Krishna Teja Chitty-Venkata

2040 Long Rd, 165H, Ames, IA, USA, 50010 ⋄ +1-515-203-5766

krishnat@iastate.edu ⋄ LinkedIn ⋄ Website ⋄ Google Scholar ⋄ DBLP

SUMMARY

- Enthusiastic PhD Student working on optimizing DNNs (Pruning, Quantization, Neural Architecture Search (NAS)) with respect to hardware (TPU-like, CPU, GPU) and DNN accelerator algorithm co-design
- Former Research Aide at Argonne National Laboratory. Former Research Intern at Intel's Graphics Processing Research group. Former Deep Learning Intern at AMD's DNN Graph optimization team

EDUCATION

Iowa State University (ISU) *Ames, Iowa, USA*
PhD in Computer Engineering. 3.55/4.0 *Aug 2017 - 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 2013 - May 2017*

ACADEMIC/PROFESSIONAL EXPERIENCE

Iowa State University (ISU) *Ames, IA, USA*
Graduate Research Assistant, Dependable Computing and Networking Laboratory *May 2018 - Present*

My research involves **(i)** Efficient Neural Network Architecture and Mixed Precision search through NAS on various devices (ASICs, GPUs), **(ii)** Co-search of CNN algorithm and hardware accelerator specifications, **(iii)** Compression (pruning and quantization) of DNNs for efficient inference on different hardware systems (CPU, GPU, ASICs), and **(iv)** Reliability of DNN accelerators. My current/previous research projects are as follows:

1. **Hardware Aware Mixed Precision Quantization Search:** The goal of the project is to develop a search algorithm to find optimal precision (bit-width) of each layer of a pretrained CNN based on the performance requirements of the underlying hardware platform (MCU, CPU, GPU, Accelerator, FPGA) (Ongoing)
2. **CNN Algorithm-Accelerator Co-search:** The aim is to develop a differentiable co-search algorithm to find a CNN-accelerator pair, where the CNN network size matches the the array size (Ongoing)
3. **Array Aware Neural Architecture Search:** Developed a search algorithm for searching efficient Convolutional Neural Network architectures for Systolic Array-based DNN accelerators (TPU, Eyeriss) by co-designing the search space with respect to the underlying size of the array. Simulator: SCALE-Sim
4. **Hardware Dimension Aware DNN Pruning:** Designed a Pruning algorithm to minimize DNN processing time on Array-based Accelerators (TPU and Eyeriss), Multi-core CPUs (Intel Skylake and i7), and Tensor Core GPUs (Volta and Turing architectures) based on the underlying hardware size (Array size, number of CPU cores, Tensor core dimension). Simulators/Programming: SCALE-Sim, OpenMP, CUDA
5. **Model Compression on Faulty DNN Accelerator:** Developed a joint pruning method on an array-based accelerator to bypass faults and compress weights for efficient inference under different faulty modes
6. **Survey Papers: i)** Reviewing latest literature and writing survey papers on Hardware-aware Pruning, Quantization, and NAS. **ii)** Surveying Systolic and Tensor Array Architectures for Deep Neural Networks

Argonne National Laboratory *Lemont, IL, USA*
Research Aide, Data Science Research Group in Leadership Computing Facility *Sept 2021 - Nov 2021*
Worked on a project titled "Searching for Sparse and Mixed Precision Quantized Networks using Neural Architecture Search for Ampere 100 Tensor Cores". In process for a paper submission in ACM HPDC 2022 Conference.

Intel Corporation *Santa Clara, CA, USA*
Research Scientist Intern, Graphics Processing Research Lab *June 2020 - Dec 2020*
Worked on Neural Architecture Search for Network design and Mixed Precision Quantization related to Image Restoration tasks and Graphics applications. The work resulted in a publication at 2021 IEEE ICIP Conference

Advanced Micro Devices (AMD) *Austin, TX, USA*
Deep Learning Intern, MIGraphX *May 2019 - Aug 2019*
Worked in the GPU graph optimization team to design compression algorithms for enhancing performance on AMD GPUs at inference run-time. Developed Post Training Quantization (PTQ) methods to convert the CNN weights from floating-point 32 format to integer precision. Benchmarks: Vgg16, ResNet50, InceptionV3, Xception

Iowa State University *Ames, IA, USA*
Graduate Teaching Assistant, Digital Logic Design *Aug 2017 - April 2018*
Responsibilities: Supervising labs, mentoring students on Verilog, FPGAs and technical projects

Research Centre Imarat, Defence R&D Organization *Hyderabad, India*
Undergraduate Technical Intern *May 2016 - June 2016*
Project Title: Design and Simulation of Ethernet Controller on FPGA

PUBLICATION(S) - SUBMITTED/ACCEPTED

1. **K. T. Chitty-Venkata** and A. Somani, "Neural Architecture Search Survey: A Hardware Perspective" in ACM Computing Surveys (Submitted; Under review)
2. **K. T. Chitty-Venkata** and A. Somani, "Hardware Dimension Aware Pruning" in ACM Transactions on Parallel Computing (Submitted; Under review)
3. **K. T. Chitty-Venkata** and A. Somani, "Array Aware Neural Architecture Search" in IEEE ASAP 2021 Conference [Paper]
4. **K. T. Chitty-Venkata**, A. Somani and S. Kothandaraman, "Searching Architecture and Precision for U-net based Image Restoration Tasks" in IEEE ICIP 2021 Conference [Paper]
5. **K. T. Chitty-Venkata** and A. Somani, "Calibration Data-Based CNN Filter Pruning for Efficient Layer Fusion" in IEEE HPCC-DSS 2020 Conference [Paper]
6. **K. T. Chitty-Venkata** and A. Somani, "Model Compression on Faulty Array-based Neural Network Accelerator" in IEEE PRDC 2020 Conference [Paper]
7. **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 [Paper]
8. **K. T. Chitty-Venkata** and A. Somani, "Impact of Structural Faults on Neural Network Performance" in IEEE ASAP Conference 2019 [Paper]

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
- **Deep Learning Frameworks/Datasets:** Pytorch, Tensorflow, Keras / CIFAR-10, ImageNet

REFERENCES/RECOMMENDATIONS

- Dr. Arun K. Somani (Doctoral Advisor): arun@iastate.edu
- LinkedIn Recommendations Section
  1. Sreeni Kothandaraman (Former Manager at Intel Corporation)
  2. Mike Vermeulen (Former Manager at Advanced Micro Devices (AMD))