Krishna Teja Chitty-Venkata

Greater Boston Area ◇ +1-515-203-5766 ◇ krishnateja95@outlook.com
LinkedIn ◇ Website ◇ Google Scholar ◇ GitHub ◇ HuggingFace

SUMMARY

- **Current Position:** Postdoctoral Researcher at Argonne National Laboratory (US Department of Energy)
- **Research Interests:** Optimizing training, inference, and fine-tuning of large language and multimodal models, including self-attention, mixture-of-experts (MoEs), and state space architectures on GPUs; pruning and mixed-precision quantization; KV cache compression; efficient model architecture design; performance bottleneck analysis; GPU and AI accelerator benchmarking; and scaling AI applications on supercomputers.

EDUCATION

Iowa State University (ISU) *Ames, Iowa, USA*
PhD in Computer Engineering, GPA: 3.55/4.0 *August 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, GPA: 8.4/10 *September 2013 - July 2017*

PROFESSIONAL EXPERIENCE

Argonne National Laboratory *Lemont, IL, USA*
Postdoctoral Researcher, Argonne Leadership Computing Facility (ALCF) *August 2023 - Present*
Managers: Murali Emani and Venkatram Vishwanath
Working in the AI/ML team of ALCF division on profiling and enhancing LLM/VLM model training, fine-tuning and inference on Supercomputers (Polaris and Aurora) and AI Accelerators. My recent research includes:

1. **PagedEviction:** Developed a structured, block-wise and attention-agnostic KV cache eviction algorithm integrated with vLLM's PagedAttention mechanism. The approach requires minimal changes to vLLM's core implementation and achieves enhanced inference efficiency for LLaMA 1B, 3B and 8B models in vLLM, while achieving similar performance on long-context LongBench datasets compared to existing methods.
2. **LExI:** Designed a novel, data-free post-training method to choose optimal number of active experts per layer in MoE models. Improved inference performance and accuracy over MoE pruning techniques on DeepSeekV2, OLMoE-1B-7B, Qwen-1.5-MoE, Mixtral, achieving up to 10% higher accuracy and 2x better vLLM throughput.
3. **MoPEQ:** Developed a novel mixed precision quantization algorithm for MoE-based Large Language and Vision Models, enabling per-expert adaptive bit-width assignment (2,3,4 bits) using Hessian-based sensitivity of each expert. Our method achieves memory reduction while maintaining accuracy on DeepSeek-VL2 tiny, small, base, and MolmoE models, outperforming baselines on nine VLMEvalKit datasets.
4. **LangVision-LoRA-NAS:** Designed a framework to integrate NAS with LoRA to optimize the finetuning efficiency of LLMs/VLMs. This approach leverages NAS to dynamically identify optimal LoRA ranks across different layers. The approach on LLaMA-3.2-11B-Vision model across diverse vision-text datasets demonstrated significant improvement in LoRA trainable parameters while preserving the baseline perplexity.
5. **LLM-Inference-Bench:** Developed a comprehensive benchmarking and profiling suite for Large Language Model Inference (LLaMA-2/3, Mistral, Qwen) on diverse accelerators (NVIDIA A100/H100/GH200, AMD MI250/MI300X, SambaNova SN40L, Habana Gaudi2) using vLLM, TensorRT-LLM, llama.cpp, and DeepSpeed-MII frameworks. Our suite provides crucial insights into the scalability and efficiency of LLMs by examining the interplay between input length, output length, and batch size, while employing various inference optimization techniques. This robust infrastructure is used for Argonne's AuroraGPT project.
6. **WActiGrad LLM Pruning:** Developed a structured LLM pruning method to prune attention heads, MLP dimension and hidden size across different layers of LLaMA and Mistral models. 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 accelerators with similar perplexity on WikiText, C4 and PTB datasets.

Argonne National Laboratory *Lemont, IL, USA (Remote)*
AI/ML Research Intern, Argonne Leadership Computing Facility *September 2021 - November 2021*

- Worked on a project titled "Searching Sparse and Mixed Precision Quantized Neural Networks for Nvidia 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 pretrained network for efficient inference on Nvidia A100 GPU using Tensor Cores.

