Sarvagya Porwal

+91-7905468618 | sarvp.srk1729@gmail.com | linkedin.com/in/sarvagya-p-1175191b5 | github.com/Sar2580P | sar2580p.github.io

Summary

Innovative AI/ML engineer with expertise in deep learning, computer vision, generative AI, and NLP. Experienced in developing diffusion models, LLMs, and retrieval-augmented generation (RAG) pipelines. Skilled in Python, PyTorch, TensorFlow, and cloud-based deployment. Published research on diffusion models and spectral attention networks. Passionate about designing scalable AI solutions, optimizing model efficiency, and leveraging advanced architectures for real-world impact.

Publications

Spectral Band Attention Network — Sep 2024 - Dec 2024
Computer Vision — Code — Γ Paper — *PyTorch, OpenCV*

- Proposed a novel **Spectral Band Attention Module (SBAM)** as a core component of a customized DenseNet. This module efficiently sampled top 15% of most informative spectral bands, achieving 87% accuracy, comparable to using the full spectral range.
- Classification of wheat seeds into 96 varieties, fine-tuned models like DenseNet-121, ResNet-50, and GoogleNet for RGB seed image classification and developed a DenseNet-121-inspired architecture for Hyperspectral data.
- Employed Regression-based Ensemble (using SVM) to combine model predictions, ensuring robustness and enhancing overall accuracy to 95%.

Smoothed Energy Guidance - Reproducibility Challenge — Jan 2025 - Feb 2025
Diffusion Models — Code — *PyTorch, Hugging Face*

- Reproduced and enhanced SEG (Smoothed Energy Guidance) from NeurIPS 2024, addressing missing ablation studies on kernel size and blurring strategies.
- Optimized energy guidance in diffusion models, generalizing the energy framework and proposing cheaper alternatives performing comparatively better on FID scores.

Tracked reverse diffusion trajectory using Frobenius Norm, Laplacian Variance, and Gradient Entropy on attention layers, ensuring better interpretability.

- The *research paper* is currently under review.

Projects

AI Agent 007: Tooling up for Success (Inter-IIT Techfest 2023) — Dec 2023 – Dec 2023
Generative AI Project — Code — *Python, Langchain, GPT-4, Hugging Face*

- Built a query-aware agent capable of allocating and reviewing tool outputs
- Focused on creating autonomous tools for efficient parameter extraction and downstream function calls
- Implemented a self-reflective ReAct style agent, curated dataset using given tool descriptions

Enriched Bots-Clever Chat — June 2024 – July 2024
Generative AI Project — Demo — *Python, Django, LlamaIndex, Hugging Face*

- Today’s bots are plain text. Our graph-based approach enriches interactions with links, pictures, and videos.
- Query is decomposed into an acyclic graph and response is generated by topologically visiting the nodes of the graph.
- While processing a node, it gets loaded with text response and enriched media in metadata.
- Then the final response is generated by topologically visiting the processed nodes.

Sentinel-2 Field Delineation July 2024 – August 2024
Computer Vision — Code *PyTorch, OpenCV, Segmentation Models*

- Developed a computer vision model for field delineation using high-resolution hyperspectral multiband images, based on Solafune’s competition dataset.
- Fine-tuned U-Net-based models (UNet++, FPN, DeepLabV3, Mask-RCNN) and applied OpenCV to identify polygons for predicted annotations by processing patched images.
- Built an ensemble model by stacking masks predicted by base models, enhancing segmentation over the patched images and achieving overall IOU = 0.96.

Experience

DeepLogic AI — Certificate July 2024 – Dec 2024
AI Engineer Intern

- Contributed to the development of a Retrieval-Augmented Generation (RAG) pipeline for enterprise search, managing email and document embeddings in a Postgres vector-store on AWS, enabling high-performance information retrieval.
- Designed normalized database schemas and optimized scalable CRUD operations for metadata-filtered searches across millions of documents.
- Developed critical components such as the Retriever, Response Generator, and Re-ranker, and implemented caching strategies to enhance chatbot integration, improving overall system interaction, efficiency, and scalability.

Education

Indian Institute of Technology, Roorkee Roorkee, India
BTech in Mechanical Engineering *Nov 2021 – Currently*

Technical Skills

Generative AI

• Langchain • Llama-Index • Hugging Face • RAG • Self-Reflection • Prompt Engineering • PEFT/LORA • DsPy • KnowledgeGraph • ReAct • LLMOps • AsyncIO • AWS • Web Scrapping • Grad.io • Fast-API • Docker • Flask • Django • Github • Linux • Shell Scripting

Computer Vision, NLP

• Pytorch • Tensorflow • OpenCV • Segmentation Models • Detectron-2 • Diffusion Models • GAN • SpaCy • Object Counting • Sentiment-Analysis • Video Captioning • Bash Scripting • Distributed Training

Certifications

- Machine Learning Certificate
- AWS Cloud Practitioner Certificate