HARSHIT TIMMANAGOUDAR

+1 (858) 405-1729 | htimmanagoudar@ucsd.edu | San Diego, CA, USA | linkedin.com/in/harshit-timmanagoudar-829b43212/ | github.com/harshitster | harshitst.netlify.app/

EDUCATION

University of California - San Diego **September 2024 - March 2026**
Master's, Computer Science *GPA: 3.67*

- Coursework: Data Systems for Machine Learning, Generative AI, Large Model Reasoning, Distributed Systems, Operating Systems, Database Principles.

PES University **December 2020 - May 2024**
Bachelor's, Computer Science *GPA: 3.81*

- Coursework: Deep Learning, Statistics for Data Science, Computer Vision, Data Structures, Cloud Computing, Natural Language Processing.

PROFESSIONAL EXPERIENCE

STABLE Lab, UC San Diego **San Diego, CA, USA**
Graduate ML Systems Researcher *May 2025 - Present*

- Architected parallelized LLM inference system using PyTorch and CUDA, with CXL 3.0 shared memory and novel vector database across GPU nodes, leveraging NCCL for distributed communication and outperforming vLLM by 1.4x on throughput measurement on concurrent user requests.
- Engineered production-grade scalable KV cache in C++/PyTorch extensions with CUDA and Triton kernels, achieving O(1) write operation complexity leading to 300x - 1600x improvement and reducing memory overhead by 16x compared to HuggingFace implementations.
- Eliminated inter-GPU communication bottleneck through CXL 3.0 shared memory architecture, enabling zero-copy tensor activations access across pipeline stages and removing network synchronization overhead in multi-node inference systems.

Peptris Technologies **Bengaluru, KA, India**
AI Research Intern *January 2024 - June 2024*

- Developed end-to-end permutation invariant Graph Attention VAE pipeline for 3D data generation, training on 150M+ structures stored in GCS buckets with RDKit-based preprocessing to produce 18K+ diverse conformations with 14% improved binding pose accuracy.
- Fine-tuned Latent Diffusion Model for pose generation, achieving 38% inference speedup (2.4s to 1.5s per molecule) through optimized sampling, deployed on AWS EC2 with Docker containers for high-throughput processing and GCS for model artifact storage in multi-cloud infrastructure.
- Evaluated 6 graph architectures (GCN, MuchGCN, PIGVAE) and diffusion techniques (score matching with Langevin dynamics, Riemannian manifolds) with MLflow experiment tracking under CTO/Chief Data Scientist supervision, establishing model selection framework for production deployment.

Peptris Technologies **Bengaluru, KA, India**
Machine Learning Intern *June 2023 - August 2023*

- Engineered protein sequence generation pipeline using transformer-based language model with sparse Mixture-of-Experts architecture in TensorFlow, reducing computational requirements by 40% while maintaining 87% structural validity validated through AlphaFold2 predictions.
- Built 3D U-Net model in TensorFlow for high-throughput binding pocket detection, processing 50+ protein structures hourly and identifying multiple potential binding sites per structure to accelerate drug discovery for molecular docking.

National University of Singapore (NUS) **Singapore**
Deep Learning Research Intern *June 2022 - July 2022*

- Led a team of 6 in developing breast cancer detection system, architecting ensemble of custom CNN, VGG19, and ResNet50 with Scikit-learn soft voting, Pandas/NumPy preprocessing and TensorFlow augmentation pipeline with Matplotlib dashboards, achieving 99.4% classification accuracy.

PROJECTS & PUBLICATIONS

Disaggregated Speculative Decoding - *Link to project* *November 2025 - December 2025*

- Accelerated autoregressive inference by 4.51x with 1.50x concurrent throughput and 52% acceptance rate by developing distributed speculative decoding using Ray to orchestrate parallel draft workers, target verification and scalable serving infrastructure in PyTorch.

DistRAG - Distributed RAG Database - *Link to project* *June 2024 - January 2025*

- Architected multi-node RAG system with PostgreSQL Citus cluster for parallel queries, LlamaIndex/Gemini for NL-to-SQL conversion via ChromaDB semantic indexing, Redis vector caching, orchestrated on Kubernetes with NGINX ingress for load balancing.
- Implemented event-driven architecture for real-time synchronization and cache invalidation, with Kubernetes StatefulSets enabling automated worker recovery and fault tolerance across node cluster.

Edge Map Extraction of High Resolution Facial Images - *Link to project* **Singapore**
First Author *February 2023 - June 2023*

- Formulated computer vision edge detection algorithm for facial images using Gaussian smoothing, directional gradients, dual-threshold detection, and Zhang-Suen thinning, outperforming Canny and HED algorithms on 10,000+ image validation set. Published at ISBM, Springer Nature 2023.

SKILLS

Languages: Python, C/C++, Golang, Rust, Bash, Java, JavaScript, SQL
Frameworks/Tools: PyTorch, Tensorflow, Keras, Ray, CUDA, NCCL, JAX, Triton, Pandas, Numpy, Scikit-learn, AWS, GCP, Git, Kubernetes, Docker, gRPC