Cheng Li

555 110th Ave NE, Bellevue, WA 98004

chengli.netlify.app chengli.cli99@gmail.com github.com/cli99 4194967797

Background

I am a senior researcher at Microsoft and am passionate about bringing AI research into production. My work has focused on understanding and optimizing inference/training of Deep Learning (DL) models, particularly on Transformers (LLMs). At Microsoft, I work on improving the performance/usability of transformer models in production (e.g. GitHub Copilot, DALL·E-2, etc.), building systematic profiling/optimization stacks for DL, and integrating SOTA system technologies into Microsoft DeepSpeed, an open-source DL optimization software suite that enables unprecedented scale and speed for training and inference.

EDUCATION

University of Illinois Urbana-Champaign **Champaign, IL**
Ph.D. in Computer Science *August 2020*
GPA: 3.95/4.0

Thesis: Performance Benchmarking, Analysis and Optimization of Deep Learning Inference

University of Michigan **Ann Arbor, MI**
M.S. in Computer Science and Engineering *May 2015*
GPA: 3.96/4.0

Shanghai Jiao Tong University **Shanghai, China**
B.S. in Electrical Engineering *August 2013*
GPA: 3.85/4.0

University of Michigan **Ann Arbor, MI**
B.S. in Computer Engineering *May 2013*
GPA: 3.63/4.0

WORK EXPERIENCE

Microsoft **Bellevue, WA**
Senior Researcher *August 2020 - Present*

- Analyzed the performance bottleneck of business critical AI models (Copilot, DALL·E-2, transformers at WebXT), and make system optimizations to improve the latency/throughput/cost. Collaborated with functional teams across Microsoft and external partners.
- Applied and integrated SOTA DL system technologies (e.g. FlashAttention, lower-bit quantization, CPU/NVME offloading) into Microsoft DeepSpeed inference and training.
- Built tools/user support (profiling, auto-tuning, Hugging Face integration etc.) to improve the usability of DeepSpeed for Microsoft 1P users, third-party customers, and the open-source community.

Alibaba Group **Sunnyvale, CA**
Research Intern *May - August 2019*

- Extended MLModelScope with automatic across-stack analysis capability.
- Used MLModelScope to benchmark and characterize public, MLPerf and AI Matrix models across systems of interest.
- Performed model/framework/system advising using the data collected, and explore its applicability in the Alibaba Cloud.

IBM Thomas J. Watson Research Center **Yorktown Heights, NY**
Research Intern *May - August 2018*

- Evaluated existing techniques for Deep Learning performance estimation on different models and systems, and understood the sources of inaccuracy.
- Developed an analysis tool that generates layer benchmarks, finds patterns within models, and performs performance prediction for Deep Learning models across hardware.

9th Programming and Tuning Massively Parallel Systems and AI School **Barcelona, Spain**
Teaching Assistant *July 2018*

- Designed GPU labs and projects for the summer school students.
- Advised the students during the summer school's hackathon.

IBM Thomas J. Watson Research Center **Yorktown Heights, NY**
Research Intern *May - August 2017*

- Developed MLModelScope a hardware/software agnostic platform for consistent benchmarking and analysis of Deep Learning inference at scale.
- Profiled and optimized the GPU-accelerated alternating least square(ALS) algorithm for Matrix Factorization.

University of Illinois Urbana-Champaign **Champaign, IL**
Lead Teaching Assistant for CS483 - Applied Parallel Programming *August - December 2016*

- Designed GPU labs, exams, and projects for a class of 200 students. Maintained the assignment and the project submission systems - **WebGPU** and **RAI**.

