Muhammad Hasnain Khan

+49-163-6479119 mhasnainkhan144@gmail.com MHasnainKhan MHasnainKhan

EDUCATION

FAST Nuces Islamabad
Masters degree in Computer Science with specialization in Deep Learning 2019 - 2021
• **Coursework:** Computer Vision, GANs, Information Retrieval, NLP, Autonomous Vehicles
• **Teaching Experience:** Natural Language Processing, Neural Networks, Discrete Structures

University of Bradford
Bachelors degree in Computer Science 2015 - 2018

EXPERIENCE

FrontNow
Lead Machine Learning Engineer Sep 2023 - Present
• Led the development and optimization of fine-tuned LLMs for a multi-lingual, conversational AI platform
• Engineered a hybrid semantic search engine using a **Retrieval-Augmented Generation (RAG)** pipeline with **FAISS** embeddings and a **knowledge graph**, boosting user engagement by over **50%**
• Architected an automated data enrichment pipeline that contextualized product data to generate SEO-optimized content, improving clients' search engine rankings
• Deployed a streaming inference framework using **RabbitMQ** and **Apache Flink** to power adaptive cross-sell agents from live user signals
• Established a robust MLOps framework for continuous model evaluation and deployment, leveraging **A/B testing** and real-time performance monitoring to iterate on user-facing models and reduce inference latency by **30%**

COMPREDICT
Machine Learning Engineer III Sep 2022 - March 2023
• Worked with automotive data acquisition systems and software such as CAN bus and OBD-II
• Predicting and diagnosing vehicle faults using virtual sensors to reduce maintenance costs and increase reliability
• Worked with cross-functional teams to integrate virtual sensor systems with other automotive systems and platforms

Google
Machine Learning Engineer II July 2021 - August 2022
• Trained large language models for extracting semantic information from natural language for purposes of visual storytelling
• Devised and deployed transformer-based semantic extraction pipelines that aggregated multilingual corpora and structured them into latent scene-embeddings for **3D GAN-driven** visual storytelling, enhancing cross-modal coherence and reducing annotation overhead by **30%**

ikeGPS
Machine Learning Engineer Sep 2019 - Dec 2021
• Focused on developing machine learning models, production deployment, testing, scaling
• Experimented a **Visual Attention-based** Model in **PyTorch** for novel Webpage Object Detection formulation
• Utilized contextual information using visual features of ordered web elements extracted using **Resnet101**
• Achieved **95% accuracy** for product Price detection (8.5% above Fast R-CNN) and interpreted Attention Visualizations

Ozi Technology
Software Engineer (AR) June 2018 - July 2019
• My responsibility in Ozi Technology (Gamerz Studio) as a Software Engineer was to create architecture and implement core features and functionality in the applications by transforming design specifications into functional apps, and establishing an effective strategy and development pipeline

PROJECT HIGHLIGHTS

Building an Autonomous Vehicle
• Collaborated in a **team of 6** to control a Polaris GEM e2 Autonomous Vehicle containing a LiDAR, Radar, and camera
• Engineered a comprehensive solution that combined sensor data fusion, perception algorithms, and decision-making processes to enable the autonomous vehicle to navigate complex environments and execute tasks without human intervention
• Programmed WASD keyboard keys to accelerate and steer vehicle using a PID controller and ROS Python commands
• Deployed pre-trained **YOLOv7** to stop vehicle on detecting a pedestrian

Multimodal Emotion Recognition System for Human-Robot Interaction
• Spearheaded a groundbreaking project focused on designing a sophisticated multimodal emotion recognition system to enable seamless human-robot interaction in various contexts
• Leveraged computer vision, speech processing, and natural language processing techniques to analyze facial expressions, speech patterns, and linguistic cues, capturing the intricate nuances of human emotions
• Employed advanced algorithms, including **convolutional neural networks (CNNs)**, **long short-term memory networks (LSTMs)**, and **attention mechanisms**, to capture and model the complex temporal and spatial dynamics of human emotional expressions

