Wei Wen

Research Scientist, Facebook AI

Ph.D., Duke University

Facebook, Inc.
1 Hacker Way
Menlo Park, CA 94025

Email: weiwen.web@gmail.com
Homepage: www.pittnuts.com

RESEARCH INTEREST

Machine learning, deep learning
Deep learning in distributed systems and edge devices
Automated machine learning, neural architecture search
Computer vision, recommender & ranking systems, natural language processing

EDUCATION

Ph.D. in Electrical and Computer Engineering, Duke University, United States, December 2019;
Advisor: Hai Li (co-chair) and Yiran Chen (co-chair)
Dissertation: efficient and scalable deep learning.

M.S. in Electronic and Information Engineering, Beihang University, China, January 2013;
Advisor: Rongke Liu
Thesis on video compression.

B.S.* in Electronic and Information Engineering, Beihang University, China, July 2010;
Thesis on low-density parity-check (LDPC) error correcting code.
** admitted to M.S. program by waiving application.*

PROFESSIONAL EXPERIENCE

Research Scientist, Facebook AI, Menlo Park, United States, Aug. 2020 – Present.
Automated machine learning, neural architecture search, recommender & ranking systems

Student Researcher, Google Brain, Mountain View, United States, Sep. 2019 – Nov. 2019.

Research Intern, Google Brain, Mountain View, United States, May 2019 – Aug. 2019.
Co-workers: Pieter-Jan Kindermans, Quoc Le, Jonathon Shlens
Automated machine learning, neural architecture search, graph neural networks

Research Intern, Facebook AI, Menlo Park, United States, May 2018 – Aug. 2018.
Mentor: Yangqing Jia
Distributed deep learning, recommender & ranking systems

Research Intern, Microsoft Research, Redmond, United States, May 2017 – July. 2017.
Mentor: Yuxiong He
Efficient deep neural networks, natural language processing

Research Intern, HP Labs, Palo Alto, United States, Jun. 2016 – Sep. 2016.
Co-workers: Cong Xu, Paolo Farabosch
Distributed deep learning and systems

Software Engineer, Agricultural Bank of China, Beijing, China, July 2013 – July 2014.
Development of online bank transaction

Research Intern, Microsoft Research Asia, Beijing, China, Apr. 2013 – Jun. 2013.
Mobile computer vision

Software Engineer Intern, Tencent, Beijing, China, July 2012 – Sep. 2012.
Advertisement product development

AWARDS, HONORS, AND RECOGNITIONS

Best Paper Award

1. *Best Paper Award*, Asia and South Pacific Design Automation Conference (ASP-DAC) for the paper entitled "Classification accuracy improvement for neuromorphic computing systems with one-level precision synapses," 2017.

Best Paper Nominations or Candidates

2. *Best Paper Candidate*, International Conference on Artificial Intelligence Circuits and Systems (AICAS) for the paper entitled "Exploration of Automatic Mixed-Precision Search for Deep Neural Networks," 2019.
3. *Best Paper Nomination*, Design Automation Conference (DAC) for the paper entitled "A New Learning Method for Inference Accuracy, Core Occupation, and Performance Co-optimization on TrueNorth Chip," 2016.
4. *Best Paper Nomination*, Design Automation Conference (DAC) for the paper entitled "An EDA Framework for Large Scale Hybrid Neuromorphic Computing Systems", 2015.

Other Paper Recognitions

5. Oral Paper, Neural Information Processing Systems (NeurIPS) for the paper entitled "TernGrad: Ternary gradients to reduce communication in distributed deep learning," 2017.

Research Adoption in Industry

6. Facebook AI, TernGrad in PyTorch and Caffe2, accessed on March 31, 2021 at https://github.com/pytorch/pytorch/.
7. Nervana Systems at Intel AI Lab, "Distiller Model Zoo," accessed on February 14, 2020 at https://nervanasystems.github.io/distiller/model_zoo.html#learning-structured-sparsity-in-deep-neural-networks.
8. Intel® AI Developer Program, "Scaling to Meet the Growing Needs of AI," accessed on October 26, 2016 at https://software.intel.com/en-us/articles/scaling-to-meet-the-growing-needs-of-ai.

