# Wei Wen

412-944-8906 | weiwen.web@gmail.com | http://www.pittnuts.com/

## EDUCATION

**Ph.D.** in Electrical and Computer Engineering, Duke University, USA 08/2014-12/2019

(Note: first three years were spent at University of Pittsburgh and then moved to Duke University with my advisors)

Research Interest: Deep Learning & Machine Learning & Neuromorphic Computing

Advisors: Dr. Hai Li & Dr. Yiran Chen. GPA: 4.00/4.00 (Duke), 3.96/4.00 (UPitt)

**M.S.** in Electronic and Information Engineering, Beihang University, China 09/2010-01/2013

**B.S.** in Electronic and Information Engineering, Beihang University, China 09/2006-07/2010

## BIO

Wei Wen obtained his Ph.D. degree from Duke University in 2019. His research is Machine Learning with focuses on efficient and scalable deep learning, and automated machine learning. He has published 31 papers in top-tier conferences and journals, receiving 1,404 citations based on Google Scholar. He received one Best Paper Award and four Best Paper Candidates. He was invited to give talks in UC Berkeley, Cornell University, Rice University, Microsoft Research and NeurIPS 2017. His achievements have been covered by several medias including Duke ECE Ph.D. program, Intel AI Developer Program, and Nervana Systems. He interned at prestigious research institutes including Google Brain, Facebook AI, Microsoft Research and HP Labs. Some of his methods have been deployed into industrial products, such as Facebook AI Infra, Intel Nervana and PyTorch/Caffe2. Dr. Wen is a reviewer of six top-tier conferences and seven journals. He is also an active open source contributor in GitHub, with 1,243 contributions.

## INDUSTRIAL EXPERIENCE

**Google Brain,** Student Researcher, Durham, NC, USA 09/2019-11/2019

Research Intern, Mountain View, CA, USA 05/2019-08/2019

Mentors: Pieter-Jan Kindermans & Gabriel Bender. Lead: Quoc Le & Jonathon Shlens.

- Automated Machine Learning (AutoML), using machine learning to design machine learning models.

**Facebook AI**, Research Intern, Menlo Park, CA, USA 05/2018-08/2018

Mentor: Yangqing Jia

- Personalization and distributed machine learning.

**Microsoft Research Redmond**, Research Intern, Redmond, WA, USA 05/2017-07/2017

Mentor: Yuxiong He

- Model compression and efficient recurrent neural networks.

**HP Labs**, Platform Architecture Group, Research Intern, Palo Alto, CA, USA 06/2016-09/2016

**Agricultural Bank of China**, Software Engineer Employee, Beijing, China 07/2013-07/2014

**Microsoft Research Asia**, Mobile and Sensing Systems Group, Research Intern, Beijing, China 04/2013-06/2013

## SELECTED HONORS & AWARDS

- Best Student Paper Finalist (3.5%), Supercomputing Conference (SC) 2019
- Best Paper Candidate, International Conference on Artificial Intelligence Circuits and Systems (AICAS), IEEE 2019
- Best Paper Award (0.56%), Asia and South Pacific Design Automation Conference (ASP-DAC), IEEE 2017
- NeurIPS Oral Paper (1.2%), Neural Information Processing Systems (NeurIPS) 2017
- Best Paper Candidate (1.83%), Design Automation Conference (DAC), IEEE 2016
- Best Paper Candidate (0.89%), Design Automation Conference (DAC), IEEE 2015

## MEDIA

- "Earn Your PhD at Duke." Duke Electrical and Computer Engineering, Accessed February 14, 2020. https://ece.duke.edu/phd.
- "Q&A: Wei Wen. Making deep learning models faster & more efficient." Duke Electrical and Computer Engineering, Accessed Feburary 14, 2020. https://ece.duke.edu/phd/students/wen.
- Dubey, Pradeep and Amir Khosrowshahi. "Scaling to Meet the Growing Needs of AI." Intel® AI Developer Program. October 26, 2016. https://software.intel.com/en-us/articles/scaling-to-meet-the-growing-needs-of-ai.
- "Distiller Model Zoo." Neural Network Distiller, Nervana Systems at Intel AI Lab. Accessed February 15, 2020. https://nervanasystems.github.io/distiller/model_zoo.html#learning-structured-sparsity-in-deep-neural-networks.

## INVITED TALKS

- Speaker, Microsoft Research Talks, “Efficient and Scalable Deep Learning”, 10/10/2019
- Poster Presenter, Microsoft Research, AI Breakthroughs Workshop, “Efficient and Scalable Deep Learning”, 9/15/2019
- Guest Lecturer, Rice University, ELEC 515 Embedded Machine Learning, “Regularization and Case Study of Unstructured Pruning: Learning Structured Sparsity in DNNs”, 10/16/2019
- Invited Speaker, UC Berkeley, Scientific Computing and Matrix Computations Seminar, “On Matrix Sparsification and Quantization for Efficient and Scalable Deep Learning”, 10/10/2018
- Invited Speaker, Cornell University, Fall 2018 Artificial Intelligence Seminar, “Efficient and Scalable Deep Learning”, 10/05/2018
- Oral Speaker, Thirty-first Conference on Neural Information Processing Systems, “TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning”, 12/6/2017
- Best Paper Candidate Presenter, 53rd ACM/EDAC/IEEE Design Automation Conference (DAC), “A new learning method for inference accuracy, core occupation, and performance co-optimization on TrueNorth chip”, 6/7/2016
- Best Paper Candidate Speaker, 52nd ACM/EDAC/IEEE Design Automation Conference (DAC), “An EDA framework for large scale hybrid neuromorphic computing systems”, 6/9/2015

