Pratishtha Gaur

(619) 953-7604 | p1gaur@ucsd.edu | PratishthaGaur | pratishtha-gaur

Education

University of California - San Diego — *San Diego, USA*

Masters of Science in Data Science (CGPA: 4/4) — *Sep 2023 - Dec 2024*

- President of Data Science Graduate Association
- Coursework: ML System, Parallel Computation, Large Language Models (LLM) , Natural Language Processing (NLP)

Vellore Institute of Technology — *Vellore, India*

Bachelors of Technology in Computer Science and Engineering (CGPA:3.98) — *Jul 2019 - May 2023*

- Coursework: Machine Learning, Artificial Intelligence, Social Information Network, Image Processing, Database System, Operating System

Professional Experience

UC San Diego — *San Diego, California*

Student Researcher — *Apr 2024 - Present*

- Improved KV cache utilization for multi-round conversations in Large Language Models by replacing the existing Least Recently Used (LRU) eviction policy to a priority-based approach, by classifying conversations as short-running and long-running tasks
- Developed a model using open source Large Language Model and Retrieval-Augmented Generation (RAG) to extract and generate academic questions from dialogues and seminar content.

Adani Wilmar — *Ahmedabad, India*

Data Science Intern — *Dec 2022 - Jun 2023*

- Implemented fuzzy matching to link menus with SKUs. Using this mapping and order distribution, developed a cost function to rank HoReCa institutes for a lead generation recommender system. This approach increased sales and reduced labor costs by 50%.

InfyU Labs — *Gandhinagar, India*

Machine Learning Intern — *May 2021 - Aug 2021*

- Utilized a pre-trained Mask-RCNN model and fine-tuned it with a custom dataset to detect and locate defects in fruits. This process led to a 30% reduction in quality control time and more efficient segregation based on defects.

Projects

Tesla T4 GPU GEMM Optimization with CUDA

CUDA, C++

- Developed CUDA kernels in C++ for GEMM operations on Tesla T4 GPUs, achieving 78% of CuBLAS's peak performance through optimizations such as tiling, shared memory, double buffering, memory coalescing, and compute coalescing.

Autograd Engine

PyTorch

- Implemented a backpropagation reverse-mode autodiff with operators like, matmul, softmax, exp over a dynamically built DAG and trained a logistic regression on top of it with PyTorch.

Transformer Experiments for Speech Attribution

PyTorch

- Implemented a transformer encoder and train it jointly from scratch with a feedforward classifier for a downstream task of predicting which politician delivered a given speech segment, experimented with different parts of the architecture such as the positional encoding, attention and pre and post normalization.

Patient Monitoring

Kafka, Amazon S3, EC2

- Applied Long Short-Term Memory (LSTM) models to analyze time series data, coordinated real-time streaming using Apache Kafka, and established a system for storing and archiving analyzed results in an S3 bucket. This system monitored hourly vital signs for a patient monitoring system, aiming to detect early signs of sepsis.

Publication

Intelligent Gym Trainer Supporting Pose Correction Using PoseNet and YoloV4 [Link] — *Chennai, India*

Presented in International Conference on Computer Science and Artificial Intelligence (ICCSAI - 22) — *Oct 2022*

- Integrated PoseNet for real-time pose detection and YOLOv4 for object detection to optimize workout effectiveness by ensuring correct body positioning and suggesting exercises based on detected equipment in the feed. Provided real-time feedback on posture correctness.

Skills

Generative AI CUDA programming, PyTorch, Large Language Model, Finetuning, Prompt Engineering, RAG, Langchain

Databases SQL, NoSQL Databases (MongoDB, Google Firestore), Apache Kafka

Languages Python, C++, R