Yixin Liu

Email: yixinliucs@gmail.com; Github; Google Scholar

Education

- **Lehigh University** — Pennsylvania, US
  PhD Student - Computer Science; GPA: 3.91/4.00 — *Sep 2022 -*
  Courses: *Advanced Programming Techniques (A), Advanced Algorithms (A), Data Mining (A), Adversarial ML (A)*

- **South China University of Technology** — Guangdong, China
  BEng. - Software Engineering; GPA: 3.82/4.00 (Top4%); Outstanding Undergraduate Thesis; Sep 2018 - July 2022
  Courses: *Principles of Compiler (98), Probability And Statistics (100), Calculus (97), Linear Algebra (94), Operating System (90), Artificial Intelligence (93), Algorithm Design and Analysis (91), Data Mining (91), Software design(91)*

Working Experience

- **Samsung Research American Inc.** — Mountain View, California
  Research Intern; Supervisor: Dr. Xun Chen — *May 2023 - Nov 2023*
  - **Preventing Unauthorized AI Training via Data Poisoning/Watermarking**: proposed a more efficient and robust approach for defensive perturbation generation; extended data protection for different data types and learning settings: diffusion model, graph data, and text data.

- **Lehigh University** — Bethlehem, Pennsylvania
  Teaching Assistant for CSE017 Java Programming — *Jan 2023 - May 2023*

- **NLP group, Baidu Inc.** — ShenZhen, China
  Algorithm Engineer; Duty: Maintenance of the PaddlePaddle RL library PARL; — *Feb 2022 - May 2022*

Selected Publications & Manuscript

- **Toward Robust Imperceptible Perturbation against Unauthorized Text-to-image Diffusion-based Synthesis** : **Yixin Liu**, Chenrui Fan, Yutong Dai, Xun Chen, Pan Zhou, Lichao Sun; *Accepted by CVPR 2024.*
- **Improving Faithfulness for Vision Transformers**: Lijie Hu*, **Yixin Liu***, Ninghao Liu, Mengdi Huai, Lichao Sun and Di Wang. (*Equal Contribution.); *Submitted to ICML'24.*
- **GraphCloak: Safeguarding Graph-structured Data from Unauthorized Exploitation** : **Yixin Liu**, Chenrui Fan, Xun Chen, Pan Zhou, and Lichao Sun; *Preprint.*
- **Watermarking Text Data on Large Language Models for Dataset Copyright Protection**: **Yixin Liu**, Hongsheng Hu, Xuyun Zhang, Lichao Sun. (*Equal Contribution.); *Preprint.* June 2022 - July 2023
- **Jailbreaking GPT-4V via Self-Adversarial Attacks with System Prompts**: Yuanwei Wu, Xiang Li, **Yixin Liu**, Pan Zhou, Lichao Sun; Submitted to ACL'24. Nov 2023
- **Stable Unlearnable Example: Enhancing the Robustness of Unlearnable Examples via Stable Error-Minimizing Noise** : **Yixin Liu**, Kaidi Xu, Xun Chen, Lichao Sun; *Accepted to AAAI 2024*; June 2023 - Aug 2023
- **Unlearnable Graph: Protecting Graphs From Unauthorized Exploitation** : **Yixin Liu**, Chenrui Fan, Pan Zhou and Lichao Sun; *NDSS'23 Poster.* Jan 2023 - July 2023
- **Securing Biomedical Images from Unauthorized Training with Anti-Learning Perturbation** : **Yixin Liu**, Haohui Ye, Lichao Sun; *NDSS'23 Poster.* Jan 2023
- **SEAT: Stable and Explainable Attention**: Lijie Hu*, **Yixin Liu***, Ninghao Liu, Mengdi Huai, Lichao Sun and Di Wang. (*Equal Contribution.); *AAAI 2023* ***Oral Presentation.*** Jun 2022 - Sep 2022
- **Conditional Automated Channel Pruning for Deep Neural Networks**: **Yixin Liu**; Yong Guo; Jiaxin Guo; Luoqian Jiang; Jian Chen; *IEEE Signal Processing Letters (SPL).* June 2021 - July 2022
- **Priority Prediction of Sighting Report Using Machine Learning Methods**: **Yixin Liu**; Jiaxin Guo; Jieyang Dong; Luoqian Jiang; Haoyuan Ouyang; *IEEE SEAI 2021.* Feb 2021 - April 2021
- **MetaTool: Deciding Whether to Use Tools and Which to Use** : Yue Huang, Jiawen Shi, Yuan Li, Chenrui Fan, Siyuan Wu, Qihui Zhang, **Yixin Liu**, Pan Zhou, Yao Wan, Neil Zhenqiang Gong, Lichao Sun; *Accepted to ICLR'24.*
- **Backdoor Attacks to Pre-trained Unified Foundation Models** : Zenghui Yuan, **Yixin Liu**, Kai Zhang, Pan Zhou and Lichao Sun ; NDSS'23 Poster. Jan 2023
- **BadGPT: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT** : Jiawen Shi, **Yixin Liu**, Pan Zhou and Lichao Sun; NDSS'23 Poster. Jan 2023

