Gaoxiang Duan

Shanghai, China
+86 17521741344
duangx0331@gmail.com

Education

University of Chinese Academy of Sciences **Sept. 2021 – Jun. 2024 (Expected)**
M.Eng. Electronic and Information Science

University of Science and Technology of China **Sept. 2013 – Jun. 2018**
B.S. Computational Mathematics

Research Interests

Human-Centric Artificial Intelligence, Natural Language Processing

Research Experience

***Graduate Researcher*, Chinese Academy of Sciences** **Sept. 2021 – present**

- Designed a distance-aware attention mechanism that reduced computational complexity while maintaining model performance by incorporating prior knowledge of locality.
- Developed an efficient attention mechanism utilizing the bitwise XOR operation for low-cost, low-precision devices. This approach avoids the costly float-point matrix multiplication by utilizing XOR to capture the "attention relationship" between tokens instead of dot products, reducing the device's requirements while maintaining model performance.
- Introduced branch prediction, a scheduling strategy in computer architecture, to arrange outputs from draft models and standard large language models, enabling faster inference without any changes to the original LLM output.
- Designed an machine learning-based PES(potential energy surface) model which is able to efficiently represent the PES of a wide variety of systems with the accuracy of quantum mechanics models.

Publications

Bitformer: An efficient transformer with bitwise operation-based attention for big data analytics at low-cost low-precision devices. **Duan, G**., Zhang, J., Zheng, X., Zhu, Y., Wang, A. (Under review) Arxiv

Fast inference of LLM from branch prediction perspective. **Duan, G.**, Zheng, X., Zhu, Y. (Under review)

Customized self-attention operator hardware acceleration on FPGA. Zhang, J. ***Duan, G*** (Under review)

An efficient transformer with distance-aware attention. IEEE Intl Conference on High Performance and Smart Computing (HPSC). **Duan, G.**, Zheng, X., Zhu, Y., Ren, T., & Yan, Y. (May 2023)

BTPA: Hardware-Software Co-design for Bitwise based Transformer with Parallelized Accelerator. The 8th International Conference on Smart Computing and Communication. Zhang, J. **Duan, G.**, Zheng, X., Zhu, Y. (Nov 2023)

Work Experience

***Founder, Organizer, Speaker & Poster Designer*, CIS Seminar** **May 2023 – Present**

- Established a biweekly deep learning seminar with over 150 researchers and students from hardware and software labs participating, promoted cross-laboratory collaboration and knowledge exchange on two areas: co-design of software and hardware for specialized transformer chips, and research on large language models. (Posters)
- Delivered lectures with a teaching component, providing foundational deep learning knowledge to attendees. (Slides Presented)

***Assistant DevOps Engineer*, Chinese Academy of Sciences Sensing and Computing Lab** **Sept. 2021 – Present**

- Managed lab servers, oversaw day-to-day maintenance, troubleshooting, and handling permissions management.
- Installed and conducted performance testing on 20 servers for the Fast National Astronomical Observatories.

Skills

Knowledge Background

Mathematics: Linear Algebra, Multivariable Calculus, Convex Optimization, Probability Theory
Artificial Intelligence: Machine Learning, Deep Learning, Brain-Computer Interface

Research Skills

Deep Learning Framework: Pytorch
Programing Languages: Python, C, Lua
Experimental Environments: Linux, Docker

General Skills

Language: Chinese(native), English(IELTS 6.5), French(beginner)
Writing: Latex, Overleaf

Volunteer Experience

***Summer School Teaching Assistant*, Zhangjiang National Lab Summer School** **June 2023**

- Instructed a group of 8 undergraduate students during a summer program focused on the compute-in-memory simulator NeuroSim under the invitation of Dr. Dongdong Li.

***Special Education Volunteer*, Fang Cao Young Volunteers Association of USTC | Everyday Progress Autism Children's Rehabilitation Center | Hefei Special Education Center** **2013 – 2018**

- Accumulated 200 hours of regular volunteer activities, including spending time with visually impaired children, engaging in play activities, reading stories, and providing assistance to children with autism during their classes.
- Organized fundraisers for school uniforms and children's slides at the *Everyday Progress Autism Children's Rehabilitation Center*.
- Served as Head of the Special Education Department in 2014.

References

Dr. Xiaoying Zheng, Chinese Academy of Sciences, Email: zhengxy@sari.ac.cn

Dr. Yongxin Zhu, Chinese Academy of Sciences, Email: zhuyongxin@sari.ac.cn

Dr. Hui Wang, Chinese Academy of Sciences, Email: wanghui@sari.ac.cn