SANDESH SHRESTHA

San Francisco, CA (510) 364-2503 | sandeshshrestha02@gmail.com | GitHub | LinkedIn

TECHNICAL SKILLS

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

Tools: VSCode, Postman, Azure, Cloudera, DBeaver, Jira, ServiceNow, Windows, Linux, Teams

Models: Llama, GPT, Mistral, SentenceBert, YOLO, SAM

EXPERIENCE

LLM Test Engineer | Applied Materials **May 2023 – Present**

LLM Infrastructure LangChain | OpenAI | Llama | Azure | Neo4j

- Served critical tooling manuals using RAG through Azure to fabrication plants, significantly reducing document lookup times from hours to seconds for 80k documents
- Developed evaluation methods to test future models and processes against current systems, capturing speed, reliability, and safety of results
- Digested LLM research daily and implemented findings, comparing to industry standard pipelines for cost, data privacy, and scaling
- Improved material science domain knowledge using PEFT finetuning on Llama13b, beating GPT-4 on the MatSciBenchmark
- Finetuned embedding and reranker models to improve retrieval on documents with industry specific language, increasing hit-rate and recall by 5%

Lead Software Engineer | Applied Materials **August 2023 – March 2024**

Embedded Computer Vision Linux | OpenCV | Embedded

- Developed firmware and software package with computer vision capabilities for semiconductor fabs globally and diagnose/calibrate precision robotics
- Reduced fab downtimes from 80+ hours to 30mins, enabling higher yields, increasing uptime and reducing calibration costs significantly
- Worked with UX team to improve onsite operator's speed, reducing expensive onsite time dramatically with additional views, setting presets, and startup simplification

Software Engineering Consultant | Bluestamp Engineering **Jan 2023 – May 2023**

Machine Learning Pose Estimation App Development TensorFlow | OpenCV | Web Scraping

- Guided client to maximize deliverables for a live classification and human pose estimation application
- Managed ML-stack development projects, including data processing and collection, increasing dataset size by 2X with augmentation techniques
- Provided ongoing technical support to clients, resolved issues on call, and optimized performance on lighter edge systems by 30% using fit smaller models

Aircraft Computer Vision Developer | Triton Unmanned Aircraft Systems **Oct 2020 - June 2022**

International Drone Competition Autonomous Surveillance PyTorch | OpenCV | Nvidia Jetson

- Architected pipeline for computer vision ML tasks on an unmanned aircraft

Continued …

- Implemented FCNN segmentation for flight navigation and object detection using VGG19 model, increasing hit-rate by 10% with lower compute draw
- Ensured accuracy with Visual SLAM, preprocessing pipeline, and data augmentation techniques
- Improved compute time by 10x on Nvidia Jetson and extracted 2x key-points with ORB feature detection

Robot Control Engineer | Triton Robotics **Oct 2020 - June 2021**
DJI Robomaster ShenZhen Competition Development C++ | ROS2 | OpenCV

- Designed computer vision pattern matching with openCV, enabling live future position prediction
- Improved low light accuracy and speed with preprocessing, with 10% earlier opponent detection and reducing hallucinations to 0
- Mentored 3 junior engineers, providing guidance on technical skills and team coordination

EDUCATION

University of California San Diego ***Bachelor of Science,*** *Mathematics - Computer Science*