Rajesh Katta

Machine Learning Engineer

kattarajesh2001@gmail.com +918309009504 India Portfolio LinkedIn Github

Medium

PROFILE

AI Engineer with 3 years of hands on experience in Generative AI, LLM finetuning, and Multimodal AI R&D and deployment. Skilled in designing, training, and optimizing deep learning architectures for Computer Vision, NLP, and cross modal understanding. Passionate about advancing AI research through practical innovation, efficient model training, and large scale deployment.

PROFESSIONAL EXPERIENCE

Machine Learning Engineer 05/2024 – Present
Vassar Labs Hyderabad, India

- Built multitask Al engine supporting classification, detection, and segmentation across domains.
- Fine-tuned Qwen-3 VL 4B on Pest/Disease datasets using LoRA/PEFT and integrated a RAG pipeline for Pest/Disease monitoring and remedy suggestions.
- Developed Generative Al models (GANs, Diffusion, VAEs) for synthetic agricultural data generation.
- Implemented Vision Transformers and UNet++ for building segmentation (9% Dice Score).
- Designed HydraChat - a multitask LLM handling chatting, summarization, translation, and QA via LoRA adapters on T5 and LLaMA2 models.
- Deployed RAGGPT - Retrieval Augmented Chatbot integrating LangChain, FAISS, and LLM contextual fusion for enterprise knowledge QA.
- Optimized Retina Net and YOLOv8 for edge deployment using GIoU + Focal Loss and quantization. Reduced model size by >90% via knowledge distillation while retaining performance (~70% mAP). Engineered Multimodal Transformers for vision language fusion and contextual reasoning on satellite imagery.

Deep Learning Engineer 03/2023 – 04/2024
Mevlana Technologies (Isaac Air – Stealth Mode Product) Gurgaon, India

- Developed autonomous navigation AI using vision + language fusion for robotic systems.
- Integrated LLMs with scene context understanding pipelines for natural language command interpretation. Implemented Diffusion Canvas - text to image generation and editing platform using Stable Diffusion + ControlNet.
- Created VisionGPT - multimodal model combining CLIP and GPT2 for image captioning and visual question answering.

Deep Learning Engineer Intern 01/2023 – 03/2023
Mevlana Technologies (Isaac Air – Stealth Mode Product) Gurgaon, India

- Developed NLP and vision-language features for Isaac's AI module.
- Built image-to-text and scene-understanding prototypes using CLIP, ViT, GPT-2.
- Implemented multilingual text processing and prompt-based interaction.

Natural Language Processing Intern 06/2022 – 08/2022
Indian Institute of Information Technology Surat
Developed seq2seq models for Hindi→English translation, including text normalization and a 200-word bilingual mapping to correct misspellings using mBART and Google Translate API. Achieved a BLEU score of 0.38 for the final machine translation system.

EDUCATION

Bachelor of Technology 06/2019 – 05/2023
Rajiv Gandhi University Of Knowledge Technologies
-Bachelor of Technology in Mechanical Engineering with a minor in Mathematics.

MEDIUM BLOG WRITER

Medium
Authored in-depth blog series exploring advanced concepts in deep learning and AI, with a focus on rigorous mathematical foundations.

SKILLS

AI/ML
- ML, DL, CNNs, ViT, LLMs, RAG, Transformers, GANs, Diffusion, Object Detection (YOLO, RetinaNet, FasterRCNN), Segmentation (Unet,UNet++, DeepLab), Vision-Language Models, Multimodal AI.

Frameworks
- PyTorch, TensorFlow, HuggingFace, LangChain, LlamaIndex, TorchGeo, OpenCV, FAISS

MLOps/Deployment
- FastAPI, Docker, Triton Server, ONNX, TorchScript, REST APIs, AWS, GCP, CI/CD, Model Optimization, Quantization, Distillation, MLflow, W&B, AWS, GCP.

Programming
- Python, SQL, Git, Linux, Bash, NumPy, Pandas, sklearn.

PROJECTS

GPT from scratch 11/2023 – 12/2023
Implemented and trained a mini-GPT Transformer architecture from scratch in PyTorch to learn tokenization, attention, and text generation mechanics.

AutoAgent 01/2024
Built autonomous Al workflow orchestrator using LangChain Agents + OpenAl Tools API for reasoning chains and tool integration.

HydraNet multi-task learning in facial attribute analysis 07/2023
Built multi-task ResNet3 for age, gender & race prediction with task-specific heads.

Auto-Pilot for Self Driving Cars, Deep Learning 03/2021
-The goal of this project is to apply Deep Learning principles to effectively teach a car to drive autonomously in a simulated environment.
-Focused on building a 11 layer - Convolution neural network in Pytorch that predicts the steering angle and speeds based on front-view.

MyTorch Framework 01/2020
Recreated a PyTorch like framework in NumPy with forward/backward propagations, loss functions, activation functions and custom optimizers from scratch.

CERTIFICATIONS

Machine Learning Specialization
DeepLearning.Ai

Deep Learning Specialization
Deep Learning.Ai

NLP Specilization
DeepLearning.Ai

Agentic AI NanoDegree
Udacity

ACHIEVEMENTS

- Ranked 512 out of 1200 in the SenNet + HOA-Hacking the Human Vasculature in 3D challenge on Kaggle, reflecting competitive expertise in medical image analysis.

- Ranked 884 out of 1874 in the RSNA Lumbar Spine Degenerative Classification Challenge on Kaggle, reinforcing capability in complex medical image classification.

- Ranked 1115 out of 2739 in the Skin Cancer Detection with 3D-TBP Challenge on Kaggle, highlighting proficiency in 3D medical imaging for dermatological applications.

- Ranked 203 out of 305 in the CGIAR Crop Damage Classification Challenge on ZINDI.