Shuai Zhang

Staff Machine Learning Researcher
Qualcomm, San Diego, CA
Email: zsimath@gmail.com

Experience

- **Staff Engineer (Machine Learning) at Qualcomm Technologies, Inc** *11/2020 - Now*
  Job description:

  Develop computer vision solutions for mobile platforms, with applications on ISP, Display, and 3A;
  Numerical optimization of AI models with Qualcomm AI pipeline for low power and low latency;
  Research and optimize leading computer vision algorithms for edge devices.

- **Senior Engineer (Machine Learning) at Qualcomm Technologies, Inc** *07/2017 - 11/2020*
  Job description:

  Research on fast computer vision algorithms on edge devices;
  Work on real-time user authentication systems (FaceID and Fingerprint) on Android devices (including face detection, anti-spoofing, eye attention detection and user adaptation);
  Work on segmentation projects, used in the Camera pipeline.

- **Achievements:**

  Fast and efficient computer vision solutions on Qualcomm mobile chips;
  Filed 6 Patents and 3 approved til now;
  Published several papers at Top conferences and Journals in the fields of applied mathematics and machine learning.

Research:

Applied Mathematics, Numerical Optimization, Machine Learning, Computer Vision.

Education

- **University of California, Irvine** Irvine, CA
  Department of Mathematics, Ph.D. in Mathematics *2012–2017*
  Department of Mathematics, M.S. in Mathematics *2012–2014*
  - Advisor: Prof. Jack Xin
  - Research areas: Compressed Sensing; Image Processing and Machine Learning

- **Shandong University** Jinan, China
  Department of Mathematics, M.S. in Computational Mathematics (GPA: 3.7/4.0) *2009–2012*
  Department of Mathematics, B.S. in Applied Mathematics (GPA: 3.83/4.0) *2005–2009*

Patents:

1. E. Mequanint, S. Zhang, Y. Qi, N. Bi. Personalized eye openness estimation. *US Patent App. 16/239, 352*.
2. S. Zhang, Y. Chen, Y. Zhou. Edge computing. *US Patent App. 16/204, 242*.
3. E. Mequanint, S. Zhang, Y. Qi, N. Bi. User adaptation for biometric authentication. *US Patent App. 16/125, 360*.

Published Papers:

1. K. Bui, F. Park, S. Zhang, Y. Qi, J. Xin. Improving Network Slimming With Nonconvex Regularization. *IEEE Access, 2021*.
2. K. Bui, F. Park, S. Zhang, Y. Qi, J. Xin. Structured sparsity of convolutional neural networks via nonconvex sparse group regularization. *Frontiers in applied mathematics and statistics, 2021*.
3. B. Shi, Y. Xu, W. Dai, B. Wang, S. Zhang, C. Li, J. Zou, H. Xiong. Tiny-Hourglassnet: An Efficient Design For 3d Human Pose Estimation. *IEEE International Conference on Image Processing (ICIP), 2020*.

