Hardik Gupta

gupt0414@umn.edu — (763) 485-6749 — GitHub — LinkedIn — hardik.page

Education

University of Minnesota - Twin Cities Minneapolis, United States
Master of Science, Robotics *Sep 2023 – Present*

Coursework: Natural Language Processing, Computer Vision, Machine Learning, Advanced Artificial Intelligence, Data Structures and Algorithms, Operating Systems

Birla Institute of Technology and Science, Pilani Pilani, India
Bachelor of Engineering, Mechanical Engineering *Aug 2018 – Jun 2023*

Skills

Languages: Python, C, C++, JavaScript, HTML, CSS

Technologies: MongoDB, MySQL, Flask, TensorFlow, PyTorch, Scikit-learn, AWS, OpenCV, Git, CUDA, Docker

Certifications: Deep Learning Specialization (Andrew Ng, Coursera)

Experience

University of Minnesota, Twin Cities Minneapolis, United States
Machine Learning Engineer - Research Assistant *Jan 2024 – Present*

- Developed an autoencoder-based image compression pipeline, achieving a 68.6% training loss reduction, 49:1 compression ratio, and 88.4% reconstruction accuracy on confocal microscopy neuron images
- Developed scalable Bayesian inference for consumer behavior analytics; optimized Transformer tokenization/attention to improve data throughput, boosting segment-level CTR by 15%

Union Bank of Switzerland Mumbai, India
Financial Data Analyst Intern *Feb 2023 – Jun 2023*

- Designed and implemented Python and VBA Macros to automate Pension IPV analysis, reducing manual effort by 25% and saving ∼30 staff hours per month (staff member saved ∼1 hour/week)
- Built robust data pipelines using Pandas and NumPy to process a total of 50K pension entries (handling 10K entries monthly), integrating with PowerBI for real-time dashboards leveraged by the Rates-FX finance team

National University of Singapore Singapore city, Singapore
Machine Learning Researcher *May 2022 – Dec 2022*

- Developed a learning-based Nonlinear Model Predictive Control (NMPC) framework using a neural network dynamics model, trained on 2,000+ hours of simulated data
- Incorporated domain randomization for improved generalizability, reducing collision rates by up to 16% in single-robot dynamic obstacle avoidance

Ericsson Gurgaon, India
Data Science Intern *Aug 2020 – Dec 2020*

- Orchestrated a robust end-to-end ML pipeline (Python, Pandas, Scikit-learn) for telecom performance data, integrating advanced feature engineering and cross-validation to improve predictive accuracy by 23%
- Implemented real-time anomaly detection with proactive alerts, identifying site issues early and saving 48 engineering hours per month

Projects

Flash Attention from First Principles with Triton Feb 2025
Python, Triton, CUDA, PyTorch *Repository*

- Developed the Flash Attention 2 algorithm entirely in Python/Triton, deriving forward/backward passes step by step, resulting in a 721-line codebase
- Achieved a 2.5x faster run over standard PyTorch attention on sequence lengths up to 4096, with a 90% reduction in GPU memory usage and a throughput of approximately 10,000 QPS, demonstrating advanced CUDA/Triton kernel optimization techniques

Retrieval-Augmented Generation for Fact Checking @ AI x Journalism Hackathon
Python, FAISS, Sentence Transformers, LLM *Repository*

- Indexed 20K+ politician statements into a FAISS database for low-latency retrieval and combined with a hosted LLaMA model for real-time fact-checking
- Achieved 2.08s average end-to-end latency (2.31s at p95) and a throughput of 0.32 QPS, peaking at 363 MiB memory usage during retrieval and generation