## OPEN SOURCE CONTRIBUTIONS

- TernGrad in PyTorch, Facebook AI, https://github.com/pytorch/pytorch.
- SkimCaffe, Intel Labs, https://github.com/IntelLabs/SkimCaffe.
- Individual projects, GitHub, https://github.com/wenwei202.

## PUBLICATIONS

***31 publications with 1,404 citations accessed on 02/16/2020 from Google Scholar***

- 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.
- 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.
- 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.
- 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.
- 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 *53nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1-6. IEEE, 2016.
- 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 *52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*, pp. 1-6. IEEE,

2015.

- Yang, Huanrui, Wei Wen and Hai Li, “DeepHoyer: Learning Sparser Neural Network with Differentiable Scale-Invariant Sparsity Measures.” In *8th International Conference on Learning Representations (ICLR)*, pp. 1-18. 2020.
- 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.
- 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.
- Lym, Sangkug, Esha Choukse, Siavash Zangeneh, Wei Wen, Sujay Sanghavi, and Mattan Erez. "PruneTrain: fast neural network training by dynamic sparse model reconfiguration." In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (Supercomputing, SC)*, pp. 1-13. 2019.
- 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.
- Yang, Qing, Wei Wen, Zuoguan Wang, and Hai Li. "Joint Regularization on Activations and Weights for Efficient Neural Network Pruning." In *18th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1-10. IEEE, 2019.
- Lym, Sangkug, Armand Behroozi, Wei Wen, Ge Li, Yongkee Kwon, and Mattan Erez. "Mini-batch Serialization: CNN Training with Inter-layer Data Reuse." In *3rd Conference on Machine Learning and Systems (MLSys)*, pp. 1-12. 2019
- 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.
- Liu, Xiaoxiao, Wei Wen, Xuehai Qian, Hai Li, and Yiran Chen. "Neu-NoC: A high-efficient interconnection network for accelerated neuromorphic systems." In *23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 141-146. IEEE, 2018.
- 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 *23rd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 534-539. IEEE, 2018.
- 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.
- 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, the VLSI Journal*, 61 (2018): 49-61.
- Wang, Yandan, Wei Wen, Beiye Liu, Donald Chiarulli, and Hai Li. "Group scissor: Scaling neuromorphic computing design to large neural networks." In *54th ACM/EDAC/IEEE Design Automation Conference (DAC),* pp. 1-6. IEEE, 2017.
- Li, Sicheng, Wei Wen, Yu Wang, Song Han, Yiran Chen, and Hai Li. "An FPGA design framework for CNN sparsification and acceleration." In *IEEE 25th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM),* pp. 28-28. IEEE, 2017.
- 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.
- Wang, Yandan, Wei Wen, Linghao Song, and Hai Helen Li. "Classification accuracy improvement for neuromorphic computing systems with one-level precision synapses." In *22nd Asia and South Pacific Design Automation Conference (ASP-DAC)*, pp. 776-781. IEEE, 2017.
- 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.

- 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 *5th International Conference on Learning Representations (ICLR),* pp. 1-12. 2017.
- 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 *2017 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 139-144. IEEE, 2017.
- 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.
- Wu, Chi-Ruo, Wei Wen, Tsung-Yi Ho, and Yiran Chen. "Thermal optimization for memristor-based hybrid neuromorphic computing systems." In *21st Asia and South Pacific Design Automation Conference (ASP-DAC),* pp. 274-279. IEEE, 2016.
- 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.
- Liu, Beiye, Xiaoxiao Liu, Chenchen Liu, Wei Wen, M. Meng, Hai Li, and Yiran Chen. "Hardware acceleration for neuromorphic computing: An evolving view." In *15th Non-Volatile Memory Technology Symposium (NVMTS)*, pp. 1-4. IEEE, 2015.
- 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.
- 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.

## REVIEW SERVICE

*Journals:*

- IEEE Transactions on Neural Networks and Learning Systems (TNNLS)
- IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)
- IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)
- IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
- International Journal of Computer Vision (IJCV)
- Journal of Parallel and Distributed Computing (JPDC)
- Neurocomputing

*International conferences:*

- 2020 International Conference on Learning Representations (ICLR)
- 2020 Conference on Computer Vision and Pattern Recognition (CVPR)
- 2019 Conference on Neural Information Processing Systems (NeurIPS)
- 2019 International Conference on Learning Representations (ICLR)
- 2019 Conference on Computer Vision and Pattern Recognition (CVPR)
- 2019 International Conference on Computer Vision (ICCV)
- 2019 IEEE International Conference on Multimedia and Expo (ICME)
- 2019 ICCV Workshop on Low-Power Computer Vision

## TEACHING

- 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