Tiejin Chen

734-210-2624 tiejin@asu.edu

Education

Arizona State University, Tempe 2023-2028 (Expected)
PhD in Computer Science *Tempe,AZ*
- **Advisor:** Hua Wei
- **GPA:** 3.89/4.00

University of Michigan, Ann Arbor 2021-2023
MS in Applied Statistics *Ann Arbor,MI*
- **GPA:** 3.98/4.00

Sichuan University 2016-2020
Bachelor in Statistics *Chengdu,Sichuan*
- **GPA:** 3.55/4.00 | 86.59/100
- **Rank:** 4/33

Paper

Tiejin Chen, Longchao Da, Huixue Zhou, Pingzhi Li, Kaizhong Zhou, Tianlong Chen, Hua Wei. *Protecting Privacy against Membership Inference Attack with LLM Fine-tuning through Flatness.*. In SIAM International Conference on Data Mining (SDM'25) (short version presented at @SetLLM Workshop ICLR 2024.)
Tiejin Chen*, Kaishen Wang*, Hua Wei. *Zer0-Jack: A Memory-efficient Gradient-based Jailbreaking Method for Black-box Multi-modal Large Language Models*. In GenSafeAI Workshop @ Neurips 2024
Tiejin Chen*, Prithvi Parag Shirke*, Bharatesh Chakravarthi, Arpitsinh Vaghela, Longchao Da, Duo Lu, Yezhou Yang, Hua Wei. *SynTraC: A Synthetic Dataset for Traffic Signal Control from Traffic Monitoring Cameras*. In Proceedings of 27th IEEE International Conference on Intelligent Transportation Systems (ITSC'24).
Ningyi Xie, Jiahua Xu, **Tiejin Chen**, Yoshiyuki Saito, Nobuyoshi Asai, Dongsheng Cai *Performance Upper Bound of the Grover-Mixer Quantum Alternating Operator Ansatz*. Physical Review A
Zicheng Wang, **Tiejin Chen**, Qinrun Dai, Yueqi Chen, Hua Wei, Qingkai Zeng *When eBPF Meets Machine Learning: On-the-fly OS Kernel Compartmentalization*. BlackHat USA 2024 Briefing.
LongChao Da, Kuanru Liou, **Tiejin Chen**, Xuesong Zhou, Xiangyong Luo, Yezhou Yang, Hua Wei *Open-TI: Open Traffic Intelligence with Augmented Language Model*. International Journal of Machine Learning and Cybernetics (IF=5.6).
Kai Ye, **Tiejin Chen**, Hua Wei, Liang Zhan *Uncertainty Regularized Evidential Regression*. In Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI' 24).

Preprint Paper

Longchao Da, **Tiejin Chen**, Lu Cheng, Hua Wei *LLM Uncertainty Quantification through Directional Entailment Graph and Claim Level Response Augmentation*. arXiv preprint arXiv:2407.00994,2024.
Tiejin Chen, Wenwang Huang, Linsey Pang, Dongsheng Luo, Hua Wei *Are Classification Robustness and Explanation Robustness Really Strongly Correlated? An Analysis Through Input Loss Landscape*. arXiv preprint arXiv:2403.06013,2024.
Tiejin Chen*, Yuanpu Cao*, Yujia Wang*, Cho-Jui Hsieh, Jinghui Chen *Federated Learning with Projected Trajectory Regularization*. arXiv preprint arXiv:2312.14380,2023.
Tiejin Chen*, Yicheng Tao* *Learning sparsity and randomness for data-driven low rank approximation*. arXiv preprint arXiv:2212.08186, 2022.

Research Projects

Reinforcement Learning From AI Feedback with Uncertainty 12/2024 – Present
- This project is funded by AWS Research Award;
- Read papers about Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAIF);

- Enhance the performance of pre-trained LLMs on various tasks including reasoning and summarization using RLAIF;
- Implement a novel algorithm that utilizes the uncertainties from conformal prediction to further enhance the performance of existing RLAIF technologies.

Jailbreak Black-box Multi-modal LLMs 05/2024 – 09/2024

- Submitted paper is under review;
- Propose a novel zeroth-order gradient-based method to jailbreak black-box MLLMs by generating malicious images;
- Propose patch coordinate descent in our method, significantly reducing memory consumption, enabling attacks on 13B models using a single GPU;
- Our method achieves high attack success rates, surpassing transfer-based methods, performing on par with white-box approaches and achieving 95% attack success rates on MiniGPT-4 ;
- Using a showcase to demonstrate that it is possible for our method to directly attack commercial MLLMs such as GPT-4o.

