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

GOOGLE SCHOLAR LINK

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

EDUCATION

01/2017 – 12/2018 Syracuse University
M.S. in Computer Engineering
Advisor: Prof. Yanzhi Wang

09/2011 – 05/2015 Saint Louis University
B.S. in Computer Engineering
Advisor: Prof. Michael H. Goldwasser

PROJECTS

Binary Deep Neural Networks

Collaborate with Tsinghua University on binarized neural network. The value of binarizing neural networks is that a binarized neural network can convert most of its computation to bitwise operations. However, an issue with existing binary quantization method is that when they are applied on first and last layer, the network will drop more than 10 points in accuracy or simply crash. However, by applying the method I propose in my work "Progressive Weight Pruning of Deep Neural Networks Using ADMM ", I can binarize all layers of neural networks with less than 6 points drop in accuracy consistently.

Structured Pruning for 3D convolutional neural networks

Structured pruning is a key method on improving inference speed for neural networks. When applied on neural networks, it essentially reduces the matrix dimension for computation, therefore, it's highly hardware friendly. I worked with researchers from DiDi Inc. to apply my work "ADAM-ADMM: A Unified Systematic Framework of Structured pruning for DNNs" on accelerating their 3D convolutional neural networks. The speed up is very promising. As a result, this project is internally selected as one of the best projects.

WORK EXPERIENCE

SenseTime Inc. Dec. 2018 - Present
Research Intern

Domain adaption and model robustness is a popular research area as most application cannot afford to have enough data and must be cautious of adversarial attacks. However, an unexplored question is

that whether model compression hurts neural networks ability for domain adaption and robustness. My leaders in SenseTime Inc. recognize my good experience in model compression and model robustness and allow me to lead this research project as an research intern.

Syracuse University Nov. 2017 – Dec. 2018

Research Assistant, Advisor: Prof. Yanzhi Wang

Proposed to use ADMM (Alternating Direction Method of Multipliers) on model compression problems, which starts a series of work that achieve state of art performance in model size reduction, inference speed up, and model efficiency (model quantization under arbitrary bits). Within one year, I first-authored 4 papers among 7 papers, including multiple oral presentations in workshops, 2 papers published in top conferences. I invented progressive ADMM, which achieves **top-1** compression rate without performance degradation among all current work when applied on weight pruning and combination of weight pruning and weight quantization. More importantly, its application on binary quantization, to my knowledge, is the only method that can quantize first and last convolution layer without severe performance 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 - 2018 System Design Contest at GPU team	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