Hao Zhang

555 Mission Rock Street, Unit 708
San Francisco, CA 94158

Phone: (412) 499-1581
Email: hao@cs.berkeley.edu
sjtu.haozhang@gmail.com
Homepage: https://cs.berkeley.edu/~hao

Research Interest

I study large-scale distributed machine learning in the joint context of machine learning and systems, concerning both *performance* and *usability*. My goal is to co-design models, algorithms, and systems to scale out ML to larger data, problems and applications, to ease the prototyping of complex ML models and algorithms, and to automate the parallelization of ML programs. My work spans across parallel ML programmability, representations of parallelisms, performance optimizations, system architectures, automatic parallelization techniques, and AutoML, with applications in computer vision, natural language processing, and healthcare.

Education

Robotics Institute, School of Computer Science, Carnegie Mellon University
Ph.D. in Robotics, 2014 - 2020
Advisor: Prof. Eric Xing
Ph.D. Dissertation: Machine Learning Parallelism Could Be Adaptive, Composable and Automated.

Department of Computer Science and Engineering, Shanghai Jiao Tong University
M.S. in Computer Science, 2011 - 2014

School of Computer Science and Engineering, South China University of Technology
B.E. in Computer Science, the Elite Class of Computer Science, 2008 - 2011

Positions

University of California, Berkeley

Postdoctoral Researcher, with Prof. Ion Stoica, June 2021 - Present

Petuum Inc, Pittsburgh

Research Scientist, May 2017 – Present.

Director of Scalable ML, December 2017 – Jan 2020.

Tech Lead, May 2017 – December 2017.

Consultant, Jul 2016 – May 2017.

Microsoft Research Asia, Beijing

Research Intern, July 2013 - January 2014.

Microsoft, Shanghai

SDE Intern, September 2012 - April 2013.

Awards and Honors

Jay Lepreau Best Paper Award, OSDI 2021.

NVIDIA Pioneer Research Award, NeurIPS 2017.

Excellent Graduates (top 5%), Shanghai Jiao Tong University, 2014.

Scholarship for Graduates, Shanghai Jiao Tong University, 2011 - 2014.

Google Excellence Scholarship, Google Inc., 2013.

Early Graduate Honor (top 1%), South China University of Technology, 2011.

Excellent Undergraduates, South China University of Technology, 2008 - 2011.

1st Class Scholarship (top 10%), South China University of Technology, 2008 - 2011.

Publications

Google Scholar Profile

Refereed Conference Papers

[1] Lianmin Zheng†, Zhuohan Li†, **Hao Zhang**†, Yonghao Zhuang, Zhifeng Chen, Yanping Huang, Yida Wang, Yuanzhong Xu, Danyang Zhuo, Joseph E. Gonzalez, and Ion Stoica. Parax: Automating Inter- and Intra-operator Parallelism for Distributed Tensor Computation. In *The 16th USENIX Symposium on Operating Systems Design and Implementation*. (**OSDI 2022**) († indicates equal contributions).

[2] Aurick Qiao, Sang Keun Choe, Suhas Jayaram Subramanya, Willie Neiswanger, Qirong Ho, **Hao Zhang**, Gregory R Ganger, and Eric P Xing. Pollux: Co-adaptive Cluster Scheduling for Goodput-optimized Deep Learning. In *The 15th USENIX Symposium on Operating Systems Design and Implementation*. (**OSDI 2021**, **Jay Lepreau Best Paper Award**).

[3] Zhuohan Li, Siyuan Zhuang, Shiyuan Guo, Danyang Zhuo, **Hao Zhang**, Dawn Song, and Ion Stoica. Terapipe: Token-level Pipeline Parallelism for Training Large-scale Language Models. In *The 38th International Conference on Machine Learning*. (**ICML 2021**).

[4] Gengwei Zhang, Yiming Gao, Hang Xu, **Hao Zhang**, Zhenguo Li, and Xiaodan Liang. Ada-segment: Automated Multi-loss Adaptation for Panoptic Segmentation. In *The 35th AAAI Conference on Artificial Intelligence*. (**AAAI 2021**).

[5] **Hao Zhang**, Christy Li, Zhijie Deng, Xiaodan Liang, Lawrence Carin, and Eric P. Xing. AutoSync: Learning to Synchronize for Data-Parallel Distributed Deep Learning. In *The 34th Conference on Neural Information Processing Systems*. (**NeurIPS 2020**).

[6] **Hao Zhang**†, Haowen Xu†, Zhiting Hu, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. AutoLoss: Learning Discrete Schedules for Alternate Optimization. In *The 7th International Conference on Learning Representations*. (**ICLR 2019**) († indicates equal contributions).

[7] Wei Dai, Yi Zhou, Nanqing Dong, **Hao Zhang**, and Eric P. Xing. Toward Understanding the Impact of Staleness in Distributed Machine Learning. In *The 7th International Conference on Learning Representations*. (**ICLR 2019**).

[8] Xiaodan Liang, Zhiting Hu, **Hao Zhang**, Liang Lin, and Eric P. Xing. Symbolic Graph Reasoning Meets Convolutions. In *The 32nd Conference on Neural Information Processing Systems*. (**NeurIPS 2018**).

