Wei WEN

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

EDUCATION

Duke University ***Durham, NC, United States*** ***08/2017-12/2019 (Expected)***

University of Pittsburgh *Pittsburgh, PA, United States* *09/2014-08/2017 (Transferred to Duke U)*

Ph.D. Electrical and Computer Engineering Supervisor: Dr. Hai Li

Research Area: Deep Learning & Machine Learning & Neuromorphic Computing

Beihang University ***Beijing, China,*** ***09/2006-07/2010 (B.S.), 09/2010-01/2013(M.S.)***

B.S., M.S. Electronic and Information Engineering

SELECTED PUBLICATION

- **Wei Wen,** Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, Hai Li, "*TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning*", the 31st Annual Conference on Neural Information Processing Systems (***NIPS***), 2017. (***Oral, 40/3240=1.2%***)
- **Wei Wen**, Yuxiong He, Samyam Rajbhandari, Minjia Zhang, Wenhan Wang, Fang Liu, Bin Hu, Yiran Chen, Hai Li, "*Learning Intrinsic Sparse Structures within Long Short-Term Memory*",the 6th International Conference on Learning Representations (***ICLR***), 2018.
- **Wei Wen,** Cong Xu, Chunpeng Wu, Yandan Wang, Yiran Chen, Hai Li, "*Coordinating Filters for Faster Deep Neural Networks*", Proceedings of the IEEE International Conference on Computer Vision (***ICCV***), 2017.
- **Wei Wen,** Chunpeng Wu, Yandan Wang, Yiran Chen, Hai Li, "*Learning Structured Sparsity in Deep Neural Networks*", the 30th Annual Conference on Neural Information Processing Systems (***NIPS***), 2016. (Integrated into Intel Nervana)
- Jongsoo Park, Sheng Li, **Wei Wen**, Ping Tak Peter Tang, Hai Li, Yiran Chen, Pradeep Dubey, "*Faster CNNs with Direct Sparse Convolutions and Guided Pruning*", the 5th International Conference on Learning Representations (***ICLR***), 2017.
- Chunpeng Wu, **Wei Wen**, Tariq Afzal, Yongmei Zhang, Yiran Chen, Hai Li, "*A Compact DNN: Approaching GoogLeNet-Level Accuracy of Classification and Domain Adaptation*", ***CVPR***, 2017.
- Yandan Wang, **Wei Wen,** Linghao Song, Hai Li, "*Classification Accuracy Improvement for Neuromorphic Computing Systems with One-level Precision Synapses* ", ***ASP-DAC***, 2017. (***Best Paper Award***)

INDUSTRY EXPERIENCE

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

Research Intern. Mentor: Yangqing Jia & co-mentor: Jongsoo Park

- Caffe2.
- Distributed Machine Learning.

Microsoft Research & Business AI*, Redmond & Bellevue, WA, USA* *05/2017-07/2017*

Research Intern. Mentors: Yuxiong He & Fang Liu

- Efficient inference methods for Machine Reading Comprehension and Recurrent Neural Networks.

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

Research Intern. Mentor: Cong Xu & supervisor: Paolo Faraboschi

- Benchmarked Distributed Deep Learning Systems.

Agricultural Bank of China*, Software Development Center, Beijing, China* *07/2013-07/2014*

Software Developer Employee. Supervisor: Lei Fan

- Developed web services for online bank transactions.

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

Research Intern. Mentor: Guobin Shen

- Computer vision on mobile devices.

Tencent Inc.*, Advertising Platform and Products Division, Beijing, China* *07/2012-09/2012*

Summer Intern. Supervisor: Yanan Zhao

- Developed MVC-framework-based advertising websites.

RESEARCH PROJECTS

Distributed Deep Learning and Large-Batch Training *01/2017-Now*

- Quantized floating gradients in SGD to overcome communication bottleneck in distributed training of Deep Neural Networks, to accelerate training speed.
- Eliminating sharp minima in large-batch size training, to improve the scalability of distributed training and to improve the generalization of converged models.

Efficient Inference Methods for Deep Learning *09/2016-03/2017*

- Worked on structurally sparse Deep Neural Networks (CNNs & RNNs) to accelerate the inference.
- Proposed methods to learn the number of filters, channels, neurons, layers and hidden sizes in Deep Neural Networks.
- Enabled regular patterns in sparse weights and obtained higher speedup than random connection pruning.
- Improve low-rank approximation methods to obtain faster deep neural networks.

AI Chip *09/2015-12/2015*

- Developed a new learning method for spiking neural networks in IBM TrueNorth chip.
- Proposed Iterative Spectral Clustering algorithm to group connections of large-scale sparse neural networks into small clusters, so that connections can be locally and densely realized by brain-inspired circuit systems.

SKILLS

- Machine learning: PyTorch, TensorFlow, Caffe
- Languages: C/C++/CUDA C, Python & numpy
- Linux, Bash Shell, git and svn
- Android Development (with Google Play publications)

SELECTED HONORS & AWARDS

- ICLR Travel Award 2018
- Graduate Student Conference Travel Fellowship, Duke ECE 2017
- NIPS Travel Award 2017
- Best Paper Award, Asia and South Pacific Design Automation Conference (ASP-DAC), IEEE 2017
- NIPS Travel Award 2016
- 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

ACADEMIC ACTIVITIES

- Paper reviewer, Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 05/2018
- Paper reviewer, Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 04/2018
- Activity volunteer, Machine Learning for Girls, FEMMES (Female Excelling More in Math, Engineering, and Science) Capstone at Duke University, 02/2018
- Paper reviewer, Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 02/2018
- Paper reviewer, IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 01/2018
- Paper reviewer, IEEE Transactions on Neural Networks and Learning Systems (TNNLS), 08/2017
- Conference volunteer, Embedded Systems Week (ESWEEK), Pittsburgh, PA, USA, 10/2016
- Paper reviewer, NIPS 2016