INVITED TALKS AND SPEECHES

1. Microsoft Research, "Efficient and Scalable Deep Learning", Redmond, United States, Oct. 2019.
2. Rice University, guest lecture at ELEC 515 Embedded Machine Learning, Oct. 2019
3. UC Berkeley, Scientific Computing and Matrix Computations Seminar, "On Matrix Sparsification and Quantization for Efficient and Scalable Deep Learning", Berkeley, United States, Oct. 2018
4. Cornell University, Artificial Intelligence Seminar, "Efficient and Scalable Deep Learning", Ithaca, United States, Oct. 2018
5. IBM T. J. Watson Research Center, AI Compute Symposium, Yorktown Heights, United States, Oct. 2018

LIST OF PUBLICATIONS

I. Google Citations (It may deviate from other records): 3641 (H-index = 22) [Update: Feb. 11, 2022]

II. Peer Reviewed Conference Publications (Total: 29)

C1. Wen, Wei, Hanxiao Liu, Hai Li, Yiran Chen, Gabriel Bender, and Pieter-Jan Kindermans. "Neural predictor for neural architecture search." *In Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

C2. Wen, Wei, Feng Yan, Yiran Chen, and Hai Li. "AutoGrow: automatic layer growing in deep convolutional networks." *In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, pp. 833-841. 2020.

C3. Wen, Wei, Yuxiong He, Samyam Rajbhandari, Minjia Zhang, Wenhan Wang, Fang Liu, Bin Hu, Yiran Chen and Hai Li. "Learning Intrinsic Sparse Structures within Long Short-Term Memory." *In 6th International Conference on Learning Representations (ICLR)*, pp. 1-14. 2018.

C4. Wen, Wei, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. "TernGrad: Ternary gradients to reduce communication in distributed deep learning." *In Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1509-1519. 2017. (**Oral paper, 1.2%**)

C5. Wen, Wei, Cong Xu, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. "Coordinating filters for faster deep neural networks." *In Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 658-666. 2017.

C6. Wen, Wei, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. “Learning structured sparsity in deep neural networks.” *In Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2074-2082. 2016.

C7. Wen, Wei, Chunpeng Wu, Yandan Wang, Kent Nixon, Qing Wu, Mark Barnell, Hai Li, and Yiran Chen. “A new learning method for inference accuracy, core occupation, and performance co-optimization on TrueNorth chip.” *In 2016 53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1-6. IEEE, 2016. (**Best Paper Candidate, 1.83%**)

C8. Wen, Wei, Chi-Ruo Wu, Xiaofang Hu, Beiye Liu, Tsung-Yi Ho, Xin Li, and Yiran Chen. “An EDA framework for large scale hybrid neuromorphic computing systems.” *In 2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1-6. IEEE, 2015. (**Best Paper Candidate, 0.89%**)

C9. Yang, Huanrui, Wei Wen and Hai Li, “DeepHoyer: Learning Sparser Neural Network with Differentiable Scale-Invariant Sparsity Measures.” *In 2020 8th International Conference on Learning Representations (ICLR)*, pp. 1-18. 2020.

C10. Inkawhich, Nathan, Wei Wen, Hai Helen Li, and Yiran Chen. “Feature space perturbations yield more transferable adversarial examples.” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7066-7074. 2019.

C11. Chen, Fan, Wei Wen, Linghao Song, Jingchi Zhang, Hai Helen Li, and Yiran Chen. “How to Obtain and Run Light and Efficient Deep Learning Networks.” *In 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 1-5. IEEE, 2019.

C12. Zhang, Jingchi, Wei Wen, Michael Deisher, Hsin-Pai Cheng, Hai Li, and Yiran Chen. “Learning Efficient Sparse Structures in Speech Recognition.” *In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2717-2721. IEEE, 2019.

