SHASHWAT SHAHI

Awarded with excellence certification in Computer Vision by OpenCV
Boston, Massachusetts | shahi.sh1028@gmail.com | 8574235879 | linkedin.com/in/shashwat-shahi/

EXPERIENCE

Novartis AG **Cambridge, MA**
Machine Learning Engineer **Jan 2025 – Present**

- Developed a **novel large vision model** for feature extraction from WSI images using **masked-image-modeling** with **masked language modeling** for a series of downstream tasks like gene expression prediction, life expectancy prediction, cancel cell detection and segmentation etc.
- **Fine-Tuned** an **LLM** with vectorized patient's data fetched from **Pinecone Vector DB** to generate the complete patient's summary used as one of the modalities to train the models for multiple downstream tasks, improving the precision of the models on an average by 30%.
- Trained multiple **deep regression** and **classification models** with **Distributed Data Processing (DDP)** and **FSDP** with learnable params up to ~1.5 billion.
- Leveraged **MIM with Contrastive Learning Approach** with architectures like **DINO v2, Vision Transformers, Resnet-50** etc. to pre-train the model, before finetuning it over domain specific data, enabling it to learn meaningful representations, improving feature extraction.
- Researched, formulated and applied multiple model training and fine-tuning strategies to train the models in **resource constrained environments** with **multi-GPU Distributed Processing**, keeping the costs minimum.

Northeastern University **Boston, MA**
Research and Course Assistant - Generative AI and LLMs **Aug 2024 – Dec 2024**

- Worked with Prof. Ramin Mohammadi to research on various applications of **Generative AI, LLMs, RAGs** along with various **prompting techniques.**
- Designed labs and assignments in python using **TensorFlow and PyTorch** for **Generative AI** coursework which is developed in partnership with Coursera.
- Worked on creating lectures on DL architectures such as **CNNs**, **RNNs, LSTMs, Transformers, BERT, GANs, RL, LLMs**, **Langchain, LangGraph** etc.

Tata Consultancy Services Ltd. **Bengaluru, India**
Machine Learning Engineer **Aug 2021 – Aug 2023**

- Designed and implemented an end-to-end system that automates the prediction of ICD-10 disease diagnostic codes (codes used by health insurance companies for disbursement) from a patient's medical records **using NLP techniques** and **Neo4J's Knowledge Graph**, with ~90% accuracy.
- Developed and deployed **microservices in Java (Spring Boot)** across **AWS (EC2)** and improved microservice communication using **Apache Kafka** and **Redis** which enabled the system to efficiently process high-frequency transactions, resulting in a 25% performance increase.
- Deployed and Built multiple DL models using **PyTorch, TensorFlow** on **AWS SageMaker and HPCs**, focusing on **NLP and CV architectures** including **Transformer-based models (BERT, GPT, T5), RNNs (LSTM, GRU), CNNs** and **Fine-tuned LLMs**, applied to tasks such as **text and image classification**, **NER, sentiment analysis etc.**
- Developed **MLOps** scalable **ETL** workflows, enhancing model training, versioning, and data handling efficiency by 30% with **Kafka, MLFlow**.
- **Invented multiple deep learning based novel architectures** which led to the **filing of four patents** across multiple geographies.

PATENTS/PUBLICATIONS

- A multimodal learning approach with Priority Weighted Graph Neural Networks **(Patent No.: 20231051755)** (*To be published in ICDMAI 2025)*
- Toward Space-Efficient Semantic Querying with Graph Databases **(Patent No.: 20231920198)** (Paper's Link)
- An Unsupervised Image Processing Approach for Weld Quality Inspection **(Patent No.: 202221033807)** (Patent's Link) (Paper's Link)
- Deep Recurrent Neural Network based audio speech recognition system (Paper's Link)

PROJECTS

Career Craft AI GitHub Repo Link
Tech Stack: *Python, NLTK, Spacy, Langchain, LLM, RAG, Neo4J, Graph Neural Network, Relation extraction and Linking, Ranking System, AWS (S3, EC2)*

- Developed an AI skill gap analysis and **recommendation system** using **Python**, **NLTK**, **Spacy**, **Langchain**, **LLM**, **RAG**, **Neo4J**, and **GNN** that identifies skill gaps with 85% accuracy and suggests personalized learning paths, deployed on **AWS** using **Flask** and **Docker**.

Support Ticket Classifier with RAG GitHub Repo Link
Tech Stack: *Python, Flask, Hugging Face Embeddings, Vector Store, LLM,* ***Llama****,* ***RAG****,* ***Langchain, Groqcloud Engine***

- Created a support ticket classification system achieving 78% accurate categorization using **Python**, **Flask**, **Hugging Face Embeddings**, **Vector Store**, **LLM**, **Llama**, **RAG**, **Langchain**, and **GroqCloud Engine** for automated ticket routing.

Text to SQL Generator by Finetuning LLMs GitHub Repo Link
Tech Stack: *Python, PyTorch, Transformers, PEFT, LoRA, QLoRA, Gradio, Hugging Face, Gretel Dataset, Quantization, SQL*

- Implemented parameter-efficient fine-tuning of LLMs for text-to-SQL translation using **PyTorch**, **Transformers**, **PEFT**, **LoRA**, **QLoRA** and **Hugging Face** tools, achieving a 16x improvement in SQL generation while reducing memory requirements by 75% on **Mistral-7B** and **Llama-3B** models and delivering results through a **Gradio** interface.

Text-to-Image Generation with Latent Diffusion Models GitHub Repo Link
Tech Stack: *Python, PyTorch, Diffusers, Transformers, Flask, Pillow, Stable Diffusion, VAE, U-Net, CLIP, Attention Mechanisms*
Fine-tuned the CLIP text encoder (clip-vit-base-patch32) on Flickr8k while freezing VAE (AutoencoderKL) and U-Net components from Stable Diffusion v1.5, implementing mixed precision training (fp16), gradient accumulation, attention slicing, reducing training loss from 0.20 to 0.06 over 9 epochs.

SKILLS

Languages and Frameworks: Python, C++, Java, CUDA, **PyTorch, TensorFlow, Keras**, NLTK, Spacy, SQL, CQL, Scala, Flask, Django
Tools: Git, Neo4J, AuraDB, **MLFlow**, **Airflow**, **Jax, Ray, Unsloth, Spark**, **AWS**, GitHub actions, Docker, Kubernetes, Huggingface, **Langchain**, **LangGraph**
Technologies: Statistical Modeling, **Machine Learning, Deep Learning**, Transformers, Attention mechanisms, **NLP**, **Generative AI**, **Computer Vision**, **Knowledge Graph,** Transfer Learning, Multimodal ML, **LLM, Vector DB, RAG, Fine Tuning**, LoRa, Q-LoRa, **PEFT**, **AI Agents**, Distributed Systems, HPC

EDUCATION

Northeastern University, Boston, MA, USA **April 2025**
Master of Science in Computer Software Engineering
Relevant Coursework: Object Oriented Design, Data Structures, **Algorithms**, Software Engineering, **Generative AI**, **Design Patterns**, **NLP**