XUKUN LIU

+1 8722392517 | e: xukunliu2025@u.northwestern.edu | Github | Personal Website

EDUCATION

Northwestern University Evanston, United States
Master of Computer Science Sept 2023 – June 2025

Southern University of Science and Technology Shenzhen, China
Bachelor of Engineering in Computer Science and Technology Sept 2019 – June 2023

WORK EXPERIENCE

Huawei Technology Shenzhen, China
Software Development Engineer June 2022 – July 2022

- Designed a neural network to recover the global beam information based on the local beam measurement.
- Responsible for model design, data processing, and improvement of model accuracy.
- Used a variety of classic GNN methods to model the problem for further development.
- Designed a neural network using the combination of co-occurrence matrix and GAT, which reached SOTA.

SELECTED AWARDS

Outstanding Graduate of Southern University of Science and Technology (SUSTech). (May 2024)
Outstanding Graduate of the Computer Science Department at Southern University of Science and Technology (SUSTech). (May 2024)
Bronze Medal in 2020 China Collegiate Programming Contest, Mianyang Site. (Oct 2020)
Bronze medal in the 2020 ICPC Asia Nanjing Regional Contest. (Dec 2020)

PUBLICATIONS

1. **XukunLiu**, BowenLie, RuqiZhang, Dongkuan Xu. *Adaptive Draft-Verification for Efficient Large Language Model Decoding, submitted to NeurIPS 2024.*
2. Dong Shu, Haoran Zhao, **Xukun Liu**, David Demeter, et al. *LawLLM: Law Large Language Model for the US Legal System, accepted at the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)*
3. BinfengXu, **XukunLiu**, et al. Gentopia. AI: A Collaborative Platform for Tool-Augmented LLMs, *The 2023 Conference on Empirical Methods in Natural Language Processing(EMNLP 2023)*
4. **XukunLiu**,, ZhiyuanPeng, DK Xu. ToolNet: Connecting Large Language Models With Massive Tools, *Submitted to 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*
5. **X. Liu**, The Utilities of Evolutionary Multi-objective Optimization for Neural Architecture Search –An Empirical Perspective, *The 17th International Conference on Bio-inspired Computing: Theories and Applications*
6. **XukunLiu**,, Haoze Lv, Chi Wang, et al. Towards Efficient Hyperparameter-Architecture Search via SynSearch: Expedited Exploration in Enormous Search Space.

TEACHING ASSISTANT EXPERIENCES

- Teaching Assistant for *Data structure and Algorithm Analysis*, Fall 2022
- Teaching Assistant for *Introduction to Python Programming*, Fall 2022
- Teaching Assistant for *Principles of Database Systems*, Spring 2022
- Teaching Assistant for *Computer Organization*, Spring 2022
- Teaching Assistant for *Introduction to Computer Programming B*, Spring 2022
- Teaching Assistant for *Data structure and Algorithm Analysis*, Spring 2021
- Teaching Assistant for *Introduction to Computer Programming A*, Spring 2021

RESEARCH EXPERIENCES

Adaptive Draft-Verification for Efficient Large Language Model Decoding West Lafayette, IN, USA
Group Leader Feb 2024 - Present

- Objective: To enhance the efficiency and speed of Large Language Model (LLM) decoding for real-time applications, reducing latency and computational demands.
- Developed a novel methodology, ADED (Adaptive Draft-Verification for Efficient LLM Decoding), which accelerates LLM decoding without requiring fine-tuning.
- Implemented an adaptive draft-verification process that evolves over time, utilizing a tri-gram matrix-based LLM representation to dynamically approximate output distributions and improve decoding efficiency.

- Designed a draft maker inspired by Monte Carlo Tree Search (MCTS), balancing exploration and exploitation to generate high-quality drafts and optimize decoding speed.
- Demonstrated through extensive experiments that ADED significantly accelerates the decoding process while maintaining high accuracy, achieving up to a 2.5X speedup in latency and a 20% improvement in acceptance rates over existing methods.

ToolNet: Connecting Large Language Models With Massive Tools Raleigh, NC, USA
Group Leader Oct 2023 - Present

- Objective: To enhance the capabilities of Large Language Models (LLMs) to perform higher-level tasks, such as following human instructions to properly use external tools (APIs)
- Developed ToolNet, a plug-and-play framework that scales up the number of tools to thousands with no performance degradation and constant token costs
- Designed a network structure where each node represents a tool and weighted edges denote transition probabilities, enabling an LLM to travel along the network by iteratively choosing the next tool from its neighbors until the task is resolved
- Demonstrated through experiments that ToolNet can achieve impressive results in complex tasks and has strong robustness against tool failures.

