Jiesong Liu

984-325-3859 | jliu93@ncsu.edu | linkedin.com/in/jiesong-liu | github

Education

North Carolina State University Raleigh, NC

Ph.D. in Computer Science *Aug. 2023 – 2028 (Expected)*

- Advisor: Dr. Xipeng Shen
- Research Interests: Efficient AI, AI optimization

Renmin University of China Beijing

Bachelor of Computer Science *Sep. 2019 – June 2023*

- GPA: 3.9/4.0
- Core Courses: Data Structures and Algorithms 96, Introduction to Computer System 96, Operating System 94, Parallel Computing 92

Research Experience

Token Reuse on Generative AI and Transformers Oct. 2023 – Jan. 2024

North Carolina State University *Raleigh, NC*

- Leveraged token similarities within a transformer block in generative AI models to efficiently eliminate computation redundancy.
- Employed the LSH algorithm to cluster tokens during a diffusion iteration and efficiently reused the clustering results across subsequent iterations.

Uncertainty Quantification-Guided Hyperparameter Optimization Aug. 2023 – Jan. 2024

North Carolina State University *Raleigh, NC*

- Examined the meaning of uncertainty, the source of uncertainty, and the impact of uncertainty on HPO.
- Quantified the uncertainty on HPO and used it to adjust resource allocation and budget allocation.
- As a general approach, the UQ-guided HPO attains a performance improvement exceeding 50% in terms of accuracy regret over the existing HPO methods.

Efficient DNN Inference on Microcontrollers via TREC Reuse Oct. 2021 – Oct. 2022

North Carolina State University *Raleigh, NC*

- Proposed the use of TREC (Transient Redundancy Elimination-based Convolution) as a new way to reduce computations in DNNs running on microcontrollers.
- Introduced a set of optimizations to mitigate the space overhead incurred by TREC.
- Empirically evaluated the effectiveness of the new solution on two models of microcontrollers, confirming the substantial benefits of the new solution (3.4-5x speedups) in enabling efficient DNNs on microcontrollers.

Generalized Reuse Patterns for Accelerating DNNs on Microcontrollers Nov. 2023 – June 2024

North Carolina State University *Raleigh, NC*

- Formulated the reuse space and derived a system of reuse patterns for reuse-based DNN optimization.
- Systematically characterized the connections between reuse patterns and data layouts in memory.
- Proposed an analytical approach to infer the implications of various reuse patterns to the performance of DNNs and provided an efficient way to identify the appropriate reuse pattern for a given DNN.

Enabling Efficient Learned Index on GPU Feb. 2021 – Feb. 2022

Renmin University of China *Beijing*

- Developed the first dynamic learned index on GPU based on the PGM-index.
- Formulated an efficient indexing strategy and harnessed shared memory optimization to enhance performance.
- Achieved an impressive 107x speedup when compared to the current state-of-the-art learned indexes, leveraging the power of the 2080Ti GPU.

Approximating Probabilistic Group Steiner Trees in Graphs Sep. 2021 – Jan. 2022

Renmin University of China *Beijing*

- Defined the problem of probabilistic group Steiner tree (PGST).
- Devised the parallel version of the pruned landmark labeling algorithm and achieved significant speedups.

Publication

[**ASPLOS'23**] "Space Efficient TREC for Enabling Deep Learning on Microcontrollers"; **Jiesong Liu**, Feng Zhang, Jiawei Guan, Hsing-Hsuan Sung, Xiaoyong Du, Xipeng Shen.

[**NeurIPS'22**] "TREC: Transient Redundancy Elimination-based Convolution"; Jiawei Guan, Feng Zhang, **Jiesong Liu**, Hsing-Hsuan Sung, Ruofan Wu, Xiaoyong Du, Xipeng Shen.

[**VLDB'24** (Revision)]"A Systematic Study on Early Stop Metrics in HPO and the Implications of Uncertainty"; Jiawei Guan, Feng Zhang, **Jiesong Liu**, Xipeng Shen.

[**TC'24**]"Enabling Efficient Deep Learning on MCU with Transient Redundancy Elimination"; **Jiesong Liu**, Feng Zhang, Jiawei Guan, Hsing-Hsuan Sung, Xiaoyong Du, Xipeng Shen.

[**TPDS'24**] "G-Learned Index: Enabling Efficient Learned index on GPU"; **Jiesong Liu**, Feng Zhang, Lv Lu, Xiaoyong Du, Guoliang Li, Dong Deng.

[**VLDB'23**] "Approximating Probabilistic Group Steiner Trees in Graphs"; Shuang Yang, Yahui Sun, **Jiesong Liu**, Xiaokui Xiao, Ronghua Li, Zhewei Wei.

[**TPDS'22**] "Exploring Query Processing on CPU-GPU Integrated Edge Device"; **Jiesong Liu**, Feng Zhang, Hourun Li, Dalin Wang, Weitao Wan, Xiaokun Fang, Jidong Zhai, Xiaoyong Du.

Honors

2023 NCSU University Graduate Fellowship

2023 NCSU Graduate Merit Awards

2020 & 2021 National Scholarships of China (highest scholarship for Chinese undergraduate)

2022 SenseTime Scholarship (30 students selected from across China)

2018 National Olympiad in Informatics (NOI)

Technical Skills

Languages: C, C++, CUDA, OpenMP, Python, SQL, Verilog
Developer Tools: Git, VS Code, Eclipse, gdb