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

Experience

Machine Learning Researcher in NLP **February 2023 – Present**
Pallas Group at UC Berkeley Artificial Intelligence Research (BAIR) Lab Berkeley, CA

- Building efficient LLM-based systems and working on a survey of AI Agents as the first author
- Contributed to Squeezed Attention, a technique to accelerate LLM inference in applications where a large portion of the input prompt is fixed. (Submitted to MLSys 2025)
- Collaborated on KVQuant (NeurIPS 2024) 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 (NeurIPS ENLSP 2023)

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

Squeezed Attention: Accelerating Long Context Length LLM Inference by Coleman Hooper*, Sehoon Kim*, **Hiva Mohammadzadeh**, Monishwaran Maheswaran, June Paik, Michael W. Mahoney, Kurt Keutzer, Amir Gholami (Submitted to MLSys 2025)

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 (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)

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

Awards

- Won Third Place at SCET's Annual Collider Cup XIII for the TensorZipper Project **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**

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