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: 19)**

AREAS OF INTEREST

- Energy-Efficient and High-Performance Deep learning and Artificial Intelligence Systems
- Stability, Robustness & Interpretability of Deep Neural Networks
- Biology-Inspired Intelligent Systems

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***, Tianyun Zhang*, Kaiqi Zhang, Jiayu Li, Jiaming Xie, Yun Liang, Sijia Liu, Xue Lin, Yanzhi Wang, *A Unified Framework of DNN Weight Pruning and Weight Clustering/Quantization Using ADMM*, New England Computer Vision Workshop 2018. **(NECV2018 Oral)**

[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] Tianyun Zhang, **Shaokai Ye**, Yipeng Zhang, Yanzhi Wang, Makan Fardad, *Systematic DNN weight pruning framework using alternating direction method of multipliers*, International Conference on Learning Representations. **(ICLR2018)**

[5] 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] Tianyun Zhang*, Kaiqi Zhang*, **Shaokai Ye***, 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 using ADMM, 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 [1] 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).

Within one year, I have 4 first-author papers among 7 co-authored papers. My released codes and compressed models have attracted broad interests from the community, with over **500** downloads in 2 months. 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

Reliable Binary Quantization for Deep Neural Networks, Tsinghua University 12/2018 - Present
Research Collaborator

The computational efficiency of neural networks can be greatly improved if weights are binarized. However, prior methods are observed to occur more than 10% accuracy drop for modern DNNs like ResNet, when all layers (including the first and last) are binarized. I proposed a quantization method using progressive ADMM (paper [1]), achieving less than 6% accuracy drop and high stability on ResNet with ImageNet data set. As far as I know, this work is currently one of the most reliable methods for binary weight quantization.

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[1]), 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 - Present
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 [1] 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

SenseTime Inc., Shenzhen *12/2018 - Present*
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*