[9] Xiaodan Liang, **Hao Zhang**, and Eric P. Xing. Generative Semantic Manipulation with Contrasting GAN. In *2018 European Conference on Computer Vision*. (**ECCV 2018**).

[10] Wei Dai, Nanqing Dong, Zeya Wang, Xiaodan Liang, **Hao Zhang**, and Eric P. Xing. SCAN: Structure Correcting Adversarial Network for Chest X-rays Organ Segmentation. In *2018 Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*.

[11] **Hao Zhang**†, Shizhen Xu†, Graham Neubig, Qirong Ho, Guangwen Yang, and Eric P. Xing. Cavs: An Efficient Runtime System for Dynamic Neural Networks. In *2018 USENIX Annual Technical Conference*. (**ATC 2018**, **Oral**, **AISys@SOSP 2017**, **MLSys@NeurIPS 2017**) († indicates equal contributions).

[12] **Hao Zhang**†, Zhijie Deng†, Xiaodan Liang, Luona Yang, Shizhen Xu, Jun Zhu, and Eric P. Xing. Structured Generative Adversarial Networks. In *The 31st Conference on Neural Information Processing Systems*. (**NeurIPS 2017,** **NVIDIA Pioneer Research Award**) († indicates equal contributions).

[13] Xiaodan Liang, Zhiting Hu, **Hao Zhang**, Chuang Gan, and Eric P. Xing. Recurrent Topic-Transition GAN for Visual Paragraph Generation. In *2017 International Conference on Computer Vision*. (**ICCV 2017**).

[14] **Hao Zhang**, Zeyu Zheng, Shizhen Xu, Wei Dai, Qirong Ho, Xiaodan Liang, Zhiting Hu, Jinliang Wei, Pengtao Xie, and Eric P. Xing. Poseidon: An Efficient Communication Architecture for Distributed Deep Learning on GPU Clusters. In *2017 USENIX Annual Technical Conference*. (**USENIX ATC 2017**, **Oral**).

[15] **Hao Zhang**, Zhiting Hu, Jinliang Wei, Pengtao Xie, Gunhee Kim, Qirong Ho, and Eric Xing. Poseidon: A System Architecture for Efficient GPU-based Deep Learning on Multiple Machines. In *2016 USENIX Annual Technical Conference*. (**USENIX ATC 2016**, **Poster**, **MLSys@ICML 2016**, **Spotlight**).

[16] **Hao Zhang**, Zhiting Hu, Yuntian Deng, Mrinmaya sachan, Zhicheng Yan, and Eric Xing. Learning Concept Taxonomies from Multi-modal Data. In *The 54th Annual Meeting of the Association for Computational Linguistics*. (**ACL 2016**, **Oral**).

[17] Zhicheng Yan, **Hao Zhang**, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic Photo Adjustment Using Deep Neural Networks. In *2016 International Conference on Computational Photography*. (**ICCP 2016**, **Invited Poster**).

[18] Henggang Cui, **Hao Zhang**, Gregory R. Ganger, Phillip B. Gibbons, and Eric Xing. GeePS: Scalable Deep Learning on Distributed GPUs with a GPU-specialized Parameter Server. In *2016 European Conference on Computer Systems*. (**EuroSys 2016**).

[19] Jincheng Mei, **Hao Zhang**, and Baoliang Lu. On the Reducibility of Submodular Functions. In *The 19th International Conference on Artificial Intelligence and Statistics*. (**AISTATS 2016**).

[20] Zhicheng Yan, **Hao Zhang**, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. HD-CNN: Hierarchical Deep Convolutional Neural Network for Large Scale Visual Recognition. In *2015 International Conference on Computer Vision*. (**ICCV 2015**).

[21] **Hao Zhang**, Gunhee Kim, and Eric Xing. Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data. In *The 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. (**KDD 2015**, **Oral**).

[22] Ye Liu, **Hao Zhang**, and Liqing Zhang. Common Spatial-Spectral Boosting Pattern for Brain-Computer Interface. In *The 21st European Conference on Artificial Intelligence*. (**ECAI 2014**).

[23] **Hao Zhang**, Ye Liu, Jianyi Liang, Jianting Cao, and Liqing Zhang. Gaussian Mixture Modeling in Stroke Patients' Rehabilitation EEG Data Analysis. In *2013 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. (**EMBC 2013**).

[24] Ye Liu, Mingfen Li, **Hao Zhang**, Junhua Li, Jie Jia, Yi Wu, Jianting Cao, and Liqing Zhang. Single-Trial Discrimination of EEG Signals for Stroke Patients: A General Multi-Way Analysis. In *2013 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. (**EMBC 2013**).

Refereed Journal Papers

[1] Zhicheng Yan, **Hao Zhang**, Baoyuan Wang, Sylvain Paris, and Yizhou Yu. Automatic Photo Adjustment Using Deep Neural Networks. In *ACM Transactions on Graphics*. (**TOG Vol 35, 2016**).

