Hiva Mohammadzadeh

Hivam.org | hiva@berkeley.edu | linkedin.com/in/hivamohammadzadeh | github.com/hivamohammadzadeh1

Education

University of California, Berkeley — **August 2021 – December 2023**
Bachelor of Science in Electrical Engineering and Computer Sciences — Berkeley, CA

Los Angeles Pierce College — **August 2019 – June 2021**
Associate of Science for Transfer in Computer Science and Programming — Woodland Hills, CA

Experience

Machine Learning Researcher in NLP — **February 2023 – Present**
Pallas Group in BAIR and SLICE Labs at UC Berkeley — Berkeley, CA

- Advisors: Prof. Kurt Keutzer and Coleman Hooper
- Building efficient LLM-based systems
- Collaborated on KVQuant which allows serving LLaMA-7B with 1M tokens on a single A100 GPU using KV Cache Quantization
- Built an architecture to accelerate generative LLM inference by 40% as co-author for SPEED paper

Modeling and Data Science Intern — **May 2022 – September 2022**
Span.io (Series B Startup) — San Francisco, CA

- Designed and implemented python software to solve Nonlinear Differential Equations to speed up analytics by 75%
- Simulated home appliance power consumption using the Span Panel data to inform next product iteration

Undergraduate Researcher — **June 2021 – October 2021**
Computational Infrastructure for Geodynamics, NSF, UCSD, NASA/JPL — CA

- Built and analyzed a model of Venus on supercomputers using Python and Fortran with Prof. Dave Stegman (UCSD)
- Found that plume-assisted tectonic subduction happens 80% faster than hypothesized while advised by Dr. Sue Smrekar
- Co-authored scientific paper in support of NASA's Venus VERITAS mission of NASA/JPL

Publications

KVQuant: Towards 10 Million Context Length LLM Inference with KV Cache Quantization by Coleman Hooper, Sehoon Kim, **Hiva Mohammadzadeh**, Michael W. Mahoney, Yakun Sophia Shao, Kurt Keutzer, Amir Gholami. (Under Review NeurIPS 2024)

SPEED: Speculative Pipelined Execution for Efficient Decoding by Coleman Hooper, Sehoon Kim, **Hiva Mohammadzadeh**, Hasan Genc, Kurt Keutzer, Amir Gholami, Sophia Shao. (NeurIPS ENLSP Workshop 2023)

Plume-Induced Delamination Initiated at Rift Zones on Venus by Andrea C. Adams, Dave R. Stegman, **Hiva Mohammadzadeh**, Suzanne E. Smrekar, and Paul J. Tackley. (Journal of Geophysical Research: Planets 2023)

Awards

- Won Third Place at SCET's Annual Collider Cup XIII — December 2023
- AnyScale's Sponsor Prize Winner from Skydeck and Cal Hacks AI Hackathon — Summer 2023
- Two-time recipient of Undergraduate Summer Fellowship award from Sky Computing Lab — 2022, 2023

Skills

Programming Languages: Python, Java, C/C++, JavaScript, SQL, MongoDB, Assembly, Fortran, MATLAB, Scheme
Developer Tools: Tmux, VS Code, Google Cloud Platform, XCode, IntelliJ, PyCharm, TI Launchpad, and Arduino
Frameworks: PyTorch, TensorFlow

Relevant Coursework

- Database Systems
- Artificial Intelligence
- Machine Learning
- Deep Learning
- Natural Language Processing
- Deep Reinforcement Learning, Decision Making and Control
- Responsible Generative AI, and Decentralized Intelligence

Projects

SnapSite | AI Hackathon 2023 by UC Berkeley Cal Hacks and Skydeck — **June 2023**

- Led the development of SnapSite, revolutionary AI tool that allows users to create websites instantly from photos of text
- Won the Sponsor's prize from AnyScale (Link to prototype)

TensorZipper Project Startup | Connected Life Challenge Lab SCET Class — **August 2023 - December 2023**

- Led the development of a novel AI model compression algorithm, leading to smaller, faster, and cheaper AI models.
- Won first place among eight class projects and third place among thirteen projects at SCET's Annual Collider Cup XIII