# SHAOKAI YE

◇ Addr: 75 Saint Alphonsus street Apt. 1616, Boston, 02120, MA, US ◇ Cell: (314)5849238 ◇ Email: shaokaiyeah@gmail.com
◇ Homepage: yeshaokai.github.io ◇ Google Scholar: googlescholar/shaokaiye **(Citations: 65)**

## AREAS OF INTEREST

Unlike human brain, neural networks are very resource consuming and lacking of robustness to environmental perturbation or algorithm-designed perturbation. My research interest is closing the gap between neural networks and human brain by studying how brain works and how we can compress neural networks and improve their robustness.

- Brain Computer Interface
- Energy-Efficient and High-Performance Deep learning and Artificial Intelligence Systems
- Stability, Robustness & Interpretability of Deep Neural Networks

## EDUCATION

- **Syracuse University, Syracuse, NY** 01/2017 - 12/2018
  M.S. in Computer Engineering
  Advisor: Prof. Yanzhi Wang
  Exchange research assistant at Northeastern University, Boston
- **Saint Louis University, St. Louis, MO** 08/2011 - 05/2015
  B.S. in Computer Engineering
  Advisor: Prof. Michael H. Goldwasser

## RESEARCH ACTIVITIES

### Publications

Authors with * signs contribute equally to the papers.

[1] **Shaokai Ye***, Kaidi Xu*, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, Xue Lin, *Adversarial Robustness vs Model Compression, or Both?*, International Conference on Computer Vision 2019. **(ICCV2019)**

[2] Tianyun Zhang*, **Shaokai Ye***, Kaiqi Zhang, Jian Tang, Wujie Wen, Makan Fardad, Yanzhi Wang, *A systematic DNN weight pruning framework using alternating direction method of multipliers*, European Conference on Computer Vision 2018. **(ECCV2018)**

[3] Ao Ren*, Tianyun Zhang*, **Shaokai Ye**, Jiayu Li, Wenyao Xu, Xuehai Qian, Xue Lin,Yanzhi Wang, *ADMM-NN: An Algorithm-Hardware Co-Design Framework of DNNs Using Alternating Direction Methods of Multipliers*, Architecture Support for Programming Languages and Operating Systems 2019. **(ASPLOS2019)**

[4] Siyue Wang, Xiao Wang, **Shaokai Ye**, Pu Zhao, Xue Lin, *Defending DNN Adversarial Attacks with Pruning and Logits Augmentation*, IEEE Signal Processing for Adversarial Machine Learning 2018. **(GlobalSIP2018 Oral)**

### Preprints

[6] **Shaokai Ye***, Tianyun Zhang*, Kaiqi Zhang*, Jiayu Li, Kaidi Xu, Yunfei Yang, Fuxun Yu, Jian Tang, Makan Fardad, Sijia Liu, Xiang Chen, Xue Lin,Yanzhi Wang, *Progressive Weight Pruning of Deep Neural Networks Using ADMM*, arXiv:1810.07378.

[7] **Shaokai Ye***, Xiaoyu Feng*, Tianyun Zhang, Xiaolong Ma, Sheng Lin, Zhengang Li, Kaidi Xu, Wujie Wen, Sijia Liu, Jian Tang, Makan Fardad, Xue Lin, Yongpan Liu, Yanzhi Wang, *Progressive DNN Compression: A Key to Achieve Ultra-High Weight Pruning and Quantization Rates using ADMM*, arXiv:1903.09769

[8] **Shaokai Ye**, Sia Huat Tan, Kaidi Xu, Yanzhi Wang, Chenglong Bao, Kaisheng Ma, *Brain-inspired reverse adversarial examples*, arXiv:1905.12171

[9] Yanzhi Wang, **Shaokai Ye**, Zhezhi He, Xiaolong Ma, Linfeng Zhang, Sheng Lin, Geng Yuan, Sia Huat Tan, Zhengang Li, Deliang Fan, Xuehai Qian, Xue Lin, Kaisheng Ma, *Non-structured DNN Weight Pruning Considered Harmful*, arXiv:1907.02124

[10] Tianyun Zhang*, **Shaokai Ye***, Kaiqi Zhang*, Jiayu Li, Jian Tang, Wujie Wen, Xue Lin, Makan Fardad, Yanzhi Wang, *ADAM-ADMM: A Unified Systematic Framework of Structured pruning for DNNs*, arXiv:1807.11091.

**Academic Talks**

[1] "Emerging use of ADMM in Deep Learning", SenseTime Inc., 01/15, 2019. **Invited Speaker.**

[2] "Reliable binary quantization for deep neural networks using progressive ADMM", Tsinghua University, 12/25, 2018. **Invited Speaker.**

## RESEARCH EXPERIENCE

**A Unified Framework of Model Compression for DNNs, Tsinghua University, Northeastern University & Syracuse University** 12/2017 - Present

*Exchange Research Assistant*

Deep Neural Networks suffer from extra-large model size and computation requirement, and effective model compression techniques are required to reduce these costs. There are multiple sources of redundancy in DNNs, such as weight redundancy, bit representation of weight, redundancy in intermediate results and bit representation of intermediate results. When all four sources of redundancy are exploited, both storage size and computational cost can be reduced effectively. However, most prior works use heuristic methods and there is a lack of a unifying framework for DNN model compression. As a powerful technique in optimization, ADMM (Alternating Direction Method of Multipliers) can effectively deal with a subset of combinatorial constraints even for non-convex problems. Observing that weight pruning and weight quantization belong to category of clustering problems, I proposed a unified framework to incorporate ADMM and utilize all aforementioned redundancy sources in DNN representation.