Gentopia.AI : A Collaborative Platform for Tool-Augmented LLMs Raleigh, NC, USA
Group Leader June 2023 – Oct 2023

- Objective: To create a collaborative platform for tool-augmented Large Language Models (LLMs)
- Contributed to the development of Gentopia, enabling flexible customization of agents through simple configurations, integrating various language models, task formats, prompting modules, and plugins into a unified paradigm
- Participated in the establishment of Gentpool, a public platform for the registration and sharing of user-customized agents, promoting the democratization of artificial intelligence
- Assisted in the design of Gentbench, a component of Gentpool, to evaluate user-customized agents across diverse aspects such as safety, robustness, efficiency, etc.

Efficient Heterogeneous Bert Redmond, DC, USA
Independent Project, jointly supervised by North Carolina University and Microsoft Research Sept 2022 – Present

- Objective: to establish a more efficient BERT model through Neural Architecture Search
- Perfected the training method of superset and proposed the “Balanced Pareto Sampling” method based on previous research, managing to improve the performance of subnets by 1%-2% in the same training time compared to the existing methods
- Applied heterogeneous search space rather than the homogeneous methods

Towards Efficient Hyperparameter-Architecture Search via SynSearch: Expedited Exploration in Enormous Search Space Raleigh, NC, USA
Independent Project, jointly supervised by North Carolina University and Microsoft Research July 2022 – Present

- Objective: to propose new neural architecture search algorithms to attain the “cost-effective” architecture
- Summarized the law by testing the performance of current mainstream NAS algorithms in different search spaces and design an algorithm to reduce the search time

EvoXbench, an All-In-One Neural Architecture Search Framework Shenzhen, China
Research Assistant to Professor Zhichao Lu May 2022– July 2022

- EvoXBench—an open-source library that integrates all technologies required for NAS algorithm development, enabling users to test or develop algorithms by simply calling python or Matlab interfaces
- Processed data, integration, and database construction: Collected most of the existing NASBench datasets, extracted the data, and curated the data using the ORM framework provided by Django
- Train the surrogate model, and oversaw the experiment process
- https://github.com/EMI-Group/evoxbench

AutoML Tools Development for Deep Learning on Edge Systems Shenzhen, China
Group Leader, Advisor: Professor Ran Cheng Sept 2021 – Jan 2022

- Aimed to design an AutoML algorithm that can be applied to a variety of devices, especially for small and low-power edge devices
- Deployed and tested a variety of neural networks on devices with different architectures, studied and analyzed their result, and supervised the algorithm and architecture design
- Applied torch to instantiate neural networks, and used celery for task sending and distributed evaluation

SELECTED PROJECT EXPERIENCES

MerryQuery for North Carolina State University(NCSU) Raleigh, North Carolina, USA

- Developed MerryQuery, an AI-powered educational assistant using retrieval-augmented generation (RAG) to enhance academic support.
- Enabled personalized student responses based on course materials and prior interactions, with robust controls for teachers to manage content.
- Successfully implemented to automate responses to common student inquiries, significantly easing teacher workloads and improving resource accessibility.

SageCube: AI Desktop Assistant on Steam Evanston, IL, USA

- Developed SageCube is an AI assistant powered by large language models, designed for desktop environments and currently available on Steam for testing.
- Features a visually appealing interface with interactive Live2D and 3D virtual avatars, which users can interact with using voice or text inputs.
- Supports a variety of voice models for text-to-speech functionality, allowing primarily voice-based interactions.
- Integrates with Steam's Workshop, enabling users to enhance customization and functionality by uploading and installing tools and acquiring new agents.

Multifunctional and Extensible Online Judge (OJ) System Shenzhen, China

- Developed a scalable online judge system to evaluate code correctness across multiple programming languages.
- Led the website design, backend construction, deployment, and development of an evaluation engine using Python's Django framework and Google's nsjail.
- Implemented system deployment with Kubernetes for automatic scaling and self-repair capabilities.
- The system passed third-party penetration testing and is now officially used by the Computer Science Department at the university, serving over 2,500 students in 13 courses.

User Profile Webpage Design for SUSTech Library Shenzhen, China

- To generate a unique school library memorial page for students
- Designed, built, and developed back-end service, and launched on the WeChat public account of the Southern University of Science and Technology Library to provide services for students
- Obtained the highest score among the teams with the same type of project

ADDITIONAL INFORMATION

Interests

- NLP, Large Language Model, Agent, Multi-modal, Efficient AI, CV

Technical Skills

- Programming Languages: Python, Rust, C/C++, JAVA, Nodejs, HTML, JavaScript, SQL

GitHub

- liuxukun2000 (Xukun Liu) (github.com)

Personal Website

- Xukunliu.com