AMIR GHOLAMI

amirgh@eecs.berkeley.edu ⋄ www.amirgholami.org ⋄ Github: amirgholami ⋄ Status: US Permanent Resident

EDUCATION

- **University of California, Berkeley** *July 2017- 2020*
  PostDoc in Berkeley AI Research (BAIR) Lab, EECS Department

- **The University of Texas at Austin** *June 2017*
  Ph.D. in *Computational Science, Engineering, and Mathematics*,
  Institute for Computational Engineering and Sciences (ICES),
  GPA 4.00/4.00 (Adviser: Prof. G. Biros)

- **Tehran Polytechnic (Amirkabir University)** *June 2011*
  Bachelor of Science in *Aerospace Engineering*,
  GPA 3.89/4.00 (Ranked #1 in the class of 2011)

EXPERIENCE

ICSI, and BAIR UC Berkeley *Spring 2020-Present*
Senior Research Scientist *Berkeley, CA*

UC Berkeley TRIPODS *Fall 2018-Fall 2020*
PostDoc member in NSF TRIPODS program at UCB *Berkeley, CA*

Simons Institute *Fall 2018*
Research Fellow in Foundations of Data Science Program *Berkeley, CA*

NVIDIA *Summer 2016*
CUDA Library/Deep Learning Software Engineer Intern *Santa Clara, CA*

Advanced Micro Devices (AMD) *Summer 2015*
Software Engineer Intern at AMD Compute Library Team *Austin, TX*

Institute for Computational Engineering and Sciences 2011-2017
Graduate Research Assistant *Austin, TX*

SELECTED AWARDS

- Recipient of **NSF FODA** PostDoctoral Fellowship *2018-2020*
- **Best Ph.D. Dissertation Award** from UT Austin *2018*
- **Outstanding Dissertation**, Institute for Computational Engineering and Sciences *2018*
- **Finalist** for **Robert J. Melosh Medal** *2018*
- **Best Student Paper**, ACM/IEEE Supercomputing conference (SC'17) *2017*
- **Gold medal** in ACM Student Research Competition at SC'15 *2015*
- **Best Student Paper finalist**, ACM/IEEE Supercomputing conference (SC'14) *2014*
- **First place** in Broader Engagement programming challenge at SC'14 *2014*
- **Second place** in 2014 TACC-BP America parallel programming contest *2014*
- **Student Employee Excellence**, UT Austin *2014*
- Graduate school's Professional Development Award, UT Austin *2013*

- **Best B.Sc. thesis of the year** in Aerospace Engineering, Tehran Polytechnic *2011*
- Graduated as **top student (out of 55)** in my undergraduate studies *2011*
- Selected as ***Outstanding Student***, Tehran Polytechnic *2010*

PUBLICATIONS

Journal/Proceedings

32. Zhewei Yao, Amir Gholami, Sheng Shen, Kurt Keutzer, and Michael W Mahoney. Adahessian: An adaptive second order optimizer for machine learning. *arXiv preprint arXiv:2006.00719*, 2020

31. Sheng Shen, Zhewei Yao, Amir Gholami, Michael Mahoney, and Kurt Keutzer. Rethinking batch normalization in transformers. *ICML (Accepted)*, 2020. [PDF]

30. Z. Dong, Z. Yao, Y. Cai, D. Arfeen, A. Gholami, M. Mahoney, and K. Keutzer. HAWQ-V2: Hessian aware trace-weighted quantization of neural networks. *CVPR'20 (arXiv:1911.03852)*, 2019. [PDF]

29. Z. Yao, A. Gholami, K. Keutzer, and M. Mahoney. PyHessian: Neural Networks through the lens of the Hessian. *arxiv:1912.07145*, 2019. [PDF]

28. P. Jain, A. Jain, A. Nrusimha, A. Gholami, P. Abbeel, K. Keutzer, I. Stoica, and J. Gonzalez. Checkmate: Breaking the memory wall with optimal tensor rematerialization. *Conference on Machine Learning and Systems (MLSys)*, 2019. [PDF]

27. S. Shen, Z. Dong, J. Ye, L. Ma, Z. Yao, A. Gholami, M. Mahoney, and K. Keutzer. Q-BERT: Hessian based ultra low precision quantization of BERT. *AAAI-20*, 2020. [PDF]

26. T. Zhang, Z. Yao, A. Gholami, K. Keutzer, J. Gonzalez, G. Biros, and M. Mahoney. ANODEV2: A coupled Neural ODE evolution framework. *NeurIPS'19 (arXiv:1906.04596)*, 2019. [PDF]

25. Z. Dong, Z. Yao, A. Gholami, M. Mahoney, and K. Keutzer. HAWQ: Hessian AWare quantization of neural networks with mixed-precision. *International Conference on Computer Vision*, 2019. [PDF]

24. L. Ma, G. Montague, J. Ye, Z. Yao, A. Gholami, K. Keutzer, and M. Mahoney. Inefficiency of K-FAC for large batch size training. *AAAI-20*, 2019. [PDF]

23. A. Gholami, Kurt K. Keutzer, and G. Biros. ANODE: Unconditionally accurate memory-efficient gradients for neural ODEs. *International Joint Conferences on Artificial Intelligence (IJCAI'19)*, 2019. [PDF]

22. K. Scheufele, A. Mang, A. Gholami, C. Davatzikos, G. Biros, and M. Mehl. Coupling brain-tumor biophysical models and diffeomorphic image registration. *Computer Methods in Applied Mechanics and Engineering*, 2019. [PDF]

21. A. Mang, A. Gholami, C. Davatzikos, and G. Biros. Claire: a distributed-memory solver for constrained large deformation diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 41(5):C548–C584, 2019

