Ankur Singh

mleankursingh@gmail.com | Ankur-singh.github.io | github.com/Ankur-singh

Summary

AI Solutions Engineer with **7 YoE** translating cutting-edge research into production AI/ML products. Combines deep expertise in **LLMs**, **multi-modal AI**, and **inference** optimization with strong **full-stack software engineering** skills. Proven track record of **leading ML teams**, **founding a startup**, and delivering results in **fast-paced** environments with evolving requirements. Multiple **Kaggle** medalist and **hackathon** winner with contributions to **open-source** projects. Excels at rapidly adopting new technologies to drive **practical**, **results-oriented** solutions.

Technologies

Languages: Python, SQL, JavaScript

Deep Learning: PyTorch, HF Transformers, TorchTune, TorchAO, Triton, Unsloth, Liger Kernel, vLLM, Llama.CPP

Other Technologies: Ray, Apache Spark, Posgres, MongoDB, Redis, Docker, Kubernetes, AWS, GCP, FastAPI

Experience

AI Solutions Engineer, Intel – Santa Clara, CA — May 2023 – Present

- **Develop**, **Profile**, **Benchmark** and **Optimize** PyTorch workloads, including Distributed Training, PeFT, Quantization, and Inference on Intel Max and Arc series GPUs using Intel's oneAPI stack
- Develop and maintain GenAI Code Samples (E2E projects and tutorials) showcasing ways to optimize LLM inference workloads on Intel hardware, across platforms (AI PC, Xeon, OpenShift, AWS)

Graduate Research Asst., SJSU Research Foundation – San Jose, CA — Sep 2022 – May 2024

- **Optimized YOLOv8** for **real-time (≈ 150 ms)** object detection and segmentation on **NVIDIA Jetson** devices
- Developed a novel **Decentralized Traffic Flow Prediction** system using **Federated Learning** (*IEEE Link* 🔗)
- Conducted comprehensive research on **Vision-Language Models (VLMs)** and **image tokenization** techniques

Machine Learning (ML) Lead, Zoop.One – Pune, India — June 2003 – Aug 2003

- Lead development of **4 Core ML services**, leveraging micro-services based architecture, housing **20+ Deep Learning models**, serving **2M+ monthly requests**. Resulted in **$1 million savings** in subscription fees, every year
- **OCR service**: Extracted info from various ID Cards by employing pipeline consisting of **7+ Computer Vision models**, achieving **6x faster** response time (≈ 1 sec) than competition, with higher accuracy, and layout support
- Constructed **Liveliness Service** for real-time face detection, recognition, matching, eye tracking, and liveliness detection with **low latency(≈ 200 ms)** to prevent identity spoofing, and frauds

CoFounder and CEO, AI Adventures – Pune, India — Aug 2018 – Sep 2021

- Led development of diverse client projects such as **Image Search**, **Receipt Digitization**, and **Visual Inspection**, resulting in automation that saved clients 100+ man hours per week

Projects / Open-Source / Competitions

Snapjobs - *AI Platform for job seekers*

- Aggregated 20K job listings from 1000+ companies using **Ray** for **distributed** and **parallel processing**
- Used LLMs to extract information from job listings and stored data in **Elasticsearch** for powerful search features
- **Fine-tuned Llama2-7B** (using **Axolotl**) to help users tailor resumes for jobs, and deployed it using **vLLM**

OPEN SOURCE

- Contributed to various projects including **torchtune**(post-training library) and **fast.ai**
- Author of ***Colab-everything*** (**40K downloads**) & ***torchserve-client***, python packages hosted on PyPI

COMPETITIONS

- **Grand prize** in AI for Social Good **hackathon** at **Intel Innovation**, 2022
- **Bronze medal** in *Shopee: Price Match Guarantee*, *Global Wheat Detection*, and *MoA* Kaggle competition

Education

San Jose State University, MS in Software Engineering — Aug 2022 – May 2024 **(3.9/4.0)**

College of Engineering Pune (COEP), BS in Information Technology — Aug 2014 – May 2018 **(7.7/10)**