Adversarial Training with Pareto-optimal Auxiliary Tasks, in International Conference of Machine Learning (ICML) 2023
Abstract: This paper investigated and used multi-tasks to enhance the adversarial robustness under the case of few training data. This technique can be used in medical domains where data have rich meaning but are in few amount.

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものは件数にのみ含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

RIKEN Ohbu award