Kristen Pereira

26pkristen@gmail.com github.com/p-kris10 linkedin.com/in/pkris10/

Education

Georgia Institute of Technology, Atlanta, GA August 2023 - May 2025
Master of Science in Computer Science, **GPA: 4.0/4.0**
Coursework: Conversational AI, Efficient ML, Social Computing, Grad Algorithms, ML, Big Data Systems, HRI

Sardar Patel Institute of Technology, Mumbai August 2019 - May 2023
B.Tech in Information Technology, **GPA: 9.56/10.0**
Coursework: AI, Computer Vision, Advanced Databases, Distributed Systems, OS

Skills

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

Experience

Software Engineer Intern, Social by Steph, Atlanta, GA May 2024 - July 2024

- Developed an AI-driven automated audience-building feature for a digital ads simulator using OpenAI assistants API and vector embeddings, achieving **90% user satisfaction** with generated tags
- Achieved a **30%** improvement in application performance by implementing **custom CUDA Kernels in C++** for existing ML models and writing pytorch bindings for the same
- Set up CI/CD pipelines and deployed models as serverless functions on **Google Cloud**, using **Pub/Sub** for asynchronous requests and containerized the system, reducing deployment time by 40%
- **Technologies:** Linux, CUDA C, FastAPI, NextJS, GCP, GitLab CI/CD, Redis, PostgreSQL, Docker, Pytest.

Software Engineer Intern, Skinzy Software Solutions, Mumbai October 2021 - June 2022

- Developed APIs for PyTorch-based vision models, handling image data preprocessing and inference. Migrated existing services to AWS Lambda, **reducing costs by 20%**.
- Collaborated with cross-functional teams and integrated stakeholder feedback to drive the refactoring of the website(React) and mobile application (Flutter).
- Optimized ML models to have **40% less storage** size and **60% less response time** using **pruning and quantization and custom CUDA kernels**.
- **Technologies:** PyTorch, ReactJS, AWS Lambda, S3, CloudWatch, ONNX, Docker, Git, Postman, Jira.

Projects

Dynamic Resolution Input for DeIT in HuggingFace Transformers

- Contributed to the HuggingFace Transformers library **(150k stars and 25k forks on GitHub)**, by adding TF and torch code to interpolate position embeddings in DeIT transformer model thus enabling dynamic input image resolutions, also wrote unit tests for both implementations.

Smart Healthcare Diagnostics Using Federated Learning

- Engineered a **full-stack web application** that enables healthcare institutions to securely collaborate on CNN model training via **federated learning**, preserving sensitive data privacy, while also supporting real-time inference and progress visualization across worker nodes. Tools used : **Flask, React, Flower, TensorFlow, WebSockets, AWS S3, AWS EC2**

Token Compression in RAGs for Inference Cost Reduction

- Developed a Python script reproducing TCRA-LLM, achieving **30% token reduction of the retrieved context in RAG systems** while maintaining the accuracy of the model thus significantly lowering operational costs when using paid LLMs. **Technologies used: LLamaIndex, HuggingFace, Python, Tonic**

Multi-threaded Data Store Implementation

- Engineered a **multi-threaded Redis-like data store**, implementing core functionalities (e.g., PING, SET, GET), enabling data replication with **99.9% synchronization accuracy** through handshake protocols and replication IDs, and optimizing for efficient client-server communication, concurrency, and performance under load using **C++, TCP/IP, multithreading**.

Full-Stack Social Media Platform

- Created a Reddit-like social media platform with features like posts, comments, upvotes using **Javascript, NextJs, Express.js, Postgresql DB, and GraphQL**. Implemented **Redis** caching and pagination for a **REST API**, causing **40% less latency** for each request. Designed robust GraphQL schemas for complex data interactions with PostgreSQL, and integrated **Docker** into the **CI/CD pipeline** to automate builds and streamline the development workflow.

Publications

- “Audio-Visual Deepfake Detection System Using Multimodal Deep Learning,” 2023 3rd International Conference on Intelligent Technologies (CONIT), Hubli, India, 2023
- “Voice Assisted Image Captioning and VQA For Visually Challenged Individuals,” 2022 IEEE 19th India Council International Conference (INDICON), Kochi, India, 2022