Cheng Zhang

Department of Electronic and Electrical Engineering, Imperial College London, London, SW7 2AZ

chengzhang98@outlook.com linkedin.com/in/chengzhang98 ChengZhang-98 (+44) 7536972519

Education

Imperial College London London, UK
PhD student in Electrical and Electronic Engineering (2nd Year) Jan. 2023 - Present
Research Interests: Efficient Machine Learning, AI Acceleration, Large Lanugage Models

University of Edinburgh Edinburgh, UK
MSc in Electronics, 72.1/100 Sep. 2021 - Aug. 2022
Project: Binarizing U-Net using Knowledge Distillation for Cell Segmentation

Beihang University Beijing, China
BEng in Automation, Top 10% Sep. 2021 - Aug. 2022
Project: Anomalous Behavior Detection in Surveillance Videos

Recent Projects

A Scalable and Modular Simulation Framework for AI Accelerator Systems ARIA
AI Accelerator, ISA Design, LLM, Pretraining, Kernel Fusion Sep. 2024 – Current

An project funded by Advanced Research and Invention Agency (ARIA) under the Scaling Compute program, in collaboration with with hardware team at University of Cambridge and compiler team at University at Edinburgh.

An Analytical Framework for Quantization Error Reconstruction Imperial College London
LLM, Parameter-Efficient Fine-Tuning, Post-Training Quantization May. 2024 – Sep. 2024

An analytical solution to quantization error reconstruction problem that benefits qLoRA-style parameter-efficient fine-tuning and post-training quantization and its computationally-efficient approximated form.

Hardware and Software Platform Inference University of Cambridge
ML Security, AI Governance, Text Generation, Image Classification April. 2024 – Nov. 2024

A classification framework capable of accurately identifying the GPU used for model inference as well as the underlying software configuration by only analyzing the numerical patterns in the model's outputs.

Experience

Rigpa AI, Technical Consultant May. 2024 – Current
LLM Inference, Software-Hardware Co-optimization for LLM Accelerator, Quantization.

Providing advice and solutions to LLM inference workload and potential optimization; Performing software-emulated LLM compression as the reference model for hardware verification

International Centre for Spatial Computational Learning, Jan. 2023 – Current
Research Student
Deep Learning for Non-Traditional Computer Architectures.

Performing software-hardware co-design for ML workload, in collaboration with experts and PhD students from Imperial College, University of Toronto, University of California Los Angeles, University of Southampton, and Industry.

Imperial College London, Teaching Assistant Jan. 2024 – Current
Advanced Deep Learning Systems

Developing teaching materials and supporting teachers of Advanced Deep Learning Systems, a module offered by EEE department on deep learning, compression, ML compiler, and custom hardware design .

Publications

Cheng Zhang, Jeffrey Wong, Can Xiao, George A Constantinides, Yiren Zhao. *QERA: an Analytical Framework for Quantization Error Reconstruction* (Under Review)

Cheng Zhang, Hanna Foerster, Robert D. Mullins, Yiren Zhao, Ilia Shumailov. *Hardware and Software Platform Inference* (Under Review)

Eleanor Clifford, Adhithya Saravanan, Harry Langford, **Cheng Zhang**, Yiren Zhao, Robert Mullins, Ilia Shumailov, Jamie Hayes. *Locking Machine Learning Models into Hardware*. The 3rd IEEE Conference on Secure and Trustworthy Machine Learning (IEEE SatML2025).

Cheng Zhang, Jianyi Cheng, George A. Constantinides, and Yiren Zhao. *LQER: Low-Rank Quantization Error Reconstruction for LLMs*. Proceedings of the 41st International Conference on Machine Learning, PMLR 235:58763-58779, 2024 (ICML2024).

Cheng Zhang, Jianyi Cheng, Ilia Shumailov, George A. Constantinides, and Yiren Zhao. *Revisiting Block-based Quantisation: What is Important for Sub-8-bit LLM Inference?* In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9988–10006, Singapore. Association for Computational Linguistics (EMNLP2023).

Zhewen Yu, Sudarshan Sreeram, Krish Agrawal, Junyi Wu, Alexander Montgomerie-Corcoran, **Cheng Zhang**, Jianyi Cheng, Christos-Savvas Bouganis, Yiren Zhao. *HASS: Hardware-Aware Sparsity Search for Dataflow DNN Accelerator*. The 34th International Conference on Field-Programmable Logic and Applications, pages 257-263, Italy (FPL2024).

Yuang Chen, **Cheng Zhang**, Xitong Gao, Robert D Mullins, George A Constantinides, Yiren Zhao. *Optimised Grouped-Query Attention Mechanism for Transformers*. Workshop on Efficient Systems for Foundation Models II at ICML2024 (ES-FoMo-II 2024)

Zixi Zhang, **Cheng Zhang**, Xitong Gao, Robert D Mullins, George A Constantinides, Yiren Zhao. *Unlocking the Global Synergies in Low-Rank Adapters*. Workshop on Efficient Systems for Foundation Models II at ICML2024 (ES-FoMo-II 2024).

Cheng Zhang, Jianyi Cheng, Zhewen Yu, Yiren Zhao. *MASE: An Efficient Representation for Software-Defined ML Hardware System Exploration*. Workshop on ML for Systems at the 37th Annual Conference on Neural Information Processing Systems (MLSys Workshop at NeurIPS2023).

Expertise

Programming Languages: Python, CUDA, C++, Verilog, Bash

Libraries: PyTorch, HuggingFace (Transformers, PEFT), CUTLASS, Triton, Pandas

Tools: Triton, NSight Compute, CMake, Git, VSCode, Verilator

Honors & Extracurricular Activities

ARIA Funded PhD programme on Scaling Compute for Machine Learning	Aug. 2024 - Current
University-level scholarship for academic excellence, Beihang University	Jun. 2021
Development Manager of Art Society at Beihang University	Sep. 2019- Sep. 2020