Krishna Teja Chitty-Venkata

Lemont, IL, USA, 50010 ⋄ +1-515-203-5766

schittyvenkata@anl.gov ⋄ LinkedIn ⋄ Website ⋄ Google Scholar

SUMMARY

- Postdoctoral Researcher at Argonne National Laboratory; currently working on Large Language Models
- PhD research on optimizing neural networks using pruning, quantization, AutoML, Neural Architecture Search (NAS) methods
- Experience working on traditional Machine Learning methods and State-of-the-art Deep Learning models, including CNNs, Transformers, Vision Transformers and Large Language Models
- Machine/Deep Learning Research Internship experience at Argonne National Laboratory, Intel, and AMD
- Iowa State University Graduate College Research Excellence award for work on efficient Deep Learning

EDUCATION

Iowa State University (ISU) *Ames, Iowa, USA*
PhD in Computer Engineering. 3.55/4.0 *Aug 2017 - July 2023*
Advisor: Dr. Arun K. Somani
Dissertation Title: Hardware-aware Design, Search and Optimization of Deep Neural Networks

University College of Engineering, Osmania University *Hyderabad, India*
Bachelor of Engineering in Electronics and Communication. 8.4/10 *Sept 2013 - May 2017*

ACADEMIC/PROFESSIONAL RESEARCH EXPERIENCE

Argonne National Laboratory *Lemont, IL, USA*
Postdoctoral Researcher, Argonne Leadership Computing Facility (ALCF) *August 2023 - Present*

Working as a Postdoc in the Data Science group of ALCF. Research on enhancing training and inference efficiency of Large Language Models (LLMs), profiling and benchmarking Neural Networks on AI Accelerators and AI for Science applications

Iowa State University (ISU) *Ames, IA, USA*
Graduate Research Assistant, Dependable Computing and Networking Laboratory *May 2018 - July 2023*

My PhD research involved designing **(i)** Neural Architecture Search algorithms to search for efficient Neural Networks for different tasks and datasets to optimize performance, latency and accuracy, **(ii)** Pruning and Quantization algorithms for Neural Network model size reduction for efficient inference. My research projects are as follows:

1. **Benchmark Design:** The project aims to build efficient Neural Architecture Search benchmarks for large-scale datasets targeting Convolutional Neural Networks, Transformers, and Vision Transformers. The Benchmarks aid Neural Architecture Search algorithms to search for neural architectures efficiently
2. **Neural Architecture Search Survey Papers:**
   (a) **Neural Architecture Search Survey:** 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, co-search methodologies of Neural algorithm and accelerator. We classified the HW-NAS methods based on Search Space (Cell, Layer-wise) and Search Algorithm (Reinforcement Learning, Differentiable, Evolutionary). The paper is published in ACM CSUR (Impact Factor: 14.32).
   (b) **Transformer NAS:** Surveyed the latest Neural Architecture Search algorithms for Transformers, BERT models, and Vision Transformer for language, speech, and vision applications. The paper is published in IEEE Access

(c) **NAS Benchmarks:** Reviewed the latest Neural Architecture Search Benchmarks, which simulate the architecture evaluation within seconds. The paper is under revision in IEEE Transactions

3. **Accelerator, Architecture and Mixed Precision Quantization Co-Search:** The goal of the project is to develop an efficient co-search algorithm to find the optimal accelerator dimensions, architecture specifications, and precision of each layer of searched network for better model performance and efficiency

4. **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

5. **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). Programming Tools: OpenMP, CUDA

6. **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 (Remote)*
Research Aide, Data Science Research Group in Leadership Computing Facility *Sept 2021 - Nov 2021*
Worked on the project "Searching Sparse and Mixed Precision Quantized Neural Networks for A100 Tensor Cores" to find efficient neural network architectures. The work was published at ACM 2022 HPDC Conference

Intel Corporation *Santa Clara, CA, USA (Remote)*
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)

1. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, & A. Somani, "Neural Architecture Search Benchmarks: Insights and Survey" in IEEE Access Journal [Paper] (Impact Factor = 3.367, Acceptance Rate = 30%)

2. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, & A. Somani, "Neural Architecture Search for Transformers: A Survey" in IEEE Access Journal [Paper] (Impact Factor = 3.367, Acceptance Rate = 30%)

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] (Acceptance Rate = 19%)

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** and A. Somani, "Array-Aware Neural Architecture Search" in IEEE ASAP 2021 Conference [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/PREPRINTS

1. **K. T. Chitty-Venkata**, Y.Bian, M. Emani, V. Vishwanath, & A. Somani, "Differentiable Neural Architecture, Mixed Precision and Accelerator Co-search" (Under Review and revision) [Manuscript]
2. **K. T. Chitty-Venkata**, S. Mittal, M. Emani, V. Vishwanath, & A. Somani, "A Survey of Techniques for Optimizing Transformer Inference" (Under review and revision) [Manuscript]

COURSE WORK (GRAD SCHOOL)

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

HONOURS/AWARDS

- Research Excellence Award by Iowa State University Graduate School, Fall 2022 [Certificate] [Certificate]
- Research Award by Graduate and Professional Student Senate (GPSS) society at Iowa State University, Spring 2023 [Certificate]
- Attended Oxford Machine Learning Summer school 2022 virtually (OxML) in ML for Health and ML for Finance tracks [Certificate] (Acceptance Rate $< 10\%$)
- Our survey paper "Neural Architecture Search Survey: A Hardware Perspective," has been identified as one of the must-read AI papers in 2022 by a group of industry experts [URL].
- HPDC 2022 Student Travel Grant Award

SKILLS

- **Programming:** C, C++, Python, Matlab, CUDA, TensorRT, OpenMP, MPI
- **Deep Learning Frameworks:** Pytorch, Tensorflow, Keras, Scikit learn
- **Deep Neural Networks:** CNNs, Transformers, Vision Transformers (ViT), BERT, GPT
- **Deep Neural Applications:** Image Classification, Object Detection, Semantic Segmentation
- **Datasets:** CIFAR-10, ImageNet, Pascal VOC, DIV2K

PEER REVIEW ACTIVITY

- **Journals:** PeerJ Computer Science (3x), Elsevier Neural Networks (3x), IEEE TNNLS, ACM CSUR, Machine Learning with Applications, Intelligent Automation Soft Computing, IEEE TCAD, MDPI Applied Sciences (2x)
- **Conferences:** DCAA'23 (AAAI), AutoML 2023 (3x),

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)