Zhewei Yao | Curriculum Vitae

City Center, Redmond, WA 98052

zheweiy@berkeley.edu; zheweiyao@gmail.com • yaozhewei.github.io
YZ

I am currently working as a Senior Researcher at Microsoft. I was a Ph.D. student in the RISELab (former AMPLab), BDD, BAIR, and Math Department at University of California at Berkeley. I was advised by Prof. Michael Mahoney and worked closely with Prof. Kurt Keutzer. My research interest lies in optimization and machine learning. Currently, I am interested in leveraging tools from randomized linear algebra to provide efficient and scalable solutions for large-scale optimization and learning problems. I apply second order methods for model compression as well as neural network analysis/optimization. I am also working on the theory and application of deep learning.

Education

- **University of California at Berkeley** — **CA, USA**
  Ph.D. in Applied Mathematics, Department of Mathematics — *Sep. 2016–May. 2021*
- **Shanghai Jiao Tong University** — **Shanghai China**
  B.S. in Applied Mathematics, Zhiyuan Honor College — *Sep. 2012–Jun. 2016*

Publications (*: equal contribution)[Google Scholar]

Conference

[1] **Hessian-Aware Pruning and Optimal Neural Implant**
S. Yu, **Z. Yao***, *A. Gholami**, *Z. Dong**, *M. W. Mahoney, K. Keutzer*
arXiv, code
Proc. WACV 2022

[2] **What's Hidden in a One-layer Randomly Weighted Transformer?**
*S. Shen**, **Z. Yao***, *D. Kiela, K. Keutzer, M. W. Mahoney*
arXiv, code
Proc. EMNLP 2021

[3] **ActNN: Reducing Training Memory Footprint via 2-Bit Activation Compressed Training**
*J. Chen, L. Zheng, **Z. Yao**, D. Wang, I. Stoica, M. W. Mahoney, J. E. Gonzalez*
arXiv, code
Proc. ICML 2021 (Oral)

[4] **I-BERT: Integer-only BERT Quantization**
*S. Kim**, *A. Gholami**, ***Z Yao***, *M. W. Mahoney, Kurt Keutzer*
arXiv, code (fairseq), code (transformers)
Proc. ICML 2021 (Oral)

[5] **HAWQ-V3: Dyadic Neural Network Quantization in Mixed Precision**
Z. Yao*, *Z. Dong**, *Z. Zheng**, *A. Gholami**, *J. Yu, E. Tan, L. Wang, Q. Huang, Y. Wang, M. W. Mahoney, K. Keutzer*
arXiv, code
Proc. ICML 2021

[6] **ADAHESSIAN: An Adaptive Second Order Optimizer for Machine Learning**
Z. Yao*, *A. Gholami**, *S. Shen, K. Keutzer, and M. W. Mahoney*
arXiv, code
Proc. AAAI 2021

[7] **A Statistical Framework for Low-bitwidth Training of Deep Neural Networks**
J. Chen, Y. Gai, ***Z. Yao****, M. W. Mahoney, and J. E. GonZalez*
arXiv, code
Proc. NeurIPS 2020

[8] **MAF: Multimodal Alignment Framework for Weakly-Supervised Phrase Grounding**
Q. Wang, H. Tan, S. Shen, M. W. Mahoney, and ***Z. Yao***
arXiv, code
Proc. EMNLP 2020

[9] **PowerNorm: Rethinking Batch Normalization in Transformers**
*S. Shen**, ***Z. Yao****, *A. Gholami, M. W. Mahoney, and K. Keutzer*
arXiv, code
Proc. ICML 2020

[10] **ZeroQ: A Novel Zero Shot Quantization Framework**
*Y. Cai**, ***Z. Yao****, *Z. Dong**, *A. Gholami, M. W. Mahoney, and K. Keutzer*
arXiv, code
Proc. CVPR 2020

[11] **PyHessian: Neural Networks Through the Lens of the Hessian**
Z. Yao*, A. Gholami, K. Keutzer, M. W. Mahoney*
arXiv, code
Proc. BigData 2020

[12] **HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks**
Z. Dong, ***Z. Yao****, Y. Cai, D. Arfeen, A. Gholami, M. W. Mahoney, K. Keutzer*
arXiv, code
Proc. NeurIPS 2020

[13] **Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT**
S. Shen, Z. Dong, J. Ye, L. Ma, ***Z. Yao****, A. Gholami, M. W. Mahoney, K. Keutzer*
arXiv
Proc. AAAI 2020.

