

Raman Shinde

raman.shinde15@gmail.com | +91 9595161238 | [LinkedIn](#) | [GitHub](#)

EXPERIENCE

- Qualcomm** Bangalore, India
Senior Lead Engineer [AI Research] Mar. 2024 - Present
 - * Implemented texture support support for OpenCL and Vulkan backends in TVM, optimizing data access patterns on Adreno GPUs to achieve a **50%** reduction in inference time.
 - * Extended TVM by adding support for new operators, layers and runtime features, enabling deployment of LLMs and Diffusion models on edge hardware.
 - * Integrated CLML acceleration into the GGML and llama.cpp frameworks, bypassing generic compute paths to deliver hardware-optimized performance for on-device inference.
- Imagination Technologies** Pune, India
Deep Learning Engineer Sep. 2021 - Mar. 2024
 - * Developed an SDK enabling neural network execution on GPUs and NNAs using TVM, simplifying and accelerating model deployment on edge devices.
 - * Optimized RelayIR by developing custom graph transforms, improving execution efficiency across diverse workloads.
 - * Contributed to the development of quantization tools (static, dynamic, and QAT) for multiple frameworks, enhancing model compression and boosting inference speed.
- Xpanxion** Pune, India
Data Scientist Jan. 2020 - Sep. 2021
 - * Developed AI/ML solutions to extract information from medical documents, including document classification and candidate extraction (e.g., benefits, rates, and drug details).
 - * Built reusable AI/ML components within the Innovation Team, including a content-based and collaborative recommendation system, NLP modules such as Named Entity Recognition (NER), QA systems, and sequence translation.
 - * Led Computer Vision projects involving object localization and detection, image segmentation, and gesture recognition.
- Siemens R&D** Pune, India
Product Development Engineer Dec. 2018 - Dec 2019
 - * Developed and enhanced an application for designing manufacturing sequences, with significant contributions to debugging and resolving issues in existing code.
 - * Accelerated feature deployment by implementing new use-cases and Proof-of-Concepts (POCs), enabling faster shipping of key features.
- TCS** Pune, India
Software Developer Dec 2015 - Nov 2018
 - * Developed monitoring and control applications for NCRA, including real-time issue monitoring, debugging, and GUI modifications based on client requirements.
 - * Provided support for financial applications such as CRD and SRD for Morgan Stanley, ensuring reliable performance and issue resolution.

TECHNICAL SKILLS

- **Languages/Frameworks** : Python, C++, C, PyTorch, TensorFlow
- **Cloud/Deployment** : AWS, GCP, Docker, Kubernetes, Git

EDUCATION

B.Tech in Electronics and Telecommunication, SGGSIE&T, Nanded with **CGPA 7.7** (2011 - 2015)