Conglong Li

conglong.li@gmail.com; conglong.li@microsoft.com
conglongli.github.io
scholar.google.com/citations?user=nXUS0gYAAAAJ

Research Interests

I'm currently a Senior Researcher at Microsoft DeepSpeed team, working on improving performance and efficiency of deep learning training and inference. In general, I work on improving performance and resource efficiency of all kinds of computer systems via experimental research, data analysis, and algorithm/policy optimizations. My broad research interests lead to experience and publications in many areas including deep learning, similarity search, distributed caching systems, networks, and computer architecture.

Education

2014–2020 **Ph.D. in Computer Science**, *Carnegie Mellon University*, Pittsburgh, PA, USA.
Advisor: David G. Andersen.
Thesis: Learned Adaptive Accuracy-Cost Optimization for Machine Learning Systems.

2013–2014 **M.S. in Computer Science**, *Rice University*, Houston, TX, USA.
Advisor: Alan L. Cox.
Thesis: GD-Wheel: A Cost-Aware Replacement Policy for Key-Value Stores.

2009–2013 **B.S. in Computer Science**, *Rice University*, Houston, TX, USA.
magna cum laude, distinction in research and creative work.
GPA 4.04, Ranked 1st in Dept. of Computer Science (Class of 2013).

Work Experience

2021–Present **Senior Researcher**, *Microsoft*, Bellevue, WA, USA.

2020–2021 **Researcher**, *Microsoft*, Bellevue, WA, USA.
Working on improving performance and efficiency of deep learning training and inference. Member of the DeepSpeed team (github.com/microsoft/DeepSpeed, microsoft.com/en-us/research/project/deepspeed) under the Web Experiences Platform Org (Search, Ads, News, Edge, Maps).

Summer 2019 **Research Intern**, *Microsoft*, Bellevue, WA, USA.
Worked on improving approximate nearest neighbor search performance. Designed ML models (GBDT, neural networks) to predict the search termination condition for each query. Achieved up to 7.1 times speedup under the same accuracy targets. Published a paper at SIGMOD 2020.

Summer 2017 **Software Engineer Intern**, *Microsoft*, Bellevue, WA, USA.
Worked on designing caching strategies for Bing Ads. Designed ML models (GBDT) to provide intelligent cache refresh decisions. Simulations on production traces demonstrate a potential 35.2 to 106.1 million dollars net profit gain in a quarter. Transferred the project to dev team to ship it in product. Published a paper at WWW 2018.

Summer 2016 **Research Intern**, *Microsoft*, Redmond, WA, USA.
Worked on designing caching strategies for Bing Ads. Designed domain-specific caching heuristics to save ads scoring cost and improve net profit. Simulations on production traces demonstrate a potential 20.7 to 70.5 million dollars net profit gain in a quarter. Published a paper at SoCC 2017.

Skills

Programming Mostly using Python and C++. Familiar with C, C#, Java.

Speaking English, Chinese (native), Japanese (JLPT N1).

Publications

arXiv preprint — DeepSpeed4Science Initiative: Enabling Large-Scale Scientific Discovery through Sophisticated AI System Technologies.
Shuaiwen Leon Song, Bonnie Kruft, Minjia Zhang, **Conglong Li** et al.
arXiv preprint arXiv:2310.04610.

arXiv preprint — DeepSpeed-VisualChat: Multi-Round Multi-Image Interleave Chat via Multi-Modal Causal Attention.
Zhewei Yao, Xiaoxia Wu, **Conglong Li**, Minjia Zhang, Heyang Qin, Olatunji Ruwase, Ammar Ahmad Awan, Samyam Rajbhandari, Yuxiong He.
arXiv preprint arXiv:2309.14327.

arXiv preprint — DeepSpeed-Chat: Easy, Fast and Affordable RLHF Training of ChatGPT-like Models at All Scales.
Zhewei Yao, Reza Yazdani Aminabadi, Olatunji Ruwase, Samyam Rajbhandari, Xiaoxia Wu, Ammar Ahmad Awan, Jeff Rasley, Minjia Zhang, **Conglong Li**, Connor Holmes, Zhongzhu Zhou, Michael Wyatt, Molly Smith, Lev Kurilenko, Heyang Qin, Masahiro Tanaka, Shuai Che, Shuaiwen Leon Song, Yuxiong He.
arXiv preprint arXiv:2308.01320.

arXiv preprint — DeepSpeed Data Efficiency: Improving Deep Learning Model Quality and Training Efficiency via Efficient Data Sampling and Routing.
Conglong Li, Zhewei Yao, Xiaoxia Wu, Minjia Zhang, Yuxiong He.
arXiv preprint arXiv:2212.03597.

arXiv preprint — Random-LTD: Random and Layerwise Token Dropping Brings Efficient Training for Large-scale Transformers.
Zhewei Yao, Xiaoxia Wu, **Conglong Li**, Connor Holmes, Minjia Zhang, Cheng Li, Yuxiong He.
arXiv preprint arXiv:2211.11586.

arXiv preprint — BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.
Teven Le Scao et al. (391 authors. I contributed to code and infrastructure to train BLOOM on the Jean Zay supercomputer as a member of the Engineering team.).
arXiv preprint arXiv:2211.05100.

