734-596-9245
saklaniayush99@gmail.com

Ayush Saklani

linkedin.com/in/ayushsaklani
github.com/ayushsaklani

EDUCATION

University of Michigan, Ann Arbor | M.S, Electrical and Computer Engineering | GPA: 3.89/4.0 Expected May 2024

Teaching Staff: EECS 598 - Foundations of LLM, EECS 442- Computer Vision, EECS 545- Machine Learning

NIT Hamirpur | B.Tech., Electronics and Communication Engineering | GPA: 3.88/4.0 May 2019

WORK EXPERIENCE

Machine Learning Engineer, Here Technologies, Chicago | CNN, AWS, Python, Inference, API May 2023 - Dec 2023

- Architected a centralized end-to-end platform for evaluating and running inference on machine learning models using docker containers on AWS to speed up evaluation and deployment
- Implemented a custom script to start and stop AWS Batch automatically based on the amount of tasks in SQS, which reduced the idle running cost of AWS ECS by 60%
- Trained and deployed computer vision models to detect roads and lanes using segmentation for automated map making
- Designed the system using AWS services, such as Batch,SQS, Lambda, and used Celery for efficient background processing

Software Engineer, Accenture | Android, Java, Embedded, Data Science Aug 2019 – July 2022

- Developed high-performance customer-facing transit payment system, serving 3 millions customers daily
- Automated remote EMV card readers configuration and firmware upgrades which decreased idle downtime of devices by 30%
- Conducted predictive analysis on the number of payments done on the transit system using XGBoost, Decision Trees providing analytics to senior management

PROJECTS

Reinforcement Learning for LLMs | Pytorch, RL Feb 2024-Present

- Implemented Proximal Policy Optimization(PPO) and Natural Language Policy Optimization(NLPO) to align Large Language models to human preference on paper summarization tasks

GenAI Chatbot [demo, github] | Pytorch, Langchain, LoRA, QLoRa Dec 2023-Present

- Built a chat app using fine tuned LLAMA using LoRA(Low Rank Approximation) where users can interact with the bot
- Enabled users to submit PDF documents, enhancing the bot's capabilities to process and extract information from uploaded files

Monocular 3D Object Detection[github] | CNN, Pytorch, Albumentations, GPU, OneCycleLr, 3D Detection Aug - Dec 2023

- Spearheaded the development of a robust 3D object detection system utilizing a monocular camera for self driving vehicles
- Implemented advanced techniques like OneCycleLr scheduler and data augmentations, such as GlassBlur, RandomFog to optimize model training and increasing robustness towards adverse weather conditions
- Ranked among the top 3 teams on the final leaderboard by successfully training and fine-tuning the system, which scored well on an untrained bad weather dataset

REDD(Real-Time Expedited Disease Detection) | CNN, Pytorch, Quantization, Segmentation, RaspberryPi, CPU Aug - Dec 2023

- Developed a functional prototype capable of Real-Time segmentation of 13 lung diseases without reliance on internet connectivity, which was deployed on small cost effective device such as Raspberry Pi
- Quantized model weights of PSPNet by making architectural changes, significantly reducing model size from 260MB to 0.55MB
- Demonstrated the feasibility of deployment on resource-constrained devices by achieving an average inference time of around 25 seconds on a Raspberry Pi for the quantized model

VisionGuard[github] | CNN, Pytorch, GPU, Object Detection, GPU Attribute Recognition, Transformer Aug - Dec 2023

- Piloted the development of VisionGuard, an innovative real-time Pedestrian Attribute Recognition system designed for recognizing intricate attributes such as clothing, accessories, age for video surveillance and security applications
- Designed a custom decoder based on the Swin Transformer architecture to predict attributes, achieving a remarkable 95% accuracy on test dataset

SKILLS

Programming Languages:	Python, C++, Java, Javascript, React.JS
Machine Learning:	Pytorch, TensorFlow, LLM, LoRA, QLoRA, Diffusion Models, MLOps
Subject Knowledge:	Deep Learning, ML, Artificial Intelligence, HMI, Reinforcement Learning, Computer Vision, NLP, SQL System Design
Cloud/Deployment:	Docker, FastApi, Celery, AWS Batch, ECS, Lambda, S3, Sagemaker, SQS, Vertex.ai, Version Control, DevOps
Others:	Arduino, Raspberry Pi