[14] **ANODEV2: A Coupled Neural ODE Evolution Framework**
*T. Zhang**, ***Z. Yao****, *A. Gholami**, *K. Keutzer, J. Gonzalez, G. Biros, and M. W. Mahoney*
arXiv, code
Proc. NeurIPS 2019

[15] **HAWQ: Hessian AWare Quantization of Neural Networks with Mixed-Precision**
*Z. Dong**, ***Z. Yao****, *A. Gholami**, *M. W. Mahoney, K. Keutzer*
arXiv, code
Proc. ICCV 2019

[16] **Inefficiency of K-FAC for Large Batch Size Training**
L. Ma, G. Montague, J. Ye, ***Z. Yao****, A. Gholami, K. Keutzer, M. W. Mahoney*
arXiv
Proc. AAAI 2020.

[17] **JumpReLU: A Retrofit Defense Strategy for Adversarial Attacks**
*N. B. Erichson**, ***Z. Yao****, *M. W. Mahoney*
arXiv
Proc. ICPRAM 2020.

[18] **Trust Region Based Adversarial Attack on Neural Networks**
Z. Yao*, A. Gholami, P. Xu, K. Keutzer, M. W. Mahoney*
arXiv, code
Proc. CVPR 2019

[19] **Hessian-based Analysis of Large Batch Training and Robustness to Adversaries**
Z. Yao*, *A. Gholami**, *Q. Lei K. Keutzer, M. W. Mahoney*
arXiv, code
Proc. NeurIPS 2018

Journal

[1] **Shallow Learning for Fluid Flow Reconstruction with Limited Sensors and Limited Data**
N. B. Erichson, L. Mathelin, ***Z. Yao****, S. L. Brunton, M. W. Mahoney, J. N. Kutz*
arXiv
Proceedings of the Royal Society A.

[2] **Inexact non-convex Newton-type methods**
Z. Yao*, P. Xu, F. Roosta-Khorasani, M. W. Mahoney*
arXiv, code
INFORMS Journal on Optimization.

[3] **A hybrid adaptive MCMC algorithm in function spaces**
Q. Zhou, Z. Hu, ***Z. Yao****, J. Li*
arXiv
SIAM/ASA Journal on Uncertainty Quantification 5 (1), 621-639

[4] **On an adaptive preconditioned Crank–Nicolson MCMC algorithm for infinite dimensional Bayesian inference**
*Z. Hu**, ***Z. Yao****, *J. Li*
arXiv
Journal of Computational Physics 332, 492-503

[5] **A TV-Gaussian prior for infinite-dimensional Bayesian inverse problems and its numerical implementation**
Z. Yao*, *Z. Hu**, *J. Li*

arXiv
Inverse Problems 32 (7), 075006 (Highlight Paper of 2016)

Book Chapter

[1] **A Survey of Quantization Methods for Efficient Neural Network Inference**
A. Gholami, S. Kim*, Z. Dong*, **Z. Yao***, M. W. Mahoney, K. Keutzer*
arXiv
Book Chapter: Low-Power Computer Vision: Improving the Efficiency of Artificial Intelligence, 2021.

