Shang Yang

Email: shangy@mit.edu Homepage: https://ys-2020.github.io/

EDUCATION

Massachusetts Institute of Technology (MIT) **Cambridge, MA**
Ph.D. Student in EECS Department *Sept. 2023 - Present*

- Advised by Prof. Song Han

Tsinghua University **Beijing, China**
Bachelor of Engineering in Electronic Information Science and Technology *Aug. 2019 - July 2023*

- Overall GPA: 3.99 / 4.0 Rank: 1 / 256

EXPERIENCE

Massachusetts Institute of Technology (MIT) **Cambridge, MA**
Research Assistant, Advised by Prof. Song Han. *July 2022 - Aug. 2023*
Topic: Efficient machine learning systems for ADAS and Large Language Models.

Tsinghua University **Beijing, China**
Research Assistant, Advised by Prof. Yu Wang *Feb. 2021 - Oct. 2022*
Topic: High-performance sparse computing kernels for graph analysis.

SELECTED PUBLICATIONS

TorchSparse++: Efficient Training and Inference Framework for Sparse Convolution on GPUs

- Haotian Tang*, **Shang Yang***, Zhijian Liu, Ke Hong, Zhongming Yu, Xiuyu Li, Guohao Dai, Yu Wang, Song Han.（*equal contributions)
- A powerful and user-friendly framework for efficient sparse convolution on GPU. Achieve an average of **1.7× end-to-end speedup** over the previous SOTA system in self-driving benchmarks, along with **2.6-7.6× faster inference speed** in graph analysis.
- **Accepted by MICRO'23. [Website] [Paper] [Presentation]**

AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration

- Ji Lin*, Jiaming Tang*, Haotian Tang†, **Shang Yang**†, Xinyu Dang, Chuang Gan, Song Han.
- A hardware-friendly algorithm for LLM low-bit weight quantization with negligible accuracy loss.
- An efficient and flexible inference framework tailored for on-device LLMs. More than **3× speedup** over the Hugging Face FP16 implementation on both desktop and mobile GPUs.
- **Accepted by MLSys'24. Significant industry and community impact**. Integrated into vLLM, FastChat, TensorRT-LLM, Hugging Face Transformers, and LMDeploy. **[Website] [Paper]**

Heuristic Adaptability to Input Dynamics for SpMM on GPUs

- Guohao Dai, Guyue Huang, **Shang Yang**, Zhongming Yu, Hengrui Zhang, Yufei Ding, Yuan Xie, Huazhong Yang, Yu Wang.
- Enlarged design space and auto-tuning technique for SpMM on GPUs. Achieve **1.3×** average speedup over NVIDIA cuSPARSE kernels, and up to **5.6×** end-to-end speedup in GNN applications.
- **Accepted by DAC'22. Best Paper Award Nominee. [Paper]**

SELECTED AWARDS AND HONORS

- **Freshmen Award** | *Tsinghua University* ***2019***
- **China National Scholarship (Top 0.2%)** | *Ministry of Education of People's Republic of China* ***2020***
- **Comprehensive Excellence Award (Highest honor, Top 3%)** | *Tsinghua University* ***2020***
- **Learning Excellence Award (Ranked 1st in 256)** | *Tsinghua University* ***2021 & 2022***