Wei Wen

412-944-8906 | weiwen.web@gmail.com | http://www.pittnuts.com/ | https://github.com/wenwei202

EDUCATION

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

Advisors: Dr. Hai Li & Dr. Yiran Chen (Note: first 3 years spent at U. of Pittsburgh and then moved to Duke with my advisors)

Research Interest: Deep Learning & Machine Learning & Neuromorphic Computing

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 SUMMARY

My research is Machine Learning with focuses on efficient deep learning, scalable deep learning, and automated machine learning. I was invited to give talks in UC Berkeley, Cornell University and NeurIPS 2017 oral. I worked with Google Brain, Facebook AI, Microsoft Research and HP Labs. Some of my research methods have been deployed into AI productions, such as Facebook AI Infra, Intel Nervana and PyTorch/Caffe2. I have authored/co-authored one Best Paper Award and three Best Paper Nominations.

INDUSTRIAL RESEARCH

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

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 Research, AI Infra and Applied Machine Learning, 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 PUBLICATIONS

34 publications with 968 citations accessed on 08/08/2019 from Google Scholar

- W. Inkawhich, **W. Wen**, H. Li, Y. Chen, "Feature Space Perturbations Yield More Transferable Adversarial Examples." ***CVPR***, 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", ***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", ***NeurIPS***, 2017. (***Oral, 40/3240=1.2%***) (Open sourced in PyTorch/Caffe2)
- **W. Wen,** C. Xu, C. Wu, Y. Wang, Y. Chen, H. Li, "Coordinating Filters for Faster Deep Neural Networks", ***ICCV***, 2017.
- **W. Wen,** C. Wu, Y. Wang, Y. Chen, H. Li, "Learning Structured Sparsity in Deep Neural Networks", ***NeurIPS***, 2016. (Adopted by Intel Nervana)
- J. Park, S. Li, **W. Wen**, P. T. P. Tang, H. Li, Y. Chen, P. Dubey, "Faster CNNs with Direct Sparse Convolutions and Guided Pruning", ***ICLR***, 2017.
- C. Wu, **W. Wen**, T. Afzal, Y. Zhang, Y. Chen, H. Li, "A Compact DNN: Approaching GoogLeNet-Level Accuracy of Classification and Domain Adaptation", ***CVPR***, 2017.

SELECTED PUBLICATIONS (CONTINUED)

- S. Lym, E. Choukse, S. Zangeneh, **W. Wen**, S. Sanghavi, M. Erez, "PruneTrain: Fast Neural Network Training by Dynamic Sparse Model Reconfiguration", International Conference for High Performance Computing, Networking, Storage and Analysis (***SC***), 2019. (***Best Student Paper Finalist***)
- Y. Wang, **W. Wen,** L. Song, H. Li, "Classification Accuracy Improvement for Neuromorphic Computing Systems with One-level Precision Synapses", ***ASP-DAC***, 2017. (***Best Paper Award***)
- **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", the 53rd IEEE Design Automation Conference (***DAC***), 2016. Acceptance Rate: 152/876=17.4%. (***Best Paper Nomination, 16/876=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", the 52nd IEEE Design Automation Conference (***DAC***), 2015. Acceptance Rate: 162/789=20.5%. (***Best Paper Nomination, 7/789=0.89%***).

SELECTED HONORS & AWARDS

- Best Student Paper Finalist, Supercomputing Conference (SC) 2019
- Best Paper Award, Asia and South Pacific Design Automation Conference (ASP-DAC), IEEE 2017
- Best Paper Nomination, Design Automation Conference (DAC), IEEE 2016
- Best Paper Nomination, Design Automation Conference (DAC), IEEE 2015
- National Scholarship (3/233), Ministry of Education China 2009
- Second Prize, National College Physics Competition China 2007

SKILLS

- Machine learning: TensorFlow, PyTorch, Caffe
- Languages: Python, C/C++
- Android Development (with Google Play publications)

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

TALKS

- Microsoft Research AI Breakthroughs invitation, 9/15/2019
- UC Berkeley, Scientific Computing and Matrix Computations Seminar, "On Matrix Sparsification and Quantization for Efficient and Scalable Deep Learning", 10/10/2018
- Cornell University, AI Seminar, "Efficient and Scalable Deep Learning", 10/05/2018
- NeurIPS 2017, TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning, 12/6/2017
- Alibaba DAMO Academy, "Deep Learning in Cloud-Edge AI Systems", SunnyVale, CA, 06/28/2018

SERVICE

- Paper reviewer, NeurIPS, ICML, ICLR, CVPR, ICCV, TPAMI, IJCV, TNNLS, TCAD, Neurocomputing, etc.
- Activity volunteer, Machine Learning for Girls, FEMMES (Female Excelling More in Math, Engineering, and Science) Capstone at Duke University, 02/2018
- Conference volunteer, Embedded Systems Week (ESWEEK), Pittsburgh, PA, USA, 10/2016