[2] Ye Liu, **Hao Zhang**, Min Chen, and Liqing Zhang. A Boosting-Based Spatial-Spectral Model for Stroke Patients' EEG Analysis in Rehabilitation Training. In *IEEE Transactions on Neural System and Rehabilitation Engineering*, 2015.

[3] Ye Liu, Mingfen Li, **Hao Zhang**, Hang Wang, Junhua Li, Jie Jia, Yi Wu, Jianting Cao, and Liqing Zhang. A Tensor-based Scheme for Stroke Patients' Motor Imagery EEG Analysis in BCI-FES Rehabilitation Training. In *Journal of Neuroscience Methods*, 2013.

Preprints and Technical Reports

[1] **Hao Zhang**. Machine Learning Parallelism Could Be Adaptive, Composable and Automated. PhD Thesis, Robotics Institute, Carnegie Mellon University, October, 2020.

[2] **Hao Zhang**, Christy Li, Zhijie Deng, Aurick Qiao, Qirong Ho, and Eric P. Xing. Oceanus: A Composable and Automated Synchronization System for Distributed Deep Learning. Preprint 2020.

[3] Zhijie Deng, Xiao Yang, **Hao Zhang**, Yinpeng Dong, Jun Zhu. BayesAdapter: Being Bayesian, Inexpensively and Robustly, via Bayeisan Finetuning. 2020. In **arXiv 2010.01979, 2020**.

[4] Aurick Qiao, Willie Neiswanger, **Hao Zhang**, Qirong Ho, Greg Ganger, and Eric P. Xing. Pollux: Co-adaptive Cluster Scheduling for Goodput-Optimized Deep Learning. In **arXiv 2008.12260, 2020**.

[5] Hao Wang, Xiaodan Liang, **Hao Zhang**, Dit-Yan Yeung, and Eric P. Xing. ZM-Net: Real-time Zero-shot Image Manipulation Network. In **arXiv 1703.07255, 2017**.

[6] Zhicheng Yan, **Hao Zhang**, Yangqing Jia, Thomas Breuel, and Yizhou Yu. Combining the Best of Convolutional Layers and Recurrent Layers: A Hybrid Network for Semantic Segmentation. In **arXiv 1603.04871, 2016**.

Professional Service and Leadership

Organizer. RISE Camp 2021.

Program Committee. AAAI 2020, UAI 2019, UAI 2018.

Reviewer. ICLR, NeurIPS, ACL, ECCV, AISTATS, ICML, NACCL HLT, CVPR, ICCV, TPAMI, SCIS, IET Computer Vision, MVAP, TCC, VLDB, etc.

Volunteer. ICML 2016, KDD 2015.

Tutorial and Invited Talks

Tutorial, *"Simple and Automatic Distributed Machine Learning on Ray"*. KDD 2021, with Zhuohan Li, Lianmin Zheng, and Ion Stoica.

Tutorial, *"Simplifying and Automating Parallel Machine Learning via a Programmable and Composable Parallel ML System"*. AAAI 2021, with Aurick Qiao, Qirong Ho and Eric Xing.

Invited Talk, *"Collective-on-Ray: High-performance Collective Communication for Distributed Machine Learning on Ray"*. Ray Summit 2021.

Invited Talk, *"Machine Learning Parallelism Could Be Adaptive, Composable, and Automated"*. UC Berkeley RISELab Seminar, 2020.

Tutorial, *"Arion: a Next-generation Distributed Deep Learning Virtual Machine"*. ICML 2019 Expo Day. .

Invited Talk. *"Cavs: A Vertex-centric Programming Interface for Dynamic Neural Networks"*. NeurIPS 2017 MLSys Workshop.

Invited Talk. *"Dynamic Topic Modeling for Monitoring Market Competition from Online Text and Image Data"*. Database Seminar, Carnegie Mellon University, 2016.

Invited Talk. *Poseidon: A System Architecture for Large-scale Distributed Deep Learning on GPU Clusters with Commodity Hardware*. Machine Learning Seminar, Carnegie Mellon University, 2015.

Teaching

Teaching Assistant in Carnegie Mellon University

10-708, Probabilistic Graphical Models, Spring 2019. Instructor: Prof. Eric Xing.

16-791, Applied Data Science, Spring 2019. Instructor: Prof. Artur Dubrawski.

10-701, Introduction to Machine Learning, Fall 2015. Instructors: Prof. Eric Xing and Prof. Ziv Bar-Joseph.

Teaching Assistant in Shanghai Jiao Tong University

Artificial Intelligence, Spring 2012. Instructor: Prof. Liqing Zhang.

References

Eric P. Xing, Professor
Machine Learning Department, School of Computer Science
Carnegie Mellon University
epxing@cs.cmu.edu

Ion Stoica, Professor
Electrical Engineering and Computer Sciences
UC Berkeley
istoica@cs.berkeley.edu

Gregory Ganger, Jatras Professor
Electrical and Computer Engineering and School of Computer Science
Carnegie Mellon University
ganger@ece.cmu.edu

Graham Neubig, Associate Professor
Language Technology Institute
Carnegie Mellon University
gneubig@cs.cmu.edu

Last updated: March 23, 2022