SELECTED PUBLICATIONS

(Refer to Google Scholar for the full list)

1. **A Comprehensive Study on Post-Training Quantization for Large Language Models** (**arXiv'23**)
   *Zhewei Yao, **Cheng Li***, Xiaoxia Wu, Stephen Youn, Yuxiong He*
2. **Understanding INT4 Quantization for Transformer Models: Latency Speedup, Composability, and Failure Cases** (**arXiv'23**)
   Cheng Li, Xiaoxia Wu*, Reza Yazdani Aminabadi, Zhewei Yao, Yuxiong He*
3. **DySR: Adaptive Super-Resolution via Algorithm and System Co-design** (**ICLR'23**)
   *Syed Zawad, **Cheng Li**, Zhewei Yao, Elton Zheng, Yuxiong He, Feng Yan*
4. **DeepSpeed Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale** (**SC'22**)
   *Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, **Cheng Li**, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, Yuxiong He*
5. **Random-LTD: Random and Layerwise Token Dropping Brings Efficient Training for Large-scale Transformers** (**arXiv'22**)
   *Zhewei Yao, Xiaoxia Wu, Conglong Li, Connor Holmes, Minjia Zhang, **Cheng Li**, Yuxiong He*
6. **The Design and Implementation of a Scalable DL Benchmarking Platform** (**IEEE CLOUD'20, Best Paper**)
   Cheng Li, Abdul Dakkak*, Jinjun Xiong, Wen-Mei Hwu*
7. **DLSpec: A Deep Learning Task Exchange Specification** (**USENIX OpML'20**)
   Cheng Li, Abdul Dakkak*, Jinjun Xiong, Wen-Mei Hwu*
8. **XSP: Across-Stack Profiling and Analysis of Machine Learning Models on GPUs** (**IPDPS'20, Best Paper**)
   Cheng Li, Abdul Dakkak*, Jinjun Xiong, Wei Wei, Lingjie Xu, Wen-Mei Hwu*
9. **Benanza: Automatic uBenchmark Generation to Compute "Lower-bound" Latency and Inform Optimizations of Deep Learning Models on GPUs** (**IPDPS'20**)
   Cheng Li, Abdul Dakkak*, Jinjun Xiong, Wen-Mei Hwu*
10. **DLBricks: Composable Benchmark Generation to Reduce Deep Learning Benchmarking Effort on CPUs** (**ICPE'20**)
    ***Cheng Li**, Abdul Dakkak, Jinjun Xiong, Wen-Mei Hwu*
11. **AI Matrix: A Deep Learning Benchmark for Alibaba Data Centers** (**arXiv'19**)
    *Wei Zhang, Wei Wei, Lingjie Xu, Lingling Jin, **Cheng Li***
12. **MLModelScope: Evaluate and Introspect Cognitive Pipelines** (**IEEE Services'19**)
    ***Cheng Li**, Abdul Dakkak, Jinjun Xiong, Wen-Mei Hwu*
13. **TrIMS: Transparent and Isolated Model Sharing for Low Latency Deep Learning Inference in Function as a Service Environments** (**IEEE CLOUD'19**)
    *Abdul Dakkak, **Cheng Li**, Simon Garcia de Gonzalo, Jinjun Xiong, Wen-Mei Hwu*
14. **Accelerating Reduction and Scan Using Tensor Core Units** (**ICS'19**)
    *Abdul Dakkak, **Cheng Li**, Jinjun Xiong, Isaac Gelado, Wen-Mei Hwu*
15. **Evaluating Characteristics of CUDA Communication Primitives on High-Bandwidth Interconnects** (**ICPE'19, Best Paper**)
    *Carl Pearson, Abdul Dakkak, Sarah Hashash, **Cheng Li**, I-Hsin Chung, Jinjun Xiong, Wen-Mei Hwu*
16. **Accelerating Reduction Using Tensor Core Units** (**HPCaML'19**)
    *Abdul Dakkak, **Cheng Li**, Jinjun Xiong, Wen-Mei Hwu*
17. **SCOPE: C3SR Systems Characterization and Benchmarking Framework** (**arXiv'18**)
    *Carl Pearson, Abdul Dakkak, **Cheng Li**, Sarah Hashash, Jinjun Xiong, Wen-mei Hwu*

LANGUAGES

Python, C/C++, CUDA, Go, JavaScript, Bash