YAZHU DONG

p: +65 80991474 | e: yazhu.dong@u.nus.edu
Github: https://github.com/l1cacheDell

EDUCATION

National University of Singapore Singapore
MSc Computer Engineering Aug 2025 – So far

- Specialisation in Computing Hardware Infrastructure (CHI)

Beijing University of Posts and Telecommunications, School of Artificial Intelligence Beijing, China
BENG in Artificial Intelligence Sept 2021 – July 2025

- GPA: 3.62/4.0
- Average score: 88.66/100

PUBLICATIONS

1. Rui Kong, Qiyang Li, Xinyu Fang, Qingtian Feng, Qingfeng He, **Yazhu Dong**, Weijun Wang, Yuanchun Li, Linghe Kong, Yunxin Liu "LoRA-Switch: Boosting the Efficiency of Dynamic LLM Adapters via System-Algorithm Co-design", Sep 2024, *arXiv preprint*;
2. Liang Mi, Weijun Wang, Wenming Tu, Qingfeng He, Kui Kong, Xinyu Fang, **Yazhu Dong**, Yikang Zhang, Yuanchun Li, Meng Li, Haipeng Dai, Guihai Chen, Yunxin Liu "Empower Vision Applications with LoRA LMM", *Eurosys 2024*;
3. **Yazhu Dong**, Yuxing Zhang, Haiyuan Li, Duanling Li, "Deep Learning-based Image Segmentation and Validation for Puncturing Robots", June 2024, *Journal of Nanjing University of Science and Technology*

RESEARCH EXPERIENCE

Beijing University of Posts and Telecommunications, School of Intelligent Engineer & Automation Beijing, China
Research Assistant to Associate Professor Haiyuan Li June 2022 – May 2024

Research on Medical Image Registration Technology and Development of Surgical Robots (A University Student Innovation and Entrepreneurship Training Program, National-level)

- ***Overview***: Developed an AI-based solution for the recognition and registration of multimodal prostate images, along with a hardware-software integrated surgical assistance system that enables real-time prediction of needle trajectory;
- Utilized Fast-SAM to address inference bottlenecks in medical image segmentation tasks and developed an interactive prompt-based segmentation interface;
- Yielded a paper: *Deep learning-based image segmentation and validation for prostate cancer surgery robots*.

Institute for AI Industry Research, Tsinghua University (AIR) Beijing, China
Research Assistant to Assistant Professor Yuanchun Li May 2023 – Apr 2024

Adapter as A Service

- ***Overview***: Reduced the batched inference latency and improved the throughput of the large language model inference system by optimizing the parallel computing process with multiple LoRA adapters;
- Took charge of the scheduling system for LoRA adapters loading in preemptive scenarios, and dynamically distributing requests across instances to ensure the lowest first-token latency and significantly boosted overall system throughput, achieving a 1.5x improvement;
- Implemented CUDA kernels to merge all LoRA adapters in a single operation, optimizing performance by reducing the fan-in and fan-out time overhead at the bottleneck, leading to a 400x speedup in inference at this critical point;
- Concluded the research into the paper *LoRA-Switch: Boosting the Efficiency of Dynamic LLM Adapters via System-Algorithm Co-design, arXiv preprint;*
- Co-authored the paper *Empower Vision Applications with LoRA LMM*, *Eurosys 2024.*

WORK EXPERIENCE

Baidu Beijing, China
High-Performance Computing R&D Intern Nov 2024 – May 2025

- **Framework Maintenance**: Maintained and optimized multiple Baidu PaddlePaddle open-source frameworks (PaddleNLP, PaddleMIX), enhancing inference efficiency and robustness for paddle-triton applications.
- **CUDA HPC Development**: Developed and optimized GPU operators, focusing on 8-bit SageAttention integrating into paddle on SM80, SM89 and SM90, speeding up 1.8x for LLM and diffusion model inference.
- **Compiling Innovation**: Upon reflection on PaddleNLP's C++/CUDA compiling bottleneck, accelerated compilation via CMake refactor, reducing build time from 60min to 25min and boosting team productivity.
- **Research Engagement**: Led bi-weekly paper sharing meetings, distilling key insights from recent AI/HPC research to support team innovation.

ChinaDaaS Beijing, China
AI Engineer Intern, AI Department July 2024 – October 2024

- Fine-tuned the BART model to handle millions of enterprise name processing, leading to the increase of data acceptance rates from approximately 42% to over 70%;
- Led the development of an internal AI software for company employees, utilizing LangChain Agent to build a software licensing and authorizing tool, incorporating self-reflection mechanism for enhanced performance;
- Developed an AI-driven solution for summarizing the company's weekly meetings by using the Whisper model for audio recognition, followed by multi-step reasoning with LLM to generate comprehensive meeting summaries, which significantly outperformed existing meeting summarization tools;
- Reproduced a Retrieval-Augmented Generation algorithm for multi-level enterprise name classification, specifically designed to handle challenging classification cases, which outperformed competitor products in terms of data accuracy, data completeness, and classification functionality.

Open-source Contributions

- **Vllm_backend (NVIDIA)**: Added comprehensive multi-LoRA serving feature support to vLLM in Triton's inference backend. This involved implementing a new local LoRA Adapters weight mapping and management solution, facilitating easier deployment of multi-LoRA model inference for developers. Contributions included documentation and CI test scripts (762 lines of code).
- **LMDeploy (Shanghai AI Lab)**: Fixed a potential key-value mapping error that could lead to system crashes during model inference. The fix enhanced the stability of InternLM when used with different Agent frameworks for tool invocation. A unit test was included to ensure robustness (39 lines of code).

Additional Information

Computer and Language Skills

- Software and tools: Docker, NVIDIA (CUDA, TensorRT), git
- Programming: C++, CUDA, Python
- Systems: Linux
- Frameworks: PyTorch, transformers, FastAPI
- Chinese (native), English (IELTS: 7)