Dav Vrat Chadha

davvrat.chadha@mail.utoronto.ca in/davvratchadha github.com/davvratchadha davvratchadha.com

Education

B.A.Sc. in Engineering Science, University of Toronto **Toronto, ON, Canada**
Major: Machine Intelligence [GPA: 3.54] *Sept 2020 – Apr 2025*

Skills & Tools

• Python • Tensorflow • PyTorch • NumPy • NLP • AI • ML • Sklearn • Keras • CUDA • Jax • Objax • spaCy • Cython • C/C++ • CI/CD • Git • MATLAB • Linux • Java • MySQL

Professional Experience

Software Engineer Intern, Memory Team **Markham, ON, Canada**
AMD - Datacenter GPU *May 2023 - Aug 2024*

- Designed Char_wizard, a memory characterization tool that reduced characterization time by 70% (650 hours) using innovative grid search algorithms. Improved runtime complexity from $O(n^2)$ to $O(n)$ and incorporated concurrency, resulting in over 2,000 hours of usage.
- Developed SuperScript, a **Python** library for memory validation and debugging, used over 100,000 times and integrated into several AMD tools to boost efficiency.
- Engineered a 4pt HBM-PHY margin testing tool for filtering **MI325** parts during manufacturing, to reduce customer RMAs.
- Designed and developed a **Python** and **C++** tool, amd-address-mapper, to translate physical addresses from ROCm workloads to failing locations in HBM, enhancing debugging efficiency and currently used internally and at customer sites.
- Contributed to AMD automation frameworks and infrastructure by developing memory automation tests for all MI300 platforms on **Linux**, enhancing testing efficiency and reliability.
- Contributed to HBM3E bring-up for MI325 product offering, providing memory validation task execution, automation, and validation tools support. Additionally, contributed to the successful HBM3 debug process during the El Capitan supercomputer bring-up phase.
- Implemented a scalable **Python** and **C** based Bit Error Rate testing tool for the HBM-PHY interface on all **MI300** platforms, enabling efficient data collection at scale.

Research Student - Computational Linguistics Lab **Toronto, ON, Canada**
University of Toronto *Sept 2024 - Present*

- Part of Prof. Gerald Penn's Computational Linguistic lab, working on a thesis project focusing on mechanistic interpretability of **large language models** using circuit discovery, model merging, and model editing, enhancing understanding of model behaviors.
- Exploring various model merging and model editing techniques to combine task-based circuits discovered via DiscoGP, for optimizing large language models for multi-task performance for resource-constrained environments, improving efficiency for on-device AI in wearables and smartphones.

ML Engineer - TalentRank (Engineering Capstone) **Toronto, ON, Canada**
Ornge, University of Toronto *Sept 2024 - Present*

- Leading a team to design and develop a **PyTorch** and **ChromaDB**-based information retrieval and **recommendation system**, successfully identifying top candidates from a large application pool.
- Implementing dense retrieval techniques to rank candidates based on job relevance, enhancing selection accuracy and reducing processing time.
- Coordinating with stakeholders to align system capabilities with hiring goals, ensuring the model effectively meets real-world recruitment needs.

ML Engineer - FINCH Satellite Mission **Toronto, ON, Canada**
University of Toronto Aerospace Team *Sept 2023 - Aug 2024*

- Implemented a novel **diffusion model** conditioned on neighboring spectral frames for **destriping** hyperspectral images resulting in PSNR = 39.2274, LPIPS = 0.2214, SSIM = 0.8817, and SAM = 0.0423, for the ICVL-HSI dataset.
- Worked on the hyperspectral data augmentation pipeline in **PyTorch**, utilizing a modified image patch extraction technique to create new images out of existing ones to increase the size of the training and validation dataset.
- Operated in an **agile** environment, contributing to continuous improvement and innovation within the team.

Teaching Assistant - CSC401/2511 Natural Language Computing **Toronto, ON, Canada**
University of Toronto *Sept 2024 - Present*

- Graded assignments, quizzes, and exams, ensuring consistent and fair evaluations, contributing to maintaining high academic standards.
- Collaborated with the Head Teaching Assistant and course instructor to uphold grading standards and policies, resulting in a streamlined grading process.

Publications

- Dav Vrat Chadha et al., *Optimizing Shmooing for AI HPC HBM Memory Characterization Using Decision Trees*, AMD Internal Publication, July 2024. Reviewed by AMD Fellows Committee.
- Ian Vyse et al., *Beyond the Visible: Jointly Attending to Spectral and Spatial Dimensions with HSI-Diffusion for the FINCH Spacecraft*, Presented in 38th Annual Small Satellite Conference, 2024. DOI: 10.48550/arXiv.2406.10724

Projects

Image Rendering Engine **Toronto, ON, Canada**
C, Performance analysis and Optimization, Linux *Sept 2024 - Oct 2024*

- Optimized a high-performance image rendering engine for real-time object manipulation based on preprocessed sensor data, achieving smooth 60 FPS rendering from 1500Hz oversampled input.
- Enhanced **C**-based performance on **Linux** by implementing efficient data structures and profiling with perf, gprof, and GCov, achieving a 700x speedup over the baseline implementation.
- Developed and optimized frame-buffer transformations (e.g., shifting, rotating, mirroring) to enable real-time, accurate image manipulation on constrained hardware.

ML Engineer - NoPunIntended **Toronto, ON, Canada**
Repository; PyTorch, Python, LLMs, NLP *Jan 2023 - Apr 2023*

- Fine-tuned an ensemble of transformer-based **LLMs** (DeBERTa and RoBERTa) using **PyTorch** to accurately detect and locate puns within text by applying contextual masking techniques and clustering with K-means for enhanced interpretability and localization.
- Advanced Amazon's research on pun detection by implementing and optimizing model architecture and training strategies, resulting in an improved method that reached a 75.58% test accuracy—a performance approaching that of GPT-4 (82.77%) on similar tasks.

Honors & Awards

- Two-time AMD Executive Spotlight Award winner - Aug 2024, Dec 2023
- AMD Intern Innovation Showcase Award - Mar 2024
- Two-time AMD Spotlight Award winner - Dec 2023, Aug 2023
- Dean's List, University of Toronto - Dec 2020