Jie Li

School of Computer Science, Nanjing University
State Key Laboratory for Novel Software Technology
Advisor: Prof. Wu-Jun Li
jie-li@smail.nju.edu.cn
+86-13373291169

Education

Nanjing University, School of Computer Science, Computer Science and Technology, *Master* 2023.9 – Present

Harbin Institute of Technology, Faculty of Computing, Computer Science and Technology, *Bachelor* 2019.9 – 2023.6

Experience

Research: Automatic Parallelism Method UniAP 2023.12 – Present

- *Description*: We propose a novel automatic parallelism method that jointly optimizes intra-layer and inter-layer parallelism. Previous methods either optimize only one category of parallel strategies or use a hierarchical optimization approach. Our method unifies both categories into an MIQP (Mixed Integer Quadratic Programming) problem to obtain a global optimum. The framework consists of three parts: performance profiling, cost model, and search algorithm.
- *Duty*: Implement memory profiling and memory cost model. Add support for Llama model and conduct experiments on large-scale clusters (up to 64 cards). Adapt the framework to Hygon DCU.
- *Deliverable*: Experiments on 4 environments and 5 Transformer-based models show up to 3.8× throughput improvement and up to 107× search speed improvement over prior state-of-the-art methods. The method has been summarized into a paper, which has been accepted by CVPR 2025 (**Oral**), with myself as **co-first author**. This paper is selected as an **Award Candidate** (only 14 out of all submissions, 0.1%). Preprint available on arXiv.

Project: Core Algorithms and Theoretical Research for Task Solving in Computing Networks 2024.9 – Present

- *Description*: Led by Academician Zongben Xu, this is a collaborative project among Nanjing University, Fudan University, and Peng Cheng Laboratory. Our team focuses on distributed training algorithms. I serve as the primary student leader on our side.
- *Duty*: Adapted our UniAP method to the Ascend NPU cluster provided by Peng Cheng Laboratory.
- *Deliverable*: On a 32-card Ascend 910a cluster, using Megatron to train Llama2-7B, UniAP-generated strategies achieve up to 1.61× throughput improvement compared to empirically manual settings.

Research: UniAP-4D, Automatic Parallelism for 4D Parallelism 2024.7 – Present

- *Description*: To cope with increasing sequence lengths in large model training, researchers introduced sequence parallelism, which can be combined with 3D parallelism to form 4D parallelism. Existing works only explore automation for 3D; our work, UniAP-4D, extends UniAP to 4D and implements a training system supporting 4D automatic parallelism.
- *Duty*: This work is basically completed by myself independently. Work includes extending the training framework to support RingAttention and DeepSpeed Ulysses sequence parallelism and their arbitrary combinations with other strategies, and extending UniAP in terms of performance profiling, cost model, and search.
- *Deliverable*: Finish training framework extension and validate correctness on 4-card A6000 with small models via stable loss curves under different strategies. Automatic parallel algorithm extensions mostly completed; further large-scale testing in progress.

Project: Chinese Medical LLM 2023.10 – Present

- *Description*: Built upon Baichuan2-13B, use Llama-factory and DeepSpeed for incremental pretraining on medical textbooks and public medical data, followed by fine-tuning and evaluation on downstream tasks.
- *Duty*: Conduct surveys on medical capabilities of open-source Chinese LLMs. Fine-tune on DISC-Med-SFT. Evaluate on PromptCBLUE and CMB benchmarks. Accelerate training by UniAP.
- *Deliverable*: Based on this model, a medical image report generation model fine-tuned on real hospital data has been deployed in Nanjing Drum Tower Hospital. Training of a multimodal model based on Llava is underway.

Project: Power Grid Anomaly Traffic Detection LLM 2024.4 – Present

- *Description*: In collaboration with the State Grid Corporation of China, we train a model to detect malicious traffic within the company’s daily network flows. The base model is CodeQwen1.5-7B.
- *Duty*: Survey LLMs in this domain and select the most effective base model. Preprocess daily traffic data into LLM-compatible format. Select public dataset for fine-tuning/evaluation. Accelerate training by UniAP.
- *Deliverable*: Still in progress. Preliminary results show an F1 score close to 0.9 on data_capec_multilabel, surpassing traditional methods (approx. 0.8). The company deems the model deployment-worthy.

Competition: ACM Programming Contest 2021.6 – 2022.4

- *Description*: Participated in one season of ACM competitions during junior year.

- *Duty*: Our team was balanced in ability; contributions were evenly distributed.
- *Deliverable*: Silver Medal, ICPC Asia Regional (Kunming).

Skills

Distributed Training: Proficient with Megatron framework and its underlying implementation; our UniAP framework is built on Megatron. Also experienced with DeepSpeed for LLM training.

Large Language Model: Familiar with mainstream LLM architectures. Experienced with Llama-factory for pretraining, fine-tuning, and inference.

Machine Learning and Deep Learning: Grasp of ML and DL algorithms.

NPU and DCU using: Experienced in distributed LLM training on Ascend NPUs and Hygon DCUs.

Programming Languages: Python and C++.

Honors and Awards

Publications:

UniAP: Unifying Inter- and Intra-Layer Automatic Parallelism by Mixed Integer Quadratic Programming (Co-first Author, CVPR 2025, Oral, Award Candidate)	2025

Competitions:

46th ICPC Asia Kunming Regional Silver Medal (Ranked 56/667)	2022
Third Prize, 14th iCAN International Contest of Innovation (China Finals)	2020
First Prize, National Olympiad in Informatics in Provinces, NOIP2017 Advanced Group	2017