20. Y. Cai, Z. Dong, Z. Yao, A. Gholami, M. Mahoney, and K. Keutzer. ZeroQ: A novel zero shot quantization framework. *Accepted in CVPR'20*, 2019. [PDF]

19. Z. Yao, A. Gholami, Q. Lei, K. Keutzer, and M. Mahoney. Hessian-based analysis of large batch training and robustness to adversaries. *NeurIPS'18*, 2018. [PDF]

18. Z. Yao, A. Gholami, P. Xu, K. Keutzer, and M. Mahoney. Trust region based adversarial attack on neural networks. *CVPR'19*, 2019. [PDF]

17. A. Gholami, A. Azad, P. Jin, K. Keutzer, and A. Buluc. Integrated model, batch and domain parallelism in training neural networks. *ACM Symposium on Parallelism in Algorithms and Architectures(SPAA'18)*, 2018. [PDF]

16. A. Gholami, S. Subramanian, V. Shenoy, N. Himthani, X. Yue, S. Zhao, P. Jin, G. Biros, and K. Keutzer. A novel domain adaptation framework for medical image segmentation. *Lecture Notes in Computer Science (LNCS)*, 2018. [PDF]

15. Z. Yao, A. Gholami, K. Keutzer, and M. Mahoney. Large batch size training of neural networks with adversarial training and second-order information. *arXiv:1810.01021*, 2018. [PDF]

14. N. Golmant, N. Vemuri, Z. Yao, V. Feinberg, A. Gholami, K. Rothauge, M. Mahoney, and J. Gonzalez. On the computational inefficiency of large batch sizes for stochastic gradient descent. *Under review*, 2018. [PDF]

13. K. Kwon, A. Amid, A. Gholami, B. Wu, K. Asanovic, and K. Keutzer. Co-design of deep neural nets and neural net accelerators for embedded vision applications. *Design Automation Conference (DAC'18)*, 2018. [PDF]

12. S. Zhao, A. Gholami, G. Ding, Y. Gao, J. Han, and K. Keutzer. Personalized emotion recognition by personality-aware high-order learning of physiological signals. *ACM Transactions on Multimedia Computing Communications and Applications (TOMM)*, 2018.

11. S. Subramanian, A. Gholami, and G. Biros. Simulation of glioblastoma growth using a 3D multispecies tumor model with mass effect. *Journal of mathematical biology*, 2019

10. B. Wu, A. Wan, X. Yue, P. Jin, S. Zhao, N. Golmant, A. Gholami, J. Gonzalez, and K. Keutzer. Shift: A zero flop, zero parameter alternative to spatial convolutions. *Computer Vision and Pattern Recognition (CVPR'18)*, 2018. [PDF]

9. A Gholami. *Fast algorithms for biophysically-constrained inverse problems in medical imaging.* PhD thesis, The University of Texas at Austin, 2017 (**Best PhD Dissertation Award**). [PDF]

8. A. Mang, S. Tharakan A. Gholami, N. Himthani, S. Subramanian, J. Levitt, M. Azmat, K. Scheufele, M. Mehl, C. Davatzikos, B. Barth, and G. Biros. SIBIA-GlS: Scalable biophysics-based image analysis for glioma segmentation. *The multimodal brain tumor image segmentation benchmark (BRATS), MICCAI*, 2017. [PDF]

7. A. Gholami, A. Mang, K. Scheufele, C. Davatzikos, M. Mehl, and G. Biros. A framework for scalable biophysics-based image analysis. *Proceedings of ACM/IEEE SuperComputing Conference (SC'17)*, 2017 (**Best Student Paper**). [PDF]

6. A. Mang, A. Gholami, C. Davatzikos, and G. Biros. PDE-constrained optimization in medical image analysis. *Optimization and Engineering*, pages 1–48, 2017. [PDF]

5. A. Mang, A. Gholami, and G. Biros. Distributed-memory large-deformation diffeomorphic 3D image registration. *Proceedings of ACM/IEEE SuperComputing Conference (SC16)*, 2016. [PDF]

4. D. Malhotra, A. Gholami, and G. Biros. A volume integral equation Stokes solver for problems with variable coefficients. *Proceedings of ACM/IEEE SuperComputing Conference (SC14)*, 2014 (**Best Student Paper Finalist**). [PDF]

3. A. Gholami, D. Malhotra, H. Sundar, and G. Biros. FFT, FMM, or MultiGrid? A comparative study of state-of-the-art Poisson solvers for uniform and nonuniform grids in the unit cube. *SIAM Journal on Scientific Computing*, 38(3):C280–C306, 2016. [PDF]

2. A. Gholami, A. Mang, and G. Biros. An inverse problem formulation for parameter estimation of a reaction–diffusion model of low grade gliomas. *Journal of mathematical biology*, 72:409–433, 2015. [PDF]

1. A. Gholami, J. Hill, D. Malhotra, and G. Biros. AccFFT: A library for distributed-memory FFT on CPU and GPU architectures. *arXiv:1506.07933*, 2015.

Workshops

4. Z. Dong, Z. Yao, D. Arfeen, Y. Cai, A. Gholami, M. Mahoney, and K. Keutzer. Trace weighted hessian-aware quantization. ***Spotlight at NeurIPS'19*** *workshop on Beyond First-Order Optimization Methods in Machine Learning*, 2019.

3. N. Mu, Z. Yao, A. Gholami, K. Keutzer, and M. Mahoney. Parameter re-initialization through cyclical batchscheduling. *Systems for ML Workshop at NeurIPS 18*, 2018. [PDF]

2. A. Gholami, K. Kwon, B. Wu, Z. Tai, X. Yue, P. Jin, S. Zhao, and K. Keutzer. SqueezeNext: Hardware-aware neural network design. *ECV Workshop at CVPR'18*, 2018. [PDF]