Intel Corporation *Santa Clara, CA, USA (Remote)*
Deep Learning Research Intern, Graphics Processing Research Lab *June 2020 - December 2020*

- Worked with the Graphics Research team to develop NAS methods to optimize convolutional networks for image restoration, denoising, and graphics applications. Applied NAS to a UNet-based architecture to determine optimal layer-wise upsampling/downsampling operations and bit width of the baseline 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 RESEARCH EXPERIENCE

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

My Ph.D. research mainly involved designing neural architecture search algorithms to search for efficient networks to optimize hardware performance and accuracy, and model compression (Pruning and Quantization) algorithms for model size reduction for efficient inference. My research projects during my PhD are as follows:

1. **ConVision Benchmark:** Developed a PyTorch-based framework standardizing CNN and Vision Transformer evaluation to address reproducibility and validation inconsistencies. Demonstrated effectiveness via training 200 models and providing rigorous analysis of metrics, including accuracy and computational efficiency.

2. **Accelerator, Architecture and Precision Co-Search:** Developed a co-joint algorithm to search hardware accelerator dimensions, neural network architecture, and per-layer precision. Optimized model performance and efficiency on MobileNetV2 CNN within the BitFusion accelerator design space

3. **Array Aware Neural Architecture Search:** Developed a search method to automatically co-design CNN architectures optimized for systolic-array-based accelerators. Unlike conventional hardware-aware NAS methods, our approach dynamically adapts search spaces based on underlying array dimensions. Demonstrated similar accuracy to baseline CNNs on CIFAR-10 using MobilenetV2 search space.

4. **Hardware Dimension Aware Pruning:** Designed a Pruning algorithm to optimize DNN/CNN 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

PREPRINTS

1. **K. T. Chitty-Venkata**, S. Madireddy, M. Emani, V. Vishwanath, "LExI: Layer-Adaptive Active Experts for Efficient MoE Model Inference", [arXiv]

2. **K. T. Chitty-Venkata**, J. Ye, B. Nicolae *et al.*, "Paged Compression: Structured and Blocked KV Cache Eviction for Efficient Large Language Model Inference" [arXiv]

3. AB Gulhan, **K. T. Chitty-Venkata**, M Emani, M Kandemir, V Vishwanath "BaKlaVa–Budgeted Allocation of KV cache for Long-context Inference" [arXiv]

4. **K. T. Chitty-Venkata**, M Emani et al. "MoE-Inference-Bench: Performance Evaluation of Mixture of Expert Large Language and Vision Models", 2024 Supercomputing PMBS Conference, [arXiv]

PUBLICATIONS

1. **K. T. Chitty-Venkata**, J. Ye, M. Emani, "MoPEQ: Mixture of Mixed Precision Quantized Experts", accepted to Binary and Extreme Quantization for Computer Vision ICCV 2025 Workshop [arXiv]
2. **K. T. Chitty-Venkata**, M. Emani, V. Vishwanath, "LangVision-LoRA-NAS: Neural Architecture Search for Variable LoRA Rank in Vision Language Models", accepted to IEEE ICIP 2025 Conference [arXiv]
3. **K. T. Chitty-Venkata**, S. Raskar *et al.* "LLM-Inference-Bench: Inference Benchmarking of Large Language Models on AI Accelerators", 2024 Supercomputing PMBS Conference [arXiv] (Acceptance Rate = 25%)
4. **K. T. Chitty-Venkata**, V.K. Sastry *et al.* "WActiGrad: Structured Pruning for Efficient Finetuning and Inference of Large Language Models on AI Accelerators", in EuroPar2024 [Paper] (Acceptance Rate = 28%)
5. 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)
6. **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)
7. **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)
8. **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%)
9. **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%)
10. **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%)
11. **K. T. Chitty-Venkata** and A. Somani, "Neural Architecture Search Survey: A Hardware Perspective" in 2022 ACM Computing Surveys [Paper] (Impact Factor: 23.8)
12. **K. T. Chitty-Venkata** and A. Somani, "Array-Aware Neural Architecture Search" in IEEE ASAP 2021
13. **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%)
14. **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%)
15. **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%)
16. **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]

HONOURS/AWARDS

- Outstanding Postdoctoral Performance Award 2025 by Argonne National Laboratory
- Research Excellence Award by Iowa State University Graduate School [Certificate] [Letter from President]
- Research Award by Graduate and Professional Student Senate society at Iowa State University [Certificate]
- Attended Oxford Machine Learning Summer school 2022 (OxML) [Certificate] (Acceptance Rate < 10%)

SKILLS (STRONG)

- **Programming:** Python, C, CUDA
- **Deep Learning Frameworks:** PyTorch, DeepSpeed, NeMo, vLLM, SGLang, TensorRT-LLM, llama.cpp
- **Networks:** CNNs, Transformers, ViT, BERT, LLMs, MoEs, State Space Models, Vision Language Models