Selected Research Experience

- **Toward Robust Imperceptible Perturbation against Unauthorized Text-to-image Diffusion-based Synthesis**: *During the internship @Samsung Research American*, we propose a transformation-robust approach for image protection against diffusion-based few-shot generation. Our method is validated across personalized generation methods (DreamBooth and Text Inversion) and an online training platform (Replicate). (Aug 2023 - Nov 2023) [PAPER] **[First Author]**
- **Stable Unlearnable Example: Enhancing the Robustness of Unlearnable Examples via Stable Error-Minimizing Noise**: *Under the supervision of Dr.Kaidi Xu@Drexel*, we propose a more efficient and robust unlearnable example generation with random perturbation, achieving 3× boosting and $\sim 20\%$ performance gain. (June 2023 - Aug 2023) **[First Author]**
- **SEAT: Stable and Explainable Attention**: *Under the supervision of Dr.Di Wang@KASUT*, we study the stability and explainability of the attention mechanism in the domain of NLP. (Jun 2022 - Sep 2022) [PAPER] **[Co-first Author]**

- **Motivations:** Recent studies show that attention is unstable against randomness and perturbations during training or testing, e.g., random seeds and slight perturbation of embeddings, which hinder it from being a faithful explanation tool. In our work, we seek to find an alternative to vanilla attention, which is more stable and faithful in explanation.
- **Methodology&Results:** we propose a new attention termed SEAT (**S**table and **E**xplainable **AT**tention) via post-hoc optimization based on the vanilla attention. Our SEAT enjoys the following advantages w/wo perturbation: i) preserving utility in prediction, ii) good interpretability under the criteria of top-k indices overlap. Extensive experiments on various datasets show that our SEAT outperforms other baseline methods under the architectures of RNN, BiLSTM and BERT, with different evaluation metrics on model interpretation, stability and accuracy.

- **Conditional Automated Channel Pruning for Deep Neural Networks**: Under the guidance of Dr. Yong Guo@SCUT, we leverage RL to solve to conduct efficient network pruning under multiple compression rates. [PAPER] **[First Author]**
  - **Motivation**: Existing works on automatic model pruning need to conduct repetitive and time-consuming searching under each specific pruning rate. Here, we propose a pruning agent conditioned on the demanded pruning rate, which can perform more efficient network pruning under different compression rates during deployment.
  - **Methodology&Results:** We formulate the layer-wise pruning as an MDP problem and design *Action-Constrained* Deep Determined Policy Gradient (DDPG) algorithm to solve it. Results suggest that our approach achieves $\sim$3 $\times$ searching speedup than the previous SOTA method.

- **Pests in the Nest, Under Arrest: A Hornet Sighting Report Evaluation System**: *Under the supervision of Dr. Han Huang@SCUT*, our hornet report system won *Finalist* Winner (2.82%) in 2021MCM. [PAPER, AWARD] (Apr. 2021)**[Leader]**
  - **Introduction**: Mobilizing the public to submit sighting reports is an important approach to controlling the Asian hornet. However, most sighting report submitters lack knowledge of the species, and there are many false reports, so identifying highly credible reports from a large number of sighting reports is an important issue. In this study, a report priority prediction model was constructed from the dissemination mechanism and report characteristics.
  - **Methodology&Results**: (i) We construct a probabilistic propagation model for the Asian hornet based on its migratory habits under the assumption of a Gaussian distribution. (ii) The problem of prioritizing sightings is modeled as a problem of prioritizing sighting reports, with a variety of rich features extracted from the reports, and a classification model based on a logistic regression algorithm is constructed to predict the credibility of the reports. (iii) Quantitative consideration of the corroboration effect between reports to finalize the prioritization of reports. Our method achieved a weighted accuracy of 83.5% on the benchmark test set.

- **Bert-based Legal Judgment Document Amount Entity Extraction Model**: *Under the supervision of Dr. Han Huang@SCUT*, I design and implement a system based on BERT to conduct amount entity extraction for legal judgment documents. [DEMO, CODE] (Feb 2021 - March 2021) **[Leader]**
- **Covid-19 Prediction Model based on Huber Regression and Hierarchical Feature**: Design hierarchical classification feature coding scheme and design algorithm based on Huber Regression. [TALK, Result, CODE] (Oct 2020) **[Leader]**
- **Anti-theft detection based on Kirchhoff's current law and CAD grid refining algorithm.** : *Under the supervision of Dr. LongJun Wang@SCUT*, I designed a Python program and GUI interface that detects potential electric theft with sensors. Besides, I studied how to refine the CAD grid diagram based on the BFS-DFS algorithm. (May 2019 - Aug 2019)

Honors and Awards

- AAAI 2023/2024 Travel Grant — Dec, 2022
- Lehigh Graduate Fellowship — Sep, 2022
- Outstanding Undergraduate Thesis (4%) — June, 2022
- National Scholarship $\times$2 (1.4%) — Sep, 2021/Sep, 2022
- Finalist at 2021 American Mathematical Contest in Modeling (2.82%) — April, 2021
- Provincial First Prize at China Contemporary Undergraduate Mathematical Contest in Modeling — Oct, 2020
- Provincial Second Prize at China National Mathematics Competition for College Students — Oct, 2020
- First Prize at the 6th China Health Information Processing Conference (4-th Track) — Nov, 2020
- Second Price at National Undergraduate Software Practice and Innovation Ability Competition — Nov, 2020
- South China University of Technology (SCUT) Premier Scholarship (4%) — Sep, 2019

Softwares

- **OpenChatPaper**: Yet another paper reading assistant based on OpenAI ChatGPT; *154 Star at Github*;
- **Arxiv2Latex**: Download the source latex code of multiple papers from Arxiv with one click. *87 Star at Github;*

Technical Skills

- **Programming Languages**: Python, C++, MATLAB, LaTeX, JavaScript, HTML, Java, WebGL
- **ML Frameworks**: PyTorch, TensorFlow, Keras