C13. Yang, Qing, Wei Wen, Zuoguan Wang, and Hai Li. “Joint Regularization on Activations and Weights for Efficient Neural Network Pruning.” *In 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1-10. IEEE, 2019.

C14. Guo, Xuyang, Yuanjun Huang, Hsin-pai Cheng, Bing Li, Wei Wen, Siyuan Ma, Hai Li, and Yiran Chen. “Exploration of Automatic Mixed-Precision Search for Deep Neural Networks.” *In 2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pp. 276-278. IEEE, 2019. (**Best Paper Candidate**)

C15. Liu, Xiaoxiao, Wei Wen, Xuehai Qian, Hai Li, and Yiran Chen. “Neu-NoC: A high-efficient interconnection network for accelerated neuromorphic systems.” *In 2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 141-146. IEEE, 2018.

C16. Li, Bing, Wei Wen, Jiachen Mao, Sicheng Li, Yiran Chen, and Hai Helen Li. “Running sparse and low-precision neural network: When algorithm meets hardware.” *In 2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 534-539. IEEE, 2018.

C17. Cheng, Hsin-Pai, Yuanjun Huang, Xuyang Guo, Feng Yan, Yifei Huang, Wei Wen, Hai Li and Yiran Chen. “Differentiable Fine-grained Quantization for Deep Neural Network Compression.” *In Compact Deep Neural Network Representation with Industrial Applications in Advances in neural information processing systems WorkShop*, pp. 1-5. 2018.

C18. Wang, Yandan, Wei Wen, Beiye Liu, Donald Chiarulli, and Hai Li. “Group scissor: Scaling neuromorphic computing design to large neural networks.” *In 2017 54th ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1-6. IEEE, 2017.

C19. Li, Sicheng, Wei Wen, Yu Wang, Song Han, Yiran Chen, and Hai Li. “An FPGA design framework for CNN sparsification and acceleration.” *In 2017 IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM)*, pp. 28-28. IEEE, 2017.

C20. Mao, Jiachen, Zhongda Yang, Wei Wen, Chunpeng Wu, Linghao Song, Kent W. Nixon, Xiang Chen, Hai Li, and Yiran Chen. “Mednn: A distributed mobile system with enhanced partition and deployment for large-scale dnns.” *In 2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, pp. 751-756. IEEE, 2017.

C21. Wang, Yandan, Wei Wen, Linghao Song, and Hai Helen Li. “Classification accuracy improvement for neuromorphic computing systems with one-level precision synapses.” *In 2017 22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 776-781. IEEE, 2017. (**Best Paper Award, 0.56%**)

C22. Wu, Chunpeng, Wei Wen, Tariq Afzal, Yongmei Zhang, and Yiran Chen. “A compact dnn: approaching googlenet-level accuracy of classification and domain adaptation.” *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5668-5677. 2017.

C23. Park, Jongsoo, Sheng Li, Wei Wen, Ping Tak Peter Tang, Hai Li, Yiran Chen and Pradeep Dubey. “Faster CNNs with Direct Sparse Convolutions and Guided Pruning.” *In the 5th International Conference on Learning Representations (ICLR)*, pp. 1-12. 2017.

C24. Cheng, Hsin-Pai, Wei Wen, Chunpeng Wu, Sicheng Li, Hai Helen Li, and Yiran Chen. “Understanding the design of IBM neurosynaptic system and its tradeoffs: a user perspective.” *In Design, Automation & Test in Europe Conference & Exhibition (DATE)*, 2017, pp. 139-144. IEEE, 2017.

C25. Wu, Chi-Ruo, Wei Wen, Tsung-Yi Ho, and Yiran Chen. “Thermal optimization for memristor-based hybrid neuromorphic computing systems.” *In 2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 274-279. IEEE, 2016.

C26. Cheng, Hsin-Pai, Wei Wen, Chang Song, Beiye Liu, Hai Li, and Yiran Chen. “Exploring the optimal learning technique for IBM TrueNorth platform to overcome quantization loss.” *In 2016 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pp. 185-190. IEEE, 2016.

