Milad Heydari

253-457-8911 | miladgheydari@gmail.com | linkedin.com/in/MiladGHeydari

Education

Georgia Institute of Technology

BS in Computer Science - AI, Math concentrations 3.86 GPA

MS in Machine Learning 4.0 GPA

Courses: Deep Learning, Deep Reinforcement Learning, Vision-Language Models, Natural Language, ML Systems, HPC Programming, Computer Vision, Robotics, Statistical ML, Machine Learning, OS

Experience

Databricks Feb 2025 -

Software Engineer — Photon *San Francisco, CA*

Databricks May - Aug 2024

Software Engineering Intern — Photon (C++ version of Spark / *paper link*) mentor *San Francisco, CA*

- Proposed novel vectorized broadcast join algorithm that was O(n) more efficient for sparse batches than SOTA. Setting a benchmarking record (85% speedup) & saving millions in compute.
- Implemented memory coalescing during batch processing of joins, improving throughput by 32x
- Optimized Spark using batching, SIMD, CUDA kernels, and zero-copying to reduce data loading time

Microsoft May - Aug 2023

Software Engineering (Applied Science / ML) Intern — Azure Hyperscale Networking mentor *Redmond, WA*

- Modeled large-scale distributed systems as real-time updating knowledge graph with devices (nodes), links (edges), and anomaly events (temporal nodes), enabling modeling spatial relationships between anomalies
- Built API that queried network/graph to detect & cluster anomalies. Fed data to LLM using RAG to diagnose issues e.g: using packet loss between datacenters to locate bad cables. 22% drop in avg time to diagnose issues

Cockroach Labs Jan - May 2023

Software Engineering Intern — Cloud Infrastructure & Distributed Systems *New York, NY*

- Created service to detect underused Kubernetes clusters and mitigate resource waste by enforcing a resource virtualized "sleep" state and predicatively scaling compute based on demand. Saving $70,000+ annually
- Enabled multi-cloud data querying/hosting via upgrading consensus algorithm for KV / database transactions

Microsoft May - Aug 2022

Applied Science Intern — Azure Networking *Redmond, WA*

- Built backtesting framework for ML models. Allow for A/B and unit testing different models and hyperparams
- Designed architecture/model for time series anomaly detection on streamed data. Performs real time signal processing on a distributed system's network data and triggering mitigation. Prevented 3 incidents in 1st week

Microsoft May - Aug 2021

Data Science Intern — Machine Learning Product Integration *Cambridge, MA*

- Worked with client, Novant Health, to develop custom NLP/ML models and data pipelines to process language data (encoder model) and monitor metrics (RNNs) in diagnoses recommendation/search tool for doctors
- Used auto-encoders to improve model efficiency and preserve privacy of data during federated learning

RR AI May - Aug 2020

Robotics & Deep Learning Research Intern — Computer Vision on Autonomous Vehicles *Washington D.C*

- GANs / programmatic synthetic data generation to supplement training of vision models
- During AV fleet movement, projected position of other vehicles into camera view of each vehicle - using positions to test/calibrate/train predicted positions from onboard vision model and assist with navigation/SLAM

Projects

Distributed Training: locking-free distributed SGD. Scaling Hogwild without atomicity guarantees
Inference Scheduler: vLLM fork with virtual token counter scheduler, batching {prefills, decodes} without stalling
LLM reasoning finetuning: improving Quiet-sTaR & sTaR chain of thought reasoning using DPO training
3D VLM editor: adding/removing objects to Gaussian splats using language. Joint CLIP/grouping training
RL autograd framework: backprop with support for randomized variables and stochastic/RL problems
Robust RLHF in LLM Training: achieving stable reward despite variance in human feedback

Skills

Languages: Python, C++, C, Golang, C#, Java, Scala, JavaScript, SQL
Tools: PyTorch, CUDA, Kubernetes, MPI, OpenMP, Spark, OpenCV, NumPy, Pandas, gRPC, HuggingFace, Wandb, LangChain, AWS, Azure, GCP