4. K. Bui, F. Park, S. Zhang, Y. Qi, J. Xin. Nonconvex regularization for network slimming: Compressing CNNs even more. *International Symposium on Visual Computing (Springer-Verlag Best Paper), 2020*.
5. J. Lyu, S. Zhang, Y. Qi, J. Xin. Autoshufflenet: Learning permutation matrices via an exact lipschitz continuous penalty in deep convolutional neural networks. *The 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD), 2020*.
6. Y. Xu, Y. Li, S. Zhang, W. Wen, B. Wang, Y. Qi, Y. Chen, W. Lin, H. Xiong. TRP: Trained Rank Pruning for Efficient Deep Neural Networks. *International Joint Conference on Artificial Intelligence (IJCAI), 2020*.
7. Y. Xu, Y. Li, S. Zhang, W. Wen, B. Wang, Y. Qi, Y. Chen, W. Lin, H. Xiong. Trained Rank Pruning for Efficient Deep Neural Networks. *The NeurIPS 2019 Workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2), 2019*.
8. E. Mequanint, S. Zhang, B. Forutanpour, Y. Qi, N. Bi. Weakly-Supervised Degree of Eye-Closeness Estimation. *IEEE ICCV Workshop on Egocentric Perception, Interaction and Computing, 2019*
9. B. Yang, J. Xin, J. Lyu, S. Zhang, Y. Qi. Channel Pruning for Deep Neural Networks via a Relaxed Groupwise Splitting Method. *IEEE The second International Conference on Artificial Intelligence for Industries (AI4I), 2019*.
10. F. Xue, J. Xin, J. Lyu, S. Zhang, Y. Qi. A Multistage Backward Differentiable Method for Constructing Light Convolutional Neural Networks. *IEEE The second International Conference on Artificial Intelligence for Industries (AI4I), 2019*.
11. Y. Xu, S. Zhang, X. Zhang, Y. Qi, J. Guo, W. Lin, H. Xiong. DNQ: Dynamic Network Quantization. *IEEE Data Compression Conference (DCC), 2019*.
12. P. Yin, J. Lyu, S. Zhang, S. Osher, Y-Y. Qi, J. Xin. Understanding Straight-through Estimator in Training Activation Quantized Neural Nets. *Seventh International Conference on Learning Representations (ICLR), 2019*
13. X. Li, S. Zhang (co-first author), B. Jiang, Y. Qi, M. Chuah, N. Bi. DAC: Data-free Automatic Acceleration of Convolutional Networks. *IEEE Winter Conference on Applications of Computer Vision (WACV), 2019*
14. P. Yin, S. Zhang (co-first author), J. Lyu, S. Osher, Y-Y. Qi, J. Xin. Blended Coarse Gradient Descent for Full Quantization of Deep Neural Networks. *Research in the Mathematical Sciences, DOI:10.1007/s40687-018-0177-6, 2019*.
15. P. Yin, S. Zhang (co-first author), J. Xin, Y. Qi. Quantization and Training of Low Bit-Width Convolutional Neural Networks for Object Detection. *Journal of Computational Mathematics, 37(3), pp. 1-12, 2019*.
16. P. Yin, S. Zhang (co-first author), J. Lyu, S. Osher, Y. Qi, J. Xin. BinaryRelax: A Relaxation Approach For Training Deep Neural Networks With Quantized Weights. *SIAM Journal on Imaging Sciences, 11(4): 2205-2223, 2018*.
17. S. Zhang, J. Xin. Minimization of Transformed L_1 Penalty: Theory, Difference of Convex Function Algorithm, and Robust Application in Compressed Sensing. *Mathematical Programming, Series B, 169(1), pp. 307-336, 2018*.
18. S. Zhang, P. Yin, J. Xin. Transformed Schatten-1 Iterative Thresholding Algorithms for Low Rank Matrix Completion. *Communications in Mathematical Sciences, 15(3), pp. 839–862, 2017*.
19. S. Zhang, J. Xin. Minimization of Transformed L_1 Penalty: Closed Form Representation and Iterative Thresholding Algorithms. *Communications in Mathematical Sciences, 15(2), pp. 511–537, 2017*.
20. H. Wang, H. Rui, S. Zhang. An Optimal-order Error Estimate for the Mass-conservative Characteristic Finite Element Scheme. *Applied Mathematics and Computation, 218(20):10271–10278, 2012*.

Invited Talks in Conferences

1. Southern California Applied Mathematics Symposium, CA US, Jun 4, 2016
2. SIAM Conference on Imaging Science, Albuquerque NM US, May 23 - 26, 2016
3. IEEE Winter Conference on Applications of Computer Vision (WACV), Hawaii US, Jan 8 - 10, 2019
4. IEEE The second International Conference on Artificial Intelligence for Industries (AI4I), Irvine CA US, Sep 21 - 23, 2019
5. The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, Vancouver BC Canada, Dec 13, 2019

Awards

Third Award in Low Latency Object Detection The IEEE Low-Power Image Recognition Challenge (LPIRC) 2019
Second Awards in Both Low Latency Image Classsification and Detection The IEEE Low-Power Image Recognition Challenge (LPIRC) 2018
Kovalevsky Outstanding Ph.D. Thesis Award University of California, Irvine 2017
SIAM Student Travel Award Society for Industrial and Applied Mathematics (SIAM) 2016
Von Neumann Outstanding Research Award University of California, Irvine 2014-2015

Conference Technical Committee and Reviews

AAAI Conference on Artificial Intelligence (AAAI), 2020 and 2021
ACM Multimedia Systems Conference (MMsys), 2019
International Joint Conference on Artificial Intelligence (IJCAI), 2019
Conference on Neural Information Processing Systems (NeurIPS), 2020
European Conference on Computer Vision (ECCV), 2020
IEEE International Conference on Image Processing (ICIP), 2018 and 2019
IEEE Winter Conference on Applications of Computer Vision (WACV), 2019
IEEE International Conference on Computer Vision (ICCV), 2019
IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020
The 5th workshop on Energy Efficient Machine Learning and Cognitive Computing (EMC2), 2019

Journal Reviews

Communications in Mathematical Sciences
IEEE Transactions on Information Forensics and Security
IEEE Transactions on Image Processing
IEEE Transactions on Vehicular Technology
IEEE Transactions on Signal Processing
International Journal of Image and Graphics
Neural Processing Letters, Springer
Neural Networks, Elsevier