Krishna Teja Chitty-Venkata
2040 Long Rd, 165H, Ames, IA, 50010 ⋄ +1-515-203-5766
krishnat@iastate.edu ⋄ www.linkedin.com/in/kt95 ⋄ https://krishnateja95.github.io

SUMMARY

- Enthusiastic researcher, working at the intersection of Deep Learning and Parallel Computing/Architecture
- Current research on optimizing DNNs with respect to special purpose (TPU-like) and general purpose hardware (CPU, GPU)
- Current Intern in Intel's Graphics Processing Research team. Former Intern at AMD's MIGraphX team.
- Experience working with Machine Learning (Regression, SVM, PCA, Clustering), Deep Learning (MLP, CNN, RNN) and Reinforcement Learning algorithms (Q-learning, SARSA).

EDUCATION

Iowa State University *Ames, Iowa, USA*
PhD in Computer Engineering. 3.55/4.0 *Aug '17 - 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 '13 - May '17*

ACADEMIC/PROFESSIONAL EXPERIENCE

Intel Corporation *Santa Clara, CA*
Research Scientist Intern, Graphics Processing Research Lab *June '20 - Present*

Iowa State University *Ames, Iowa*
Graduate Research Assistant, Dependable Computing and Networking Laboratory *May '18 - Present*

My research involves optimization (compression) of DNNs for efficient inference on hardware and applying to different problems (eg: fault tolerance). My research projects are as follows:

1) Array Aware Pruning/Training: Designed a Pruning algorithm and a Hyper-parameter tuning method for CNN, MLP networks to minimize the number of computation cycles of DNN forward pass on a systolic array based Neural Network accelerator.

2) Studying Structural Faults and Pruning model on Faulty DNN Hardware: Examined the impact of row and column faults on a systolic array on the performance of DNNs and proposed mitigation strategies. (Presented a short version at ASAP '19;).

3) CPU-GPU Aware Pruning: Developing a novel combined symmetric Pruning, Quantization and layer fusion algorithm to compress Neural Networks and accelerate on SIMD device (GPU) and multi-core CPUs to achieve high speedup during inference.

Advanced Micro Devices (AMD) *Austin, Texas*
Machine Learning Intern, MIGraphX *May '19 - Aug '19*
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, Xception.

Iowa State University *Ames, Iowa*
Graduate Teaching Assistant *Aug '17 - April '18*
TA for Digital Logic Design course in fall '17 and spring '18 semesters. Responsibilities: Supervising labs, mentoring students on Verilog HDL, FPGAs, course assignments and technical projects.

Undergraduate Research Experience
1) Bachelor Thesis: Development of Embedded System based Device for Detection of Fluorine in Water.
2) Part of a team to develop signal processing algorithms to extract modulation parameters like carrier frequency, bandwidth, code rate, cycles per phase and number of phases of Low Probability of Intercept (LPI) radars.

Research Centre Imarat, Defence R&D Organization *Hyderabad, India*
Undergraduate Technical Intern *May '16 - June '16*
Worked on a project titled "Design and Simulation of Ethernet Controller on FPGA". Developed a Transmitter and Receiver of the Ethernet controller using Verilog HDL according to the IEEE 802.3.

Bharat Dynamics Limited, (A Govt. Of India Enterprise) *Hyderabad, India*
Undergraduate Technical Intern *Dec '15*

PUBLICATION(S)

K. T. Chitty-Venkata and A. Somani, "Array Aware Training/Pruning: Methods for Efficient Forward Propagation on Array-based Neural Network Accelerators"

K. T. Chitty-Venkata and A. Somani, "Impact of structural faults on neural network performance," in 2019 IEEE 30th International Conference on Application-specific Systems, Architectures and Processors (Poster).

Metuku Shyamsunder, Kakarla Subbarao, Bharath Regimanu, CVSSD Krishna Teja (2017) "Estimation of modulation parameters for LPI radar using Quadrature Mirror Filter Bank," IEEE UPCON 2017 (Best paper award).

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
Deep Learning Networks/Data sets: MLP, CNN, RNN, MNIST, Cifar10, Imagenet
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.

Checkpoint-based Fault Mitigation Techniques for GPUs: Proposed a new fault tolerant algorithm to aid Check-pointing which attempts to reduce the communication overhead between CPU and GPU.

TEAM WORK EXPERIENCE

- Part of Dependable Computing and Networking Laboratory (DCNL) group consisting of six researchers
- Executive Committee member at Indian Students' Association (ISU) for 2018-19 academic year

RECOMMENDATION(S)

- Dr. Arun K. Somani (Doctoral Advisor): arun@iastate.edu
- LinkedIn Recommendations Section