

Yang Sui

732-310-0866 | ys764@scarletmail.rutgers.edu | [Yang Sui's Homepage](#) | [Github](#) | [LinkedIn](#) | [Google Scholar](#) |

RESEARCH INTERESTS

Efficient learning and inference of deep neural networks on computer vision tasks:

- **Model pruning and sparsification.**
- **Low-rank approximation.**
- **Model quantization.**
- **Efficient Vision Transformer.**
- **Neural Image Compression.**

EDUCATION

Rutgers University

Ph.D. candidate in the Dept. of Electrical and Computer Engineering

- GPA: 4.0/4.0

New Brunswick, NJ

01/2020 - Expected 12/2023

Jilin University

Master of Science in the Dept. of Electronic Science and Engineering

Changchun, Jilin

09/2016 - 07/2019

Jilin University

Bachelor of Engineering in the Dept. of Electronic Science and Engineering

Changchun, Jilin

09/2012 - 07/2016

PUBLICATIONS

*Authors with * signs contribute equally to the papers.*

- **In submission**

Yang Sui, Wanzhao Yang, Miao Yin, Yu Gong, Bo Yuan.

“Co-Exploring Sparsification and Low-Rank Decomposition for Compact DNNs.”

- **In submission**

Yang Sui, Ding Ding, Xiang Pan, Xiaozhong Xu, Shan Liu, Bo Yuan, Zhenzhong Chen.

“Corner-to-Center Long-range Context Model for Efficient Learned Image Compression.”

- **[AAAI 2023 (Oral)]**

Huy Phan, Miao Yin, Yang Sui, Bo Yuan, Saman Zonouz.

“CSTAR: Towards Compact and STructured Deep Neural Networks with Adversarial Robustness.”

- **[AAAI 2023 (Oral)]**

Jinqi Xiao, Chengming Zhang, Yu Gong, Miao Yin, Yang Sui, Lizhi Xiang, Dingwen Tao, Bo Yuan.

“HALOC: Hardware-Aware Automatic Low-Rank Compression for Compact Neural Networks.”

- **[AAAI 2023 Workshop (Best Paper Runner-Up Award)]**

Yang Sui, Wanzhao Yang, Miao Yin, Yu Gong, Bo Yuan.

“Towards Sparse and Low-rank Neural Networks with Hybrid Compression.”

- **[AAAI 2023 Workshop]**

Yang Sui, Miao Yin, Bo Yuan.

“Training Low-Rank CNNs with Orthogonality From Scratch.”

- **[IEEE TC]**

Yu Gong, Miao Yin, Lingyi Huang, Chunhua Deng, Yang Sui, Bo Yuan.

“Algorithm and Hardware Co-Design of Energy-Efficient LSTM Networks for Video Recognition With Hierarchical Tucker Tensor Decomposition.”

- **[CVPR 2022]**

Miao Yin, Yang Sui, Wanzhao Yang, Xiao Zang, Yu Gong, Bo Yuan.

“HODEC: Towards Efficient High-Order DEcomposed Convolutional Neural Networks.”

- **[NeurIPS 2021]**

Yang Sui, Miao Yin, Yi Xie, Huy Phan, Saman Zonouz, Bo Yuan.

“CHIP: CHannel Independence-based Pruning for Compact Neural Networks.”

- **[ICCAD 2021]**
Boyang Zhang*, Yang Sui*, Lingyi Huang, Siyu Liao, Chunhua Deng and Bo Yuan.
“Algorithm and Hardware Co-design for Deep Learning-powered Channel Decoder: A Case Study.”
- **[ISCA 2021]**
Chunhua Deng, Yang Sui, Siyu Liao, Xuehai Qian and Bo Yuan.
“GoSPA: An Energy-efficient High-performance Globally Optimized SParse Convolutional Neural Network Accelerator.”
- **[CVPR 2021]**
Miao Yin, Yang Sui, Siyu Liao, Bo Yuan.
“Towards Efficient Tensor Decomposition-Based DNN Model Compression With Optimization Framework.”

WORKING EXPERIENCE

Tencent Americas

Research Intern

Internship

05/2022 - Present

- Efficient Vision Transformer.
- Efficient Neural Image Compression.

JD

Algorithm Engineer

Full-time

07/2019 - 01/2020

- Face verification and recognition over tens of millions of data.
- Provide the face-ID verification algorithms and models for banks of Thailand.

Baidu

Research & Development Engineer

Internship

03/2018 - 12/2018

- **Initiator** of Paddle-Lite (Deep learning framework on mobile, embedded, and IoT devices) with **6.4k stars**.
- Implement and efficiently fuse the convolutional layers, normalized layers, activation functions, etc. in Paddle-Lite with C++ and assembly based on neon instructions without third-party libraries.
- Design specialized calculation process of various types of convolutional operations.

REVIEWER ACTIVITIES

PC members and Reviewer :

- **KDD**, ACM SIGKDD Conference on Knowledge Discovery and Data Mining.
- **NeurIPS**, Conference on Neural Information Processing Systems.
- **ECCV**, European Conference on Computer Vision.
- **ICML**, International Conference on Machine Learning.
- **CVPR**, International Conference on Computer Vision and Pattern Recognition.
- **AAAI**, AAAI Conference on Artificial Intelligence.
- **ISCAS**, The IEEE International Symposium on Circuits and Systems.
- CVPR Workshop on Transformers for Vision.
- **TNNLS**, IEEE Transactions on Neural Networks and Learning Systems.

TEACHING ASSISTANT

Rutgers University-New Brunswick

- 2020 Fall: 14:332:351 Programming Methodology II.