Shaokai Ye

Syracuse University
223 Link Hall, Syracuse, NY 13244

Cell: +1 (314) 584-9238
Email: shaokaiyeah@gmail.com

AREAS OF INTEREST

1. Energy-Efficient and High-Performance Deep Learning and Artificial Intelligence Systems
2. Stability, Robustness and Interpretability of Deep Neural Networks
3. Biology-Inspired Intelligent Systems

ACADEMIC PAGE

https://yeshaokai.github.io

GOOGLE SCHOLAR LINK

https://scholar.google.com/citations?user=Gky1L_gAAAAJ&hl=en

GITHUB

https://github.com/yeshaokai

EDUCATION

01/2017 – 12/2018 Syracuse University

M.S. in Computer Engineering
Advisor: Prof. Yanzhi Wang
Exchange research assistant at Northeastern University, Boston

09/2011 – 05/2015 Saint Louis University

B.S. in Computer Engineering
Advisor: Prof. Michael H. Goldwasser

RESEARCH PROJECTS

Binary Quantization for Deep Neural Networks

- Advantage of binarized neural networks: eliminating multiplications as weights are binarized values
- Limitation in prior methods: more than 10% accuracy drop for modern DNNs like ResNet, when all layers (including the first and last) are binarized
- I propose a progressive weight quantization method using ADMM (Alternating Direction Method of Multipliers), achieving < 6% accuracy drop and high stability on ResNet for ImageNet data set
- Currently the highest performance worldwide

Structured Pruning for 3D Convolutional Neural Networks on Activity Detection

- Structured pruning can effectively improve inference speed of DNNs while maintaining accuracy and regularity in DNN structure
- In collaboration with researchers from DiDi Inc., I propose and apply ADMM-based structured pruning to accelerate 3D DNNs on activity detection
- Promising speedup has been achieved, and this project is internally selected as one of the best

projects in DiDi AI Research

WORK EXPERIENCE

SenseTime Inc. Dec. 2018 - Present

Research Intern

- Domain adaption and model robustness are key research area, due to (i) limitation in training data for many applications, and (ii) vulnerability of DNNs to adversarial attacks
- I am investigating the effect of model compression on DNNs for domain adaption and robustness
- Currently leading this research project at SenseTime Inc.

Syracuse University Nov. 2017 – Dec. 2018

Research Assistant, Advisor: Prof. Yanzhi Wang

- Proposed to adopt ADMM (Alternating Direction Method of Multipliers) on model compression problems of DNNs, achieving state-of-the-art performance in model size reduction, and speedup in inference phase.
- Within one year, I have 4 first-author papers among 7 co-authored papers
- Invented progressive ADMM-based weight pruning/quantization for DNNs, which achieve Top-1 compression rate worldwide without accuracy degradation on DNN weight pruning/quantization
- The only binary quantization that can quantize the first and last layer with acceptable accuracy degradation

Geonumerical Solutions Inc., Saint Louis June 2015 – Oct. 2016

Software Developer

- Built continuous integration server using Node.js during internship period. This server makes sure that no broken build gets merged into the master branch.
- Configured and installed software for cluster environment. Configured parallel debugger and parallel image renderer for our software. Also worked on making our cluster meet clients' security requirements.
- Customized simulation software. Used Python to customize the pipeline of using our software for data collection and data analysis. That work greatly reduces workload for everyone involved.
- Wrote the front end and configured backend at Azure, as an attempt to transit our scientific software to software as a service.

COMPETITION AND AWARD

Top 10 - System Design Contest (GPU Competitions) at Design Automation Conference 2018
Graduate Award from Syracuse University 2017-2018
Cognition scholarship from Saint Louis University 2011-2015

PUBLICATION

* Equal Contribution

1. 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**)

2. 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**)
3. Tianyun Zhang, Shaokai Ye, Yipeng Zhang, Yanzhi Wang, Makan Fardad, " Systematic Weight Pruning of DNNs using Alternative Direction Method of Multipliers" International Conference on Learning Representations Workshop, 2018.(**ICLR 2018, workshop track**)
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 (**Oral**). (**GlobalSIP2018**)

PREPRINTS

1. Shaokai Ye*, TianyunZhang*, 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" arXiv1810.07378
2. Shaokai Ye*, Tianyun Zhang*, Kaiqi Zhang, Jiayu Li, Jiaming Xie, Yuan 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(**Oral**).
3. 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