Jingzhi Fang (方竟志)

✉ jfangak@connect.ust.hk

Education

2019 – 2024	🔖	**The Hong Kong University of Science and Technology**
		Ph.D. of Computer Science and Engineering
2015 – 2019	🔖	**Beihang University**
		B.Sc. of Computer Science, SHENYUAN Honors College

Research Interest

I am interested in accelerating the inference of AI models. I have done some work to speed up the deep neural network (DNN) inference with compiler optimizations (i.e., improving the inference efficiency without changing the model output) on the computation graph level and the operator level. I am now working on improving the inference throughput of large language models (LLMs).

Research Publications

1. J. Fang, Y. Shen, Y. Wang, and L. Chen, "Stile: Searching hybrid sparse formats for sparse deep learning operators automatically," *Proceedings of the ACM on Management of Data*, vol. 2, no. 1, pp. 1–26, 2024.

2. Y. Shen, L. Chen, J. Fang, X. Zhang, S. Gao, and H. Yin, "Eficient training of graph neural networks on large graphs," *To appear in: Proc. VLDB Endow.*, vol. 17, no. 12, pp. 4237–4240, 2024. 🔗 DOI: 10.14778/3685800.3685844.

3. J. Fang, Y. Shen, Y. Wang, and L. Chen, "ETO: accelerating optimization of DNN operators by high-performance tensor program reuse," *Proc. VLDB Endow.*, vol. 15, no. 2, pp. 183–195, 2021. 🔗 DOI: 10.14778/3489496.3489500.

4. J. Fang, Y. Shen, Y. Wang, and L. Chen, "Optimizing DNN computation graph using graph substitutions," *Proc. VLDB Endow.*, vol. 13, no. 11, pp. 2734–2746, 2020. 🔗 URL: http://www.vldb.org/pvldb/vol13/p2734-fang.pdf.

Skills

Languages	🔖	English and Chinese.
Coding	🔖	Python, C, C++, CUDA, …
Misc.	🔖	Willing to face challenges and passionate about study.

Miscellaneous Experience

Awards and Achievements

2017	▌	Honorable Mention in Mathematical Contest In Modeling.
2018	▌	Second prize of The 28th Beihang University "Fengru Cup" Competition.
2022	▌	HKUST RedBird Academic Excellence Award for Continuing PhD Students in 2021/22.

Teaching

2020	▌	Teaching assistant for COMP4331 Data Mining in HKUST.
	▌	Teaching assistant for COMP1022P Introduction to Computing with Java in HKUST.
2021	▌	Teaching assistant for MSBD5002 Data Mining and Knowledge Discovery in HKUST.
2022	▌	Teaching assistant for COMP1021 Introduction to Computer Science in HKUST.
	▌	Teaching assistant for COMP5631 Cryptography and Security in HKUST.

Intern

| 2021 | ▌ | Research Intern in Huawei, working on the joint optimization from both the graph level and the operator level for higher DNN inference efficiency. |

Journal Reviewer

TKDE.