

Krishna Teja Chitty-Venkata

2040 Long Rd, 165H, Ames, IA, 50010 ♦ +1-515-203-5766
krishnat@iastate.edu ♦ [Linkedin](#) ♦ [Website](#) ♦ [Google Scholar](#)

SUMMARY

- Enthusiastic researcher, working at the intersection of Deep Learning and Systems
- Current research on optimizing DNNs (Pruning, Quantization, Neural Architecture Search) with respect to hardware (TPU-like, Multi-core CPU, GPU) and DNN accelerator algorithm co-design
- Former Deep Learning Research Intern at Intel's Graphics Processing Research team. Former Deep Learning Intern at AMD's DNN Graph optimization team

EDUCATION

| | |
|----------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------|
| Iowa State University PhD in Computer Engineering. 3.55/4.0 Advisor: Dr. Arun K. Somani | <i>Ames, Iowa, USA</i> <i>Aug '17 - Present</i> |
| University College of Engineering, Osmania University Bachelor of Engineering in Electronics and Communication. 8.4/10 | <i>Hyderabad, India</i> <i>Sept '13 - May '17</i> |

ACADEMIC/PROFESSIONAL EXPERIENCE

| | |
|-----------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------|
| Iowa State University Graduate Research Assistant, Dependable Computing and Networking Laboratory | <i>Ames, IA, USA</i> <i>May '18 - Present</i> |
|-----------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------|

My research involves optimization of DNNs for efficient inference on different hardware systems, co-design, and reliability of DNN accelerators. My current/previous research projects are as follows:

1. **Array Aware Architecture Search:** Design of a joint search algorithm for Architecture, Quantization and Array Accelerator supporting different precision (In progress). Simulators: SCALE-Sim and Bitfusion
2. **Review Paper:** A Comprehensive review of Pruning, Quantization, Neural Architecture Search from a hardware perspective (In progress for a publication in ACM Computing Survey Journal)
3. **Array Aware Pruning/Training:** Designed a Pruning algorithm and a Hyperparameter tuning method for CNN, MLP networks to minimize computation cycles of DNN forward pass on Array-based Neural Network Accelerators (TPU, Eyeriss) based on the hardware size. Simulator Used: SCALE-Sim
4. **CPU, GPU Dimension Aware Pruning:** Developed a combined Node Pruning, Symmetric Quantization, and layer fusion method based on Multi-core CPUs and Tensor Cores GPUs for efficient inference
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

| | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------|
| Intel Corporation Research Scientist Intern, Graphics Processing Research Lab Worked on Neural Architecture Search for Network design and Mixed Precision Quantization related to Image Restoration tasks (Super Resolution and Denoising) and Graphics applications | <i>Santa Clara, CA, USA</i> <i>June '20 - Dec '20</i> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------|

| | |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|
| Advanced Micro Devices (AMD) Machine Learning Intern, MIGraphX Worked in the MIGraphX (GPU graph optimization) team to design compression algorithms for enhancing performance on AMD GPUs at inference run-time. Developed quantization techniques to convert the weights of CNN from floating-point to integer precision on CNN benchmarks like Vgg16, ResNet50, Inception | <i>Austin, TX, USA</i> <i>May '19 - Aug '19</i> |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|

| | |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|
| Iowa State University Graduate Teaching Assistant, Digital Logic Design Responsibilities: Supervising labs, mentoring students on Verilog, FPGAs and technical projects | <i>Ames, IA, USA</i> <i>Aug '17 - April '18</i> |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|

Research Centre Imarat, Defence R&D Organization
Undergraduate Technical Intern
Project Title: Design and Simulation of Ethernet Controller on FPGA

Hyderabad, India
May '16 - June '16

Bharat Dynamics Limited, (A Govt. Of India Enterprise)
Undergraduate Technical Intern

Hyderabad, India
Dec '15

PUBLICATION(S) - SUBMITTED/ACCEPTED

1. **K. T. Chitty-Venkata** and A. Somani, "Array Aware Neural Architecture Search" in IEEE ASAP 2021 Conference (Submitted; Under review)
2. **K. T. Chitty-Venkata** and A. Somani, "Hardware Dimension Aware Pruning" in IEEE Transactions on Computers Journal (Submitted; Under review)
3. **K. T. Chitty-Venkata**, S. Kothandaraman and A. Somani, "Searching Architecture and Precision for U-net based Image Restoration Tasks" in IEEE ICIP 2021 Conference (Accepted)
4. **K. T. Chitty-Venkata** and A. Somani, "Calibration Data-Based CNN Filter Pruning for Efficient Layer Fusion" in IEEE HPCC-DSS 2020 Conference [[Paper](#)]
5. **K. T. Chitty-Venkata** and A. Somani, "Model Compression on Faulty Array-based Neural Network Accelerator" in IEEE PRDC 2020 Conference [[Paper](#)]
6. **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](#)]
7. **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
- **Machine Learning Frameworks:** Tensorflow, Pytorch, Keras, Scikit Learn
- **Other:** Linux, Shell Scripting, Verilog HDL, FPGA, Gem5 and ZSim Simulators, HTML

RELEVANT ACADEMIC PROJECTS

- **Reinforcement Learning using Neural Networks:** Designed and implemented Q-learning algorithm using DNNs as function approximator for acrobat-v1, an environment taken from OpenAI gym
- **High Performance Cache Simulation on GPU:** Developed a CPU-GPU based cache simulator and compared it with traditional CPU-only simulation. Developed the simulator using C (for CPU-only) and CUDA C (for CPU-GPU). CPU-GPU cache simulation performed better than CPU-only simulation

REFERENCES/RECOMMENDATIONS

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