Nilesh Nayan

983-405-1097 | nilesh.nayan42@gmail.com | in/nilesh-nayan-b01ba1162 | nnn007 | Portfolio

Education

University of Massachusetts, Amherst — **Aug 2024 - Present**
Master of Science in Computer Science (MSCS) — *GPA: -*

Birla Institute of Technology and Science (BITS) Pilani, India — **Aug 2017 - May 2021**
Bachelor of Engineering in Electrical and Electronics (EEE) — *GPA: 8.74/10*

Technical Skills

- **Skills**: Data Structures and Algorithms, OOP, Design Patterns, Software System Design, Cloud Computing, Architecture, Data Science, AI, Statistics, Machine Learning, Deep Learning, Database Systems
- **Tools and Frameworks**: Python, C++, Java, MATLAB, TensorFlow, PyTorch, PySpark, Ray, FastAPI, Apache Airflow, AWS, Postgres, MongoDB, Redis, Kubernetes, Docker
- **Courses**: Computer Programming, Microprocessor & Interfacing, Operating Systems, Foundations of Data Science, Optimization, Machine Learning, Neural Network & Fuzzy Logic, Information Retrieval, Probability & Statistics, Applied Statistical Methods
- **MOOCs**: deeplearning.ai specializations, O'Reilly's ML System Design and MLOps

Work Experience

ML Engineer 3, Applied AI | Comcast — **Jul 2021 - Aug 2024**

- Designed and developed AI for Operations project backend in microservices-based architecture hosted on AWS using Kubernetes with Airflow DAGs and real-time inference using state-of-the-art (SOTA) time-series models like Anomaly Transformers, TimesNet, N-HiTS, N-BEATS and Prophet.
- Developed auto-scalable APIs using FastAPI and handled production server deployment with a scale of over 1 million calls per day at an optimal point considering performance and cost.
- Built the event-driven architecture for proactive anomaly alerting and system dependency graph-based Root Cause Analysis (RCA) using Dynamic Time Warping (DTW) distance.
- Led the R&D of the log data mining pipeline using the Drain3 algorithm integrated with features like log trend anomaly alerting, log RCA and Q&A triage bot using Large Language Models (LLMs).

Data Science Research Intern | Jupiter — **Jul 2020 - Dec 2020**

- Designed data structures for Redis cache in Voice Annotation Platform, crucial for training Speech Models on Indian accents. Developed a generic Singleton Class module, for broader caching usage.
- Performed the research work on the mobile SMS NER project using the Flair model and improved on the existing regex-based approach by ∼20% and further reduced inference latency by 75% from 1.2 seconds to 300ms per text using embedding optimization and quantization.
- Worked on advanced SQL and Airflow using Spark engine on Big Data for early data drifts.

Projects

LLM based Talent Acquisition (TA) helper framework and bot | *Git* | Demo — **Mar 2022 - May 2022**

- Implementation of LLM Agent using LangChain and RAG-based retrieval system utilizing GPT-3.5 for the text data present in the Job Description and Candidate's profile.
- Used FAISS (FaceBook AI Similarity Score) on text embeddings to provide the relevant results based on queries to help the TA team optimize their normal filtering process, speeding up around 50 times.

Smart Advertising based on Hyperlocal Factors using IoT and CV | *Git* | Demo — **Jan 2020 - May 2021**

- Created a system focused on users for targeted advertising, utilizing IoT sensors, VGGNet-19 for extracting features, AWS for processing data and the XGBoost algorithm for precise advertisement predictions based on extracted metadata to recommend the most relevant ads from dataset.
- Improved the accuracy of baseline ad predictions by ∼22% by incorporating collaborative filtering

Text Summarization - NLP | *Git* | Demo — **Oct 2019 - Nov 2019**

- Designed a system that generated text captions for the given input image replicating a SOTA paper.
- Used ResNet for Computer Vision application to extract the most relevant information from images with an attention mechanism and fed to LSTM for the generation of text captions. Achieved a BLEU-4 score of 29.1% on the Microsoft COCO benchmark.

Publications

- Praveen Manoharan, **Nilesh Nayan**, Aaditya Sharma, Aravindakumar Venugopalan. "Building a scalable real-time ML inference platform for AIOps". In Proceedings of 2023 Machine Learning Developers Summit (MLDS), Bengaluru, India. *ADaSci*
- Hongcheng Wang, Praveen Manoharan, **Nilesh Nayan**, Aravindakumar Venugopalan, Abhijeet Mulye, Tianwen Chen, Mateja Putic. "AI for IT Operations (AIOps) - Using AI/ML for Improving IT Operations". In Proceedings of Society of Cable Telecommunications Engineers (SCTE) 2022 Fall Technical Forum, Philadelphia, United States of America. *NCTA*