SHAOKAI YE

◇ Addr: HuangjinZhongxin, Nanping, Fujian 353000, CN ◇ Cell: (86) 1501917603 ◇ Email: shaokaiyeah@gmail.com
◇ Homepage: yeshaokai.github.io ◇ Google Scholar: googlescholar/shaokaiye

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

[1] **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.

[2] 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

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 sizes and computation requirement, and effective model compression technique is required to reduce the model size and workload. I proposed to use ADMM(Alternating Direction Method of Multipliers) to apply combinatorial constrains on models' weights. This idea spawns a series of work that leverage various redundancies in DNNs, such as weight redundancy, bit representation of weight redundancy and redundancy in intermediate results. In addition to the framework, I also propose the progressive ADMM method to largely improve the feasibility and quality of our framework, achieving the **top-1** compression rate worldwide. As far as I know, when the same method is applied on binary quantization, it also achieves the highest performance when all layers (including the first and last) are binarized. Within one year, I have 4 first-author papers among 7 co-authored papers. My released code and compressed models have attracted more than **500** downloads in the community. 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 and intermediate results 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 propose a quantization method using progressive ADMM(Alternating Direction Method of Multipliers), achieving less than 6% accuracy drop and high stability on ResNet for ImageNet data set. As far as I know, this work currently has the highest performance worldwide.

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

Structured pruning can effectively improve inference speed of DNNs. However, prior methods have trouble maintaining the accuracy when applied on 3D Deep Neural Networks as 3D DNNs are highly sensitive to pruning methods. By applying progressive ADMM, the feasibility and quality of pruning are guaranteed. Without performance degradation, the accelerated model can perform action recognition as fast as 1 second per action. The project is selected for internal presentation.

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

In hardware design of Deep Neural Networks' inference engines, there is a trade off between weights' sparsity and quantization level of weights. It is generally difficult to obtain sparsity and quantization at the same time without performance degradation. In collaboration with Tsinghua University, as far as I know, I achieve highest compression rate among existing work, with combination of weight pruning and weight quantization without performance degradation.

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