Ankur Singh

☎ +1 (408) 941 5818 ✉ mleankursingh@gmail.com ankur-singh.github.io

EDUCATION

San Jose State University, CA (USA) **CGPA: 3.82/4.0**
Master of Science in Software Engineering *Aug 2022 – May 2024*

College of Engineering Pune, Pune (India) **CGPA: 7.67/10**
Bachelor of Engineering in Information Technology *Aug 2014 – May 2018*

EXPERIENCES

Total **4 years of experience** in Python, SQL, Data Science, Machine Learning, Deep Learning, and MLOps.

AI Solutions Intern, *Intel* - San Jose, CA **May 2023 – Present**

- Implementing code samples to showcase state-of-the-art techniques, including **QLoRA** and **RAG** on Intel's hardware.
- Developed end-to-end workflows, such as **distributed training**, **fine-tuning LLMs**, and **INT4/8 inference** on CPU.
- **Technologies**: PyTorch, AI Toolkit, HuggingFace, Transformers, Datasets, Accelerate, Kubernetes, Kubeflow, AWS

Graduate Research Assistant, *SJSU Research Foundation* - San Jose, CA **Sep 2022 – May 2023**

- Under Dr. Wu, working on Traffic Flow Prediction(TFP) using **Federated Learning**, while preserving **user privacy**.
- Under Dr. Liu, helped **optimize**, **benchmark** and **deploy** various detection & segmentation models on **edge devices**.
- **Technologies**: PyTorch, Docker, ONNX, OpenVINO, NVIDIA Triton, TorchServe, NVIDIA Jetson

Machine Learning (ML) Team Lead, *Zoop.one* - India **Sep 2021 – Jul 2022**

- Successfully launched **four ML services**, using 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 Identity Cards using pipeline consisting of **7+ deep learning models**, still achieving **6x faster** response time **(≈1 sec)** than competition, higher accuracy, and multi-line field support.
- Developed **Document extractor**: **Heatmap Regression-based** model for ID card or document auto-cropping, exported the model to **TFLite (4.4 MB)** for on-device inference. **Liveliness service**: Real-time face detection, recognition, matching, and **liveliness** detection with **super-low latency (≈200 ms)** to prevent spoofing.
- **Technologies**: FastAPI, TorchServe, MLflow, WandB, AirFlow, Label Studio, K8s, ELK stack, Prometheus, Grafana

CoFounder and CEO, *AI Adventures LLP* - India **Aug 2018 – Sep 2021**

- Led development of diverse client projects such as Jewellery **Image Search**, Receipt Digitization, Smart Attendance.
- Developed **five** comprehensive **courses** covering Machine Learning, from basics to model deployment, and assisted **800+ individuals** in initiating their journey in AI/ML.

PROJECTS / OPEN SOURCE

Snapjobs: AI Powered Job assistant | *Python, vLLM, ElasticSearch, MongoDB, AirFlow* **Present**

- Developing a one-stop solution for tailoring resumes to specific job descriptions using LLMs, streamlining application tracking, providing real-time job openings with enhanced search capabilities to facilitate job seekers' success.

ChatSpartan | *LangChain, Airflow, ElasticSearch, llama_cpp, Gradio, Pinecone* **Dec 2023**

- User-friendly chatbot for college website to assist visitors navigate and efficiently search for information on the site.

Open Source **Ongoing**

- Author of ***Colab-everything*** (**36K+ downloads**) & ***torchserve-client***, python packages hosted on PyPI.
- Contributed to packages including *LazyPredict*, *fastai*, *category_encoders*, *YOLOv5*, & AI samples in **oneAPI Samples**.

COMPETITIONS

Targeted Pest Control | *Intel Innovation, 2022* **Grand Prize Winner**

- Develop a **compact CNN model** to differentiate weeds from crops, enabling efficient deployment on drones.

Shopee - Price Match Guarante | *Kaggle Code Competition* **Bronze Medal**

- Built an ensemble of **multi-modal NN** with **ArcFace Loss** and **Representation Learning** to determine if two products are similar based on their images, description and other meta data.

Global Wheat Detection | *Kaggle Code Competition* **Bronze Medal**

- Finetuned **EfficientDet** and **YOLOv5** models with advanced techniques (MixUP, Mosaic, Pseudo Labelling, TTA, WBF) for precise wheat head detection in **noisy** outdoor field images.

Mechanisms of Action | *Kaggle Code Competition* **Bronze Medal**

- Leveraged **TabNet** with specialized **feature extraction** from gene expression and cell viability data to classify drugs based on their biological activity in a **multi-label** problem with 207 labels.

SKILLS

- **Languages & Databases:** Python, SQL, MongoDB, Postgres, ElasticSearch, Pinecone, ChromaDB, Faiss
- **LLM Stack:** HuggingFace, PeFT, vLLM, LlamaCPP, LangChain, LlamaIndex, Ollama, OpenAI, LoRAX