Rishav Raj

Dallas, TX | rishavr@smu.edu | +1-469-961-3462 | https://rraj.me

Education

Southern Methodist University — **Graduation date:** May 2024
MS Computer Science (AI/ML Specialization)

Chandigarh University — **Graduation date:** May 2019
B.E Computer Science & Engineering

Work Experience

AT&T Center for Virtualization
Dallas, TX

MLOps Engineer (Graduate Research Assistant) — March 2023 - May 2024

- Co-authored research paper on Generative AI & Synthetic Data to create customized Face Recognition Dataset using Stable Diffusion and Control Nets, offering enhanced guidance. Dataset featured on BiometricUpdate.
- Led the development of a PyTorch training and inference pipeline, achieving a 90% decrease in time consumption by training large models in a distributed environment on HPC.
- Accelerated inference on DGX-A100 by utilizing quantization and multiprocessing for parallel inference, achieving up to 5x faster inference times while maintaining model accuracy.
- Built a Serverless data ingestion pipeline on AWS, facilitating seamless data ingestion for the project, using file chunking, and AWS S3 Acceleration, AWS Lambda, allowing users upload large files in S3, with ease.

Research Paper: SIG - A Synthetic Identity Generation Pipeline for Generating Balanced Datasets for Face Recognition*

Boltzmann Labs (Startup)-AI for Drug Discovery
Bangalore, India

AI/ML Engineer — Nov 2019 - May 2022

- Performed feature engineering, selection utilizing techniques such as RFE and Ridge Regression on high-dimensional multimodal large dataset.
- Built machine learning and deep learning models like XGBoost, Variational Autoencoders (VAEs), Attention Networks etc. optimizing hyperparameters with Optuna.
- Designed architecture training pipelines on the cloud (Google Cloud Platform) using MLFlow, for experiment tracking and model monitoring.
- Created a CI/CD pipeline using Jenkins to streamline model deployment, integrated with efficient testing and versioning.
- Streamlined deployment process leading a 3-person team, to deliver ML Models as containerized services using Docker and Kubernetes.
- Collaborated with colleagues to develop and deploy RESTful APIs, enabling seamless model accessibility over the internet, which further helped secure funding of $110,000 for the company.
- Engineered high-performance Message Management System with RabbitMQ and Celery, accelerating task ingestion and boosting productivity metrics, by lowering manual effort by 80%.

Projects

Debate Bot
https://code-the-vote.devpost.com/

- Led my team and WON 1st prize in the hackathon, by designing gamified digital experiences to combat voter suppression and expand voting access.
- Built an interactive chatbot using LLM. Currently scaling it and finetuning the LLM model to increase response accuracy & efficiency, along with trying out RAG for the task.

Exploring Models for Supervised Hypernym Discovery | Python, Pytorch, NLP

- Conducted in-depth research on Hypernym discovery strategies to drive project success.
- Explored and implemented models for supervised hypernym discovery, incorporating CRIM and SPON architectures with novel methods, leading to a 15% improvement in precision and recall over baseline models.

VR: Capture & Replay in Unity | Unity, C#

- Developed a robust tool capable of capturing VR device data and replaying it deterministically within Unity.
- Main objective is to enhance the development process by eliminating the need to constantly rebuilding.

Skills

Languages: Python, C/C++

Web Technologies: JavaScript, ReactJS, Typescript, RESTful, GraphQL,

Frameworks: Pytorch, Pytorch-Lightning, Spring Boot, Spring3, jUnit, Tensorflow, HuggingFace, Serverless, FastAPI

Database: SQL, MongoDB, Redis, DynamoDB

Developer Tools: Vim, Git, RabbitMQ, Celery, Docker, Kubernetes, CI/CD, GCP, Jenkins, AWS

Additional: LLMs, RAG, DevOps, MLOps, CUDA, Debugging