YIRAN DING

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

RESEARCH INTEREST

MLsys; HPC; CV; Parallel Computing; Distributed System; Computer Architecture

EDUCATION

School of Electronics & Information, Hangzhou Dianzi University Sep. 2021 - Jun. 2024
School of Mathematics, Hangzhou Dianzi University Sep. 2020 - Jun. 2021
GPA: 3.8/4.0 (90/100, 5%)

RESEARCH EXPERIENCE

LLM Inference Optimization March. 2023 - Now

- Developed a novel **diagonal 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.
- These techniques offer **greater batch size**, significantly increasing maximum throughput and have promising applications in delay-insensitive scenarios, such as **long sequences tasks**.

Medical Image Processing: 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

[**LLM inference in Edge Device**]: Developed an **offline** large language model based on the **7B Alpaca model** to address privacy and security concerns with cloud deployment. Implemented **Chinese Q&A** and dialogue functions, tested against similar models, 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.

[**Integrated Forest Management**] (MCM/ICM 2022 E): Use mathematical modeling to create optimal forest management plans. Considers 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, and the model is applied to a specific forest for effectiveness.

[**Optimizing Ride-Sharing Services**]: Analyzes the problem of matching customers and suppliers in a large-scale ride-hailing service using **greedy** and **simulated annealing** algorithms. Develop an **online model** that considers various factors, such as customer satisfaction, availability, and route optimization. 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.