YIRAN DING (丁羿然)

yiran.ding2@gmail.com ◇ Hangzhou, China ◇ Website ◇ Blog

RESEARCH INTEREST

LLM; MLsys; HPC; CV; Computer Architecture

EDUCATION

School of Electronics & Information, Hangzhou Dianzi University (HDU) Sep. 2021 - Jun. 2024
School of Mathematics, Hangzhou Dianzi University Sep. 2020 - Jun. 2021
GPA: 3.8/4.0 (90/100, 3%)

RESEARCH EXPERIENCE

LLM Sequence Extension ***MSRA*** July 2023 - Now

- Pioneered a groundbreaking **interpolation** technique for RoPE, significantly extending the sequence length of the Llama model to 32K with flash-attention, all without the need for fine-tuning.
- Successfully conducted evaluations on various downstream tasks, including **Passkey Retrieval** and **Quality**(reading comprehension).

LLM Inference Optimization ***HDU*** March 2023 - Now

- Developed a novel **block schedule** method by granularizing batches into layers, which has the potential to theoretically improve throughput and latency by **2x** compared to current best block schedules.
- **Compressed** weights, KV cache, and activation into **4 bits** without significant accuracy loss through **clustering, reordering**, and using **sparse attention** to reduce memory consumption.

Medical Image Processing: ***HDU*** Nov. 2021 - Aug. 2022

- Led and designed the project of automatically evaluating finger tapping videos of Parkinson' s disease patients.
- Developed **LSTM-FCN** based model to classify patients. The result has 83.7% accuracy, which in dataset of this paper defeats the state-of-the-art results in literatures.
- **Utilized**: Pose estimation (Mediapipe Hands), RIFE algorithm (Time Series Interpolation), LSTM, FCN.

OTHER EXPERIENCE

[**An Open Reproduction of LLaMA2**](GitHub): Participate in the development of an industrial-grade LLaMA2 SFT/RLHF training framework utilizing DeepSpeed and Ray for distributed training.

[**LLM inference in Edge Device**]: Developed an **offline** LLM based on the **7B Alpaca model**. Implemented **Chinese Q&A** and dialogue functions, and deployed on an 8GB edge device with 16Tops computing power in int8. Expanded the Chinese vocabulary, **fine-tuned** the model with Chinese instruction data and utilized **int4** quantization to compress the model, significantly improving its understanding and execution of Chinese instructions.

[**DGEMM**] (Report): Implemented and optimized various matrix multiplication techniques for improved performance, including **block-wise**, **recursive**, and **cache-oblivious** approaches, reducing computation time by up to **82%**. Improved data access by reordering matrix data in **Z-morton pattern** for better cache utilization.

[**Forest Management**] (MCM/ICM 2022 E): Use mathematical modeling to create optimal **forest management** plans, considering factors such as carbon sequestration, tree growth rates, and economic value to maximize the forest's integrated value. Techniques include **logistic regression**, **Monte Carlo simulation**, and single-objective planning.

[**Optimizing Ride-Sharing Services**]: Develop an **online model** that considers matching customers and suppliers in a large-scale ride-hailing service, using **greedy** and **simulated annealing** algorithms. The models achieve **high satisfaction** rates and demonstrate **strong stability** and **scalability**.

AWARDS AND ACTIVITIES

Scholarship: The First Prize Scholarship (Four semesters), Award rate 5%. Scholarship of Provincial Government, Award rate 5%

Activities: Taught new students about programming skills such as Python, Matlab, etc. Instructed them to solve NP-hard Graph Theory Problems with Heuristic Algorithms, and Time Series Forecasting Problems with LSTM Neural Networks.