Wei Wen

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)

Dissertation: Efficient and Scalable Deep Learning

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

RESEARCH INTERESTS

Machine learning and deep learning, including automated machine learning, efficient neural networks and distributed machine learning, with applications to computer vision, natural language processing, and recommender & ranking systems.

INDUSTRIAL EXPERIENCE

Facebook AI, Research Scientist, Menlo Park, CA, USA 08/2020-Now

- Automated machine learning and neural architecture search.
- Large-scale deep learning.
- Recommender and ranking systems.

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

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

Mentor: Pieter-Jan Kindermans. 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

- AI personalization and machine learning fundamentals.

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

Mentor: Cong Xu

- Distributed deep learning.

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

SELECTED PUBLICATIONS

- **W. Wen**, H. Liu, H. Li, Y. Chen, G. Bender, P.-J. Kindermans, "Neural Predictor for Neural Architecture Search", *European Conference on Computer Vision (**ECCV**)*. 2020
- **W. Wen**, F. Yan, Y. Chen, H. Li, "AutoGrow: Automatic Layer Growing in Deep Convolutional Networks", *SIGKDD Conference on Knowledge Discovery and Data Mining (**KDD**)*. 2020. [Research Track: 216/1279=16.8%]
- H. Yang, **W. Wen**, H. Li, "DeepHoyer: Learning Sparser Neural Network with Differentiable Scale-Invariant Sparsity Measures." In *International Conference on Learning Representations (**ICLR**)*. 2020.
- N. Inkawhich, **W. Wen**, H. Li, Y. Chen. "Feature space perturbations yield more transferable adversarial examples." In *Computer Vision and Pattern Recognition (**CVPR**)*. 2019.
- S. Lym, E. Choukse, S. Zangeneh, **W. Wen**, S. Sanghavi, M. Erez. "PruneTrain: fast neural network training by dynamic sparse model reconfiguration." In *International Conference for High Performance Computing, Networking, Storage and Analysis (**SC**)*. 2019. [**Best Student Paper Finalist, 3.5%]**
- S. Lym, A. Behroozi, **W. Wen**, G. Li, Y. Kwon, M. Erez. "Mini-batch Serialization: CNN Training with Inter-layer Data Reuse." *In Conference on Machine Learning and Systems (**MLSys**)*. 2019
- **W. Wen**, Y. He, S. Rajbhandari, M. Zhang, W. Wang, F. Liu, B. Hu, Y. Chen, H. Li. "Learning Intrinsic Sparse Structures within Long Short-Term Memory." In *International Conference on Learning Representations (**ICLR**)*. 2018.
- **W. Wen**, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, H. Li. "TernGrad: Ternary gradients to reduce communication in distributed deep learning." In *Advances in neural information processing systems (**NeurIPS**)*. 2017. **[Oral, 1.2%]**
- **W. Wen**, C. Xu, C. Wu, Y. Wang, Y. Chen, H. Li. "Coordinating filters for faster deep neural networks." In *Proceedings of the IEEE International Conference on Computer Vision (**ICCV**)*. 2017.
- Y. Wang, **W. Wen**, L. Song, H. Li. "Classification accuracy improvement for neuromorphic computing systems with one-level precision synapses." In *Asia and South Pacific Design Automation Conference (**ASP-DAC**)*, 2017. [**Best Paper Award, 0.56%**]
- C. Wu, **W. Wen**, T. Afzal, Y. Zhang, Y. Chen, H. Li. "A compact dnn: approaching googlenet-level accuracy of classification and domain adaptation." In *Computer Vision and Pattern Recognition (**CVPR**)*. 2017.
- S. Park, S. Li, **W. Wen**, P. T. P. Tang, H. Li, Y. Chen, P. Dubey. "Faster CNNs with Direct Sparse Convolutions and Guided Pruning." In *International Conference on Learning Representations (**ICLR**)*. 2017.
- **W. Wen**, C. Wu, Y. Wang, Y. Chen, H. Li. "Learning structured sparsity in deep neural networks." In *Advances in neural information processing systems (**NeurIPS**)*. 2016.
- **W. Wen**, C. Wu, Y. Wang, K. Nixon, Q. Wu, M. Barnell, H. Li, Y. Chen. "A new learning method for inference accuracy, core occupation, and performance co-optimization on TrueNorth chip." In *Design Automation Conference (**DAC**)*. 2016. [**Best Paper Candidate, 1.83%]**
- **W. Wen**, C.-R. Wu, X. Hu, B. Liu, T.-Y. Ho, X. Li, Y. Chen. "An EDA framework for large scale hybrid neuromorphic computing systems." In *Design Automation Conference (**DAC**)*. 2015. [**Best Paper Candidate, 0.89%**]

INVITED TALKS

- Speaker, Microsoft Research Talks, "Efficient and Scalable Deep Learning", 10/10/2019
- Guest Lecturer, Rice University, ELEC 515 Embedded Machine Learning, 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, Artificial Intelligence Seminar, "Efficient and Scalable Deep Learning", 10/05/2018

MEDIA

- "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.

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

SKILLS

- PyTorch, TensorFlow, Caffe2, Python, C/C++, CUDA