Risan Raja

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

PROFILE SUMMARY

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   **Jun 2024 - Mar 2025**

- Led the development of Automated Image Registration and Stacking for histopathological Images. Repurposed various **SOTA MRI** based registration pipelines for histopathological images by optimizing it to process large histopathological images (200GB+/image) using **deep learning based optimization(LDDM)** to improve the registration accuracy by over **70%** and sped up annotation speed by **3x**.
- Formulated a novel optimized pipeline for finding errors within the annotated anatomical structures by adopting the best practices from **geospatial data processing**. Reduced the time taken from **5 minutes to under 15 seconds**.
- Developed an end to end ETL pipeline using **Dask and Pytorch** which leveraged the existing anatomical data stored in geospatial format to stream asynchronously **100K patches/s** of large histopathological for training **GANs and Stable Diffusion networks** for image generation and segmentation.
- Augmented the development of **RAG based QA** system by reducing the latency of the ORM to query and interact with the knowledge base by **21%**.
- Strategized and implemented inference serving using in house **DGX A100 cluster(5 Nodes)** using **NVIDIA Triton Inference Server**. **Apache Airflow** was used as the main orchestrator.
- Assisted in developing the in-house code suggestion tool using **AST based syntax generation** for the research team using Mistral 70B. Supplemented it with subject based knowledge graph to enhance the code suggestion. Worked with the front end team to develop the chrome based plugin to be used in JupyterHub.

IBG CONSULTING..

Business Analyst   **Aug 2020 - May 2024**

- Contributed to strategic decision-making by developing machine learning models leveraging **PyTorch and NIXTLA**. These models continuously 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%**.
- Finetuned **T5-based LLM** using PEFT to eliminate jargon within the published analyst reports, in return enhancing the available corpus more streamlined for downstream tasks improving the pipeline efficiency by approx **30%**.

EDUCATION

B.S. Data Science and Programming, IIT Madras   **2021-2024**

- TA for the course C Programming

B.Sc Information Technology, SMU   **2015-2018**

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

Domains   NLP, Non-Stationary Time Series Modeling, Computer Vision, Computational Geometry, Image Registration, Non-Linear Optimization, Bayesian Optimization

Awards

1ˢᵗ Place, PixelMind AI Hackathon **2023**

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

3ʳᵈ 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 alpha generation** to meet the competition requirements.The guided search algorithm refines the alpha by using **SGD** as one of its heuristics.

Technical Projects

Pointwise Temporal Fusion Transformer..

- Revamped the Temporal Fusion Transformer from scratch to support multi-horizon predictions in non-stationary financial datasets. Its key innovation is operating as a deep time index model for time series forecasting. Custom training loop was developed to ingest 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. Converted the project from **Tensorflow** to **Pytorch Lightning** to leverage DDP for distributed training.

ONDC Indexing System..

- Fine tuned existing **JINA** LLM embedding model to semantically index 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. It was also additionally optimized using attention layer fusion to optimize for large scale data.Additionally, created a custom Docker image for 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.