Hao Kang

my github
google scholar

personal link
my email

Education

Georgia Institute of Technology Aug. 2023 – Present
PhD. Student in Computer Science *Advisor: Tushar Krishna*
Zhejiang University Aug. 2019 – June. 2023
Bachelor in Computer Science with CKC Honor

Experience

Research Intern at MSR May. 2024 – Aug. 2024
efficient LLM and model compression *Mentor Srikant Bharadwaj*

- a paper accepted by MLsys 2025

Graduate Researcher at GT Aug. 2023 – Now
Efficient machine learning and LLM agent *Advisor Prof. Tushar Krishna*

Undergrad Researcher at UCLA Aug. 2022 – Mar. 2023
dataset distilling *Advisor Prof. Baharan Mirzasoleiman*

- a paper accept by ICML 2024

Undergrad Researcher at MIT Feb. 2022 – Aug. 2022
model compression and edge ml *Advisor Prof. Song Han*

- a 4k+ star Github repo
- Deploy model on cell phone with TVM android and pytorch mobile

Publications

Win Fast or Lose Slow: Balancing Speed and Accuracy in Latency-Sensitive Decisions of LLMs
LLM agents, efficient ml
Hao Kang, Qingru Zhang, Han Cai, Weiyuan Xu, Tushar Krishna, Yilun Du, Tsachy Weissman
Neurips 2025 Spotlight

TURBOATTENTION: EFFICIENT ATTENTION APPROXIMATION FOR HIGH THROUGHPUTS LLMS
efficient ml, hardware
Hao Kang, Srikant Bharadwaj, James Hensman, Tushar Krishna, Victor Ruehle, Saravan Rajmohan
Mlsys 2025

GEAR: An Efficient KV Cache Compression Recipe for Near-Lossless Generative Inference of LLM
model compression, efficient ml
Hao Kang*, Qingru Zhang*, Souvik Kundu, Geonhwa Jeong, Zaoxing Liu, Tushar Krishna, Tuo Zhao
NIPS ENLSP 2025 Best Paper Candidate

Effectively and Efficiently Combining Language Models
efficient ml, hardware
Chenyu Wang*, Zishen Wan*, **Hao Kang***, Zhiqiang Xie, Vijay Janapa Reddi, Tushar Krishna, Yilun Du
In submission

Towards Sustainable Learning: Coresets for Data-efficient Deep Learning
dataset distilling, efficient ml
Yu Yang, **Hao Kang**, Baharan Mirzasoleiman
ICML2024

AI Metropolis: Scaling Large Language Model Agent Interaction with Out-of-order Execution
LLM agents, efficient ml
Zhiqiang Xie, **Hao Kang**, Ying Sheng, Tushar Krishna, Kayvon Fatahalian, Christos Kozyrakis
Mlsys 2025

Privatar: Enabling Privacy-preserving Real-time Multi-user VR via Secure Outsourcing
efficient ml, ai security
Jianming Tong, Hanshen Xiao, **Hao Kang**, Edward Suh, Tushar Krishna
Mlsys 2025

Lvlm-compress-bench: Benchmarking the broader impact of large vision-language model compression
Ml efficiency, Benchmark
Souvik Kundu, Anahita Bhiwandiwalla, Sungduk Yu, Phillip Howard, Tiep Le, Sharath Nittur Sridhar, David Cobbley, Hao Kang, Vasudev Lal
Mlsys2024

Open-source Projects

THOP: PyTorch-OpCounter
a pytorch operator profiler which has over **4.8k** stars

GEAR
KV cache compression which has over **140** stars

Extracurricular

Research Interests
My research interests focus on making large models more efficient, including post-training compression techniques and structural design innovations. I can handle problems ranging from CUDA kernel development to high-level algorithm design. Recently, I have been particularly interested in improving the efficiency of Mixture of Experts (MOE) models and multi-agent systems. My goal is to ensure that my research brings real benefits to both academia and industry, bridging the gap between cutting-edge technology and practical applications.