ANKUR SINGH

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

Summary

With 5+ years in Product Management, Deep Learning, Computer Vision, and Software Development, I'm passionate about solving business problems, building ML products, and contributing to open source. Currently pursuing a Masters in Software Engineering at SJSU. In my previous role, at Zoop.One, built and led a high-performing team, deploying 4 ML services in 10 months, yielding impressive results.

Skills

Languages & Databases	Python, Java, SQL, Postgres, SQLite3, MongoDB, Redis
API Development & Web apps	Flask, FastAPI, Rest API, gRPC, Web Socket, Postman, Streamlit, Gradio
Software Development	GitHub, GitHub Actions, Python Packaging, Design Patterns, Unit Testing
Machine / Deep Learning	Pandas, Sklearn, XGBoost, Keras, Pytorch, Fastai, HF Transformers, LangChain
DevOps & MLOps	Docker, Kubernetes, Kubeflow, Terraform, ONNX, TorchServe, TFlite

Work Experience

AI Solution Intern, @ Intel   Remote (May. 2023 - Present)

- Intel Cloud Optimization Modules (ICOM) for AWS: Finetune **GPT-2 Small** model on OpenWebText dataset.
- Training with PyTorch's Distributed Data Parallel (DDP) on CPUs using IPEX and oneCCL.
- **Technologies**: Pytorch, HF Transformers, Datasets, Accelerate, Kubernetes, Kubeflow, AWS

Graduate Research Assistant, @ SJSU   San Jose, CA (Sept. 2022 - May 2023)

- Worked on improving Knowledge Distillation performance for Traffic Flow Prediction.
- Deployed various detection and segmentation models on Edge devices for real-time inference.
- **Technologies**: Pytorch, Docker, Jetson, ROS2 Humble, TorchServe, ONNX, NVIDIA Triton, TFLite

ML Team Lead, @ Zoop.one   Pune, India (Sept. 2021 - July 2022)

As a Founding Member of the ML team, built a strong team and scaled ML usage company wide. Enhanced MLOps with annotation tools, model registry, monitoring, training, and other best practices implementation.

- Developed **OCR service** for Identity Cards with 7+ deep learning models, achieving 6x faster speed, higher accuracy, and multi-line field support.
- Created a Heatmap Regression-based **document extractor service** for auto-cropping documents/ID cards, with a lightweight TFLite model for edge deployment (4.4 MB).
- Led development of a real-time **liveliness service** with features like face detection, recognition, matching, achieving low latency (~150 ms) for real-time video feed inference.

Technologies: Pytorch, MLFlow, OpenCV, Label Studio, Docker, FastAPI, ONNX, Wandb, FiftyOne, TorchServe

CoFounder and CEO, @ Ai Adventures LLP   Pune, India (Aug. 2018 - Sept. 2021)

- Built the tech-stack required for generation, deployment & distribution of all courses.
- Lead development of various client projects like Image Search, Smart Attendance, Parking management.

Education

- **Masters in Software Engineering** (spl. Data Science) @ SJSU   CGPA: **3.9**/4.0 (Aug. 2022 - May 2024)
- **BTech in Information Technology** @ COEP   CGPA: **7.67**/10 (Aug. 2014 - May 2018)

Competitions

- **Targeted Pest Control** @ Intel Innovation, 2022   Grand Prize
- **Shopee - Price Match Guarantee** @ Kaggle   Bronze Medal (207th / 2426)
- **Global Wheat Detection** @ Kaggle   Bronze Medal (191st / 2245)
- **Mechanisms of Action** @ Kaggle   Bronze Medal (318th / 4373)

OSS: Author of colab-everything (32K+ downloads) and torchserve-client python packages published on PyPI.