Vima Gupta

GTID: XX6459

linkedin.com/in/vima-gupta

+1–4703349450 vimagupta.github.io

Education

Georgia Institute of Technology Atlanta, GA

PhD Computer Science, specializing in Systems for ML and advised by Dr. Anand Iyer and Dr. Ada Gavrilovska

M.S. Computer Science, specializing in Computing Systems (thesis track), *Jan'21-May'23* *GPA: 3.80/4.0*

Relevant Coursework: Systems for Machine Learning, Advanced Operating Systems, Statistical Machine Learning

Birla Institute of Technology and Science (BITS), Pilani *2014-2018*

Bachelors of Engineering in Electrical & Electronics Engineering, *GPA: 8.07/10* Pilani, India

Relevant Coursework: Neural networks and Fuzy Logic, Quantum Info Computing

Publications

- **V. Gupta**, A. Austin, E. Pinto, J. Young and T. Conte, "Effective qubit mapping routing and scheduling for Trapped-Ion shuttling architectures" (Under review at ISCA)
- **V. Gupta** and S. Varma, "Understanding Infinity: Neural Network Models of Becoming a 'Cardinal Principle Knower'" (Under Review for AAAI'23) [Paper]
- **V. Gupta** and S. Varma, "Learning to count: a neural network model of the successor function" Proceedings of the Annual Meeting of the Cognitive Science Society, vol. 44, 2022. [Poster]
- **V. Gupta** and R. Singhal, "Performance analysis of a visible light vehicle-to-vehicle wireless communication system," 2019, International Conference on Microwave Integrated Circuits, Photonics and Wireless Networks (IMICPW). IEEE, 2019 **(Best Paper Award)** [Paper]

Research Experience

Navigating the landscape of small language models *Aug'23 – Dec'23*

Advisor: Dr. Anand Iyer

- Modeling the pareto-frontier curve for accuracy (quality of output generated across tasks) versus system footprint (latency, throughput and KV cache memory) trade-offs for small language models and large language models.

Improving LLM inference on LLaMA 2 70B through speculative decoding *Aug'23 – Dec'23*

Research Advisors: Dr. Kexin Rong, Dr. Alexey Tumanov

- Designing scheduling policies for speeding up LLM inference for lower latency request and while gaining higher throughput.
- Understanding how small models can serve as auto-complete for LLM deployment by increasing its semantic awareness.

Improving ranking consistency of One-shot Neural Architecture Search techniques *Aug'23 – May'23*

Research Advisor: Dr. Alexey Tumanov

- Special Problems: A novel combination of curriculum learning with priors guiding the training process of NAS, aimed at improving the gap between ranking of weight-shared and independent training approaches for CNNs on Image-mini dataset.

Solving the qubit mapping and routing problem for shuttling-based quantum computers *Jan'22 – May'23*

Research Advisor: Dr. Thomas Conte (Center for Research into Novel Computing Hierarchies) *Master's Thesis (GRA)*

- Extending MaxSat techniques to improve the robustness and latency of SOTA algorithms for finding a feasible mapping from logical to physical qubits in a shuttling-based trapped ion quantum computer with dynamically evolving connectivity.

Teaching neural networks how to count from a cognition standpoint *Jan'22 – Dec'22*

Research Advisor: Dr. Sashank Varma

- Understanding becoming a "Cardinal Principle Knower", through learning the successor function by simulating human learning environment in lightweight MLP's through latent representation analysis.

Work Experience

Cerebras Systems *May 2022 – July 2022*

ML Frameworks Intern, Backend Atlanta, GA

- Converted the block sparse attention graph in BigBird, an NLP transformer which extends upon BERT, to match with existing highly optimized full attention kernel, from Tensorflow to MLIR lowering, at compile time for improved performance.
- The transformation was implemented through an MLIR graph match and rewrite pattern, which was automated in C++.

PACE: Physical Activity and Care for Everyone *May 2021 – Dec 2021*

Part-time co-founder, CREATE-X Atlanta, GA

- Developed an exercise library for Android application to enable remote physical training using Google's Mediapipe to give real-time feedback through pose detection.
- Conducted market research and designed the product website, and contributed towards iOS and Android application development towards our demo for CREATE-X, start-up incubator.

Arm Embedded Technologies *May 2018 – Dec 2020*
Design Engineer Bengaluru, India

- Led a sub-team of three interns to design an IoT subsystem for the open-source ecosystem. Synthesis, floorplanning and PnR for high performance cores, ultra low power machine learning accelerators and octa-core clusters in a customer facing role.
- Youngest engineer selected consecutively to present innovative work on system design at Arm's Global Engineering Conference.

Course Projects

Manifold mixup based regularization in federated learning *StatML: Jan'22 – May'22*

- Proposed a novel algorithm to increase regularization in the personalization layers of FedPer to improve generalizability.

Graph Convolutional Neural Networks for optimal circuit partitioning *Jan'21 – May'21*

- Achieved lower min-cut size compared to baseline (KL algorithm) using an unsupervised loss function on GCN through transfer learning by training on larger circuits (10^5 vertices)and performing inference on smaller circuits (10^3 vertices).

Designed infrastructure for Map-Reduce applications using gRPC in C++ *Advanced OS: July'21 – Dec'21*

- Designed a multithreaded program dynamically assigning map/reduce tasks in client-server architecture with file sharding.

Virtual CPU scheduler with memory coordinator for KVM based hypervisor *Advanced OS: Sep'21 – Oct'21*

- Implemented a vCPU scheduler and a memory coordinator using libvirt APIs to dynamically manage the resources assigned to each guest machine and collect statistics using hypervisor calls while satisfying variable workloads including egde cases.

Physical Design Aware NoC design for DNN Inference Accelerator, MAERI. *IC Networks: Jan'21 – May'21*

- Designed an H-tree inspired connectivity structure (layout) for the MAERI architecture for lower inference runtimes.

Skills and Teaching Experience

Programming skills – C++ (DSA and OOPs), Python, C, OpenMP, OpenMPI, Assembly, MATLAB, Agile practices

Python Libraries and software suites – PyTorch, Numpy, Matplotlib, Tensorflow, Streamlit, Qemu, Libvirt, Vtune

Graduate Teaching Assistant – Computer Vision (OMSCS 6476): Designed and graded assignments for a class of 500+ students.

Achievements and Extra-curricular activities

- Awarded the **EDIC fellowship** at EPFL, Lausanne, one among fifty candidates selected across the world.
- Awarded the **Adobe Research Women in Technology scholarship** 2022 from candidates across North America
- Student Organizations: Secretary at Quantum computing Association (2021), India Club Finance Leader (2021), English Drama Club Co-ordinator (2016-2017), Logistics head at Department of Controls (2015-2017).
- Awarded bronze medal for basketball in Bits Open Sports Meet, 2015
- Awarded 'Most Outgoing Student of the Year' in high school, 2012
- Secured All India 3rd rank a national level quizzing competition, 'Kaho What's My Idea' hosted by Derek'O'Brien, 2011