• Integrated the emotion recognition system with a robotic platform, enabling the robot to perceive and respond appropriately to the emotional cues exhibited by humans, facilitating more intuitive and engaging human-robot interaction

Enhancing Mathematical Reasoning in LLMs via Self-Supervised Fine-Tuning
• Fine-tuned a **Qwen 2.5-32B** model on a curated dataset of 1,000 mathematical problems to improve its multi-step reasoning capabilities, leveraging a novel technique of forcing the model to generate "Wait" tokens to extend its reasoning process and improve accuracy
• Utilized a self-supervised learning approach where the model generates its own **chain-of-thought** reasoning steps, which are then used as training data for subsequent fine-tuning iterations
• Implemented a robust evaluation framework using the **AIME 2024** and **MATH 500** benchmarks to track the model's performance, achieving a **56.7%** accuracy on AIME 2024, a significant improvement over the baseline model
• The fine-tuned model was deployed on **Google Cloud Run**, with a focus on optimizing inference latency and throughput for real-time applications

Knowledge-Augmented Reasoning Engine via Fine-Tuned LLM
• Architected a novel framework for enhancing factual reasoning by fine-tuning a decoder only LLM with a domain specific **Knowledge Graph**
• Employed a hybrid methodology integrating **Parameter-Efficient Fine-Tuning (PEFT)** with a Retrieval-Augmented Generation (RAG) pipeline, which dynamically injected relevant sub-graphs into the model's context window
• Optimized the model's reasoning pathways using **Chain-of-Thought (CoT)** prompting on complex, multi-hop queries, resulting in a measurable decrease in factual hallucination and a significant improvement in zero-shot domain-specific question-answering tasks

Instruction-Tuned Multimodal LLM for Enhanced Scene Understanding
• Architected and instruction-tuned a **Multimodal Large Language Model (MLLM)** for complex visual reasoning tasks by integrating a pre-trained **Vision Transformer (ViT)** with a **decoder-only LLM**
• Implemented a lightweight **projection module** to align visual features with the LLM's word embedding space, followed by end-to-end **instruction fine-tuning** on a diverse dataset of image-text pairs to unlock conversational and zero-shot VQA capabilities
• Enhanced the model with **referring expression generation** and **multimodal grounding** capabilities, allowing it to not only describe image content but also to localize and reference specific objects within the visual scene in its textual output

Real-Time Multi-Sensor Fusion for Autonomous Perception
• Engineered a unified perception pipeline that leverages camera, **LiDAR** and **radar modalities** via a **cross-modal BEV-centric fusion architecture combining attention-based token-level** alignment and 3D detection heads to boost object recall by 17% in edge weather scenarios
• Architected a **temporal self-attention** module that ingests sequential sensor frames, enabling end-to-end joint detection + localization in a single deep network and reducing runtime overhead by over **20%** compared to task-separate baselines
• Deployed the system in a real-world autonomous vehicle testbed, implemented **model quantization** and mixed-precision inference on embedded GPU platforms to ensure **sub-50 ms** latency for real-time decision making

Audio-Visual Fusion for Dynamic Pedestrian Awareness
• Built a **self-supervised** audio-visual fusion system leveraging ambient footstep-sound signals and camera imaging to detect and predict pedestrian motion in real time
• Architected an attention-based multimodal network that dynamically weights audio and visual embeddings, enabling robust detection under heavy occlusion and low-light conditions achieving **LIDAR comparable** performance at significantly lower cost
• Deployed the solution on an embedded GPU platform (**Jetson Orin Nano**) integrating asynchronous sensor pipelines and quantised inference, paving the way for scalable surveillance and safety-critical robotics applications

SKILLS

• **Programming:** Python, C#, C, C++, LaTex, SQL, NoSQL, Java
• **Other Technologies:** PyTorch, TensorFlow, Keras, Scikit-learn, GIT, PySpark, Tableau, Hadoop, Unity3D, CoreML, OpenCV, MLOPS, CICD, AWS, GCP, Google Datastudio, Docker, Databricks, FastAPI, Pytest, GraphQL, Jenkins, RabbitMQ, LangChain, LlamaIndex, LangGraph