jaijoshi031020@gmail.com

Jai Joshi

https://www.linkedin.com/in/jj0310
https://www.github.com/JaiJoshi123

Education

Georgia Institute of Technology Aug 2023 – May 2025
Master of Science, Computer Science, Atlanta, Georgia ***GPA: 3.88/4.00***
- **Coursework**: Conversational AI, Efficient ML, Human and Machine Learning, Big Data Systems, Computer Vision

Sardar Patel Institute of Technology Aug 2019 – May 2023
Bachelor of Technology, Information Technology, Mumbai, India ***CGPA: 9.72/10***
- **Coursework**: Data Structures, Algorithms, Advanced Database Management Systems, AI, Distributed Systems

Skills

Tools & Languages: Python, C++, Java, JavaScript, SQL, Dart, TypeScript, Git, AWS, Docker, Google Cloud, Apache Spark
Frameworks & Libraries: React, Node, Express, Django, Flask, CUDA, FastAPI, Redis, PyTorch, TensorFlow, Scikit-learn

Work Experience

MicroStrategy Tysons Corner, Virginia
Software Engineer Intern *May 2024 – August 2024*
- Implemented auto-completion feature for search engine within MicroStrategy's analytics platform utilizing **vector space embeddings** and efficient **in-memory indexing** (using Faiss) to accelerate and refine query suggestions
- Enhanced **search response times** by **20x** using a caching mechanism to preemptively retrieve and validate SQL from semantically similar past inquiries, ensuring accuracy and faster responses
- Developed a Cube recommendation engine using **Retrieval-Augmented Generation (RAG)** for efficient metadata management, facilitating precise Cube identification to power MicroStrategy's Auto Dashboard and BI features
- Authored an API in Spring Boot for the telemetry service, enabling secure data retrieval between two microservices (**PostgreSQL**)
- **Technologies:** PostgreSQL, Faiss, TypeScript, Spring Boot, Azure, LLMs, AWS

PricewaterhouseCoopers LLP Mumbai, India
Software Engineer Intern *Jan 2022 – Jun 2022*
- Spearheaded Oracle ERP implementation projects, demonstrating adept management of BI reporting systems for generating and scheduling reports. Streamlined Oracle HCM extracts to facilitate seamless auto data migration
- Engineered intricate technical OIC integrations, resulting in a remarkable 10% surge in automation of client-side processes
- Assessed crucial performance metrics on an ad-hoc basis, using cutting-edge big data analytics tools such as PowerBi
- **Technologies:** Java, MySQL, Oracle ERP Implementation Tools

Projects

Multi LLM Agent Debate Network — *vLLM (Inference Optimization), LangGraph, Parallel Computing*
- Designed multi-agent collaboration using LangGraph to improve LLM decision-making and coordination to predict cognitive presence in MOOC forums and achieved **90% accuracy**, boosting instructor feedback quality
- Optimized inference through **dynamic few-shot prompting**, **quantization**, **multi-GPU parallelism**, and **vLLM-based enhancements**, achieving a **10x** speedup.

Global-Dynamic Filter Pruning — *PyTorch, CUDA, Pruning, CNN Model Compression*
- Optimized CNN models by reducing storage size by **70%** and response time by **60%** through global & dynamic unsalient filter pruning scheme, quantization, custom **CUDA** kernels, and **PyTorch bindings**, maintaining testing accuracy

KGInPaint: Image Inpainting with Interactive Scene Graphs — *Huggingface Transformers, Scene Graph Generation (SGG)*
- Architected an interactive dashboard for KGInPaint, allowing image uploads, scene graph interaction, and object removal or replacement with in-painted results
- Designed a lightweight **Relation Transformer (RelTR)** for efficient triplet detection, outperforming traditional Scene Graph Generation (SGG) models
- Integrated a **DETR-inspired encoder-decoder** for scene graph generation and combined Meta's SAM with **HuggingFace's inpainting model** for high-quality image restoration

Visual Question Answering AI ChatBot — *React, Multi-Modal, ResNet-50 + BERT, Co-Attention Networks*
- Engineered a **React** dashboard featuring a **LLaMA-2 powered chatbot** capable of answering questions about uploaded images using a multi-modal attention model. Integrated **ResNet-50** for image feature extraction and **BERT** for text encoding, employing **Parallel Co-Attention** for joint image-question reasoning.

DeCluttering Research Assistant Tool — *NLP Algorithms, Collaborative Filtering, Browser Extension*
- Created a browser extension, incorporating the REST framework, which operates on 3 core principles: **BERT** algorithm for information summarization, **Latent Dirichlet Allocation algorithm (NLP)** for text classification, and **Collaborative filtering**, leveraging predictive modeling, for recommending relevant articles to expand the user's knowledge

Research Publications

1. Cataract Detection by Leveraging VGG-19 Classification Model on Retinal Images (2022, 13th ICCCNT)
2. Sign Language Certification Platform with Action Recognition using LSTM Neural Networks (2022, IC3SIS)