Ankur Singh

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

Summary

AI Solutions Engineer with **7 YoE** (5 industry & 2 academia) delivering production-grade AI solutions by translating SOTA research into business impact. Specialized in **LLMs**, **multimodal AI**, and **inference optimization** with strong **full-stack** skills. Proven track record of **leading ML teams**, **founding a startup**, and rapidly shipping high-impact ML products in **fast-paced**, ambiguous environments. **Kaggle Competitions** and **hackathon** winner, and open-source contributor.

Technologies

Python, SQL, JavaScript || Numpy, Scipy, PyTorch, Transformers (HF), DeepSpeed, Triton, PyTorch Lightning, vLLM || Ray, Spark, Postgres, MongoDB, ElasticSearch || Docker, Kubernetes, CI/CD || AWS, GCP || FastAPI, Django || LLM inference, Distributed Training, Multimodal pipelines.

Experience

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

- **Accelerate GenAI model performance by up to 3x** through profiling and optimizing PyTorch workloads – including distributed training, **PeFT**, **quantization**, and **inference** – using Intel's oneAPI stack and latest hardware (Xeon, AI PC)
- **Drove adoption of Intel hardware for LLM workloads** by designing and maintaining end-to-end GenAI code samples and tutorials, enabling developers to deploy optimized inference pipelines across AI PCs, OpenShift clusters and AWS

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

- **Enabled real-time edge inference** by optimizing YOLOv8 for object detection and segmentation on NVIDIA Jetson (latency $\approx$ 150 ms), supporting autonomous driving use-case
- **Pioneered a Decentralized Traffic Flow Prediction system** using **Federated Learning**, improving privacy preserving forecasting across nodes, published in *IEEE* 🔗)

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

- **Saved $1M+ annually** by leading the design and deployment of 4 core ML microservices housing 20+ deep learning models; scaled system to serve over **2M monthly requests** through a modular, cloud-native architecture
- **Delivered industry-leading OCR pipeline** (7+ CV models) for ID card parsing with **6× faster response time ($\approx$1s)** and superior layout-aware accuracy—beating commercial APIs in both latency and precision
- **Built a real-time liveliness detection service** with face recognition, matching, eye tracking, and spoof detection; achieved **sub-200ms latency**, strengthening anti-fraud protection across customer onboarding workflows

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

- **Delivered automation solutions saving 100+ manual hours/week** by leading development of diverse ML projects, including **image-based product search**, **receipt digitization**, and **visual defect inspection** for enterprise clients

Projects / Open-Source / Competitions

Snapjobs - *AI Platform for job seekers*

- **Scraped and indexed 20K+ job listings** from 1,000+ companies using Ray and LLM-based extraction; enabled semantic search with Elasticsearch
- **Fine-tuned and deployed LLaMA2-7B** to generate personalized resumes, enhancing relevance and job targeting

OPEN SOURCE

- Contributed to high-impact, PyTorch native ML libraries including **torchtune**(LLM post-training) and **fast.ai**
- Authored Python packages ***Colab-everything*** (**40K+ downloads**) & ***torchserve-client***, simplifying workflows for ML prototyping and model inference

COMPETITIONS

- **Grand Prize Winner**, AI for Social Good **hackathon @ Intel Innovation** 2022
- **3x Kaggle Bronze Medalist –** *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)**