Daize Dong

Personal Page | 🎓 Google Scholar | ✉ dzdong2019@gmail.com

Personal Information

I am a first-year Ph.D. student in Computer Science at Rutgers University, New Brunswick, advised by Prof. Hongyi Wang. My research focuses on improving the efficiency of neural networks across algorithmic and system levels. I am interested in the following areas:

1. **Efficient Architectures:** Mixture of Experts, Sparse Activation Models.
2. **Model Compression:** Pruning, Quantization.
3. **High-Performance Computation:** System Optimization, Efficient Kernels.

Education

Rutgers University, New Brunswick *Sep. 2025 – Present*
Ph.D. in Computer Science

University of Electronic Science and Technology of China *Sep. 2019 – Jul. 2023*
B.E. in Computer Science & Mathematics

Work Experience

Zhipu AI – *Junior Researcher* *Sep. 2024 – Jun. 2025*
Mixture of Experts, LLM Pretraining

Shanghai Artificial Intelligence Laboratory – *Research Assistant* *Jul. 2023 – Aug. 2024*
Mixture of Experts, Sparse Activation, Large Language Models

Westlake University – *Research Assistant* *Apr. 2023 – Aug. 2024*
Graph Transformers, Molecule Generation, AI for Biology

JD Explore Academy – *Research Intern* *Feb. 2022 – Oct. 2022*
Network Pruning, Quantization, Model Compression

Publications

1. **Towards Efficient Mixture of Experts: A Holistic Study of Compression Techniques.** [Paper]
Shwai He*, **Daize Dong***, Liang Ding, Ang Li.
Transactions on Machine Learning Research (TMLR).

2. **DLO: Dynamic Layer Operation for Efficient Vertical Scaling of LLMs.** [Paper]
Zhen Tan*, **Daize Dong***, Xinyu Zhao, Jie Peng, Yu Cheng, Tianlong Chen.
First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models (ICLR 2025 Workshop).

3. **Dynamic Data Mixing Maximizes Instruction Tuning for Mixture-of-Experts.** [Paper]
Tong Zhu, **Daize Dong**, Xiaoye Qu, Jiacheng Ruan, Wenliang Chen, Yu Cheng.
2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025).

4. **A Graph is Worth K Words: Euclideanizing Graph using Pure Transformer.** [Paper]
Zhangyang Gao*, **Daize Dong***, Cheng Tan, Jun Xia, Bozhen Hu, Stan Z. Li.
The 41st International Conference on Machine Learning (ICML 2024).

* Equal Contribution

5. **iDAT: inverse Distillation Adapter-Tuning.** [Paper]
   Jiacheng Ruan, Jingsheng Gao, Mingye Xie, **Daize Dong**, Suncheng Xiang, Ting Liu, Yuzhuo Fu.
   2024 IEEE International Conference on Multimedia and Expo (ICME 2024). ***(Oral)***

6. **PAD-Net: An Efficient Framework for Dynamic Networks.** [Paper]
   Shwai He, Liang Ding, **Daize Dong**, Boan Liu, Fuqiang Yu, Dacheng Tao.
   Proceedings of The 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023).

7. **SD-Conv: Towards the Parameter-Efficiency of Dynamic Convolution.** [Paper]
   Shwai He, Chenbo Jiang, **Daize Dong**, Liang Ding.
   IEEE/CVF Winter Conference on Applications of Computer Vision, 2023 (WACV 2023).

8. **SparseAdapter: An Easy Approach for Improving the Parameter-Efficiency of Adapters.** [Paper]
   Shwai He, Liang Ding, **Daize Dong**, Miao Zhang, Dacheng Tao.
   Findings of The 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022).

Projects

LLaMA-MoE: Building Mixture-of-Experts from LLaMA with Continual Pre-training. [Paper] [Code]
The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024).

LLaMA-MoE v2: Exploring Sparsity of LLaMA from Perspective of Mixture-of-Experts with Post-Training. [Paper] [Code]
ArXiv Preprints.

Deep Learning Skills

Development Envs: Linux, Docker, Slurm.

DL Tools & Libraries: PyTorch, Transformers, DeepSpeed.

Training Frameworks: Megatron, Torch Lightning.

Inference Frameworks: vLLM, SGLang.