Yasin Mazloumi

Senior ML Research Engineer · Kempner Institute, Harvard University

Boston, MA · abbas_mazloumi@harvard.edu · abbasmazloumi.com · LinkedIn · Google Scholar

ML research engineer building systems for large-scale training and inference of deep-learning, LLM, and multimodal models across multi-node GPU clusters. Ph.D. in Computer Science with a decade of work spanning distributed high-performance computing, GPU systems, and graph analytics. NVIDIA DLI University Ambassador; HiPC'20 Best Paper.

EXPERIENCE

Senior ML Research Engineer — Kempner Institute, Harvard University Jun 2024 – Present

- Build distributed-training systems and scalable ML workflows for large-scale foundation-model research on the institute's GPU clusters.
- Contribute to KempnerForge, an open-source PyTorch-native framework for fault-tolerant distributed training (FSDP2, tensor/expert/pipeline parallelism, FP8, and Mixture-of-Experts) at 125M–70B parameters.
- Train multimodal vision–language models with diverse encoders (CLIP, SigLIP2, AIMv2, DINOv2) and fusion strategies (cross-attention, joint decoders, mixture-of-transformers) to study cross-modal transfer.
- Design and teach distributed-training workshops (DDP, FSDP, tensor & pipeline parallelism), including the LLM Distributed Training workshop and the NVIDIA DLI "Data Parallelism" course at the Kempner NeuroAI symposium.

Lecturer in Computer Science — UC Riverside Sep 2023 – Jun 2024

CS 005: Introduction to Computer Programming.

Lecturer in Computer Science — San Diego State University Aug 2023 – Feb 2024

CS 635: Advanced Object-Oriented Programming (graduate).

Graph AI Software Engineer — Katana Graph Oct 2021 – Dec 2022

Built high-performance distributed graph-analytics systems (Graph AI team).

Associate Instructor — UC Riverside Apr 2021 – Sep 2021

CS 153: Operating Systems · CS/EE 147: GPU Computing & Programming.

Teaching / Research Assistant — UC Riverside Jul 2017 – Sep 2023

Teaching / Research Assistant — University of Tehran Sep 2012 – Aug 2016

EDUCATION

Ph.D., Computer Science — UC Riverside 2023

Dissertation: Distributed Evaluation of Batches of Iterative Graph Queries (MultiLyra, BEAD, SimGQ / SimGQ+ — HiPC'20 Best Paper); advisor Prof. Rajiv Gupta.

M.Sc., Computer Science — UC Riverside 2020

M.Sc., Computer Architecture — University of Tehran 2014

B.Sc., Computer Engineering — University of Mazandaran 2009

HONORS & AWARDS

NVIDIA University Ambassador & Instructor Jan 2025

NVIDIA DLI Certification — Data Parallelism (multi-GPU training) Nov 2024

HiPC'20 Best Paper Award — *SimGQ* Dec 2020

Dean's Distinguished Fellowship, UC Riverside 2016–2020

PUBLICATIONS

PREPRINTS & UNDER REVIEW

› Understanding the Design Space and Cross-Modality Transfer for Vision-Language Models. Under review, 2026.

› Task Relevance Is Not Local Replaceability: A Two-Axis View of Channel Information. Under review, 2026.

› The Emergence of Complex Behavior in Large-Scale Ecological Environments. Under review, 2026.

PEER-REVIEWED

› DIRT: The Distributed Intelligent Replicator Toolkit. ALife, 2025.
› ExpressWay: Prioritizing Edges for Distributed Evaluation of Graph Queries. BigGraph, 2023.
› SimGQ+: Simultaneously Evaluating Iterative Point-to-All and Point-to-Point Graph Queries. JPDC, 2022.
› BEAD: Batched Evaluation of Iterative Graph Queries with Evolving Analytics Demands. BigData, 2020.
› SimGQ: Simultaneously Evaluating Iterative Graph Queries. HiPC, 2020. ★ **Best Paper Award.**
› MultiLyra: Scalable Distributed Evaluation of Batches of Iterative Graph Queries. BigData, 2019.
› Enabling Faster Convergence in Distributed Irregular Graph Processing. BigData, 2019.
› Border Gateway Protocol Anomaly Detection Using Neural Network. BigData, 2019.
› Fast Data Delivery for Many-Core Processors. IEEE TC, 2018.
› Parallel Forwarding for Efficient Bandwidth Utilization in Networks-on-Chip. ARCS, 2017.
› High-Performance Hybrid-Switched Network-on-Chip Using Shortcut Paths. ICEE, 2016.
› Dynamic Resource Sharing for High-Performance 3-D Networks-on-Chip. IEEE CAL, 2016.
› A Hybrid Packet/Circuit-Switched Router to Accelerate Memory Access in NoC-Based CMPs. DATE, 2015.
› Integrated Circuit-Packet Switching NoC with Efficient Circuit-Setup Mechanism. Journal of Supercomputing, 2015.

TECHNICAL SKILLS

Languages C/C++, Python, Bash

ML & distributed systems PyTorch (DDP, FSDP2, TP/PP/EP), FP8 & Mixture-of-Experts, Torch Profiler, NVIDIA Nsight, CUDA; multi-node graph analytics on GPU/CPU clusters (SLURM)

Practices Test-driven development, object-oriented design, extending large open-source projects

RESEARCH INTERESTS

– Generative AI; large-scale distributed training & inference
– Multimodal learning & vision–language models
– Distributed graph analytics & graph-based ML
– High-performance & parallel computing
– Computer & GPU architecture and programming