Raman Shinde

https://github.com/Raman-Raje
Email : raman.shinde15@gmail.com
Mobile : +91 9595161238

Career Summary

- Senior Lead Engineer specializing in AI inference, compute, and performance optimization for Deep Learning and LLMs, with strong expertise in TVM, MLC-LLM, and Llama.cpp.
- Enhanced neural network performance by optimizing code generation, operator schedules, and runtimes specifically for edge and mobile devices.
- Expertise in developing Deep Learning architectures including Transformers, LLMs, CNNs, and LSTMs, bridging the gap between model design and hardware execution.
- Proficient in scalable ML deployments using AWS, GCP, Docker, and Kubernetes, backed by a strong foundation in classical ML algorithms (Random Forest, SVM, GBDT).

Experience

- **Qualcomm** Banglore, 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.
  - Extended TVM by adding support for new operators, layers and runtime features, enabling deployment of LLMs and Diffusion models on edge hardware.
  - Integrated Qualcomm 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.
  - Implemented LSTM/RNN support in the Neural Compute SDK for PyTorch, TensorFlow and ONNX, ensuring broader compatibility and accelerated performance.
  - 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) using libraries like Fonduer, Tesseract, and OpenCV.
  - 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.
  - ∗ Built applications for various clients using Python and C++, contributing to diverse projects and delivering tailored software solutions.

Technical Skills

- **Languages:** Python, C++
- **Database:** MySql, MongoDB
- **Data Analysis:** Pandas, Numpy, Matplotlib, Seaborn, openCV
- **ML/DL Toolkit:** Keras, scikit-learn , tensorflow, pytorch, TVM

Certifications/Internship

- Applied Machine Learning course at Applied AI. (Jan 2018 to May 2019)
- Completed Standford Statistical Learning (Self-Paced) course.
- Completed Deep Learning Specialization course from Coursera
- Internship at IARE, Aurangabad on Industrial automation. (May 2014 - Jun 2014)

Education

- **B.Tech** in Electronics and Telecommunication from SGGSIE&T, Nanded with **CGPA 7.7** (2011 - 2015)
- Class Xll (HSC), form Maharashtra State Board of Education with **83.33%** (2009 - 2011)
- Class X (SSC), form Maharashtra State Board of Education with **90.92%** (2008 - 2009)