Kaizhao Liang

(650)7720082 | kyleliang919@gmail.com | https://github.com/kyleliang919

Education

University of Illinois, Urbana Champaign — Urbana, IL
Bachelor of Science in Computer Science (Highest Honors) — *Aug. 2016 – May 2020*

Experience

Research Scientist (ML) — Spring 2025 – Now
Meshy LLC — *Austin, TX*

Senior Principal Software Engineer (ML) — May 2024 – Spring 2025
SambaNova Systems — *Austin, TX*

- Composition of Experts (Samba-1)
- Speculative Decoding for super fast inference
- Memory efficient Optimizers
- Long context reasoning

Principal Software Engineer (ML) — March 2022 – May 2024
SambaNova Systems — *Palo Alto, CA*

- Developing and testing software frameworks for large scale neural network training, such as large language models (GPT3, GPT-Neox, Bloom and Llama), text-image pretraining(CLIP, Stable diffusion) as well as other vision models (ViT and ConvNext).
- Designing novel and efficient algorithms on specialized hardwares(Dataflow). (Structured sparsity for fast inference).
- Leading a cross-functional team ranging from compiler to data collection for research(publications) as well as product development

Senior Software Engineer (ML) — March 2021 – March 2022
SambaNova Systems — *Palo Alto, CA*

- Mega High Resolution Convolutions Algorithms on Dataflow (100K x 100K)
- Sparsity is all you need (Pixelated Butterfly)

Software Engineer (ML) — June 2020 – March 2021
SambaNova Systems — *Palo Alto, CA*

- Numerical testing and verification of basic modules for deep learning
- Model building and beta testing compiler stack

Undergraduate Research Assistant — June 2018 – May 2020
Secure Learning Lab advised by Prof. Bo Li — *Urbana Champaign, IL*

- Investigating various aspects of deep neural networks, including adversarial training, adversarial robustness and attack, transfer learning
- Exploring improving intrinsic bias in convolution neural networks.

Publications

Cautious optimizers: Improving training with one line of code [in submission]
Kaizhao Liang*, Lizhang Chen, Bo Liu, Qiang Liu

Composition of Experts on the SN40L Reconfigurable Dataflow Unit [IEEE Micro]
Raghu Prabhakar, Ram Sivaramakrishnan, Darshan Gandhi, Yun Du, Mingran Wang, Xiangyu Song, Kejie Zhang, Tianren Gao, Angela Wang, Karen Li, Yongning Sheng, Joshua Brot, Denis Sokolov, Apurv Vivek, Calvin Leung, Arjun Sabnis, Jiayu Bai, Tuowen Zhao, Mark Gottscho, David Jackson, Mark Luttrell, Manish K. Shah, Edison Chen, Kaizhao Liang*, Swayambhoo Jain, Urmish Thakker, Dawei Huang, Sumti Jairath, Kevin J. Brown, Kunle Olukotun

Memory-Efficient LLM Training with Online Subspace Descent [NeurIPS 2024]
Kaizhao Liang*, Bo Liu, Lizhang Chen, Qiang Liu

Communication Efficient Distributed Training with Distributed Lion [NeurIPS 2024]
Bo Liu, Lemeng Wu, Lizhang Chen, Kaizhao Liang*, Jiaxu Zhu, Chen Liang, Raghuraman Krishnamoorthi, Qiang Liu

Lion Secretly Solves Constrained Optimization: As Lyapunov Predicts [ICLR 2024]
Lizhang Chen, Bo Liu, Kaizhao Liang*, Qiang Liu

Simulating Disease Progression via Progressive Image Editing
Kaizhao Liang*, Xu Cao, Kuei-Da Liao, Tianren Gao, Zhengyu Chen, Tejas Nama

Uncovering the Connections Between Adversarial Transferability and Knowledge Transferability [ICML 2021]
Kaizhao Liang*, Jacky. Y. Zhang*, Boxin Wang, Zhuolin Yang, Sanmi Koyejo, Bo Li

Unrestricted Adversarial Examples via Semantic Manipulation [ICLR 2020]
Anand Bhattad*, Min Jin Chong*, Kaizhao Liang, Bo Li, David A. Forsyth

Adversarial Mutual Information for Text Generation [ICML 2020]
Boyuan Pan, Yazheng Yang, Kaizhao Liang, Bhavya Kailkhura, Zhongming Jin, Xian-Sheng Hua, Deng Cai, Bo Li

Pixelated Butterfly: Simple and Efficient Sparse training for Neural Network Models [ICLR 2022]
Beidi Chen, Tri Dao, Kaizhao Liang, Jiaming Yang, Zhao Song, Atri Rudra, Christopher Re

A Unified Knowledge Distillation Framework for Deep Directed Graphical Models [CVPR 2023]
Yizhuo Chen, Kaizhao Liang, Zhe Zeng, Shuochao Yao, Huajie Shao

MAPLM: A real-world large-scale vision-language dataset for map and traffic scene understanding [CVPR 2024]
Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James M Rehg, Chao Zheng

Nl-augmenter: A framework for task-sensitive natural language augmentation
Kaustubh D Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Kaizhao Liang and et al.

A survey on multimodal large language models for autonomous driving
Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, Chao Zheng

Patents (US)

Lossless Tiling In Convolution Networks Published

- Backward Pass (US11934343B2)
- Data Flow Logic (US20220309323A1)
- Graph Metadata Generation (US12001936B2)
- Materialization of Tensors (US20220309028A1)
- Padding and Re-Tilling at Section Boundaries (US20220309318A1)
- Padding Before Tiling, Location-Based Tiling, and Zeroing-Out (US11263170B1)
- Read-Modify-Write in Backward Pass (US11250061B1)
- Resetting Overlap Factor to Zero at Section Boundaries (US20220309325A1)
- Section Cuts (US20220309322A1)
- Section Boundary (US11227207B1)
- Tiling Configuration (US11195080B1)
- Tilling at Section Boundaries (US20220309319A1)
- Tiling Configuration between Two Sections (US20240168913A1)
- Tiling Configuration for A Sequence of Sections of A Graph (US11995529B2)
- Weight Gradient Calculation (US11232360B1)