Narain Pattabhiraman

Linkedin Github Portfolio narain13579@gmail.com

EDUCATION

Arizona State University, *M.S, Computer Engineering, Computer Systems* *Aug '23 – May '25*

SKILLS AND CERTIFICATIONS

Programming Languages: Python, C/C++, Spark, SQL, CUDA, Triton, LLVM.
Machine Learning and Data Science: MLOps, Computer Vision, NLP, GenAI, LLMs, Airflow, Sagemaker.
Cloud Technologies AWS, GCP, Docker, DevOps, Git, Terraform, Jenkins, Kubernetes, Kafka.
Frameworks and Tools: REST, Websocket, Pytorch, Clang, LLVM, GDB, VLLM, Autogen, Unsloth, Vector Dabases, RAG.

WORK EXPERIENCE

AI Engineer (Full-Time), ASU Enterprise Technology, Tempe, USA *May '25 – Present*
MLOps & AI Development Engineer, ASU Enterprise Technology, Tempe, USA *Jan '24 – May '25*

- Built a **dynamic agentic DAG workflow engine** that supports heterogeneous **LLMs** and **tools**, allowing on-the-fly composition of complex, multi-step workflows with adaptive routing and tool selection.
- **Built LLM infrastructure** on ASU's SOL Cluster using Kubernetes, enabling dynamic resource scaling for inference.
- Trained **Embedding Models** for Domain specific documents using **MLM Pretraining**, with **DeBERTa** architecture.
- Developed the backend for **CreateAI**, an **LLM-powered** chat assistant featuring **multi-model support**, **rate monitoring**, **RAG**, **agentic evaluation** to prevent hallucinations and **semantic caching**.
- Engineered a indexing system for **multimodal question answering**, integrating **text, video, image** representations for data retrieval.
- Implemented **serverless inference of ML models** on AWS Lambda using **GGML**, significantly improving scalability and flexibility by enabling deployment of custom models at scale and at reduced cost compared to established model providers.

Software Engineer, Fidelity Investments, Chennai, India *Apr '22 – Jul '23*

- Developed and implemented robust **Infrastructure and Pipelines** for the deployment of **Machine Learning Models** on **AWS**.
- Engineered **Online Feature Stores** to support real-time data and optimized batch inferencing to handle large-scale data.
- Contributed to the development and deployment of diverse machine learning models, including **Document Layout Models**.
- Led the deployment and performance optimization of **Large Language Models**, including **BLOOM** and **Flan-T5**, using **Tensor and Model Parallelism** to accelerate token generation and enhance model performance.
- **Collaborated in deploying scalable machine learning workflows**, ensuring efficient model and data lifecycle management.

AI Researcher, QPIAI Technologies, Bangalore, India *May '21 – Apr '22*

- Implemented a **Graph Convolutional Network (GCN)** to train graph contrastive learning models for different circuit layouts, achieving high recall in identifying similarities between different circuit layouts.
- Introduced a custom **Ranking Algorithms** for product recommendations supporting automated price estimation for sales.
- Involved in developing **Safety Monitoring Systems** in warehouses using cameras to ensure compliance with social distancing.
- **Developed an AutoML platform for ML models**, enabling automatic hyperparameter tuning and model selection.
- Contributed to the hardware/software co-design process focusing on algorithmic optimizations to maximize inferencing throughput.

Data Analyst, Thoughtware Analytics, Bangalore, India *Jan '20 – May '21*

- Conducted **Demand Forecasting** for manufacturing and service-oriented businesses, focusing on **Supply Chain Improvements**.
- Developed **Computer Vision** based for quality assurance in manufacturing processes, reducing material waste.
- **Conducted operational research to optimize logistics**, utilizing advanced algorithms to improve route planning and resource allocation, reducing costs and delivery times.

Open Source Contributions

- **Pytorch, Pytorch-lightning, Albumentations, Kornia**.

ACHIEVEMENTS

- **Winner, AMD Synthetic Data Hackathon** – Fine-tuned LLM reasoning via reinforcement learning for a question-answering agent.

PROJECTS

Robotic Arm Maneuvering Using Deep Reinforcement Learning *B.Tech Thesis Project, Amrita Vishwa Vidyapeetham*

- Designed and implemented a simulation environment for a robotic arm in **Unity3D** with the **Pytorch**, and fabricated a physical prototype using **3D-printed components** and **servo motors** for real-world experimentation.
- Developed and trained agents using advanced RL algorithms: **DDPG**, **A2C**, and **PPO** for sample-efficient on-policy learning.
- Engineered continuous control tasks with (**>30 DoF**), leveraging neural network policies to process raw sensor data and image.
- Demonstrated successful transfer of neural network control policies trained in simulation to the real-world robotic arm.

Compiler System for Graph Optimization and Op-Fusion for Deep Learning

- Engineered a compiler system targeting mobile backends with heterogeneous computer architecture.
- Implemented a kernel mapping strategy leveraging **TVM** and **TASO**, enhancing performance.
- Designed **high performant SIMD kernels** for **Edge Devices (ARM NEON)** for **Computer Vision algorithms**.
- Developed **Custom CUDA kernels** for **Jetson Orin**, achieving speedups benchmarked against OpenCV C++ for **ARM Devices**.