Workshop

[1] **Parameter Re-Initialization through Cyclical Batch Scheduling**
N. Mu, **Z. Yao***, A. Gholami, K. Keutzer, M. W. Mahoney*
arXiv
Proc. MLSYS Workshop at NeurIPS 2018

[2] **An Empirical Exploration of Gradient Correlations in Deep Learning.**
*D. Rothchild, R. Fox, N. Golmant, J. Gonzalez, M. W. Mahoney, K. Rothauge, I. Stoica, **Z. Yao***
Integration of Deep Learning Theories, NeurIPS 2018

Preprint and Technical Report

[1] **Inexact Newton-CG Algorithms With Complexity Guarantees**
***Z. Yao**, P. Xu, F. Roosta, S. J. Wright, M. W. Mahoney*
arXiv

[2] **How Much Can CLIP Benefit Vision-and-Language Tasks?**
*S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K. Chang, **Z. Yao**, K Keutzer*
arXiv, code

[3] **MLPruning: A Multilevel Structured Pruning Framework for Transformer-based Models**
***Z. Yao**, L. Ma, S. Shen, K. Keutzer, M. W. Mahoney*
arXiv, code

[4] **Q-ASR: Integer-only Zero-shot Quantization for Efficient Speech Recognition**
*S. Kim, A. Gholami, **Z. Yao**, A. Nrusimha, B. Zhai, T. Gao, M. W. Mahoney, K. Keutzer*
arXiv

[5] **Benchmarking Semi-supervised Federated Learning**
Z. Zhang, **Z. Yao***, Y. Yang, Y. Yan, J. E. Gonzalez, and M. W. Mahoney*
arXiv, code

[6] **Residual Networks as Nonlinear Systems: Stability Analysis using Linearization**
*K. Rothauge, **Z. Yao**, Z. Hu, and M. W. Mahoney*
arXiv

[7] **On the Computational Inefficiency of Large Batch Sizes for Stochastic Gradient Descent**
*N. Golmant, N. Vemuri, **Z. Yao**, V. Feinberg, A. Gholami, K. Rothauge, M. W. Mahoney, J. Gonzalez*

arXiv

[8] **Large batch size training of neural networks with adversarial training and second-order information**
Z. Yao*, *A. Gholami**, *K. Keutzer, M. W. Mahoney*
arXiv, code

Research Experiences

- **Microsoft** **WA, USA**
  Senior Researcher *Jun. 2021–Present*
  - Design efficient training and inference algorithms for extreme large model
  - Optimize system for machine learning training

- **University of California at Berkeley** **CA, USA**
  Ph.D. Researcher at RISELab, BAIR, and BDD *Sep. 2016–May. 2021*
  - Developed second order methods for machine learning and optimization
  - Designed efficient training and inference algorithms for deep learning

- **Facebook** **CA, USA**
  Software Engineer *May. 2020–Aug. 2020*
  - Tried Gauss-Newton method for deep learning
  - Investigated different variants of Gauss-Newton methods for computer vision tasks and recommendation systems

- **Amazon AWS AI** **CA, USA**
  Applied Scientist *May. 2019–Aug. 2019*
  - Applied machine learning algorithm to explore very large scale configurations problems
  - Investigated transfer learning and exploration of TVM computation configuration generation with different batch sizes and GPUs
  - Investigated reinforce learning to explore fast database query answering, particularly on the Materialized View Update and Vacuum frequency.

- **Lawrence Berkeley Notional Laboratory** **CA, USA**
  Researcher intern at NERSC *May. 2018–Aug. 2018*
  - Implemented CPU Parallelization of PyTorch to train large climate dataset (over 400 Gb)
  - Tested robustness on models trained with scientific datasets

- **Shanghai Jiao Tong University** **Shanghai, China**
  Undergraduate Researcher *Sep. 2014–Jun. 2016*
  - Considered MCMC algorithm in infinite-dimensional space
  - Designed a TG-prior with better edge-preserving property and two new adaptive algorithms

Others

- **Programming Languages:** Python, Pytorch, Tensorflow, C++, Java, Matlab,
- **Reviewer for:** NeurIPS 2018/2020/2021, ICLR 2019/2020/2022, ECCV 2020, ICML 2019/2021,

CVPR 2021, ICCV 2021, NLPCC2021, AAAI2022, WACV2022, JMLR, Machine Learning (Springer Netherlands), Journal of Systems Architecture

- **Teaching:**
  - **Stat 89A: Linear Algebra for Data Science** — **UC Berkeley**
    Graduate Student Instructor — *Spring 2018*
  - **Math 16A: Analytic Geometry and Calculus** — **UC Berkeley**
    Graduate Student Instructor — *Spring 2017 & Fall 2016*