

SUMMARY

- Enthusiastic researcher, working at the intersection of Deep Learning (CV) and Parallel Computing/Architecture
- Current research on optimizing DNNs (Pruning, Quantization, Neural Architecture Search) with respect to hardware (TPU-like, Multi-core CPU, GPU) and DNN accelerator algorithm co-design
- Current Deep Learning Research Intern at Intel's Graphics Processing Research team. Former Deep Learning Intern at AMD's DNN Graph optimization team.

EDUCATION

Iowa State University

PhD in Computer Engineering. 3.55/4.0

Advisor: [Dr. Arun K. Somani](#)

Ames, Iowa, USA

Aug '17 - Present

University College of Engineering, Osmania University

Bachelor of Engineering in Electronics and Communication. 8.4/10

Hyderabad, India

Sept '13 - May '17

ACADEMIC/PROFESSIONAL EXPERIENCE

Intel Corporation

Research Scientist Intern, [Graphics Processing Research Lab](#)

Santa Clara, CA

June '20 - Present

Working on Neural Architecture Search for Neural Network design and Mixed Precision Quantization related to Image Restoration tasks and Graphics applications.

Iowa State University

Graduate Research Assistant, Dependable Computing and Networking Laboratory

Ames, Iowa

May '18 - Present

My research involves optimization of DNNs for efficient inference, co-design and reliability of DNN accelerators. My current/previous research projects are as follows:

- 1) Array Aware Pruning/Training: Designed a Pruning algorithm and a Hyper-parameter tuning method for CNN, MLP networks to minimize the number of computation cycles of a DNN forward pass on a systolic array-based Neural Network accelerator.
- 2) Model Pruning on Faulty DNN Accelerator: Examined different faulty scenarios on a systolic array based accelerator and developed a pruning method for faulty DNN hardware to bypass faults and compress weights.
- 3) CPU-GPU Aware Pruning: Developed a combined Node Pruning, Quantization and layer fusion method based on number of Multi-core CPUs and dimension of Tensor Cores (GPU) to achieve high speedup during inference.
- 4) Joint Architecture Search for co-design of Quantized Neural Network and Mixed Precision supporting array-based hardware (In progress)
- 5) A Comprehensive review of Pruning, Quantization, Neural Architecture Search from a hardware perspective (In progress for a publication in ACM Computing Survey Journal)

Advanced Micro Devices (AMD)

Machine Learning Intern, MIGraphX

Austin, Texas

May '19 - Aug '19

Worked in the MIGraphX (GPU graph optimization) team to design compression algorithms for enhancing performance on AMD GPUs at inference run-time. Developed quantization techniques to convert the weights of CNN from floating point to integer precision on CNN benchmarks like Vgg16, ResNet50, Inception, Xception.

Iowa State University

Graduate Teaching Assistant, Digital Logic Design

Ames, Iowa

Aug '17 - April '18

Responsibilities: Supervising labs, mentoring students on Verilog, FPGAs and technical projects.

Research Centre Imarat, Defence R&D Organization
Undergraduate Technical Intern
Project Title: Design and Simulation of Ethernet Controller on FPGA

Hyderabad, India
May '16 - June '16

Bharat Dynamics Limited, (A Govt. Of India Enterprise)
Undergraduate Technical Intern

Hyderabad, India
Dec '15

PUBLICATION(S) - SUBMITTED/ACCEPTED

K. T. Chitty-Venkata, S. Kothandaraman and A. Somani, "Tensor Core-Aware Tuning and Pruning Methods" in IEEE IPDPS 2021 Conference (Submitted - Under Review)

K. T. Chitty-Venkata and A. Somani, "Calibration Data-Based CNN Filter Pruning for Efficient Layer Fusion" in IEEE HPCC-DSS 2020 Conference (Accepted)

K. T. Chitty-Venkata and A. Somani, "Model Compression on Faulty Array-based Neural Network Accelerator" in IEEE PRDC 2020 Conference (Accepted)

K. T. Chitty-Venkata and A. Somani, "Array Aware Training/Pruning: Methods for Efficient Forward Propagation on Array-based Neural Network Accelerators" in IEEE ASAP 2020 Conference (Accepted)

K. T. Chitty-Venkata and A. Somani, "Impact of Structural Faults on Neural Network Performance" in IEEE ASAP Conference 2019 (Accepted)

COURSE WORK (GRAD SCHOOL)

Hardware: Computer System Architecture, Applications of Parallel Computing (CS267 UC Berkeley), Design and Analysis of Algorithms, Fault Tolerant Computing, Real Time Systems, Communication Systems

Machine Learning: Probabilistic Methods, Statistics Theory for Research, Deep Learning, Machine Learning, Statistical Methods for Machine Learning

SKILLS

Programming: C, C++, Python, Matlab

Parallel Programming: CUDA, OpenMP, working knowledge of MPI

Machine Learning Frameworks: Tensorflow, Pytorch, Keras, Scikit Learn

Deep Learning Networks/Data sets: MLP, CNN, RNN, MNIST, Cifar10, Imagenet

Other: Linux, Shell Scripting, Verilog HDL, FPGA, Gem5 and ZSim Simulators, HTML

RELEVANT ACADEMIC PROJECTS

Reinforcement Learning using Neural Networks: Designed and implemented Q-learning algorithm using DNNs as function approximator for acrobat-v1, an environment taken from OpenAI gym.

High Performance Cache Simulation on GPU: Developed a CPU-GPU based cache simulator and compared it with traditional CPU-only simulation. Developed the simulator using C (for CPU-only) and CUDA C (for CPU-GPU). CPU-GPU cache simulation performed better than CPU-only simulation.

Checkpoint-based Fault Mitigation Techniques for GPUs: Proposed a new fault tolerant algorithm to aid Check-pointing which attempts to reduce the communication overhead between CPU and GPU.

TEAM WORK EXPERIENCE

- Part of Dependable Computing and Networking Laboratory ([DCNL](#)) group consisting of six researchers
- Executive Committee member at Indian Students' Association (ISU) for 2018-19 academic year

RECOMMENDATION(S)

- Dr. Arun K. Somani (Doctoral Advisor): arun@iastate.edu
- Mike Vermeulen (Former Manager at AMD) - [LinkedIn Recommendations Section](#)