Sandesh Shrestha

San Jose, CA (510) 364-2503 | sandeshshrestha02@gmail.com | github.com/sbsx | LinkedIn

Technical Skills

Technologies: Python, Tensorflow, Pytorch, OpenCV, OpenAI, Langchain, Neo4j, Milvus, ChromaDB, Huggingface, Llama-Index, Clickhouse, Git, Bash, FastAPI, MongoDB

Tools: VSCode, DVC, Postman, Azure, Databricks, DBeaver, Jira, ServiceNow, Linux

Experience

LLM/ML Engineer | Applied Materials **May 2023 – Present**
LLM Infrastructure LangChain | OpenAI | Llama | Azure | Neo4j

- 81% F1-score in predicting wafer manufacturing completion time, outperforming manual predictions by 15%
- Implemented U-shaped normalizing flow for wafer manufacturing anomaly detection using pretrained segmentation models
- Fine-tuned PEFT Llama2-13b which beat much larger GPT-4 on MaterialScience Benchmark
- Optimized document retrieval for 80k machine manuals, reducing lookup from hours to seconds using GPT/RAG with secure, personalized access
- Comparative evaluation of open-source LLMs and vector databases against Azure solutions
- Implemented vLLM local model hosting for enhanced data security in material science research

Lead Software Engineer | Applied Materials **August 2023 – March 2024**
Embedded Computer Vision Linux | OpenCV | Embedded

- 160x faster wafer position calibration times, automating wafer placement with computer vision
- Expanding product security by developing multi-protocol connectivity that meets high-security fabrication plant requirements
- Improved system processing speed by 4x and extending battery life by 2x with lightweight, upgradeable OS/firmware
- Optimized operator workflows, dramatically reducing onsite diagnostics and calibration time

Software Engineering Consultant | Bluestamp Engineering **Jan 2023 – May 2023**
Machine Learning Pose Estimation App Development TensorFlow | OpenCV | Data Collection

- Increased dataset size by 2x with data augmentation, improved data ingest pipeline.
- Reduced GPU requirements by 30% using retrained smaller models for edge devices
- Provided ongoing technical support to clients, resolved issues on call, maximized deliverables for a live classification and human pose estimation application

Computer Vision Developer | Triton Unmanned Aircraft Systems **Oct 2020 - June 2022**
Autonomous Drone Competition PyTorch | OpenCV | Nvidia Jetson

- 5x faster inference on Nvidia Jetson and extracted 2x key-points with ORB feature detection
- 10% increased hit rate with FCNN segmentation and VGG19 object detection
- Ensured accuracy with Visual SLAM, preprocessing pipeline, and data augmentation techniques

Computer Vision Developer | Triton Robotics **Oct 2020 - June 2021**
DJI Robomaster ShenZhen Competition C++ | ROS2 | OpenCV

- 10% faster opponent detections in low light environments with 95% true positive classifications
- Pattern matching through C++ OpenCV to enable enemy position prediction
- Mentored 3 junior engineers, providing guidance on technical skills and team coordination

Education

University of California - San Diego **B. Sc., *Mathematics - Computer Science 2022***