HEXU ZHAO

+1 9296897240 · hz3496@nyu.edu · Personal Homepage

Firm faith in analytical thinking, hard work, and consistent self-improvement

Education

Courant Institute, New York University New York, USA
Ph.D in ML system, advised by Prof. Jinyang Li and Prof. Aurojit Panda 2023 - present

Honored Yao Class, Tsinghua University Beijing, China
Bachelor of Computer Science 2019 - 2023

Publications

On Scaling Up 3D Gaussian Splatting Training
Hexu Zhao, Haoyang Weng, Daohan Lu, Ang Li, Jinyang Li, Aurojit Panda, Saining Xie
*Homepage, Github Code **(350+ Stars),** twitter, 量子位. Adopted by gsplat etc*
Under Submission to ICLR2025

On Optimizing the Communication of Model Parallelism
Yonghao Zhuang*, **Hexu Zhao*(co-1st author)**, Lianmin Zheng, Zhuohan Li, Eric P. Xing, Qirong Ho, Joseph E. Gonzalez, Ion Stoica, Hao Zhang
MLSys2023

Fully Hyperbolic Neural Networks
Weize Chen, Xu Han, Yankai Lin, **Hexu Zhao**, Zhiyuan Liu, Peng Li, Maosong Sun, Jie Zhou
ACL2022

Development of a Doctor-in-the-loop Interpretation Framework for Insulin Titration in Diabetes: a Proof-of-concept Study
Haowei He*, Zhen Ying*, Biao Li, Yujuan Fan, Ping Wang, Jiaping Lu, Liming Wu, **Hexu Zhao**, Xiaoying Li, Yang Yuan, Ying Chen
In submission to *The Lancet Diabetes & Endocrinology Journal*

Research Experience

System for 3D Gaussian Splatting Research at NYU 12/2023 - present
Advised by Prof. Jinyang Li and Aurojit Panda; collaborate with Prof. Saining Xie and Ang Li

- Build the first multi-GPU training framework for 3D gaussians splatting. Our design has been adopted by gsplat, TikTok 4D reconstruction team, many 3DGS research labs etc. We get 350+ stars on Github and well received on twitter and other medias.
- We get state-of-the-art PSNR on 4K resolution Rubble and MatrixCity Small City by distributing 40 Millions gaussians over 16 GPUs. Visualization (choose 4K).
- Our ongoing projects can train even larger scene (>50k images; Matrixcity LargeCity) on 128 GPU to SOTA PSNR. We apply much deeper system techniques and engineering considerations. Reconstruct Ithaca 365 Visualization.

Research Intern at Microsoft DeepSpeed 05/2024 - 8/2024
Advised by Guanhua Wang, Olatunji Ruwase

- I worked on Fusing GEMM and Collective Communication into single kernel for Fined-grained Overlapping at GPU warp-level, by leveraging Tile-based GEMM design and GPU-initiated communication including IPC handle, NVSHMEM etc.
- After carefully hand-optimizing the GEMM CUDA kernel and the allreduce CUDA kernel, our fused version of the code can achieve up to a 19% speedup compared to the non-fused baseline.

Research Assistant at MBZUAI 02/2022 - 10/2022
Advised by Prof. Eric Xing and Prof. Hao Zhang and Lianmin Zheng

- Contributed to the Distributed Machine Learning System Alpa Project.
- Develop an improved approach for cross-mesh resharding communication pattern in distributed machine learning, achieving up to a 10x speed increase; For GPT-3 and U-Transformer end-to-end training, we improve throughput by 10% and 50%, respectively. It's accepted by MLsys2023.

Research Intern at Shanghai Andrew Yao Research Institute 02/2021 - 11/2021
Advised by Prof. Yang Yuan

- Participated in a Diabetes Management Project, leveraging AI to determine insulin dosage.
- I worked on the entire AI healthcare development cycle: conducting needs assessments in hospital, preprocessing data, model design and training, model deployment and etc.

Research Intern at Tsinghua NLP Lab 10/2020 - 02/2021
Advised by Prof. Zhiyuan Liu

- Developed a system for the distributed training of knowledge graph embeddings, leveraging

Megatron-LM's communication primitives.
- Participated in a hyperbolic neural network project, applied its hyperbolic algorithm on common knowledge graph embedding models. It’s accepted by ACL2022.

Engineering Experience

Engineering intern at Polyhedra 02/2023 - 08/2023

- Implemented Zero-Knowledge Proof protocol: GKR protocol.
- Accelerated computation primitives in ZKP, including speed up NTT with GPU.

Engineering intern at Haihua Research Institute 06/2020 - 09/2020

- Participated in AI+healthcare startup Qianfang Medical led by *Prof. Yang Yuan*.
- Code with Scala and learned engineering skills: Type-Safe, OOP/FP, microservices(Akka) and etc.

Other Selected Experience

China National Olympiad of Informatics Gold Medalist 08/2018

- I taught myself, overcame great challenges and risks, and became my hometown’s first Olympiad Gold Medalist, making history in my hometown city of 500 Millions.

Teaching Olympiad of Informatics 2018-2022

- Taught over 2000 senior high school students algorithms and data structures for the Olympiad of Informatics.

AI-track Investment Analyst Intern at ZhenFund Venture Capital 06/2021 - 09/2021

TA for Type-Safe Front-end and Back-end System at Tsinghua University 07/2022 - 08/2022

Outstanding Scholarship for Social Works at Tsinghua University 2020 - 2022

- 2019 Yao Class Monitor, Student Adviser of the 2022 Yao Class and etc.

Academic Referees

Prof. Jinyang Li: jinyang@cs.nyu.edu **Prof. Aurojit Panda:** apanda@cs.nyu.edu