Srijay KOLVEKAR

Machine Learning Engineer

linkedin.com/in/srijayjk github.com/srijayjk
+49 1764 3255 739 srijayjk@gmail.com
Munich, Germany

An enthusiastic Machine Learning Engineer passionate about solving engineering problems. Expertise includes object detection, tracking, classification, and Embedded AI using 2D and 3D data. possesses a solid understanding of the fundamentals of machine learning, embedded AI, signal processing, and adheres to clean programming principles to ensure the development of robust, efficient and production-grade solutions.

Skills

Programming	**C**, **C++**, **Python**, Matlab
ML Frameworks	**Pytorch**, Sklearn, Tensorflow, Keras
ML Operations	**ML Flow**, **Hydra**, Docker, Data Version Control
ML Deployment	**FASTAPI**, **FLASK**, **Google Cloud Platform**, Gunicorn
AI Embedded	**model pruning**, **quantization**, **RTOS**, **QNX**, **snapdragon q-drive**
LLM	**Langchain**, **Chromadb**, **Agents**, **RAG**, **TinyLlama-1B**, **ConversationBufferMemory**

Experience

current – Feb 2023 | **Machine Learning Engineer, MAGNA ELECTRONICS, Munich**

- Engineered advanced object detection and classification models utilizing state-of-the-art deep learning architectures.
- Spearheaded the design and implementation of an end-to-end machine learning framework for radar object classification and environment detection, managing the entire lifecycle from data ingestion to SoC deployment.
- Successfully deployed optimized deep learning models on embedded System-on-Chips (SoCs) TI C66 and Arm R5. Specifically, achieved a 20% reduction in model size and a 10% improvement in runtime performance on TI SoCs with small optimized models with Knowledge distillation
- Deployed larger, complex models on TDA4 and Snapdragon Ride platforms within a QNX real-time operating system
- Conducted research in machine learning for radar signal processing, focusing on low-level complex ADC data for ADAS applications

ML | DL | Radar | Lidar | Signal Processing | Embedded AI | Deployment | Quantization | Data Analysis | Deep Learning Accelerator

Nov 2022 – Sept 2021 | **Machine Learning (Internship and Work Student) , PHENO-INSPECT GMBH, Stuttgart**

- Developed and deployed advanced computer vision models, including semantic and instance segmentation, object, and keypoint detection, for phenotyping solutions such as weed and plant detection.
- Designed and implemented robust data processing and parsing pipelines utilizing GraphQL queries to optimize data flow for deep learning models.
- Consistently demonstrated strong teamwork and a proactive approach to project responsibilities

Semantic Segmentation | Key-Point Detection | ML-Ops | Object Detection | Computer Vision

Education

Oct 2022 – Oct 2019 | **MSc Electrical Engineering, UNIVERSITY OF STUTTGART, Stuttgart**

- Advanced Mathematics, Detection and Pattern Recognition, Deep Learning, Machine Learning, Computer Vision
- Signal Processing, Communication Engineering

ML | Software Engineering | Software Architecture Design | Presentation | Communication Skills

Language

English (c1) ●●●●●
German (B1) ●●●○○

LLM Agent Development, Fine Tuning and Quantization (LoRA) 2025

GitHub Link

Developed multiple AI agents, including a data analysis agent and a weather application agent. Focused on prompt engineering and creating custom tools for Large Language Models (LLMs) using frameworks like LangChain. Implemented Retrieval Augmented Generation (RAG) with ChromaDB for efficient information retrieval and utilized AgentExecutor for orchestrating agent actions. Employed RecursiveCharacterTextSplitter for robust text processing and integrated TinyLlama (1B parameters) as the underlying language model.

LLM | AI Agents | LangChain | RAG | ChromaDB | Prompt Engineering | TinyLlama

Nano GPT 2024

GitHub Link

Implemented nanoGPT to explore GPT fundamentals and attention mechanisms. Trained the model on a large dataset, gaining practical experience with language model development.

GPT | LLM | Generative Model

AI-Assisted Semi-Automatic Labeling of Eye Contact Data Using Contrastive Learning 2022

Thesis

Created a semi-automatic labeling method for eye contact data using contrastive learning and active learning. Enhanced feature space embeddings by distinguishing similar and dissimilar eye contact clusters.

Contrastive Learning | Pose Estimation | Key-Point Detection | Active Learning

Interpretation of Activation Maps in Generative Modeling 2021

GitHub Link Paper

Developed an XAI method using Grad-CAM to interpret activation maps in VAEs and CVAEs. This revealed latent space information, enhancing understanding of generative model decisions. Improved model transparency by visualizing critical feature influences.

VAE | CVAE | Grad-CAM

Deep Learning Based Decomposition of High-Density Surface EMG Signal Using GRU Neural Network 2020

GitHub Link Presentation

Developed a GRU-based deep learning model to decompose high-density EMG signals, predicting individual motor unit action potentials.

EMG | Time Series | GRU | LSTM