Irfan Hamid

irfanhamid19@gmail.com | Portfolio | LinkedIn | GitHub | +44 7471069088

EDUCATION

The University of Edinburgh, Edinburgh, UK | Grade: Merit   September 2023-November 2024
- *Master of Science in Artificial Intelligence*
- *Relevant Courses:* Applied Machine Learning, Machine Learning practical, Probabilistic Modelling and Reasoning, Machine Learning Theory, Image and Vision Computing, Accelerated Natural Language Processing, Advanced Robotics

Vellore Institute of Technology, Vellore, India | CGPA: 9.28/10   June 2017-June 2021
- *Bachelor of Technology in Electrical and Electronics Engineering*
- *Relevant Courses:* Signals and System, Neural Networks and Fuzzy Control, Data Structures and Algorithms, Digital Signal Processing, Robotics and Control, Advanced Control Theory, Applied Linear Algebra

EXPERIENCE

Forest Research (Northern Research Station), Edinburgh, UK   March 2024 - August 2024
- Conducted an industry-partnered machine learning research with Forest Research (the research agency of the Forestry Commission, UK government) for my MSc dissertation, focusing on the classification of tree species in the Forest of Dean using high-resolution multispectral satellite imagery from Planet Labs' SuperDove 8 satellites
- Implemented and trained deep learning models, including **ResNet-34**, **DenseNet-40** and **Vision Transformers (ViT)** to perform species classification. Utilized **QGIS** for geospatial preprocessing, spatial analysis, and visualization of labelled tree data
- Performed a comparative evaluation of the models and analyzed classification accuracy across various tree species. Additionally, examined species spectral curves to understand and explain model predictions

Wipro Limited, *SAP Business Warehouse (BW) Consultant*, Chennai, India   July 2021 - June 2023
- Implemented SAP BW/4HANA data provisioning, developing Advanced DataStore Object (aDSO) to perform ETL (Extract, Transform, Load) processes. Designed and optimized SAP BW process chains for Nomad Foods Europe Limited to automate these processes, reducing manual intervention and ensuring timely, reliable data updates
- Developed and customized SAP BW queries accessing data from aDSOs via Composite Providers, enabling real-time data analysis aligned with KPIs. Facilitated the generation of SAP BW reports for informed decision-making, ensuring accurate reporting for strategic planning and providing insights to optimize business performance
- Monitored the entire reporting pipeline to ensure optimal performance, addressing issues for timely and accurate data delivery

PROJECTS

Deep Learning Pipeline for Football Analysis: Developed a deep learning pipeline for analyzing football match videos. Fine-tuned two **YOLOv8** models: one for object detection and the other for pitch keypoint detection, and integrated ByteTrack for tracking. Used **SigLIP** embeddings, **UMAP**, and **KMeans** for team classification, and applied perspective transformations to map player and ball positions onto a radar-style top-down pitch view, visualizing team positioning and pitch control with Voronoi diagrams.

LLM Character Based Chatbot: Fine-tuned Meta's **LLaMA 2** model using **LoRA** and **4-bit quantization** to build a character-based chatbot. The resulting adapter and training dataset were published to Hugging Face for open access and reproducibility. Additionally, designed an interactive Gradio interface to enable real-time conversations with the model.

Retrieval-Augmented Generation (RAG) Pipeline for Textbook Search: Built a **Retrieval-Augmented Generation (RAG)** system to answer textbook-based queries using **Large Language Models (LLMs)**. Extracted and embedded text from PDF textbooks to design a vector-based retriever for fetching relevant context. Retrieved content was augmented with user queries to construct context-aware prompts for generating responses using Google's **Gemma-7B-it** model.

Non-Self-Referential Attention in Transformer Models: Explored modifications to Transformer architecture and developed a method called Non-Self-Referential Attention, which attenuates self-attention scores along the diagonal of the attention matrix using a tunable factor. This method reduces a token's focus on itself, promoting richer attention across input tokens and improving translation quality, resulting in a **2.12% BLEU** score improvement when applied on 'en-pt' translation subset of the opus_books dataset.

TECHNICAL SKILLS

- *Programming Languages, Libraries & MLOps:* Python, PyTorch, NumPy, Pandas, scikit-learn, SQL, spaCy, LlamaIndex Transformers (Hugging Face), Ultralytics, OpenCV, RoboFlow, AWS, Docker, Git, GitHub Actions, DVC, MLflow, Kubernetes
- *Machine Learning:* Deep Learning Architectures (Transformers, CNNs, RNNs, VAEs), Bayesian Inference, Approximate Inference, Computer Vision, Natural Language Processing (NLP), LLMs, LLM Fine-Tuning (including PEFT methods like LoRA), RAG, LLM Compression