Yixin Liu

Email: yixinliucs@gmail.com

Homepage: liuyixin-louis.github.io; Github: github.com/liuyixin-louis; Google Scholar;

Education

- **Lehigh University** — Pennsylvania, US
  PhD Student - Computer Science; GPA: 3.91/4.00 — *Sep 2022 -*
- **South China University of Technology** — Guangdong, China
  BEng. - Software Engineering; GPA: 3.82/4.00; Ranking: Top4%; — *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

- **NLP group, Baidu Inc.** — ShenZhen, China
  Algorithm Engineer; — *Feb 2022 - May 2022*
  Maintaining the RL Library PARL; Surveying on the topic of Black-box Optimization; Auto-testing framework development

Selected Publications

- **SEAT: Stable and Explainable Attention**: Lijie Hu*, **Yixin Liu***, Ninghao Liu, Mengdi Huai, Lichao Sun and Di Wang. (*co-first author); Accepted by 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; Accepted by 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; Accepted by IEEE SEAI. Feb 2021 - April 2021

Other Publications&Preprint&Submissions

- **Securing Biomedical Images from Unauthorized Training with Anti-Learning Perturbation** : **Yixin Liu***, Haohui Ye*, Lichao Sun; (*co-first author); Accepted to NDSS23 Poster. Jan 2023
- **Unlearnable Graph: Error-Minimizing Structural Poisoning for Protecting Knowledge Inside Graphs From Unauthorized Exploitation** : **Yixin Liu**, Chenrui Fang, Lichao Sun; Short version accepted by NDSS23 Poster; Full version submitted to IJCAI 2023. Jan 2023
- **Enhancing Robust Unlearnable Example with Stable Error-Minimizing Noise**: **Yixin Liu**, Kaidi Xu, Lichao Sun; Preprint. July 2022 - Nov 2022
- **Improving Faithfulness for Vision Transformers** : Lijie Hu*, **Yixin Liu***, Ninghao Liu, Mengdi Huai, Lichao Sun and Di Wang; Submitted to ICML 2023. Nov 2022 - Jan 2023
- **A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to ChatGPT** : Ce Zhou, Qian Li, Chen Li, Jun Yu, **Yixin Liu**, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, Lichao Sun; Preprint. Feb 2023
- **Backdoor Attacks to Pre-trained Unified Foundation Models** : Zenghui Yuan, **Yixin Liu**, Kai Zhang, Pan Zhou and Lichao Sun ; Accpected to NDSS23 Poster. Jan 2023
- **BadGPT: Exploring Security Vulnerabilities of ChatGPT via Backdoor Attacks to InstructGPT** : Jiawen Shi, **Yixin Liu**, Pan Zhou and Lichao Sun; Accpected to NDSS23 Poster. Jan 2023
- **Data Ownership Identification via Membership Inference under Backdoor**: Hongsheng Hu*, **Yixin Liu***, Xuyun Zhang, Lichao Sun; Preprint. May 2022 - July 2022
- **Adversarial attack and defense on graph data: A survey**: Lichao Sun, Yingtong Dou, Carl Yang, Kai Zhang, Ji Wang, **Yixin Liu**, Philip S. Yu, Lifang He, and Bo Li; Preprint. -

Researches

- **Enhancing Robust Unlearnable Example with Stable Error-Minimizing Noise**: We identify the instability issue in the current robust unlearnable example and introduce random perturbation terms for improving stability. [UNDER REVIEW] **[First Author]**
  - **Motivations:** Existing poisoning-based data protection techniques fail to tackle the train-test perturbation misalignment problem, which hinders the delusive performance of added noise.
  - **Methodology:** In our work, we identify such an issue stems from the misalignment between the target model and the source model, where the defensive noise trained against worst-case perturbation becomes ineffective during evaluation. To migrate it, we propose Stable Error-Minimizing (SEM) noise that trains defensive noise against random perturbation.
  - **Results:** Extensive experiments show SEM achieves a new SOTA performance. Notably, SEM achieves 3.91× speedup and $\sim 30\%$ testing accuracy protection performance gain on CIFAR-10 under adversarial training with $\epsilon = 4/255$.
- **SEAT: Stable and Explainable Attention**: We study the stability and explainability of the attention mechanism in the domain of NLP. [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.

- **Data Ownership Identification via Membership Inference under Backdoor** : We study how to identify unauthorized usage of some private data for training and how to protect sensitive information from being leaked. [UNDER REVIEW] **[Co-first Author]**
  - **Motivations:** Feasibility of extracting sensitive information, e.g., email addresses and phone numbers, from large pre-trained language models poses privacy risks for each individual. To migrate it, we creatively propose a backdoor-based membership inference approach for identifying and preventing unauthorized usage of private data.
  - **Methodology&Results:** To protect different types of private fields, we propose three levels of trigger for masking the private field. After masking, to identify whether private data is used or not, we poison the protected data by creating a backdoor mapping between the trigger and target label and leverage hypothesis testing for verification.

- **Conditional Automated Channel Pruning for Deep Neural Networks**: We study the conditional pruning problem to conduct efficient network pruning under multiple compression rates. [PAPER, CODE] **[First Author]**
  - **Motivation**: Current 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 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**: We designed a hornet sighting report system, and the paper got *Finalist* Winner (2.82%) in 2021MCM. [PAPER, AWARD] (Apr '21) **[Leader]**
  - **Introduction**: Mobilizing the public to submit sighting reports is an important approach to controlling the Asian hornet. However, most sighting report submitter 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 modelled 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**: Design and implement a system based on BERT to conduct amount entity extraction for legal judgment documents. [DEMO, CODE] (Feb '21 - March '21) **[Leader]**
  - **Introduction**: Extracting money fields such as principal and interest in legal loan judgments is usually done manually, which is time-consuming and laborious. In this project, we propose a BERT-based model for extracting monetary entities from legal lending judgments and automating the extraction of ten types of monetary entities.
  - **Methodology&Results**: (i) We design the TF-IDF-Bayesian approach for processing the judgment text. (ii) We adopt regular expressions for the extraction of monetary entities and propose context-based features with the BERT feature extractor. (iii) Results show that our algorithm achieves a weighted accuracy rate of 93.5%. And we completed the testing of the module, packaged the algorithm as docker and deployed it to the server.

- **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 '20) **[Leader]**

Honors and Awards

- Lehigh Graduate Fellowship — Sep, 2022
- Outstanding Undergraduate Thesis (PAPER) [**12/271≈4%**] — June, 2022
- National Scholarship $\times 2$ [**1/71≈1.4%**] — Sep, 2021/Sep, 2022
- *Finalist* at 2021 American Mathematical Contest in Modeling [**284/10061≈2.82%**] — April, 2021
- *Provincial First Prize* at Contemporary Undergraduate Mathematical Contest in Modeling — Oct, 2020
- *Provincial Second Prize* at National Mathematics Competition for College Students — Oct, 2020
- *Champion* at the 6th China Health Information Processing Conference (Evaluation 4) — Nov, 2020
- *Runner-up* at National Undergraduate Software Practice and Innovation Ability Competition — Nov, 2020
- SCUT Premium Scholarship — Sep, 2019

Softwares

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