Anish Kumar Maurya

M.Tech: Cyber Physical System (IIT Jodhpur)

anishkumarmaurya12@gmail.com | github.com/Ansih12kumar | linkedin.com/in/anishkmaurya | +91-9554593062

Profile Summary

- ML Engineer with 1.6+ years of experience and an M.Tech in Cyber-Physical Systems (IIT Jodhpur). Specialized in LLM fine-tuning, quantization, and OpenVINO/GGUF optimization for high-performance CPU deployment. Designed scalable FastAPI–Celery–Docker ML pipelines for legal/medical NLP, OCR, and Document AI, backed by expertise in Python, PyTorch, Hugging Face, Elasticsearch, and Dgraph.

Education

M.Tech. (CPS) Indian Institute of Technology (IIT), Jodhpur (Jun 2022–Jun 2024), CGPA: 7.72
B.Tech. (EE) Institute of Engineering and Rural Technology(IERT) (Aug 2013– July 2017), CGPA: 7.5%

Experience

•Datafoundry Private Ltd *June 2024 – Present*
ML Engineer Bengaluru, On-site

- Designed and deployed a fully Dockerized, event-driven ML pipeline using FastAPI, Celery, and LLMs to perform NER on 40K+ Supreme Court legal judgment HTML documents; asynchronously extracted entities, indexed outputs in Elasticsearch, and auto-synchronized new entries to Dgraph by generating optimized knowledge graphs achieving query latency under 2 seconds.
- Curated and preprocessed medical NER datasets for supervised fine-tuning of Gemma2-2B-IT, Gemma3 (1B, 3B), and Qwen3 (0.6B, 1.7B, 4B) models using parameter-efficient LoRA with 4-bit quantization via BitsAndBytes. Converted the optimized models to OpenVINO IR and GGUF formats, achieving up to 33% improvement in inference performance and enabling high-efficiency, CPU-optimized deployment for clinical NLP applications.
- Built multilingual NLP pipelines for bidirectional Hindi-English translation, NER, and Q&A on long scanned PDFs and HTML using EasyOCR for Hindi text extraction and GGUF-quantized models for efficient inference.
- Quantized LLMs across vision, audio, and text modalities using OpenVINO Toolkit, ONNX, and GGUF formats to optimize memory usage and enable faster, GPU-free inference on CPUs; generated OpenVINO-compatible tokenizers and detokenizers for efficient text processing.
- Built an OCR pipeline for scanned PDFs integrating Tesseract, EasyOCR, and YOLO for checkbox detection and structured text extraction; executed data annotation and fine-tuned models to enhance downstream computer vision model accuracy in document analysis tasks.
- Fine-tuned BERT for sentiment analysis on Prime Minister's speeches, achieving 95.8% precision and 96.5% recall, outperforming Random Forest and Logistic Regression. **GitHub**

•Indian Institute of Technology Jodhpur *June 2022 – May 2024*
Teaching Assistant Jodhpur, India

- Teaching Assistant, Data Science & ML Lab: Supported labs.
- Teaching Assistant, Embedded Systems & Computer Vision Lab: Assisted in image processing.

Projects

•Automatic Image Classification for Surveillance based CPS Applications *Computer Vision*
M.Tech Project

– Developed real-time image detection model using Python, OpenCV, and VisDrone dataset, applying deep learning for effective training and validation.Deployed deep learning model on Raspberry Pi with Google Coral TPU, achieving 28 FPS, showcasing edge computing for advanced surveillance.

Technical Skills

- **Programming:** Python , C/C++,HTML/CSS, JavaScript, SQL
- **Tools & OS:** Git, Linux, Windows, Visual Studio Code
- **Frameworks:** FastAPI, Celery, LangChain, LangGraph, Streamlit, TensorFlow, Keras, PyTorch, Rasa
- **Libraries:** Pandas, NumPy, Scikit-learn, Hugging Face,Ollama, FAISS, Chroma, Seaborn, Matplotlib
- **Cloud & DevOps:** Microsoft Azure (Functions, Container Instances), AWS (LLM Deployment, EC2, S3), Docker, Linux, CI/CD

Achievements

•GATE (EE) 2022 Cracked GATE with 97.48 percentile (AIR-1756)

•GATE (IN) 2022 Cracked GATE with 97 percentile (AIR-601)