Risan Raja

risan.raja@icloud.com | 963-332-6718 | github.com/risan-raja | linkedin.com/in/risanraja

Objective: Seeking a Research-Focused Machine Learning Engineer position to utilize 4+ years of experience and a strong aptitude for problem-solving in developing and implementing innovative solutions within a dynamic and challenging environment.

Work Experience

Center for Computational Brain Research, IIT Madras

ML and Computer Vision Researcher **2024-2025**

- Developed Automated Image Registration and Stacking for Histological Images. Repurposed various **SOTA MRI** based registration pipelines for histological images by optimzing it to handle large histological images (200GB+/image).
- Developed a novel optimzed pipeline for finding errors within the annotated anatomical structures by adopting the best practices from geospatial data processing.
- Implemented an end to end ETL pipeline using **Dask and Pytorch** which leveraged the existing anatomical data stored in geospatial format to stream asynchronusly **100K patches/s** of large histological for training the deep learning models.
- Contributed to the development of **RAG based QA** system for Neuroscience based projects.
- Strategised and implemented model serving using in house DGX A100 cluster using **NVIDIA Triton Inference Server** for the developed models.
- Developed automated CI/CD pipeline to integrate with the frontend using Jenkins and Docker
- Contributed in deploying in-house code suggestion tool using **AST based code generation** for the research team using Mistral 7B. Supplemented it with subject based knowledge graph to further enhance the code suggestion. Worked with the front end team to develop the chrome based plugin for the code suggestion tool.

IBG Consulting

Business Analyst **2021-2024**

- Contributed to strategic decision-making by developing a machine learning model leveraging **PyTorch and NIXTLA**. This model analyzed commodity options and futures market trading patterns to assess the viability of sales and export demand in international markets relevant to the company's raw material procurement and improved the turn around time of the warehouse stock by 20%.
- Fine Tuned **T5** based model using PEFT to eliminate jargon within the published analyst reports to make available corpus more streamlined for downstream tasks improving the pipeline efficiency by approx 30%.

Coderstrust

Digital Marketing Strategist **2017**

Nielsen Sports

Digital Analyst **2016**

Awards

First Place, PixelMind AI Hackathon **2023**

- Created an AI Agent using **RL** by leveraging the MAXIM and SPLINET for automatically enhancing high-resolution photgraphy images.

Third Place, DSA Challenge, IITM **2023**

9th Place, WorldQuant Alpha Challenge **2023**

- Developed Simulated Annealing based Genetic Algorithm which used the PnL generated by the Alpha as a heuristic. The automated alpha generation used AST based code generation to meet the competition requirements.The guided search algorithm further refines the alpha using **SGD** based optimization.

Education

B.S. Data Science and Programming, IIT Madras **2024**

- TA for the course C Programming

BSc Information Technology, SMU **2018**

Technical Projects

Pointwise Temporal Fusion Transformer

- Redesigned Temporal Fusion Transformer from ground up to handle multi horizon prediction in non-stationary financial data. The novelty is in the ability to perform like a deep time index model for time series forecasting tasks. Custom training loop was developed to handle the large training and utilized 8 A100 GPUs to train the model in 3 days. Finally outperforming the best solution published in the original kaggle challenge. Rewrote the code from **Tensorflow** to **Pytorch Lightning** to leverage FSDP for distributed training.

ONDC Indexing System

- Retrained existing **JINA** LLM embedding model to handle indian categorical data for the ONDC project. The model was optimized to handle the large scale data and was deployed using Google Kubernetes Engine(GCP) and also Vertex AI endpoint using GRPC. The model was also further optimized using attention layer fusion to optimize for large scale data. Further also created custom docker image to handle the model serving using NVIDIA Triton Inference Server.

Sparse Embedding Transformer Model Optimization

- Optimized the existing **NAVER SPLADE** model based on BERT to embed categorical information using max-pooling. Reused weights from a model which was trained on a contrastive learning task to further optimize the model. Main optimization objective was to keep the VRAM usage to a minimum.
  Engineered the model in two parts to handle both document and query embeddings to further optimize the model for the large scale data.

Skills

Languages Python, JavaScript, C

Frameworks TensorFlow, PyTorch, Scikit-learn, Pytorch Lightning, Git, Flask, Django, ONNX, GCP, AWS, Kubernetes, Docker, Dask, PySpark, NIXTLA, JINA, TF Serving, NVIDIA Triton Inference Server