Krishna Teja Chitty-Venkata

Greater Boston Area ◇ +1-515-203-5766
LinkedIn ◇ Portfolio ◇ Google Scholar ◇ GitHub

SUMMARY

- Postdoctoral Researcher at Argonne National Laboratory, working on optimizing finetuning and inference of Large Language Models on GPUs and AI Accelerators
- PhD from Iowa State University, focused on enhancing the performance of neural networks on CPUs, GPUs and AI accelerators using pruning, quantization, AutoML, neural architecture search (NAS) methods

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 [Dissertation]

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

ACADEMIC/PROFESSIONAL EXPERIENCE

Argonne National Laboratory *Lemont, IL, USA*
Postdoctoral Researcher, AI/ML Group, Argonne Leadership Computing Facility *August 2023 - Present*
Working in the AI/ML group of the ALCF division on (i) enhancing fine-tuning (PEFT and Model Alignment) and inference efficiency of LLMs and VLMs, (ii) profiling and benchmarking models on GPUs and AI Accelerators, and (iii) AI for science applications under the supervision of Murali Emani and Venkatram Vishwanath. My current and past research projects are as follows:

1. **Blocked KV Cache Eviction (ongoing):** Developing a Blocked KV Cache eviction method to prune KV tokens that do not contribute to generating new output tokens during LLM/VLM inference process.
2. **LLM Inference Bench:** Lead a wide scale benchmarking of llama-style models (LLaMA-2, LLaMA-3, Mistral and Qwen) on different hardware platforms (Nvidia A100, H100, AMD MI250, Intel PVC, SambaNova SN40L and Habana Gaudi2) using different inference frameworks (vLLM, TensorRT-LLM, llama.cpp, Deepspeed-MII). The goal is to understand the impact of input length, output length and batch size on the hardware performance using several inference optimization techniques. This infrastructure will be used for AuroraGPT project.
3. **LLM Pruning:** Developed WActiGrad, a structured pruning method to remove heads, MLP dimension and hidden size in LLMs. We applied the methodology to LLaMA and Mistral models. We integrated the pruned LLMs into different inference frameworks and achieved enhanced performance on Nvidia A100 GPU, Cerebras CS-2, Groq LPU and Graphcore Bow Pod64.

Argonne National Laboratory *Lemont, IL, USA (Remote)*
Research Intern, Data Science Research Group, Leadership Computing Facility *Sept 2021 - Nov 2021*

- Worked on the project titled "Searching Sparse and Mixed Precision Quantized Neural Networks for A100 Tensor Cores." The goal is to search for an optimal neural network configuration (CNN kernel and channel sizes) and compression strategy (sparsity pattern and parameter precision) for every layer in the network for efficient inference on Nvidia A100 GPU using Tensor Cores.

Intel Corporation *Santa Clara, CA, USA (Remote)*
Research Scientist Intern, Graphics Processing Research Lab *June 2020 - Dec 2020*

- Worked in the Graphics research team on developing Neural Architecture Search methods to find an optimal Convolution network for Image Restoration tasks and Graphics applications. We applied the methodology on Unet to find the best upsampling and downsampling operation and precision for each layer in the model.

Advanced Micro Devices (AMD) *Austin, TX, USA*
Deep Learning Intern, MIGraphX *May 2019 - Aug 2019*

- Worked in the GPU optimization team to design compression algorithms for enhancing inference runtime on AMD GPUs. Developed Post Training Quantization (PTQ) methods to lower the CNN weights from floating-point 32 format to integer 8 precision. Implemented the quantization algorithms and analysis on Vgg16, ResNet50, InceptionV3, Xception benchmark, resulting in negligible accuracy loss on Imagenet.

