Ravi Ghadia

rghadia@utexas.edu • ghadiaravi13.github.io/ • ghadiaravi13

Skills: Large Language Models | GPU Architecture | Deep Learning | Combinatorial Optimization | Statistics | Distributed Systems | Reinforcement Learning | LLVM | Compilers

Education

The University of Texas at Austin **Aug'24 – Present**
Masters of Science in Electrical and Computer Engineering, CGPA: 4.0/4.0
Focus: Computer Architecture, Computer Systems and Embedded Systems

Indian Institute of Technology, Kharagpur **July'17 – June'21**
Bachelor of Technology in Electronics and Electrical Comms. Engg, CGPA: 9.35/10
Minor in Computer Science and Engineering

Industry Experience

GPU Kernels Intern, AMD **Sept'25 - Dec'25**
Manager: Bragadeesh Natarajan | GEMM Kernels, ROCm, HipBLAS

- Working on the Tensile library team for generating efficient GEMM kernels, optimized at assembly level
- Built an instruction level profiler to capture GPU runtime information during kernel execution.
- Captures registers/shmem values for every instruction, allowing a global view for debugging at scale – for instance visualizing memory accesses when using swizzled layouts

Summer Research Intern, Together.ai **May'25 – Aug'25**
Manager : Max Ryabinin | Large Scale Distributed Training, Long-Context LLMs

- Enabling training of LLMs with context >1M tokens on distributed GPU clusters via context parallelism
- Proposed a head-wise pipelining technique that significantly reduces activation memory during attention computation
- Allows training of Llama-8B on a 8×H100 node with maximum context length of **5M tokens**; 25% higher than SOTA

GPU Architect, NVIDIA **July'21 - Jul'24**
Manager: Sivakumar Anandan | Machine Learning, GPU Architecture, Statistical Analysis

- Developed high-fidelity statistical Debug system at unit level to understand GPU inefficiencies
- Built predictive ML models to project performance for compute workloads: HPC, LLMs, Recommender Systems
- Designed statistical Monte-Carlo simulation environments for DGX-class systems to enable management/marketing with product road-map decisions on datacenter systems
- Improved the runtime & resource complexity of highly intensive algorithms: MaxQ (using RL) and Bin-Optimization (using CNNs), enabling rapid turn-around for data requests

Architecture Intern, NVIDIA **April'20 – July'20**
Mentor: Sivakumar Anandan | Reinforcement Learning, GPU Architecture

- Used Reinforcement Learning to solve combinatorial optimization of the most energy-efficient GPU configuration
- Developed proof-of-concept promising massive runtime benefits compared to the traditional brute-force solution
- Follow-up work accepted as an Oral presentation at NTECH 2023, NVIDIA's internal Global Tech Conference

Research Experience

Graduate Research Assistant, The University of Texas at Austin **Aug'24 – Present**
Advisor : Prof. Poulami Das | Large Language Models, GPU Architecture, Machine Learning

- Optimizing **inter-token generation latency** during LLM inferencing by tuning **KV cache** size in real-time
- Leveraging Sparse attention to identify key tokens of importance to manage KV cache more economically

Research Assistant, H2Lab, University of Washington **Nov'22 – Aug'24**
Advisor : Prof. Prithviraj Ammanabrolu | Large Language Models, Reinforcement Learning

- Used **freeform linguistic feedback** to train large language models via **reward function learned on the feedback**
- Designed **vectorized policy framework** with localized rewards for incorporating multi-faceted feedback
- Experimental runs showed better alignment with the feedback: **improved Alignscore** on Question-Answering

Undergraduate Research Assistant, CNeRG CSE, IIT Kharagpur **Sept'20 – Jun'21**
Advisor : Prof. Pawan Goyal | Large Language Models, Reinforcement Learning, Statistics

- Critically analyzed the **shortcomings of Cross-Entropy** Loss for training **generative dialog models**
- Hypothesized the **model artifact theory**: For different inputs, model generates the same output response (which has almost no semantic relation to the input)
- Proposed **ranking based reward function** to train LLMs in an RL setting - mitigates the shortcomings of CE-Loss

