Sunil Acharya

Phone: +91-7406893061 | Email: sun1lach@outlook.com
GitHub: github/sun1lach| Linkedin: linkedin/sun1lach | Portfolio: sun1lach.me

A **senior machine learning engineer** with 8+ years of experience in building end-to-end AI solutions, specializing in distributed LLM training, open-source contributions, retail edge AI, and performance optimization.

Experience

Intel Corporation June 2019 – Present
AI Software Solutions Engineer *Bengaluru, IN*

- **LLM Enablement & Distributed Training:** Enabled and scaled llama 3.1 (8B, 70B, 405B) LLM training on Intel Gaudi 2 & 3 accelerators with various distributed parallel training strategies; productionized distributed checkpoint conversion feature with Hugging Face compatibility to external customers.
- **Optimized Edge AI:** Designed and deployed Intel Automated Vision Checkout at retail sites in India, achieving low-latency (70ms) deep learning inference and scalability on Intel's integrated GPU, supporting 100+ daily transactions at the store front.
- **Federated AI:** Contributed to Linux Foundation's 'securefederatedai/openfl'' by introducing JAX/FLAX support, federation long and short lived component timeout feature, and interactive examples for secure, efficient federated learning on private medical records.
- **LLM Fine-Tuning as a Service:** Designed a ZenML/cnvrg.io-based fine-tuning framework for internal use as a premium offering.
- **Performance Profiling:** Experience in accelerators compute/memory profiling, device & host trace analysis to optimize distributed training workloads and application level optimization for edge inference.
- **Data Pipelines:** Built an event-driven, cost-optimized AWS data pipeline, reducing costs by 8x and publishing product data in 5 minutes (down from 6 hours), supporting 40M API hits monthly.
- **Outreach:** Presented demos and delivered technical talks on cloud cost optimization and product innovations at Intel India Innovation day, Intel ConnectiON and Intel India tech talk series (Cloud Community of Practice) events.

Western Digital Apr 2017 – May 2019
Software Engineer *Bengaluru, IN*

- Led the development of an ML-based storage validation framework using a random forest classifier for predicting microSD card failures during test bench execution across 300+ Android devices in parallel, resulting in a 40% reduction in validation time and early detection of 80% of potential card failures.
- Developed a cross-platform distributed client server based performance studio application for performance profiling of SD/uSD, USB, and enterprise SSDs.

Bharat Sanchar Nigam Limited (BSNL) Jan 2015 – Dec 2015
Engineering Trainee/Intern *Bengaluru, IN*

- Operational testing of wireless equipment. Configured routers, modems, and mainline distribution frames. Examined optical fiber functioning at various stages.
- A hands-on approach to fusion splicing of optical fiber and monitoring various signal parameters using OTDR devices.

Education

International Institute of Information Technology Aug 2023 - Sept 2023
7th Summer School on Machine Learning *Hyderabad, IN*

Liverpool John Moores University Nov 2020 - Aug 2022
Master of Science in Machine Learning & Artificial Intelligence (76%) *Liverpool, UK*

- **Research Thesis:** Evaluate the effectiveness of various knowledge distillation methods on smaller neural network architectures.

International Institute of Information Technology Nov 2020 - Nov 2021
Post graduate diploma in Machine Learning & Artificial Intelligence (GPA: 3.60 / 4.00) *Bengaluru, IN*

Visvesvaraya Technological University Aug 2012 - May 2016
Bachelor of Engineering (BE), Electronics & Communication Engineering (74.32%) *Bengaluru, IN*

Projects & Certifications

Certification on "Accelerators for Deep Learning" - IIT Roorkee | *certificate link*

- Executive certificate on Accelerators for Deep Learning covering deep learning algorithms and computer architecture with an emphasis on AI acceleration on various computing systems, such as FPGAs, mobile/desktop GPUs, smartphones, ASICs, DSPs and CPUs.

Custom & Efficient CNN architectures from Scratch | *project link*

- Designed and Implemented 25k, 143k, 340k, 600k & 1M parameter efficient custom CNN architectures.
- Upon knowledge distillation on these custom CNN architectures, model accuracy surpasses ResNet-18/34/152 baselines with 10-20x less model parameters on FMNIST & CIFAR-10.

Intel AI Everywhere Conference | *Hackathon*

- Feature-engineered using VIF, RFE, and PCA. Built a hyperparameter-tuned logistic regression, random forest, and XGBoost model to predict the outcome (pass or fail) of new turn-ins using historical records of turn-ins.
- Selective filtering was done based on the classification outcome of the turn-ins to save execution time in the DevOps pipeline in the context of HW design validation.

GradCam Visualization of CIFAR-10 dataset with Albumentations. | *project link*

- Focuses on building and training a ResNet-18 model on the CIFAR-10 dataset.
- Implemented data augmentation using the Albumentations library, a custom dataset loader, plotting train and test loss curves, GradCam visualization of randomly sampled misclassified images, and visualization of misclassified images with labels and appropriate legends.

Style Transfer using Generative Adversarial Network (GAN) | *Project*

- Built a Generative adversarial model(modified U-Net Architecture) which can generate artificial T1 to T2 and vice-versa MRI images of different contrast levels from existing MRI scans.

Technical Skills

Programming languages: Python, C++, Java
Libraries/Frameworks: PyTorch, Tensorflow, JAX/FLAX, LitGPT, OpenFL, Transformers, Deepspeed, Megatron-LM, ZenML (MLOps), Llama-Factory, Vue.js, FastAPI, perfetto, fio
Technologies, Platform & Practices: Clean Architecture (Domain Driven Design), Hugging Face, Git/Gerrit (code reviews), AWS, microservices, containers, Jenkins, VS Code.