Krishna Teja Chitty-Venkata

2040 Long Rd, 165H, Ames, IA, USA, 50010 ⋄ +1-515-203-5766
krishnat@iastate.edu ⋄ LinkedIn ⋄ Website ⋄ Google Scholar

SUMMARY

- Enthusiastic PhD Student working on optimizing DNNs using Pruning, Quantization, Neural Architecture Search (NAS) with respect to hardware (ASIC, 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 for ASIC and GPU devices **(ii)** Co-search of a network and hardware accelerator, **(iii)** Pruning and Quantization for efficient inference on CPU, GPU, ASIC hardware platforms, and **(iv)** Reliability of DNN accelerators.

My ongoing research projects are as follows:

1. **Accelerator Aware Mixed Precision Quantization Search:** The goal of the project is to develop an efficient search algorithm to find the optimal precision (bit-width) of each layer of a CNN model based on the performance requirements of the underlying accelerator supporting mixed-precision multiplication
2. **NAS Benchmark Design:** The project aims to develop efficient Neural Architecture Search benchmarks on a large-scale dataset for Convolutional Neural Networks and Transformers. The NAS Benchmarks are further used for the co-search algorithm to find a network-accelerator pair automatically

My previous research projects are as follows:

1. **Neural Architecture Search Survey Papers:**
   (a) **Hardware-aware Neural Architecture Search:** Reviewed State-of-the-art literature on hardware-aware NAS methods specific to MCU, CPU (mobile and desktop), GPU (Edge and server-level), ASIC, FPGA, ReRAM, DSP, and VPU, followed by the co-search methodologies of Neural algorithm and accelerator. We classified the HW-NAS methods based on Search Space (Cell, Layer-wise), Search Algorithm (RL, Differentiable, Evolutionary). The paper is published in ACM Computing Surveys.
   (b) **NAS Benchmarks:** Surveyed the latest NAS Benchmarks Dictionaries, which simulate the architecture evaluation within seconds. The paper is under review at IEEE Transactions on Neural Networks
   (c) **Transformer NAS:** Surveyed latest NAS algorithms targeting Transformers, BERT models, and Vision Transformer for language, speech, and vision applications. The paper is under review at IEEE Access
2. **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

3. **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
4. **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

Argonne National Laboratory *Lemont, IL, USA*
Research Aide, Data Science Research Group in Leadership Computing Facility *Sept 2021 - Nov 2021*
Worked at the intersection of Pruning, Quantization, and NAS for the project titled "Searching Sparse and Mixed Precision Quantized Networks for A100 Tensor Cores". The work was published at ACM 2022 HPDC

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 for Image Restoration tasks and Graphics applications. The work resulted in a publication at IEEE 2021 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 lower the CNN weights from floating-point 32 format to integer precision. Benchmarks: Vgg16, ResNet50, InceptionV3, Xception

PUBLICATION(S) - SUBMITTED/ACCEPTED

1. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, & A. Somani, "Neural Architecture Search for Transformers: A Survey" (Under review at IEEE Access)
2. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, & A. Somani, "Neural Architecture Search Benchmark Dictionaries: A Survey of Past and Future Work" (Under review at IEEE TNNLS)
3. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, and A. Somani, "Efficient Design Space Exploration for Sparse Mixed Precision Neural Architectures" in ACM HPDC 2022 Conference [Paper]
4. **K. T. Chitty-Venkata** and A. Somani, "Neural Architecture Search Survey: A Hardware Perspective" in ACM Computing Surveys (2021 Impact Factor: 14.32) [Paper]
5. **K. T. Chitty-Venkata** *et al.*, "Array-Aware Neural Architecture Search" in IEEE ASAP 2021 [Paper]
6. **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]
7. **K. T. Chitty-Venkata** and A. Somani, "Calibration Data-Based CNN Filter Pruning for Efficient Layer Fusion" in IEEE HPCC-DSS 2020 Conference [Paper]
8. **K. T. Chitty-Venkata** and A. Somani, "Model Compression on Faulty Array-based Neural Network Accelerator" in IEEE PRDC 2020 Conference [Paper]
9. **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]
10. **K. T. Chitty-Venkata** and A. Somani, "Impact of Structural Faults on Neural Network Performance" in IEEE ASAP Conference 2019 [Paper]

PUBLICATIONS UNDER PROGRESS

1. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, & A. Somani, "Fast Accelerator Aware Differentiable Mixed Precision Quantization Search" (In progress for submission to IEEE TCAD)
2. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, & A. Somani, "Efficient Neural Search Benchmark Design on Large Scale Dataset for CNN and ViT" (In progress for submission to ACM TECS)

HONOURS/AWARDS

- Accepted to Oxford Machine Learning Summer school 2022 (OxML) in ML for Health and Finance tracks
- HPDC 2022 Student Travel Grant Award

SKILLS

- **Programming:** C, C++, Python (Pytorch, Tensorflow), Matlab, CUDA, OpenMP, MPI
- **Deep Neural Networks:** CNNs, Transformers, Vision Transformers

REFERENCES/RECOMMENDATIONS

- Prof. Arun K. Somani (Doctoral Advisor): arun@iastate.edu
- LinkedIn Recommendations Section
  1. Dr. Murali Emani (Manager while working at Argonne National Laboratory)
  2. Sreeni Kothandaraman (Manager while working at Intel)
  3. Mike Vermeulen (Manager while working at AMD)