Yiran Ding

✉ yiran.ding2@gmail.com | github.com/Yiyi-philosophy | Website | Blog | google scholar | Yiran Ding

EDUCATION

Hangzhou Dianzi(Electronic and Technology) University 2021/09 - 2024/06

Bachelor of Engineer Electronics & Information Hangzhou, Zhejiang, China

- GPA: **3.8**/4.0 (**90**/100, Top 3%)
- The First Prize Scholarship (Four semesters), Award rate 5%. | Scholarship of Provincial Government, Award rate 2%

RESEARCH INTERESTS & SKILLS

- LLM:
  - NLP: Evaluation, Data Engineering, SFT
  - MLSys: Inference Optimization, Finetuning
  - Architecture: Transformer, Mamba
- Skills
  - Python(Pytorch), C/C++, Matlab
  - OpenMP, MPI, CUDA,
  - Git, Shell, Docker, Conda | Verilog

PUBLICATIONS

- **LongRoPE: Extending LLM Context Window Beyond 2 Million Tokens. Y. Ding**, L. L. Zhang, C. Zhang, Y. Xu, N. Shang, J. Xu, F. Yang, M. Yang. (2024). *Forty-first International Conference on Machine Learning* (**ICML**). [Paper]

RESEARCH EXPERIENCE

LLM Sequence Extension: LongRoPE 2023/06 - 2024/07

Intern, Microsoft Research Asia (MSRA), advised by Li Lyna Zhang Beijing, China

- Extends the context window of pre-trained LLMs(Llama, Mistral) to **2048k** tokens with up to only **1k fine-tuning steps** at 256k training lengths, maintaining original performance.
- Exploits **non-uniformities in positional interpolation** for better fine-tuning initialization, uses a progressive extension strategy, and **readjusts** LongRoPE to **recover short context** window performance.
- Supported fine-tuning of **Phi-3**(mini, small) to **128k contexts**: Phi-3 Model, Phi-3 Report
  - Prepare and clean 128k-length datasets from different sources to finetuning, and methods to recover short context (4k) performance.

LLM Inference Optimization, advised by Prof. 2023/03 - 2023/07

HDU, advised by Prof. Zheng Miao Hangzhou, China

- 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 2023/03 - 2023/07

HDU, advised by Prof. Zhu Li Hangzhou, China

- Led and designed the project of automatically evaluating finger tapping videos of Parkinson's disease patients. item 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. item **Utilized**: Pose estimation (Mediapipe Hands), RIFE algorithm (Time Series Interpolation), LSTM, FCN.

OTHER EXPERIENCE

LLM inference in Edge Device 2023/07 - 2023/09

- 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) 2023/07 - 2023/09

- 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.

Last Updated on August 28, 2024