Pingwei SUN

(+86)18641202035 | psunah@connect.ust.hk | polarispw.github.io/ | github.com/polarispw

Education

The Hong Kong University of Science and Technology *Hong Kong SAR, China*

Master of Science in Big Data Technology *Aug 2023 - Jan 2025*

- **Topics of interest:** Efficient LLM, Compression and acceleration of LLM

Northeastern University *Liaoning, China*

Bachelor of Engineering in Artificial Intelligence *Sep 2019 - Jul 2023*

- **GPA: 90.23/100**
- **Courses:** Probability Theory and Statistics, Optimization Theory, Data Structure and Algorithm, Computer System, Pattern Recognition, Machine Learning, Deep Learning, Natural Language Processing
- **Extracurricular Activities:** Leader of the official visual studio of the Northeastern University

Skills

Languages Python, C/C++, Verilog

Tools Pytorch, Deepspeed, MNN, TNN, llama.cpp, Git, Docker, Huggingface, Vivado

Experience

iSING Lab, Department of CSE, HKUST *Hong Kong SAR, China*

Research Intern *Sep 2023 - Now*

- **Research topic: APWSVD**: Efficiently Compress LLM with Activation and Parameter Weighted SVD
- **Contributions:** It reduces more than 50% of GPU memory footprint and 40% of compression time while ensuring the post-compression performance is competitive with **LLM-Pruner** and **SliceGPT**.

Tensor Lab, OPPO Research Institute *Beijing, China*

Algorithm Engineer (Intern) *Mar 2023 - Jul 2023*

- **High-performance deployment:** Optimize the performance bottleneck of BERT-like models on the end-side due to sliced accesses, and port a GEMM kernel for self-developed NLP inference frameworks (ARM CPU)
- **LLM on edge side:** Investigate LLM end-side deployment solution and build a demo based on LlaMA, covering model quantization and inference framework (it is now released in the latest ColorOS).

School of CSE, Northeastern University *Liaoning, China*

Teaching Assistant of Computer System (22fall) *Sep 2022 - Jan 2023*

- **Contributions:** Lab environments setup, delivering lectures, and grading for undergraduates.

Projects

Fine-tuning vs Prompting, Can Language Models Understand Human Values? *Hong Kong SAR, China*

Department of CSE, HKUST *Sep 2023 - Nov 2023*

- Based on the Sem-eval2023 human values dataset, methods such as classifier fine-tuning and end-to-end fine-tuning (Prompt, CoT) are involved in this project to validate the comprehension capabilities of PLMs in understanding human values at different scales (125M to 7B) and structures.

High-performance CPU Design and AI Application Based on MIPS ISA *Liaoning, China*

School of CSE, NEU *Jan 2023 - Jun 2023*

- A dual-issue six-stage pipeline CPU based on MIPS was built, and it cooperates with a CNN acceleration core to process the MINIST dataset. All designs are coded in **Verilog** and implemented on **FPGA** board. It is also my graduation thesis, which received excellent reviews from professors.

BERT Based Sentiment Analysis *Liaoning, China*

NEU NLP-Lab *Jun 2022 - Aug 2022*

- Based on the BERT model, this project enhances the performance of the model on extremely unbalanced data sets through a variety of fine-tuning methods and a **contrastive learning** task.

Honors (selected)

2020 **National Scholarship**, Ministry of Education of the PRC *Liaoning, China*

2020 **First-class Scholarship**, Northeastern University *Liaoning, China*

2021 **Second-class Scholarship**, Northeastern University *Liaoning, China*

2022 **Third Prize in "Loongson Cup"**, National Student Computer System Capability Challenge *Beijing, China*