Kristen Pereira

kristenp1123@gmail.com p-kris10.github.io linkedin.com/in/pkris10/ github.com/p-kris10

Education

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

Sardar Patel Institute of Technology, Mumbai **GPA: 9.56/10.0**
B.Tech in Information Technology,
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 models and APIs for 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
- Set up CI/CD pipelines and deployed models as serverless functions on **Google Cloud**, using **Pub/Sub** for asynchronous requests and containerized the system, optimizing deployment time by **40%**
- Transformed a legacy codebase into a NextJS app, cutting development time by **30%** and increasing user retention by **20%** through frontend enhancements
- **Technologies:** Linux, FastAPI, NextJS, GCP, GitLab CI/CD, Redis, PostgreSQL, Docker, Pytest.

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

- Designed APIs for PyTorch-based vision models, handling image data preprocessing and inference
- Optimized ML models to have **40% less storage** size and **60% less response time** using **pruning and quantization and custom CUDA kernels and pytorch bindings**
- Led the migration of key backend services to **AWS Lambda, cutting infrastructure costs by 20%** while ensuring scalability and high availability
- Integrated **AWS CloudWatch** for real-time performance monitoring and automated alerting, ensuring system reliability and prompt issue resolution
- Reduced deployment time **35%** via **GitLab CI/CD** optimization and enhanced test automation, improving system reliability
- **Technologies:** PyTorch,CUDA C, ReactJS, AWS Lambda, S3, CloudWatch, ONNX, Docker, Git, Postman, Jira.

Projects

Dynamic Resolution Input for DeIT in HuggingFace Transformers ↗

- Enhanced Vision AI models in HuggingFace library **(150k stars and 25k forks)** through open source contribution

Token Compression in RAGs for Inference Cost Reduction ↗

- **Architected** a Python implementation of TCRA-LLM using **LLamaIndex**, **HuggingFace**, and **Tonic**, achieving a 30% token reduction in RAG systems while maintaining model accuracy and optimizing operational costs for paid LLMs by reducing retrieved context

Dynamic Quantization of Large Language Model ↗

- Extended Meta's Fairseq library to support CPT ↗ and implemented both post-training quantization and **quantization-aware fine-tuning** on RoBERTa model
- Optimized multi-GPU communication and pipeline parallelism, improving training throughput and minimizing memory overhead. Technologies : **Meta's Fairseq, PyTorch Profiler, NVIDIA Visual Profiler**

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
- Built support for real-time inference and progress visualization across worker nodes. Tools used: **Flask, React, Flower, TensorFlow, WebSockets, AWS S3, AWS EC2**

Multilingual Text-based Image Search

- Built a stock image search platform that achieved **85% CTR** on the first page, using **multilingual knowledge distillation and cosine similarity** for cross modal retrieval. Tech stack : **React, TensorFlow, ONNX, Flask, Heroku, Docker**

Publications

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