AWANTIKA SRIVASTAVA

Computer Vision Engineer | Applied AI (Vision & GenAI)

+91-8920482037 | sawantika81@gmail.com | LINKEDIN | Github

PROFILE SUMMARY

Computer Vision Engineer with 2+ years of hands-on experience building and deploying real-time computer vision and deep learning systems. Strong expertise in CNN-based vision models, image pipelines, dataset curation, model optimization, and low-latency inference. Practical exposure to generative vision models, vision–language architectures, and prompt-based workflows, with a strong interest in Stable Diffusion, LoRA fine-tuning, and personalized generative AI experiences. Experienced in taking models from research to production, optimizing for performance, mobile inference, and cost efficiency, and collaborating closely with product, backend, and design teams.

CORE TECHNICAL SKILLS

- **Programming & Data Science:** Python, C++, SQL, Pandas, Numpy, Scikit-learn, Jupyter Notebook.
- **Statistics & Mathematics:** Statistical Modeling, Descriptive Statistics, Hypothesis Testing, Probability, Sampling, Scenario Analysis.
- **Machine Learning:** Supervised & Unsupervised Learning, Regression, Classification, Clustering, Random Forest, Decision Trees, SVM, KNN, K-Means, XGBoost, Model Evaluation Metrics (Accuracy, Precision, Recall, F1-score, ROC-AUC).
- **Deep Learning & AI:** Neural Networks, CNN, RNN, LSTM, Transformers (BERT), Model Fine-tuning, Time-series Forecasting, ARIMA, SRIMA, TensorFlow, PyTorch
- **Computer Vision:** Image Classification, Object Detection (YOLO, SSD, MobileNet, ResNet), Real-time Video Analytics, Image Pipelines, Face Embeddings & Similarity Scoring, Image Consistency & Feature Extraction, Paddle OCR, OpenCV
- **Generative Vision & Vision-Language:** Stable Diffusion, LoRA, Prompt Engineering, RAG, Multi-Agent AI Workflows, Embeddings & Vector Search, TinyLLaMA / Transformer Models
- **NLP:** Text preprocessing, Tokenization, Chunking, Sentiment Analysis, Topic Modeling (LSA, LDA), Transformer-based models, NLTK, TF-IDF
- **Model Optimization & Deployment:** Low-latency Inference, Edge AI, Quantization, Mobile Optimization, GPU-based Inference (basic)
- **Cloud & Platforms:** Hugging Face Spaces, Replicate, AWS EC2, S3
- **Tools & Collaboration:** Git, GitHub, Documentation, Docker, CI/CD, REST APIs, FastAPI/Flask

EXPERIENCE

Computer Vision Engineer | PPS International Pvt. Ltd. **January 2024-Present**

- Designed and deployed **real-time computer vision systems** using **CNN-based object detection models.**
- Built **end-to-end image and video pipelines**, including preprocessing, inference, and post-processing.
- Optimized deep learning models for **fast inference, low latency, and constrained hardware environments**.
- Applied **model quantization and TensorFlow Lite** to enable efficient **edge and mobile deployments**.
- Worked with **large-scale image and video datasets**, supporting dataset curation and quality checks.
- Collaborated closely with **backend and product teams** to productionize vision models.
- Maintained **production-ready Python code**, model versioning, and inference workflows.
- Monitored model **performance** using **dashboards** and **logs**, supporting **debugging** and iterative improvement.
- **Collaborated** closely with senior data scientists, ML engineers, and platform teams to ship production **AI** features.

PROJECTS

Railway Driver Assistance System (RDAS) | Enterprise ML Project

- **Designed and deployed a** real-time computer vision–based ML system **for unsafe** driver behavior detection using **CNN-based SSD MobileNet** model.
- Trained and optimized models on large-scale video datasets, achieving **20–25 FPS** real-time processing with **<150 ms inference latency**.
- Implemented **end-to-end ML pipelines** for data ingestion, preprocessing, model training, evaluation, and production inference.
- Deployed optimized models using **TensorFlow Lite** on edge/production environments for continuous monitoring.
- Built a **Flask-based web dashboard** to visualize detections and automatically record **30-second event clips**, reducing manual review effort.

Chatbot Using LLM & RAG | Applied ML Project

- Built a Lightweight **LLM-Powered chatbot** using **Tiny LLaMA** to answer user queries over content.
- Implemented a **Retrieval-Augmented Generation (RAG) pipeline** to retrieve relevant resume sections for contextual question answering.
- Selected **Tiny LLaMA** to ensure **low memory footprint and fast inference**, making the solution suitable for resource-constrained environments.
- Applied **prompt engineering techniques** to improve response relevance and consistency.
- Deployed the chatbot as an interactive **Streamlit web application** for real-time user interaction.

YouTube Comments Sentiment Analyzer | link-https://youtube-ai-analyzer-ndzqo6r2mepjrtsdjmwaxl.streamlit.app/

- Deployed transformer-based **NLP models** as **production-ready** services with **REST APIs**.
- **Fine-tuned** and served a **DistilBERT**-based **sentiment** classification model for large-scale text inference.
- **Built** and deployed an interactive **streamlit web application** to perform real-time **sentiment analysis** on YouTube comments.
- **Processed** high-volume text **data** with sub-second inference **latency** for real-time sentiment analysis.

CERTIFICATION

- IBM Data Science & AI Certification
- AWS Generative AI with Large Language Models
- OpenCV Computer Vision Certification

EDUCATION

IMS Engineering College, Ghaziabad **September - 2020**

Bachelor of Technology