MOHAMMED KASHIF

Bangalore, India | kashifmohammed04@gmail.com | +91 9156026054 | LinkedIn | GitHub

Summary

Built and scaled GPU inference systems processing **80M+ images** at **2.5M inferences/day**; reduced LLM token costs by **60–80%** and lifted search CTR by **6%** through image relevancy and multi-modal model orchestration at scale.

Education

Indian Institute of Technology (IIT), Kharagpur Jul 2021 – May 2023
M.Tech in Computer Science and Data Processing ***CGPA: 9.11/10***
Published research in an IEEE international conference on instrumentation and intelligent systems.

Experience

JustDial Jan 2024 – Present
AI/ML Engineer *Bangalore, India*

Inference Infrastructure & ML Systems

- Architected a **distributed GPU-Accelerated Inference platform** using Kafka-based orchestration using CPU/GPU execution tiers, enabling scalable processing of **80M+ images** across 15 business verticals with independently scalable inference workers. Enabling scalable multi-LoRA inference for live processing while reducing memory footprint by ∼92%.
- Productionized high-throughput multimodal inference pipelines for 25K+ categories using SigLIP2 deep learning ONNX Runtime, on GPUs, sustaining **2.5M image** inferences/day with Milvus-backed retrieval and optimized batch scheduling.
- Fine-tuned 6 SigLIP2 classification models using LoRA-based PEFT training on 4K-6K image-text pairs to build a multimodal classification system across business verticals. Acheiving classification accuracy in the range of **86% to 98%**.
- **Optimized LLM/VLM serving pipelines** using **TensorRT-LLM** and Triton Inference Server through FP16 quantization, KV-cache reuse, and concurrent request batching, achieving **74% higher inference throughput** under multi-user GPU workloads.

Retrieval & LLM Systems

- Implemented a production **hybrid retrieval architecture** combining Elasticsearch lexical retrieval with Qdrant-based semantic ANN search, reducing LLM token consumption by **60–80%** while improving retrieval precision and response relevance in conversational workflows.
- Developed embedding-driven image retrieval and ranking systems validated through A/B testing, contributing to a **7–8% QoQ CTR uplift** and enabling zero-shot category onboarding through semantic clustering architectures.
- Architected and productionized Justdial AI Search using **multi-model LLM orchestration** and **semantic retrieval pipelines** to transform ambiguous queries into structured search intents, improving intent classification accuracy from **83% to 95%**, reducing zero-result searches by 14%, and increasing AI-search CTR/lead conversion by 6%.

Media Processing Infrastructure

- Designed and deployed an event-driven distributed **HLS transcoding** platform using RabbitMQ and FFmpeg, processing **106K videos/month** into adaptive bitrate streams with asynchronous media processing pipelines.
- Engineered scalable content-adaptive video encoding infrastructure processing **4.5M minutes/month**, **replacing AWS MediaConvert** with an on-premise distributed transcoding system to improve infrastructure control and reduce external processing dependency.

Cropin Technologies Nov 2023 – Jan 2024
Data Engineer Intern *Bangalore, India*

- Developed multi-class **crop disease classifier** using **PyTorch (ResNet9/18)** and **Hugging Face ViT** with a **two-stage hierarchical architecture**: Crop identification and crop-specific disease classification; achieved **∼99% accuracy** on PlantVillage crop classification and **89% average accuracy** on downstream disease classification tasks using ViT.

Projects

AI-Powered Multimodal Content Platform Jun 2023 – Nov 2023
Core Architect and Developer

- Architected and deployed a **zero-to-production** modular database schema and **25+ core REST APIs** in Flask; implemented secure HTTPS exposure, real-time analytics, and structured preference routing for live traffic.
- Engineered an automated data pipeline processing **3K+ daily articles**, integrating LLM summarization, multi-class categorization, and **embedding-driven image retrieval** to reduce processing overhead by **80%**.
- Built automated **GitHub Actions CI/CD** pipelines to deploy containerized **Docker** services onto **AWS** infrastructure; integrated **Prometheus** monitoring to guarantee a strict **sub-250 ms P95 latency** SLA.

Technical Skills

AI Infrastructure & Inference: TensorRT-LLM, Triton Inference Server, ONNX Runtime, GPU Inference, Quantization, KV Cache Optimization, High-Throughput Serving
Retrieval & Search Systems: Hybrid Search, RAG Systems, vector databases Elasticsearch, Dense Vector Retrieval
ML Frameworks: PyTorch
Distributed Systems: Kafka, RabbitMQ, Redis, Event-Driven Architectures
Languages: Python, C++, Go, SQL
Infrastructure & Tooling: Docker, FastAPI, Flask, AWS, FFmpeg, Git