Gaoxiang Duan

Shanghai, China
+86 17521741344
duangx0331@gmail.com

Education

M.Ed. **Shanghai Advanced Research Institute, Chinese Academy of Science** — **Sept. 2021 – Jun.2024(Expected)**
Electronic and Information Science
Shanghai, China

B.A. **University of Science and Technology of China (USTC)** — **Sept. 2013 – Jun. 2018**
Computational Mathematics
Hefei, Anhui, China

Research Experience (at the graduate level)

Efficient Transformer Mechanism Design

- Proposed *a distance-aware attention mechanism*, motivated by the fact that existing attention mechanisms do not incorporate prior knowledge of locality. This reduces computational complexity while maintaining model performance. (Publication on IEEE HPSC 23)
- Proposed *a bitwise operation-based attention mechanism* 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 product. This reduces the device's requirements while still maintaining comparable model performance. (Under review on IEEE Transaction on Big Data)

Fast Inference in Large Language Model

- Introduced branch prediction, a scheduling strategy in computer architecture, to arrange outputs from the draft model and the standard LLM. This enables faster inference without any changes to the original LLM output. (Writing Paper)
- Redesigning the decoder to facilitate structured output in the language generation model, as opposed to sequential output. This alignment with human writing patterns concurrently accelerates the inference process. (In progress)

Co-design of Software and Hardware for Specialized AI Chips

- Quantified the requirements of precision of Transformer during inference.
- Designing the device friendly version of Transformer to fit the constraint of chip. (In progress)

Publications

Publications in Progress, Under Review

Under Review. *Bitformer: An efficient Transformer with bitwise operation-based attention for Big Data Analytics at low-cost low-precision devices.* Currently under review by the Journals of IEEE Transaction on Big Data (TBD). **Duan, G.**, Zhang, J., Zheng, X., Zhu, Y., Wang, A.

In progress. *Fast inference of LLM from branch prediction perspective.* **Duan, G.**, Zheng, X., Zhu, Y.

In progress. *Customized self-attention operator hardware acceleration on FPGA. Zhang, J.* ***Duan, G***

Publications

2023, May. 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.

Professional Service

Founder, Organizer & Speaker of CIS Seminar, May 2023 – Present

这里改成我们的 seminar 主页()

- Established a recurring Deep Learning Seminar, uniting researchers and students from various labs, promoting cross-laboratory collaboration and knowledge exchange.
- Leveraged prior experience in delivering informative presentations to offer early sessions with a teaching component, providing foundational deep learning knowledge to attendees.
- Spearheaded the initiative to delve into specialized sub-topics, the seminar now encompasses two distinct focus areas: Co-design of Software and Hardware for Specialized AI Chips and the research about the Large Language Model (LLM) inference.
- -The seminar attracted participants from both hardware and software-oriented research groups within our laboratory. Notably, I designed and implemented a new hardware-friendly attention mechanism, which subsequently underwent hardware optimization by members of the hardware team.

Refereeing Activities

- Co-Reviewer, International Conference on Data Mining (ICDM), 2023
- SubReviewer, The 18th IEEE Asia Pacific Conference on Circuits and Systems (IEEE APCCAS), 2022

Assistant DevOps Engineer in Lab, Sept. 2021 – Present

Sensing and Computing Lab at Chinese Academy of Science

- Managed lab servers, overseeing 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, Mathematical Statistics

Artificial Intelligence: Machine Learning, Deep Learning (mainly in NLP)

Research Skills

Deep Learning Framework: Pytorch

Programing Language: Python, C, Lua

Experiments Environment: Linux, Docker

General Skills

Language: Chinese(native), English(skilled), French(beginner)

Adobe Photoshop: graphic design

Final Cut Pro: video editing

Volunteer Experiences

COVID-19 Pandemic Volunteer, April 2022

Shanghai

- Conduct telephone communications with patients and their families seeking assistance and provide guidance on purchasing medication and seeking medical attention.
- Coordinate potential medical resources and treatment solutions.
- Organize and maintained information from help requests on social media platform like Weibo.

Special Education Volunteer, 2013 - 2018

Fang Cao Young Volunteers Association of USTC │ Everyday Progress Autism Children's Rehabilitation Center │ Hefei Special Education Center

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

Summer School Educational Volunteer, June 2023

Zhangjiang Laboratory Summer School

- Conduct presentations to introduce summer program content to undergraduate students.
- Provide guidance on using the *Neurosim* simulator for simulating DNN model.
- Facilitate learning and discussions among students.