Viraj Singh

+91-8433426632
f20171099h@alumni.bits-pilani.ac.in
https://linkedin.com/in/virajsingh3782ba1b4
https://github.com/viraj-singh1998

EDUCATION

Birla Institute of Technology and Science (BITS) Pilani, Hyderabad Campus
B.E. Computer Science + M.Sc. Economics, August 2017 - July 2022 CGPA: 8.09

EXPERIENCE

Fidelity National Financial India (FNFI), Bangalore
Data Scientist, June 2023 - Present
Overhauled the document intelligence pipeline holistically, improved the accuracy and reduced operational costs for an LLM-enabled service that processes over 1.5 million multi-page PDF documents per month.

- Introduced the use of LLMs for entity extraction, which led to an increment of ≈20% in average extraction accuracy and ≈15% in both precision and recall; leveraged PEFT (Parameter Efficient Fine Tuning) methods such as QLoRA for memory efficient supervised fine-tuning.
- Optimized text generation throughput by batching inference requests and using Paged Attention (vLLM), increasing the effective generated tokens/sec by 8x at roughly 66% of the compute costs.
- Built and deployed a pipeline consisting of a mixture of fine-tuned open source (such as the instructions tuned versions of Mistral and LLama on in-house data) and closed source (OpenAI GPT-4o, Claude Sonnet 3.5) LLMs hosted on Amazon Bedrock and invoked by Lambda functions to asynchronously carry out OCR, inference and ingestion of model responses into DynamoDB.
- Replaced text-only input based GRUs carrying out entity extraction and document classification with fine-tuned multimodal (text + image input) models (LayoutLM).

Lenscorp AI, Gurugram
Computer Vision Research Engineer, June 2022 - Jan 2023
Worked on one of the world's fastest and most accurate fingerprint matching algorithms, setting new benchmarks in matching accuracy and time.

- Improved **T**rue**P**ositive**R**ate@**F**alse**P**ositive**R**ate (for FPRs up to 10^{-3}, 10^{-5} and 10^{-7}) by implementation of extensive literature research and experimentation with a variety of convolutional neural networks (CNN), vision transformers (ViT) and autoencoder architectures.
- Refined base neural network models with channel & spatial attention mechanisms, re-scoring algorithms and auxiliary loss functions; implemented image preprocessing techniques and mechanisms to make the network invariant to scale, translation and rotation.
- Integrated a new static deployment framework with Intel OpenVINO, for packaging the model and inference code as an SDK which sped up the matching by over 350%.
- Optimized the data loading and preprocessing pipeline by over 50% using concurrent programming.

Clari Copilot (FKA Wingman), Bangalore
Machine Learning Intern, July 2021 - June 2022
Integrated a topic tagging feature in a call notetaker application which exhibited high user engagement. Used machine learning models to produce actionable insights for users.

- Built and deployed an end-to-end topic model as a containerised microservice using Python, MongoDB, ElasticSearch that assigns topics to turns of call transcripts by clustering their context-rich vector representations generated by sentence transformer neural networks. Boosted user engagement of 'Topics' section by over 20%.
- Developed the pipeline to query text conversations from a NoSQL database, preprocess and perform inference asynchronously using an AWS SageMaker endpoint.
- Designed a "Next Steps" intent classification model by fine-tuning a transformer encoder to categorize utterances within sales calls as cues for a follow up, with an F1 score of ≈95%.

QUALIFICATIONS & SKILLS

- **Interests:** NLP - NLU (Natural Language Understanding) & NLG (Natural Language Generation), Large Language Models (LLMs), AI Agents, Recommender Systems, Retrieval Augmented Generation(RAG), Computer Vision, MLOps, Distributed Systems, Machine Learning, Software Development, System Design
- **Software and frameworks:** PyTorch, Tensorflow, HugginFace (transformers), LangChain, LangGraph, AWS Sagemaker, AWS Lambda, AWS Bedrock, AWS DynamoDB, Azure Cloud, Flask, FastAPI, Scikit-learn, Numpy, Pandas, Docker, Apache Spark, MongoDB, ONNX, Openvino, Git, MLFlow, ElasticSearch
- **Programming Languages:** Python, C, C++