

# Wei WEN

412-944-8906 | [weiwen.web@gmail.com](mailto:weiwen.web@gmail.com) | <http://www.pittnuts.com/> | <https://github.com/wenwei202>

## EDUCATION

---

|  |  |  |
|--|--|--|
| <b>Duke University</b>   | <b>Durham, NC, United States</b>       | <b>08/2017-12/2019 (Expected)</b>                    |
| University of Pittsburgh   | Pittsburgh, PA, United States          | 09/2014-08/2017 (Transferred to Duke U)              |
| <u>Ph.D.</u>   | Electrical and Computer Engineering    | Supervisor: Dr. Hai Li                               |
| Research Area: Deep Learning & Machine Learning & Neuromorphic Computing |  |  |
| <b>Beihang University</b>  | <b>Beijing, China,</b>                 | <b>09/2006-07/2010 (B.S.), 09/2010-01/2013(M.S.)</b> |
| <u>B.S., M.S.</u>  | Electronic and Information Engineering | Rank: 10/170 (B.S.)                                  |

## SELECTED PUBLICATION

---

- **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, 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**, 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.
- 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

---

|   |                        |
|---|------------------------|
| <b>Facebook Research, Menlo Park, CA, USA</b>                                   | <b>05/2018-08/2018</b> |
| Research Intern, Supervisors: Yangqing Jia & Jongsoo Park                       |                        |
| • Caffe2.   |                        |
| <b>Microsoft Research, AI &amp; Research, Redmond, WA, USA</b>                  | <b>05/2017-07/2017</b> |
| Research Intern, Supervisors: Yuxiong He & Fang Liu                             |                        |
| • Machine Reading Comprehension and Recurrent Neural Networks.                  |                        |
| <b>HP Labs, Platform Architecture Group, Palo Alto, CA, USA</b>                 | <b>06/2016-09/2016</b> |
| Research Intern, Supervisors: Dr. Paolo Faraboschi and Dr. Cong Xu              |                        |
| • Benchmarked Distributed Deep Learning Systems.                                |                        |
| <b>Agricultural Bank of China, Software Development Center, Beijing, China</b>  | <b>07/2013-07/2014</b> |
| Software Developer Employee, Supervisor: Mr. Lei Fan                            |                        |
| • Developed web services for online bank transactions.                          |                        |
| <b>Microsoft Research, Mobile and Sensing Systems Group, Beijing, China</b>     | <b>04/2013-06/2013</b> |
| Research Intern, Supervisor: Dr. Guobin Shen                                    |                        |
| • Computer vision on mobile devices.  |                        |
| <b>Tencent Inc., Advertising Platform and Products Division, Beijing, China</b> | <b>07/2012-09/2012</b> |
| Summer Intern, Supervisor: Mr. 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 generalization of converged models

### *Model Compression and Acceleration*

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), 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