Jinyu He

17671747126 tommy514@foxmail.com
24 Wuhan, Hubei
Intended position: Quant Developer, Algorithm Engineer (AI Infra, HPC direction)

EDUCATION

Master of Computer Science, Xiamen University 985 211 Double 1st-Class 2022.09 - 2025.06

- Supervisor and Laboratory: Professor Ji Rongrong, Associate Professor Zheng Xiawu; Media Analysis and Computing Laboratory (Ministry of Education Key Laboratory)
- Major Research Areas: **Model Compression** (Quantization, Pruning), **AutoML** (Meta-Learning, Hyperparameter Search)
- Other Research Experience: Efficient Fine-Tuning, Large Language Models (LLMs), Multimodal Learning
- GPA:3.9/4.0 **Ranking: 3/118** Outstanding Student Scholarship of Xiamen University **TOEFL: 103**

Bachelor of Computer Science, Hohai University 211 Double 1st-Class 2018.09 - 2022.06

- Student work experience: as **ACM captain** of innovation laboratory; Participate in maintaining OJ and organize ICPC training for the whole team; Responsible for the school competition twice, and the number of participants exceeded **100.**
- Algorithm Competition Ability: Codeforces **Rating: 2115** LeetCode Rating: 2257 (**Top1.4%**) CCF-CSP 390 (**Top 0.7%**)
- GPA:4.7/5.0 **Ranking: 5/242** **National Scholarship** CET-6: 555 CET-4: 565

INTERNSHIP

Kendall Square Captital (Beijing) - Quantitative Algorithm Development Intern 2024.06 - Present

- Summary: Accelerated inference and training for LightGBM models. Optimized hyperparameter search to improve LightGBM model performance.
- Inference acceleration: Developed CUDA operators to accelerate the inference process of LightGBM models. Leveraged GPU's shared memory and texture memory in CUDA kernel functions to achieve maximum acceleration. Achieved a **1000x speedup** in inference time compared to single-core CPU, reducing inference time from **42 seconds to 42 milliseconds**.

NIO (Shanghai) Autonomous Driving Perception Group - AI Infra Intern 2024.02 - 2024.05

- Summary: Study the model structure, training framework and delivery process in the optimization part, **improve** training **efficiency** and quasi-model **performance**
- Model structure optimization, participate in od task head reconstruction, use torch profiler analysis and reduce cuda sync, improve training parallelism and SM utilization (**15% -> 25%**)
- Training efficiency optimization, responsible for the operator requirements of the **linear sum distribution problem** (linear assignment), torch implements the **Hungarian algorithm** to solve the problem; The design of cuda operator scheme uses **auction algorithm** to improve parallel efficiency. model side merging batch uses cuda operator calculation to improve model training throughput (**7.9 -> 12.2** samples/s)
- Model delivery optimization, based on effective objectives to optimize the **training quantization** calibration set, improve the multi-task model quantization accuracy, tld task drop points significantly reduced (**-3% -> -1%**)

PUBLICATION

GreedyAgent: Crafting Efficient Agents for MetaLC via Greedy Algorithm Selection. **(ICIC 2024 Oral) First author**

- The paper is based on the **learning curve for the meta-learning** competition @ AutoML Conf. 2023 **winner**, I am the **competition captain**, the code has been open source
- Problem: the training curve for meta-learning, in the intelligent body-environment design algorithm selection strategy, focus on training cost limited conditions of the model all-time performance
- Methods: The original problem is summarized as part of the condition of unknowable 0-1 knapsack problem, based on greedy thinking, according to the cost-effective index design agent algorithm selection strategy
- Results: In the Codalab final evaluation stage of 30 data sets, the average ALC index is **much higher than all participating teams (0.32 -> 0.39)**

Towards Generalized and Parameter Efficient Network Pruning in Transfer Learning in **(ECCV 2024 in voting) Co-first author**

- This paper proposes a novel approach that combines **structured pruning** with **efficient fine-tuning**, significantly enhancing the performance of **pruned pre-trained large models**.
- Method: Introduced **reconstruction loss (65.4 -> 70.2)** and **iterative pruning (70.2 -> 72.8)** into the structured pruning process.

Combined these techniques with the SSF efficient fine-tuning method.

- Experiments: Conducted extensive comparative and ablation experiments on VTAB-1k, FGVC, and ScienceQA datasets. Covered transfer learning tasks in computer vision (CV) and multimodal (MM) tasks.
- Results: Achieved accuracy improvement on CV tasks with ViT-B/16 compared to existing pruning methods **(70.9 -> 72.8)**. Maintained accuracy while achieving a 30% pruning rate on MM tasks with LaVIN-7B.

COMPETITION

Kaggle Competition-LLM Science Exam (Silver, Top 3%) 2023.07 - 2023.10

- Introduction: Fine-tuning LLM, reasoning on single-choice questions in scientific contexts, public data sets LLM Science MAP @ 3 points **0.905 (Top 3%)**
- Complementary crawling of the 50k Wikipedia dataset, LoRA fine-tunes the llama2-13b model, and uses it as a teacher model to distill knowledge of the Deberta model. **(0.792 -> 0.895)**
- Deploy the Deberta model trained in three different configurations for inference, and use the voting mechanism to obtain the output after feature fusion. **(0.895 -> 0.905)**

AWARDS / SKILLS

- ACM Awards: **ICPC Asian Regional Finals Bronze** Award, ICPC Nanjing/Shenyang/Yinchuan/Kunming Regional Bronze Award, **CCPC Weihai Bronze** Award, **CCPC Jiangsu Division Silver Award Second**, **GPLT Ladder Individual First Prize**, Blue Bridge Cup C ++ (Group A) National Second Prize, CCSP National Finals Bronze Award, **CCSP East China Division Gold** Award
- Academic Awards: Meta-learning Competition from Learning Curve @ AutoML Conf. 2023 **World Champion**, Kaggle Competition **Silver Medal (Top 4%)**
- Programming skills: common languages C ++, Python; Familiar with PyTorch; Familiar with Linux, Git basic operation; Understand CUDA, TensorRT