Publications and Preprints

Dialogue Without Limits: Constant-Sized KV Caches for Extended Responses in LLMs
Ravi Ghadia, Avinash Kumar, Gaurav Jain, Prashant Nair, Poulami Das
Accepted to ICML 2025 | Arxiv [preprint]

- Implemented MorphKV: a dynamic token selection algorithm with constant KV cache usage for LLM inference
- Our approach saves **53% more memory** while improving accuracy by 18% on avg. compared to SOTA methods

MaxQ Optimization using Reinforcement Learning
Ravi Ghadia, Vamsi Krishna, Karthik Prakash, Sivakumar Anandan, Raghavendra Bhat
Accepted for Oral Presentation at NTECH US 2023 (NVIDIA-Internal Global Conference, acceptance 18%)

- Implemented RL based solution achieving the MaxQ configuration of a GPU (optimal configuration with best Performance per Watt)
- Delivered ∼**2000x** runtime and resource benefits as compared to the conventional brute force approach

CORAL: Contextual Response Retrievability Loss Function for Training Dialog Generation Models
Bishal Santra, Ravi Ghadia, Manish Gupta, Pawan Goyal
Bachelor's Thesis | Arxiv[preprint]

- Proposed retriever based loss function that considers context to assign loss for the generated output
- Achieved state-of-the-art on relevance metrics like MauDe/DeB against several strong pretrained baselines

Perf Activity Driven Instantaneous Power Projection
Ravi Ghadia, Sivakumar Anandan, Raghavendra Bhat
Accepted to NTECH India 2022 (NVIDIA-Internal Conference, acceptance rate 22%)

- Built a framework that allowed high precision energy analysis and helped isolate inefficient regions in the graphics pipeline

Teaching Experience

Teaching Assistant: Advanced Machine Learning **Aug'24 – Present**
Instructor : Prof. Joydeep Ghosh, McCombs School of Business, UT Austin

- Advanced course on ML covering Predictive modeling for practical business applications scaling on big data
- Facilitated interactive teaching sessions for a class with a highly diverse set of students from different backgrounds

Certified Instructor | NVIDIA Deep Learning Institute **Dec'21 – Jul'24**

- Served as an instructor for courses on **Transformer based NLP applications** and **LLM Inference Optimizations**
- Conducted sessions during NVIDIA GTC and assisted other instructors as TA during related courses

Technical Strengths

- **Programming Languages:** Python, C/C++, MATLAB, LC-3b
- **Frameworks:** Pytorch, Tensorflow, LLVM, Django, Streamlit
- **Libraries:** HuggingFace, OpenAI Gym, RL4LMs, stable-baselines
- **Profilers:** NVIDIA Nsight, Radeon Graphics Profiler
- **Utilities:** Docker, Perforce, Git, Bash, Linux

Selected Academic Projects / Competitions

Maverick 2.0 Hackathon | AB InBev **April'21 - May'21**
National Finalists (top 8 out of 750+ teams Pan India) | Machine Learning

- Developed an application to recommend customized discounts basis product data across various sectors
- Implemented pipelining for real-time request processing | Applauded by the panelists for outstanding design

Secure Authentication via user-behaviour **Aug'20 - Nov'20**
Advisor : Dr. Sudipta Mukhopadhyay | Machine Learning, Statistics

- Authenticated users based on their usage profile for mouse activity | click-time, pause-time, cursor-velocity etc.
- Used self-organizing maps for feature-extraction | Prevented unauthorised access with an 83% recall

Extracurricular Activities

- **Volunteered as a Mentor** at Mentor Together, a Non-Profit Organization aiming to assist underprivileged young-minds in their student-to-professional transition
- Served as the Hall Alumni Committee head, orchestrating alumni funds to initiate annual donation drive for Ambassadors Children Home, an orphanage near the IIT Kharagpur campus