

SUMMARY

AI/ML enthusiast and system architect driven by a strong passion for advancing artificial intelligence through both professional and personal projects. I focus on designing, implementing & deploying large language models (LLMs) to tackle complex challenges, expanding the potential of AI applications in real-world settings. Outside of my professional work, I devote considerable time to studying and experimenting with open-source machine learning models, building them from the ground up. My continuous exploration of AI-driven solutions inspires me to innovate and develop impactful, scalable systems.

EDUCATION

MS, CS, Arizona State University | GPA 3.92 *August 2017 - December 2019*
BTech, CS, Maulana Abul Kalam Azad University of Technology | GPA 8.67/10 *July 2012 - July 2016*

TECHNICAL SKILLS & COURSEWORK

Languages: Python, GOLANG, C++, SQL, Bootstrap
ML Technologies: Pytorch, Huggingface, vLLM, Numpy, TensorFlow, OpenCV, Pandas, Scikit-learn, Matplotlib
Cloud DevOps: Docker, Kubernetes, Helm, AWS
Databases: MySQL, DynamoDB, OCI NoSQL, PostgreSQL

WORK EXPERIENCE

AI/ML Lead, Full-Stack Backend Developer | A10Networks - Incedo | Pune *June 2023*

- Leading the R&D team at A10Networks to drive the company's AI vision by creating innovative solutions for building and deploying LLM guardrails to safeguard applications against Cyber attacks.
- Designing a containerized solution for seamless deployment of our guardrails on any on-prem Kubernetes cluster using Helm.
- Designed and implemented a highly scalable system capable of retrieving up to 10K attacked IP addresses from a data warehouse containing 65M entries in as little as 13 secs. The system allows users to apply filters based on attack categories such as botnets, reflectors, and command and control (C2) servers, allowing for the targeted retrieval of relevant data.
- Implemented an enhanced rate-limiting solution in Golang to efficiently manage incoming requests. By processing a certain number of requests and queuing the remaining ones, the solution resulted in significant improvements over the previous Flask implementation, including a 45% reduction in CPU usage, a 38% reduction in memory usage, and a 70% improvement in response time for searching attack entries based on a specific IP address.

Technology Stack: Python, GOLANG, Flask, FastAPI, PyTorch, vLLM, DynamoDB, Docker, Helm, Kubernetes

Research Engineer, Lead | DiDi Labs | California *Nov 2021 - April 2023*

- Architected & Implemented a Graph-Based Neural Network to forecast the heading direction of pedestrians within a scene to aid our car's decision-making process.
- Boosted the efficiency of the model from 65% to 72%.
- Optimize and deploy models using techniques such as model compression, pruning, and quantization to improve performance and reduce computational cost.
- Constructed different metrics to measure the accuracy of our algorithms.

Technology Stack: C++, Python, PyTorch

Backend Software Developer | AmazonGO | Washington *Feb 2020 - Sept 2021*

- Designed Java APIs based on SQS to receive requests from clients and utilize vision-based algorithms for predicting the ultimate shopping events.
- Collaborate with data scientists and software engineers to take machine learning models from research and development to production.

Technology Stack: Python, Java, Numpy, Scikit-learn, OpenCV, SQS, Guice

AI Engineer Intern | Invitae | Massachusetts *May 2019 - August 2019*

- Developed and deployed an ML-based PHI scrubbing model using an API, reducing the risk of unauthorized access to sensitive healthcare data by removing PHI elements while retaining data utility for analysis and research.
- Designed and built an end-to-end object detection [pipeline](#) for scanned images using OpenCV, OCR, and ML models.
- Performed various data augmentation methods to generate a large custom dataset.

Technology stack: Python, Numpy, Pandas, NLTK, Spacy, OpenCV, Sklearn, Flask, JS