Sunil Acharya

740-689-3061 sun1lach sun1lach sun1lach sun1lach

Education

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*

Vidyanidhi Jr. College of Science June. 2010 - Mar. 2012
HSC (74.83%) *Mumbai, IN*

St. Xaviers High School June. 2000 - Mar. 2010
SSC (74.14%) *Mumbai, IN*

Experience

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

- Currently working on enabling LLMs (Llama 3.1) on Gaudi 2 & 3 Accelerators through Deepspeed and Megatron-LM. Productionized Megatron-LM llama 3.1 & mistral distributed checkpoints to Huggingface format conversion utility tool to external customers. Experience with compute, memory profiling & trace analysis of model, host and device operations to identify bottleneck with various distributed parallel strategies.
- Designed & implemented (POC) a ZenML/cnvrg.io based LLM fine-tuning as a service, to be offered as a premium feature for the internal users.
- Designed & implemented an Intel Automated Checkout (ACE) application at the edge using a Domain-Driven design (Clean Architecture) approach, deployed it, and did successful pilots at various retailer sites in south of India. Currently clocking revenue at store front of selected retailers.
- Enabled & optimized deep learning model training and inference on a point-of-sale (POS) edge device, leveraging an integrated GPU (iGPU) and a memory-aware batched training strategy to update model parameters in place.
- Profiled the edge stack to identify bottlenecks using *cProfile* and optimized the code to achieve ~70ms inference and ~500ms batched(*bs=64*) image training latency on a low compute (Intel core i3/i5) iGPU.
- Linux Foundation's Secure Federated AI OpenFL project contributor - Extended the capability of OpenFL federated learning by enabling federation using the JAX/FLAX framework and introduced an interactive notebook example for the same with the CIFAR-10 dataset. Also, implemented a federation timeout feature to safely tear down short- and long-lived components during runtime.
- Presented technical demos at Intel India Innovation Day and Intel ConnectiON events showcasing faster F&V checkouts with the Intel ACE stack. Delivered a technical talk on cost-optimized, production-ready data pipelines built on AWS at the Intel India Tech Talk Series (Cloud Community of Practice).
- Designed the architecture and implemented a production-ready event-driven data pipeline on AWS. Implemented AWS Glue Jobs with Scala, a native Apache Spark language instead of Python, along with optimizations led to an ~8x reduction in job execution time. Execution time optimization led to a cost reduction from ~$3000/year to ~$750/year for 112 job runs per day in both the pre-production and production environments. Enabling authenticated downstream applications like ARK, MARK, CARE, and intel.com to access product catalog data through microservice API endpoint data via the APIGEE gateway, clocking ~40 million API hits monthly.

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

- Experience building cross-platform CLI and GUI tools for performance profiling of SD/uSD, USB, and enterprise SSD devices using open source tools like FIO, VDBench, etc.
- Successfully designed, developed, and deployed an E2E Distributed Client Server Application utilized by several teams to execute 350+ mobile devices in parallel to test, validate, and certify uSD cards.
- Experience in configuring and chaining 48/24 ports Brocade Network switches & tune power class ports to support 80+ PoE surveillance cameras for automated uSD/SD card testing.

sketchmyroom.com (Rhythm of space) Aug. 2016 – Dec. 2016

Software Engineer Trainee *Bengaluru, IN*

- Interiors and architecture design-related experience with Full Stack Web Application Development. Primarily responsible for the development of a backend application that exposes in-house portfolios to clients via APIs, giving them access to a variety of Architecture designs and an online customization option.

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.

Projects, Case Studies & Hackathon

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

Languages: Python, C++, Java
Libraries/Frameworks: PyTorch, TF/Keras, 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), HuggingFace, Git/Gerrit (code reviews), AWS, microservices, containers, Jenkins, VS Code.