Rohan Ippalapally

irohan90839@gmail.com · (469) 592-2472 · Linkedin · GitHub

Profile

Results-driven Machine Learning Engineer with **4 years of experience**, specializing in **Computer Vision**, **NLP**, **Model Optimization**, **MLOps**, and **LLMs**. Seeking challenging opportunities across **Machine Learning & AI** to leverage expertise in developing and deploying **production-grade** AI solutions that drive business value.

Work Experience

Machine Learning Engineer - Cyber Dive Corp., Mesa, AZ May 2022 – Jan 2026

Cyber Dive is a technology company specializing in the development of the Aqua One smartphone, designed to help parents monitor and manage their children's online activities in real time.

- Led the development of an **on-device content moderation system** utilizing **YOLO, SSD-MobileNet, and EfficientDet**, achieving **94% precision**; accelerated training with multi-GPU (**4x NVIDIA**) parallelization.
- Enhanced model **detection accuracy by 20%** and **reduced false positives by 50%** using data augmentation, fine-tuning, and transfer learning strategies.
- Improved the detection system's **inference speed by 20%** and reduced **model size by 50%** through quantization and pruning for real-time mobile execution.
- Designed and deployed an **end-to-end MLOps sampling pipeline** on AWS using **ECS, DynamoDB Streams, and Lambda triggers** to automatically process customer sessions, surface low-confidence detections, and improve production model performance.
- Collaborated with Android engineers to integrate ML pre-processing and post-processing pipelines, improving **APK response time by 25%** and **inference accuracy by 50%**.
- Developed an **on-device NLP model** leveraging **BERT, MobileBERT, and DistilBERT** to detect online grooming behaviors, achieving **92% accuracy**.
- Optimized tokenization and pre-processing pipelines, reducing **inference latency by 30%** and **memory usage by 40%** using **TensorFlow-Lite** for edge devices.
- Designed a feature to detect Suspicious Contacts using **AWS Step Functions, Lambda, and the Twilio API** for real-time verification and alerts, achieving **30% faster response time**.
- Assisted in developing an **OCR text recognition system** using **CRAFT** and **PARSeq** text models, achieving **85% accuracy** and improving **processing speed by 40%** through efficient frame selection and model optimizations.
- Refined **video classification** models using a **CNN-RNN hybrid** architecture for content moderation, improving **accuracy by 80%**.

Skills

Programming Languages: Python, Kotlin, JavaScript, SQL, TypeScript, R, C

ML Frameworks: Scikit-Learn, TensorFlow, PyTorch, ONNX, OpenCV, PIL, NLTK, Keras

Model Optimization & Deployment: Model Quantization, Pruning, TensorFlow-Lite, ML-Kit, MediaPipe, PyTorch Mobile

LLMs & Generative AI: AI Agents, RAG, Prompt Engineering, Hugging Face Transformers, LangChain, FAISS, Agentic Workflows, Multi-Agent Systems (Google ADK)

Cloud & Deployment Tools: AWS (Rekognition, SageMaker, Bedrock, Lambdas, DynamoDB (NoSQL), AppSync, Kinesis, ECS, SQS, CloudWatch), Docker, FastAPI, GCP, Elasticsearch

MLOps & Monitoring: CometML, Databricks, Deepchecks, Roboflow, Neptune, MLflow, Datadog

Developer Tools: Git, Jira, Postman, GraphQL, Android Studio, VS Code, RStudio, CI/CD Pipeline

Publications & Patents

Patent: US-20260019493 - *USPTO Patent Application* Jan 2026

Object Detection using Thermal Imaging - *IEEE Publication* Dec 2020

Education

Master of Science in Computer Engineering (Computer Systems) July 2021 - May 2023
The University of Texas at Dallas, Richardson, TX

Bachelor of Technology in Electronics and Communication July 2016 - Aug 2020
Amrita School of Engineering, Bengaluru, India

Certifications

Mathematics for Machine Learning - *Coursera, Imperial College London* Oct 2020

Deep Learning Specialization - *Coursera, DeepLearning.ai* Aug 2020

SQL for Data Science - *Coursera, UC Davis* May 2020