ACADEMIC EXPERIENCE

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

My PhD research mainly 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 during my PhD are as follows:

1. **ConVision Benchmark:** Developed a benchmarking suite to train, validate, infer and measure the performance of 200+ Convolutional Neural Networks and Vision Transformers on multiple test datasets. We analyzed the impact of each network on different datasets, validation metrics and their hardware performance on A100 GPU, followed by publishing our wide-scale analysis in the MDPI AI Journal.
2. **Accelerator, Architecture and Mixed Precision Quantization Co-Search:** Developed an efficient joint co-search algorithm to find the hardware accelerator dimensions, neural network architecture and precision of each layer for optimal model performance and efficiency. We applied our method on MobileNetV2 CNN and BitFusion accelerator search space and obtained better accuracy-latency tradeoff models
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
4. **Hardware Dimension Aware 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 and accelerator simulator
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:**
   (a) **Hardware-aware Neural Architecture Search:** Published review papers 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).
   (b) **Transformer NAS and efficient Inference:** Published review papers on transformer (vanilla transformer, BERT, GPT and ViT) optimization techniques such as knowledge distillation, pruning, quantization, NAS and lightweight network design at the algorithmic level and novel hardware accelerators at the hardware level.

PUBLICATION(S)

1. **K. T. Chitty-Venkata***, S. Raskar*, B. Kale, F. Ferdaus, A. Tanikanti, K. Raffenetti, V. Taylor, M. Emani. V. Vishwanath, "LLM-Inference-Bench: Inference Benchmarking of Large Language Models on AI Accelerators", Accepted for PMBS SC 2024 Conference
2. **K. T. Chitty-Venkata**, V.K. Sastry, M. Emani, V. Vishwanath, S. Shanmugavelu and S. Howland, "WActiGrad: Structured Pruning for Efficient Finetuning and Inference of Large Language Models on AI Accelerators", in EuroPar2024 Conference [Paper]

3. S. B. Vijayakumar*, **K.T. Chitty-Venkata***, K. Arya, A. Somani, "ConVision Benchmark: A Contemporary Framework to Benchmark CNN and ViT Models", MDPI AI 2024 [Paper] (Impact Factor = 3.3)
4. **K. T. Chitty-Venkata**, Y.Bian, M. Emani, V. Vishwanath, & A. Somani, "Differentiable Neural Architecture, Mixed Precision and Accelerator Co-search" in IEEE Access [Paper] (Impact Factor = 3.9)
5. **K. T. Chitty-Venkata**, S. Mittal, M. Emani, V. Vishwanath, & A. Somani, "A Survey of Techniques for Optimizing Transformer Inference" in Journal of System Architecture [Paper] [arXiv] (Impact Factor = 4.5)
6. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, & A. Somani, "Neural Architecture Search Benchmarks: Insights and Survey" in IEEE Access Journal [Paper] (Impact Factor = 3.9, Acceptance Rate = 30%)
7. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, & A. Somani, "Neural Architecture Search for Transformers: A Survey" in IEEE Access Journal [Paper] (Impact Factor = 3.9, Acceptance Rate = 30%)
8. **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 [Paper] (Acceptance Rate = 19%)
9. **K. T. Chitty-Venkata** and A. Somani, "Neural Architecture Search Survey: A Hardware Perspective" in ACM Computing Surveys (2022 Impact Factor: 16.6) [Paper]
10. **K. T. Chitty-Venkata** and A. Somani, "Array-Aware Neural Architecture Search" in IEEE ASAP 2021 Conference [Paper]
11. **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] (Acceptance Rate = 45%)
12. **K. T. Chitty-Venkata** and A. Somani, "Calibration Data-Based CNN Filter Pruning for Efficient Layer Fusion" in IEEE HPCC-DSS 2020 Conference [Paper] (Acceptance Rate = 20%)
13. **K. T. Chitty-Venkata** and A. Somani, "Model Compression on Faulty Array-based Neural Network Accelerator" in IEEE PRDC 2020 Conference [Paper] (Acceptance Rate = 40%)
14. **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]
15. **K. T. Chitty-Venkata** and A. Somani, "Impact of Structural Faults on Neural Network Performance" in IEEE ASAP Conference 2019 [Paper]

*Equal Contribution

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]
- Selected for Oxford Machine Learning Summer school 2022 (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].

SKILLS

- **Programming:** C, C++, Python, Matlab, CUDA, TensorRT, OpenMP, MPI
- **Deep Learning Frameworks:** Pytorch, Tensorflow, Deepspeed, Huggingface Transformers Library