Uncertainty Quantification of LLMs 03/2024 – 09/2024

- Read papers about uncertainty quantification of LLMs;
- Previous works utilize semantic similarity to compute the uncertainty of LLMs' answers. Semantic similarity is a symmetric measure, which ignores the directional similarity behind different answers while we try to build a directional graph to utilize the ignored information;
- Combining the original answers with augmented claim-based answers to help compute the directional similarity without bias and missing information;

Privacy Benchmark of Multi-Modal LLMs 03/2024 – 06/2024

- Read papers about Multi-modal LLMs (MLLMs) and safety aspects of Multi-modal LLMs including jailbreak and defending methods of Multi-modal LLMs;
- Build a dataset that evaluates the potential privacy leakages of Multi-modal LLMs. The dataset contains different tables as images and evaluates whether MLLMs will output private information in the images or private information in their memory induced by the input images;
- Evaluate the influence of different tasks for potential privacy leakages under various MLLMs including GPT-4o, Idefics2, etc. The results show that different tasks have different risks of privacy leakage.
- Write the paper that is currently under review by ARR.

Ambiguity Related Tasks Evaluation of LLMs 02/2024 – 08/2024

- Get familiar with datasets on ambiguity and its corresponding tasks including ambiguity detection and clarification question generation;
- Build a new dataset on ambiguity based on the articles from different areas of science from different tests using LLMs. The dataset is challenging and contains a new task of ambiguity-type classification.
- Evaluate the ambiguity dataset with different closed-source LLMs such as Gemini, GPT and Claude by their APIs. The results show that even the most state-of-the-art LLMs fail in the ambiguity-related tasks;
- Evaluate the dataset with different open-source LLMs and fine-tune LLMs on the dataset to improve the ability of ambiguity-related tasks of LLMs.
- Write a paper that is currently under review by ARR.

Privacy Fine-tuning for LLMs 09/2023 – 02/2024

- Submitted paper is accepted by SDM 2025;
- Explore the weight loss landscape of differential private (DP) models and discover that DP-trained model has a sharper weight loss landscape;
- Come up with three methods from the perspective of cross-layers, within-layers and cross-models to flatten the weight loss landscape during differential private training;
- Experimental results on both black-box and white-box settings show that the methods can bridge the performance gap between privacy LLMs and normally trained LLMs;

Relationship between Classification Robustness and Explanation Robustness 10/2023 – 03/2024

- Submitted paper is under review of KDD;
- Obtain models with different levels of explanation robustness through adversarial training;
- Visualize the input loss landscape w.r.t explanation loss with models with different level of explanation robustness;
- Come up with a flat-aware training algorithm that manually adjusts the input loss landscape w.r.t explanation loss;

- Experimental results show that the previous conclusion that there is a strong correlation between classification robustness and explanation robustness might be wrong.

Dataset Condensation 04/2022 – 02/2023

- Supervised by Prof.Jinghui Chen at Pennsylvania State University
- Research about Dataset Condensation which aims to create a much less dataset than original one and network trained on this new dataset can have similar performance with networks trained on the original dataset;
- Explore method which aims to have state-of-the-art performance; Try to combine Dataset Condensation with continual learning method such as AGEM;
- Research about utilizing dataset condensation to extract global information under federated learning and using global information to reduce the influence of Non-I.I.D federated learning.

Algorithm Competition: Adversarial Robustness of Deep Learning Based on ImageNet 08/2022 – 11/2022

- Attended the algorithm competition sponsored by Pazhou Lab, Guangzhou, which aims to get high average accuracy on ImageNet under different white box attacks such as AutoAttack with different radius of perturbation;
- Replaced ReLU in Wide-ResNet with a more smoothing activation function such as SiLU to make the loss landscape smoother which is beneficial to the robustness of the deep learning model;
- Added Non-local means denoising filters to ResNet, which can reduce the effect of perturbation from white box attacks;
- Adversarially trained several ResNet and EfficientNet under AutoAttack with different radius on ImageNet, and trained an ensemble model with all models and a pertain Swin Transformer to get a final model;
- Ranked 5th among all participants and won a prize of about 6000 dollars.

Experience

Points Technology 03/2021 – 08/2021

Algorithm Intern *Shanghai*

- Get to learn federated learning. Reproduce the vertical logistic regression in federated learning way by numpy. Learn some basic knowledge of homomorphic encryption and secret sharing;
- Research about the recommendation system. Reproduce the SVD,FM,FunkSVD,BiasSVD algorithm with numpy, reproduce AutoRec. Denoisy AutoRec,NFM,AFM,AFN,NFM,FiBiNet,DeepFm etc. deep learning recommendation algorithm by Pytorch;
- Design a vertical DeepFm algorithm. Work with team to realize the vertical DeepFm.

Technical Skills

Technologies: Python, R, Amazon AWS, PyTorch, Transformers, Large Language Models, Reinforcement Learning
Hobbies: Mystery Novels, Oscar Predication
Extracurricular Activities: Deputy director of Reasoning Association for organizing mystery games and organizing Sichuan University to join in the national BBS mystery contest.