Sarvagya Porwal

☎ ✉ in 🐙 🌐 🎓

SUMMARY

AI/ML Engineer and Researcher with expertise in deep learning, computer vision, generative AI, and NLP. Experienced in designing and implementing diffusion models, large language models, and retrieval-augmented generation systems. Skilled in building scalable AI pipelines using Python, PyTorch, and TensorFlow, with hands-on experience in cloud deployment and distributed training. Published research on diffusion models and spectral attention networks. Passionate about bridging research and production by translating cutting-edge AI methods into efficient, real-world solutions.

PUBLICATIONS

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 smoothing in diffusion models, replacing redundant operations with EMA and BoxBlur, improving efficiency without loss in quality.
- 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.

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

- Classification of wheat seeds into 96 varieties, fine-tuned models like DenseNet-121, ResNet-50, and GoogleNet for RGB seed image classification (max 78% accuracy) and developed a DenseNet-121-inspired architecture for Hyperspectral data.
- Applied Spectral Band Attention Module (SBAM) for band selection, ranking bands by their contribution to the classifier's accuracy, achieving 87% accuracy on Hyperspectral images.
- Employed Regression-based Ensemble (using SVM) to combine model predictions, ensuring robustness and enhancing overall accuracy to 95%.
- The research paper has been accepted (view here).

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

Avathon — SDE-1 **July 2025 – Present**

→ *avathon.com*

- Architected and implemented distributed task processing infrastructure using Celery with message queuing, enabling asynchronous job execution, automatic retries with exponential backoff, and real-time task monitoring across microservices, significantly improving system throughput and reducing blocking operations.
- Contributed to the development of Neon, a proprietary Scala-based graph database designed for supply chain solutions, implementing semantic search capabilities on customer data distributed across graph nodes and enabling efficient querying of complex relational data structures.
- Developed and maintained Django REST APIs and comprehensive test suites for the Central Asset Manager service, ensuring robust functionality, data validation, and compliance with system reliability standards.

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

→ *Certificate*

- 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 – July 2025*

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