

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 2021 Curriculum Learning: A Regularization Method for Efficient and Stable Billion-Scale GPT Model Pre-Training.

Conglong Li, Minjia Zhang, Yuxiong He.

arXiv preprint arXiv:2108.06084.

arXiv 2021 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.

arXiv preprint arXiv:2104.06069.

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 *Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.*

SIGMOD 2020 Improving Approximate Nearest Neighbor Search through Learned Adaptive Early Termination.

Conglong Li, Minjia Zhang, David G. Andersen, Yuxiong He.

In *Proceedings of the 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 *Proceedings of 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 *Proceedings of the 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 *Proceedings of the 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.
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 *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*. Nominated for Best Paper.
- EuroSys 2015 GD-Wheel: A Cost-Aware Replacement Policy for Key-Value Stores.
Conglong Li, Alan L. Cox.
In *Proceedings of the Tenth European Conference on Computer Systems*.
- ACM TACO 2013 Reducing DRAM Row Activations with Eager Read/Write Clustering.
Vol. 10(4) Myeongjae Jeon, **Conglong Li**, Alan L. Cox, Scott Rixner.
In *ACM Transactions on Architecture and Code Optimization*.

Professional Service

- MLSys 2020 External Reviewer (1 paper) for *Proceedings of Machine Learning and Systems 2020*.
- PACT 2019 External Reviewer (2 papers) for *28th International Conference on Parallel Architectures and Compilation Techniques*.
- Middleware 2018 External Reviewer (1 paper) for *Proceedings of the 19th International Middleware Conference*.
- ICAC 2018 External Reviewer (1 paper) for *2018 IEEE International Conference on Autonomic Computing*.
- IEEE CLOUD 2018 External Reviewer (1 paper) for *2018 IEEE 11th International Conference on Cloud Computing*.