Rohan Vitrouthu

AI, ML, Data Engineer

Baltimore, MD, United States | +1 443 338-3378 | rohanmdus172@gmail.com | GitHub

AI/ML Engineer with experience in developing deep learning models, building scalable ML pipelines, and deploying AI solutions using TensorFlow, PyTorch, and NVIDIA GPUs. Skilled in data pipelines (Spark, Kafka, Airflow), model optimization (pruning, quantization, distributed training), and MLOps on cloud platforms (AWS, GCP, Azure). Passionate about applying AI to solve real-world problems in computer vision, NLP, and generative AI.

Experience

AI / ML Engineer | Genentech | USA | Sept 2023 - Present

- Built and deployed **AI agents** using **LangChain, Hugging Face Transformers, and PyTorch**, containerized with **Docker/Kubernetes**, to automate knowledge retrieval, workflow orchestration, and decision-making in research and enterprise applications.
- Designed and implemented **RAG pipelines** with **Haystack, FAISS, and ElasticSearch**, integrating domain-specific biomedical data sources to improve contextual accuracy and knowledge grounding.
- Fine-tuned **large language models (LLMs)** using **Hugging Face Transformers, PyTorch Lightning, and scikit-learn**, applying LoRA/PEFT on proprietary datasets to enhance scientific literature analysis and clinical insights extraction.
- Optimized inference pipelines with **PEFT, DeepSpeed, and Hugging Face Accelerate**, deploying via **NVIDIA Triton Inference Server** and **CUDA/cuDNN** to enable scalable and cost-effective production inference.
- Researched and integrated state-of-the-art **generative AI and multi-agent frameworks** (e.g., **AutoGen, LangChain Agents**) on **NVIDIA GPUs**, delivering proof-of-concept solutions.

Data Engineer | Neysa | Mumbai, IN | June 2020 – July 2022

- Designed and implemented **data pipelines** for AI/ML workflows using **Apache Spark (PySpark), Kafka Streams, and Apache Airflow**, ensuring high-throughput and low-latency data ingestion.
- Built **ETL/ELT pipelines** with **Python (Pandas, NumPy, OpenCV, spaCy)**, **Informatica IICS**, and **Azure Data Factory** to preprocess structured and unstructured datasets (images, text, sensor data) for ML model training.
- Developed **feature engineering pipelines** with **dbt, PySpark, scikit-learn**, and managed **feature stores (Feast)** to enable consistent and reusable ML features across teams.
- Automated **model retraining pipelines** using **MLflow, GitHub Actions, Jenkins, Docker, and Kubernetes**, enabling CI/CD for evolving datasets in production.

Education

- Master of Science, Data Science, University of Maryland, Baltimore County, Baltimore, MD | CGPA: 3.74/4.0
- Bachelor of Technology, Information Technology, Jawaharlal Nehru Technological University, Hyderabad, India | CGPA: 8.81/10

Certifications

- Microsoft Azure AZ-900: Azure Fundamentals
- NVIDIA Deep Learning Institute: Accelerating End-to-End Data Science Workflows
- Databricks Lakehouse Fundamentals
- Informatica Cloud Data Integration Developer

Skills

Machine Learning & Deep Learning

- **Frameworks & Libraries:** PyTorch, TensorFlow, Hugging Face Transformers, scikit-learn, OpenCV, NLTK, spaCy
- **Generative AI & Agents:** LangChain, AutoGen, Haystack, FAISS, ElasticSearch, Chroma
- **Model Optimization:** LoRA, PEFT, Quantization, Pruning, Distributed Training (**DeepSpeed, Hugging Face Accelerate**)
- **Experiment Tracking:** MLflow, Weights & Biases

MLOps & Data Pipelines

- **Workflow Orchestration:** Apache Airflow, Docker, Kubernetes, GitHub Actions, Jenkins, Azure DevOps
- **Big Data & Streaming:** Apache Spark (PySpark), Kafka Streams
- **ETL/ELT & Feature Stores:** Informatica IICS, Azure Data Factory, dbt, Feast

Cloud Platforms & Deployment

- **AWS:** SageMaker, S3, EC2, Lambda
- **Azure:** Data Factory, Synapse, DevOps
- **GCP** (general services)
- **Inference & Acceleration:** NVIDIA CUDA/cuDNN, Triton Inference Server, Jetson Orin Nano

Databases & Querying

- Oracle, PostgreSQL, MySQL, MongoDB, T-SQL

Core Tools & Version Control

- Git, GitHub, Linux/Windows environments

Projects

- **Generative AI Chatbot**
  - Developed and fine-tuned an **LLM (Hugging Face Transformers)** for domain-specific Q&A, integrating it with a **retrieval-augmented generation (RAG)** pipeline for accurate responses.
- **Music Recommendations based on Human Emotions**
  - Employed SVM, Random Forest, and CNN algorithms to predict emotions from facial images, achieving 85% accuracy in emo- tion detection.
- **Sectoral Stock Analysis of NIFTY-50 Stocks**
  - Built and trained an RNN model using LSTM layers to forecast closing prices of automobile stocks in the NIFTY 50 index where I utilized historical stock data from January 1, 2015, to April 30, 2021, achieving an RMSE of 889.40 and MAE of 601.69 on the test dataset.