Dayou DU

ddu487@connect.hkust-gz.edu.cn | DD-DuDa

RESEARCH INTERESTS

- Hardware-software co-design for efficient deep learning system
- Model compression (pruning, quantization, knowledge distillation)

EDUCATION

The Hong Kong University of Science and Technology, Guangzhou, China — Sept.2022 - Present
Master of Philosophy in Data Science and Analytics — GPA: 3.97/4.00
Supervisor: Prof. Xiaowen Chu

South China University of Technology, Guangzhou, China — Sept.2018 - July 2022
Bachelor of Engineering in Intelligent Science and Technology — GPA: 3.63/4.00
Graduation project: Design and implementation of automatic annotation platform based on multi-target tracking. [Code]

ACADEMIC EXPERIENCE

Microsoft Research Asia — Beijing, China
Research Intern in Systems Research Group — July 2023 - Dec. 2023

PROJECTS

Research on Model Quantization and Knowledge Distillation for Large Language Models

- Developed a framework integrating Quantization-Aware Training (QAT) with Knowledge Distillation to enhance Large Language Models (LLMs) efficiency at ultra-low precisions (below 4-bit), achieving state-of-the-art results in language understanding and complex reasoning benchmarks.
- Designed an innovative distillation objective function and integrated tailored quantization techniques, leading to faster convergence and enhanced model performance.
- Implemented a low-bit LLM inference kernel based on Triton and CUDA, resulting in over 3× memory reduction and 4× speed up than the FP16 implementation.

Research on model quantization for Vision Transformer

- Conducted a comprehensive literature review on model quantization techniques applicable to vision transformers.
- Evaluated various quantization approaches for their effectiveness in enhancing accuracy and accelerating inference of quantized models.

Research on Benchmarking and Dissecting the Nvidia Hopper GPU Architecture

- Leveraged the Transformer Engine to explore FP8 precision in Hopper's Tensor Core, focusing on benchmarking matrix multiplication acceleration, transformer layers, and inference performance of Large Language Models.

Accelerating Deep Neural Networks with Sparse Tensor Core

- Developed a deep learning framework integrating GPU acceleration using PyBind11.
- Executed structured pruning and enhanced sparse matrix multiplication using Sparse Tensor Core on Ampere architecture.

PUBLICATIONS

- **Dayou Du**, Yijia Zhang, Shijie Cao, Jiaqi Guo, Ting Cao, Xiaowen Chu, Ningyi Xu, "BitDistiller: Unleashing the Potential of Sub-4-Bit LLMs via Self-Distillation," under review. [Paper], [Code]
- **Dayou Du**, Gu Gong, Xiaowen Chu, "Model Quantization for Vision Transformer: A survey". (In preparation)
- Yijia Zhang, Sicheng Zhang, Shijie Cao, **Dayou Du**, Jianyu Wei, Ting Cao, Ningyi Xu, "AFPQ: Asymmetric Floating Point Quantization for LLMs," under review. [Paper], [Code]
- Weile Luo, Ruibo Fan, Zeyu Li, **Dayou Du**, Qiang Wang, Xiaowen Chu, "Benchmarking and Dissecting the Nvidia Hopper GPU Architecture," IPDPS 2024. [Paper]

AWARDS

- Outstanding Graduation Thesis Award, 2022.
- Bronze Medal, Kaggle: TensorFlow - Help Protect the Great Barrier Reef, Object Detection, 2022.
- Silver Medal, Kaggle: CommonLit Readability Prize, Text Classification, 2021.
- The First Prize Scholarship (2021), The Third Prize Scholarship (2020), The SDL Scholarship (2019)

SEVICES

- TA for Advanced Machine Learning, HKUST(GZ) DSAA5013.

SKILLS

- **Programming:** Python, C++, Shell, Docker
- **Deep Learning:** CNN, RNN, Vision Transformers, Large Language Models
- **Platform:** PyTorch, TensorRT, Deepspeed, OpenPPL
- **System:** Openai Triton, CUDA, CuBLAS, cuSPARSELt, OpenMP
- **Language:** Mandarin, English, Cantonese