Aravind Pradeep

Entry-Level ML Engineer | Edge ViT Systems & LLMs

Cottbus
03046
Germany
+49 176 67314504
aravindpradeep001@gmail.com
linkedin.com/in/aravind-pradeepmadathinal
github.com/aravindpradeep

Summary

Entry-level **Machine Learning Engineer** specializing in **Content-Aware Vision Transformer (ViT) Systems** and deployment of ML models on edge and cloud infrastructure. Combines an M.Sc. in Artificial Intelligence with multiple years of coding experience, production work at Perinet GmbH, and a solid foundation in **full-stack development** (React, TypeScript) from bachelor's studies. Experienced in building end-to-end ML systems, RAG pipelines, and containerized services using Python, PyTorch, Docker/Kubernetes, and Golang APIs.

Education

Oct 2023 – Est. Dec 2025
M.Sc. Artificial Intelligence (Research Profile), *Brandenburg University of Technology*, Cottbus, Germany, GPA: 2.6 (92 ECTS)
- **Thesis:** "Content-Aware Vision Transformer Systems on Edge Devices" – Optimizing ViT architectures for real-time inference on resource-constrained hardware.
- Key modules: Image Processing & Computer Vision, Neuromorphic Computing, Research Module AI, Explainable ML, Data Mining.

Jul 2018 – Mar 2021
B.Sc. Computer Application, *BVM Holy Cross College*, Kottayam, India
Core coursework in software engineering and web development, including projects using **React** and **TypeScript**.

Experience

Jun 2024 – Present
Working Student – AI Firmware & Edge ML, *Perinet GmbH*, Cottbus, Germany
- Co-developed **RAG-based conversational AI** using vector embeddings and semantic search, integrated with real-time sensor streams (MQTT) for IoT/automotive scenarios.
- Designed **anomaly detection systems** on distributed edge nodes using optimized ML models, enabling privacy-preserving real-time inference on device data.
- Containerized (Docker/LXC) and orchestrated **ViT/ML services** with Kubernetes for ARM-based edge devices, achieving production-grade stability.
- Built scalable **Golang** backend APIs for AI service integration and contributed to CI/CD pipelines (GitLab) for automated build, test, and deployment of ML components.
- Optimized neural networks, including transformer-based models, for low-latency inference on ARM hardware via quantization and efficient data pipelines.

Oct 2021 – Aug 2022
Software Engineer Trainee, *Cognizant Technology Solutions*, India
- Worked in a large-scale enterprise environment, collaborating in cross-functional agile teams on production systems.
- Gained experience with legacy systems and databases, strengthening debugging, reliability, and software engineering fundamentals.

Key Projects

Content-Aware ViT Systems on Edge Devices, *M.Sc. Thesis*
- Researching **Vision Transformer (ViT)** optimization for content-aware processing on resource-constrained edge hardware.
- Implementing model compression techniques (pruning, quantization, distillation) to enable real-time ViT inference with minimal latency/memory usage.
- Developing an evaluation framework for ViT performance across edge scenarios, focusing on accuracy–latency trade-offs and deployment robustness.

AI Agent & RAG Pipeline for QA, *Perinet*

- Engineered a lightweight **AI agent** with a RAG pipeline using **Gemma-2B** and **Phi-2** via **LangChain** and Hugging Face.
- Implemented a **Golang** backend to orchestrate retrieval, model calls, and chain-of-thought style workflows; deployed via containerized LXC for IoT stability.

ML Anomaly Detection & Error Analysis, *University Research Project*

- Built ML models for **real-time anomaly detection** and performed systematic error analysis on misclassifications.
- Proposed data preprocessing and architecture changes that improved model robustness on validation data.

Full-Stack Web App (React / TypeScript), *Personal Project*

- Developed a small full-stack application with a **React + TypeScript** frontend and RESTful backend, focusing on clean UI, component design, and robust API integration.

Technical Skills

AI / ML	Vision Transformers (ViT), Edge AI, Content-Aware Systems, Supervised Learning, Deep Learning, Computer Vision, RAG, LLMs, Anomaly Detection, Model Compression (Pruning/Quantization).
Frameworks	PyTorch, TensorFlow/TensorFlow Lite, Hugging Face Transformers, LangChain, LlamaIndex, scikit-learn, OpenCV.
Programming	Python (Advanced), Golang, SQL, C++, C.
Frontend	React, TypeScript, JavaScript, HTML, CSS.
MLOps / Systems	Docker, Kubernetes, LXC, MQTT, GitLab CI/CD, REST/gRPC APIs, ARM optimization.
Edge / Cloud	Azure, Databricks, MLflow, PyTest, Git/GitHub, Edge hardware (ARM/Coral TPU).

Languages

English	Fluent (C1) – Technical communication, documentation, and presentations.
German	B1 – Actively improving for German tech workplaces.
Malayalam	Native.