C27. Liu, Beiye, Xiaoxiao Liu, Chenchen Liu, Wei Wen, M. Meng, Hai Li, and Yiran Chen. “Hardware acceleration for neuromorphic computing: An evolving view.” *In 2015 15th Non-Volatile Memory Technology Symposium (NVMTS)*, pp. 1-4. IEEE, 2015.

C28. Wang, Yandan, Wei Wen, Hai Li, and Miao Hu. “A novel true random number generator design leveraging emerging memristor technology.” *In Proceedings of the 25th edition on Great Lakes Symposium on VLSI*, pp. 271-276. 2015.

C29. Liu, Beiye, Wei Wen, Yiran Chen, Xin Li, Chi-Ruo Wu, and Tsung-Yi Ho. “Eda challenges for memristor-crossbar based neuromorphic computing.” *In Proceedings of the 25th edition on Great Lakes Symposium on VLSI*, pp. 185-188. 2015.

III. Refereed Journal Articles (Total: 2)

J1. Chen, Yiran, Hai Helen Li, Chunpeng Wu, Chang Song, Sicheng Li, Chuhan Min, Hsin-Pai Cheng, Wei Wen, and Xiaoxiao Liu. “Neuromorphic computing's yesterday, today, and tomorrow–an evolutional view.” *Integration* 61 (2018): 49-61.

J2. Hu, Miao, Yandan Wang, Wei Wen, Yu Wang, and Hai Li. “Leveraging stochastic memristor devices in neuromorphic hardware systems.” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 6, no. 2 (2016): 235-246.

IV. Publicly Accessible Technical Reports (Total: 2)

T1. Wen, Wei, Yandan Wang, Feng Yan, Cong Xu, Chunpeng Wu, Yiran Chen, and Hai Li. “Smoothout: Smoothing out sharp minima to improve generalization in deep learning.” arXiv preprint arXiv:1805.07898, 2018.

T2. Park, Jongsoo, Sheng R. Li, Wei Wen, Hai Li, Yiran Chen, and Pradeep Dubey. “Holistic sparsecnn: Forging the trident of accuracy, speed, and size.” arXiv preprint arXiv:1608.01409 1, no. 2, 2016.

PROFESSIONAL SERVICE ACTIVITIES

Journal and Conference Referee

1. Neural Information Processing Systems (NeurIPS)
2. International Conference on Machine Learning (ICML)
3. International Conference on Learning Representations (ICLR)
4. Computer Vision and Pattern Recognition (CVPR)
5. International Conference on Computer Vision (ICCV)
6. European Conference on Computer Vision (ECCV)
7. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
8. International Journal of Computer Vision (IJCV)
9. IEEE Transactions on Neural Networks and Learning Systems (TNNLS)
10. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)
11. ACM Transactions on Embedded Computing Systems (TECS)

12. IEEE Transactions on Knowledge and Data Engineering (TKDE)
13. IEEE Transactions on VLSI Systems (TVLSI)
14. IEEE International Conference on Multimedia (ICME)
15. Journal of Parallel and Distributed Computing (JPDC)
16. Neurocomputing
17. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)
18. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)

Panelist

1. Panelist of National Science Foundation (NSF) Workshop on Machine Learning Hardware Breakthroughs, 2020

Professional Event Volunteer

1. Machine Learning for Girls, FEMMES (Female Excelling More in Math, Engineering, and Science) Capstone, Duke University, Feb. 2018
2. Embedded Systems Week (ESWEEK), Pittsburgh, United States, Oct. 2016

MEDIA REPORTS

Duke Electrical and Computer Engineering Ph.D. program cover page, "Q&A: Wei Wen. Making deep learning models faster & more efficient," accessed on February 14, 2020 at https://ece.duke.edu/phd/students/wen.

TEACHING ACTIVITIES

Teach Assistant, CEE 690/ECE 590: Introduction to Deep Learning, Duke University, Fall 2018

Teach Assistant, STA561/COMPSCI571/ECE682: Probabilistic Machine Learning, Duke University, Spring 2019