Krishna Teja Chitty-Venkata
2040 Long Rd, 165H, Ames, IA, 50010 ⋄ +1-515-203-5766
krishnat@iastate.edu ⋄ Linkedin ⋄ Website

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 special purpose (TPU-like) and general purpose hardware (Multi-CPU, GPU)
- Current Intern in Intel's Graphics Processing Research team. Former Intern at AMD's MIGraphX team.

EDUCATION

Iowa State University *Ames, Iowa, USA*
PhD in Computer Engineering. 3.55/4.0 *Aug '17 - Present*
Advisor: Dr. Arun K. Somani

University College of Engineering, Osmania University *Hyderabad, India*
Bachelor of Engineering in Electronics and Communication. 8.4/10 *Sept '13 - May '17*

ACADEMIC/PROFESSIONAL EXPERIENCE

Intel Corporation *Santa Clara, CA*
Research Scientist Intern, Graphics Processing Research Lab *June '20 - Present*

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

Iowa State University *Ames, Iowa*
Graduate Research Assistant, Dependable Computing and Networking Laboratory *May '18 - Present*

My research involves optimization (compression) of DNNs for efficient inference on hardware and applying to different problems (eg: fault tolerance). 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 DNN forward pass on a systolic array based Neural Network accelerator.

2) Model Pruning on Faulty DNN Accelerator: Proposed and examined different faulty scenarios on a systolic array based hardware and developed a pruning method on faulty hardware.

3) CPU-GPU Aware Pruning: Developing a combined symmetric Pruning, Quantization and layer fusion algorithm to compress Neural Networks and accelerate on Tensor Cores (GPU) and multi-core CPUs 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 publication in ACM Computing Survey)

Advanced Micro Devices (AMD) *Austin, Texas*
Machine Learning Intern, MIGraphX *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 *Ames, Iowa*
Graduate Teaching Assistant, Digital Logic Design *Aug '17 - April '18*
Responsibilities: Supervising labs, mentoring students on Verilog, FPGAs and technical projects.

Research Centre Imarat, Defence R&D Organization *Hyderabad, India*
Undergraduate Technical Intern *May '16 - June '16*
Project Title: Design and Simulation of Ethernet Controller on FPGA

Bharat Dynamics Limited, (A Govt. Of India Enterprise) *Hyderabad, India*
Undergraduate Technical Intern *Dec '15*

PUBLICATION(S)

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 (Submitted - Under Review)

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
- LinkedIn Recommendations Section