However, there is a challenge in the direct application of ADMM in DNN training due to the non-convex objective function. Therefore, I proposed progressive ADMM to guarantee solution feasibility, solution quality and shorter training time. The proposed framework achieved the **highest** weight pruning rates for DNNs in benchmark datasets (By the time when paper [2] was released) as well as the **highest** compression rate when weight pruning and weight quantization are combined (By the time when paper [6] was released).

In addition to the pursuit of compression rate, I've realized that model compression might have sacrificed performance of neural networks, such as robustness. While large capacity benefits robustness, model compression tries to reduce models' capacity and this seems to be a dilemma. In the work [1], I study the intriguing property of weight pruning in adversarial setting and proposed a concurrent adversarial training and model compression framework.

Within two years, I have 5 papers as a first author and 2 of them are accepted by top conferences, among the total of 10 co-authored papers.

My released codes and models have attracted broad interests from the community, with over **800** downloads and **65** citations from Google Scholar. I am also responsible for leading a wide range of collaborations with universities and labs such as Tsinghua University, Peking University, Northeastern University, DiDi AI Lab, and MIT-IBM Watson AI Lab.

## RESEARCH PROJECTS

**Muse Chip Development** Tsinghua University

*4/2019 - Present*

Muse architecture is the first in both academic and industrial to support multiple granularities of sparsity in neural networks as well as multiple scheme of quantization. In this project, I am responsible for leading the algorithm team for model compression. So far, our team provides model compression scheme such as irregular pruning, kernel pruning, channel pruning, group pruning, fixed point quantization, power of 2 quantization and mixed of power of 2 quantization.

**Understanding brain as encoder and decoder using FNIRS** Tsinghua University

*7/2019 - Present* *Project Lead*

Near-infrared spectroscopy is a relatively new tool of brain computer interface and in this project, we use FNIRS for understanding how brain works as a encoder and decoder. We ask 20 subjects for watching 50 short films labeled with negative and positive emotions. The project is divided into 3 stages: In the first stage (we are currently here), we built a simple classifier that decodes the brain data to the stimulus' label. We especially rely on data from

prefrontal lobe to study how individuals regularize their emotion after watching the films. In the second stage, we will built a neural network that learns how brain encodes stimulus into observed fnirs data. In the third stage, we will use the learnt neural network to create optimal stimulus that can affect human's recognition and emotion.

**Structured Pruning for 3D DNNs on Action Detection, DiDi AI Lab** 11/2018 - 12/2018
*Research Collaborator*

Convolution computations in DNNs are commonly transformed to matrix multiplications that are handled by GEMM (general matrix multiplication), which is highly optimized for parallel computation leveraging GPUs. Structured pruning, a special form of weight pruning, reduces the matrix dimensions so that GEMM computation efficiency can be greatly improved. Therefore, structured pruning is friendly to devices that use GPUs. However, prior methods have trouble maintaining the accuracy when applied on 3D Deep Neural Networks as 3D DNNs are very sensitive to pruning methods. By applying progressive ADMM (I proposed in paper[6]), the feasibility and quality of pruning are guaranteed. Without performance degradation, the compressed model achieved state-of-art action recognition speed in 3D settings.

**Hardware-Algorithm Co-Design of DNNs engine, Tsinghua University** 05/2018 - 12/2018
*Research Collaborator*

For Deep Neural Networks to be deployed in resource limited platforms such as FPGA and ASICs, a sophisticated framework that efficiently utilizes available hardware resources is needed. In collaboration with Tsinghua University, we proposed our framework ADMM-NN that, for the first time, jointly optimizes weight pruning, weight quantization, and activation quantization as a holistic DNN optimization framework. My paper [2] and paper [6] provide the fundation for ADMM-NN framework. So far, I am very much involved in the massive development and testing of ADMM-NN framework across a broad range of applications and hardware.

## WORK EXPERIENCE

**Institute for interdisciplinary,Tsinghua University** *4/2019 - Present*
Full-time Researcher
**SenseTime Inc., Shenzhen** *12/2018 - 4/2019*
Research Intern
**Geonumerical Solutions Inc., St. Louis** *06/2015 - 10/2016*
Software Developer

## SCHOLARSHIP & COMPETITION

**System Design Contest(GPU Competitions)** (Top 10/80), Design Automation Conference *2018*
**Graduate Award** (Top 30%), Syracuse University *2017-2018*
**Cognition scholarship** (Top 25%), Saint Louis University *2011-2015*

## REFERENCES

**Dr. Kaisheng Ma**
Assistant Professor in Tsinghua University
Institute for Interdisciplinary Information Sciences
Email: kaisheng@mail.tsinghua.edu.cn

**Dr. Yanzhi Wang**
Assistant Professor in Northeastern University
College of Engineering
Email: yanz.wang@northeastern.edu