ICLR 2023 — Maximizing Communication Efficiency for Large-scale Training via 0/1 Adam.
Yucheng Lu, **Conglong Li**, Minjia Zhang, Christopher De Sa, Yuxiong He.
In *Eleventh International Conference on Learning Representations*.

HiPC 2022 Best Paper — 1-bit LAMB: Communication Efficient Large-Scale Large-Batch Training with LAMB's Convergence Speed.
Conglong Li, Ammar Ahmad Awan, Hanlin Tang, Samyam Rajbhandari, Yuxiong He.
In *29th IEEE International Conference on High Performance Computing, Data, and Analytics*.

NeurIPS 2022 — The Stability-Efficiency Dilemma: Investigating Sequence Length Warmup for Training GPT Models.
Conglong Li, Minjia Zhang, Yuxiong He.
In *Thirty-sixth Conference on Neural Information Processing Systems*.

NeurIPS 2022 Oral Paper — XTC: Extreme Compression for Pre-trained Transformers Made Simple and Efficient.
Xiaoxia Wu, Zhewei Yao, Minjia Zhang, **Conglong Li**, Yuxiong He.
In *Thirty-sixth Conference on Neural Information Processing Systems*.

NeurIPS 2022 — ZeroQuant: Efficient and Affordable Post-Training Quantization for Large-Scale Transformers.
Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, **Conglong Li**, Yuxiong He.
In *Thirty-sixth Conference on Neural Information Processing Systems*.

ICML 2022 — DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power Next-Generation AI Scale.
Samyam Rajbhandari, **Conglong Li**, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, Yuxiong He.
In *39th International Conference on Machine Learning*.

ICML 2021 — 1-bit Adam: Communication Efficient Large-Scale Training with Adam's Convergence Speed.
Hanlin Tang, Shaoduo Gan, Ammar Ahmad Awan, Samyam Rajbhandari, **Conglong Li**, Xiangru Lian, Ji Liu, Ce Zhang, Yuxiong He.
In *38th International Conference on Machine Learning.*

SIGMOD 2020 — Improving Approximate Nearest Neighbor Search through Learned Adaptive Early Termination.
Conglong Li, Minjia Zhang, David G. Andersen, Yuxiong He.
In *2020 ACM SIGMOD International Conference on Management of Data.*

MLSys 2019 — Scaling Video Analytics on Constrained Edge Nodes.
Christopher Canel, Thomas Kim, Giulio Zhou, **Conglong Li**, Hyeontaek Lim, David G. Andersen, Michael Kaminsky, Subramanya R. Dulloor.
In *Machine Learning and Systems 2019.*

WWW 2018 — Better Caching in Search Advertising Systems with Rapid Refresh Predictions.
Conglong Li, David G. Andersen, Qiang Fu, Sameh Elnikety, Yuxiong He.
In *2018 World Wide Web Conference.*

SoCC 2017 — Workload Analysis and Caching Strategies for Search Advertising Systems.
Conglong Li, David G. Andersen, Qiang Fu, Sameh Elnikety, Yuxiong He.
In *2017 Symposium on Cloud Computing.*

ANCS 2017 — Using Indirect Routing to Recover from Network Traffic Scheduling Estimation Error.
Conglong Li, Matthew K. Mukerjee, David G. Andersen, Srinivasan Seshan, Michael Kaminsky, George Porter, Alex C. Snoeren.
In *2017 ACM/IEEE Symposium on Architectures for Networking and Communications Systems.*

CoNEXT 2015 — Scheduling Techniques for Hybrid Circuit/Packet Networks.
Best Paper Nominee — He Liu, Matthew K. Mukerjee, **Conglong Li**, Nicolas Feltman, George Papen, Stefan Savage, Srinivasan Seshan, Geoffrey M. Voelker, David G. Andersen, Michael Kaminsky, George Porter, Alex C. Snoeren.
In *11th ACM Conference on Emerging Networking Experiments and Technologies.*

EuroSys 2015 — GD-Wheel: A Cost-Aware Replacement Policy for Key-Value Stores.
Conglong Li, Alan L. Cox.
In *Tenth European Conference on Computer Systems.*

ACM TACO 2013
Vol. 10(4) — Reducing DRAM Row Activations with Eager Read/Write Clustering.
Myeongjae Jeon, **Conglong Li**, Alan L. Cox, Scott Rixner.
In *ACM Transactions on Architecture and Code Optimization.*

Professional Service

Conferences — Reviewer for *ICLR 2024*, *NeurIPS 2023*, *ICML 2022*, *MLSys 2020*, *PACT 2019*, *Middleware 2018*, *ICAC 2018*, *IEEE CLOUD 2018*.

Journals — Reviewer for *Elsevier Neural Networks*, *ACM Transactions on Storage*, *ACM SIGMOD Record*.