Dacheng Li

(858)465-0617 | dacheng177@berkeley.edu
Homepage: dacheng-li.info

EDUCATION

University of California, Berkeley Jun 2023

Ph.D. in Computer Science

- Advisors: *Prof. Ion Stoica* and *Prof. Joseph Gonzalez*. Research: Machine Learning and distributed systems.

Carnegie Mellon University Dec 2021 - Feb 2023

- Research Assistant at Machine Learning Department; Advisors: *Prof. Eric P. Xing* and *Prof. Hao Zhang*.
- Research: *MPCFormer: fast, performant and private Transformer inference with MPC*.

Carnegie Mellon University Aug 2020 - Dec 2021

Master of Science in Machine Learning

- GPA: 3.95/4,0; Advisors: *Prof. Eric P. Xing* and *Prof. Hao Zhang*.
- Research: *AMP: Automatically Finding Model Parallel Strategies with Heterogeneity Awareness*.

University of California, San Diego Sep 2016 - Mar 2020

Bachelor of Science in Computer Science

- GPA: 4.0/4.0, Advisor: *Prof. Zhuowen Tu;* Research: *Dual Contradistinctive Autoencoders*.

Publication

- **Li, Dacheng***, Rulin Shao*, Anze Xie, Eric P Xing, Joseph E Gonzalez, Ion Stoica, Xuezhe Ma, Hao Zhang. "LightSeq: sequence level parallelism for distributed training of long context transformers", Under submission to ICLR 2024.
- **Li, Dacheng***, Rulin Shao*, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalez, Ion Stoica, Xuezhe Ma, and Hao Zhang. "How Long Can Context Length of Open-Source LLMs truly Promise?", Under submission to Neurips 2023 workshop.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, **Dacheng Li**, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica. "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena." (**NeurIPS 2023**)
- **Li, Dacheng***, Rulin Shao*, Hongyi Wang*, Han Guo, Eric P. Xing, Hao Zhang, "MPCFormer: fast, performant and private Transformer inference with MPC." (**ICLR 2023, spotlight**)
- **Li, Dacheng**, Hongyi Wang, Eric P. Xing, and Hao Zhang. "AMP: Automatically Finding Model Parallel Strategies with Heterogeneity Awareness." (**NeurIPS 2022**)
- Bian, Song*, **Dacheng Li***, Hongyi Wang, Eric P. Xing, Shivaram Venkataraman. "Does compressing activations help model parallel training?" (Under submission to **NeurIPS 2023**)
- Parmar, Gaurav*, **Dacheng Li*,** Kwonjoon Lee*, and Zhuowen Tu. "Dual contradistinctive generative autoencoder." (**CVPR 2021**) * denotes equal contribution

INDUSTRY EXPERIENCE

LLM evaluation and improvement Aug 2023 - Present

Student researcher Google

- Developed evaluation benchmarks for chatbot evaluation, including multi-turn capability and long-context ability.
- Developed an automatic pipeline to measure chatbot ability, provide feedback using real-world conversations.

OPEN-SOURCE CONTRIBUTIONS

Member of Large Model Systems Organization (lmsys)

- Core contributor to *FastChat* and *MT-bench*, a system for training, serving and evaluating LLM-based chatbots (**28.6 K** stars).
- Developed *FastChat-T5*, a compact and commercial friendly chatbot (**520K** download).
- Developed *LongChat* and *LongEval*, a series of long-context chatbots and evaluation benchmark (**400** stars).

Awards

Amazon Research Awards (Proposal and Project Lead) Dec 2022

A Faster and More Accurate Secure Model Serving Framework on the Cloud (PI: *Eric P. Xing*, Award funding: $80000)