Vijay Murari Tiyyala

✉ vtiyyal1@jh.edu | ☎ +12405711678 | vijaymuraritiyyala | iMvijay23 | Baltimore MD

Objective: A highly motivated Master's student specializing in Machine Learning, Data Science, and Natural Language Processing. Aiming to contribute to research and development roles that challenge my problem-solving and technical skills.

Technical Skills

- Programming Languages: Python, Java, R, C++
- Model Training and Deployment: PyTorch, TensorFlow, HuggingFace, SLURM, Deepspeed, Spark, Docker, kubernetes
- Data Management and Operating Systems: SQL, Apache Solr, Airflow, Linux, Shell Scripting, Git, Tableau
- Web and Cloud Technologies: HTML/CSS, PHP, Google Cloud Platform, Azure, AWS

Education

Master's in Computer Science - ***Johns Hopkins University,*** *Baltimore* Aug 2022 – Dec 2023
Focus: Machine Learning, Data Science, NLP, Database, Information Retrieval, Statistics GPA: 3.6/4.00

Work Experience

NLP Research Assistant - ***Center for Language and Speech Processing*** Aug 2023 – Present

- Developed an empathic medical advisory chatbot utilizing LLAMA2 to assist healthcare professionals in patient interaction.
- Employed Apache Solr for data indexing and trained LLAMA2 with deepspeed on multi-GPU cluster for efficient model fine-tuning.
- Fine-tuned LLAMA2 using Direct Preference Optimization(DPO) an RLHF technique, for evaluating responses generated by trained model.

NLP Research Assistant - ***Center for Language and Speech Processing*** Jun 2023 – Aug 2023

- Collaborated with Prof. Mark Dredze to architect and implement a Retrieval Augmented Generation (RAG) chatbot. Utilized LLM's to allow real-time querying across 1M+ articles.
- Integrated Apache Solr Cloud for data indexing, achieving a 2x increase in query performance, critical for the chatbot's users.
- Pioneered data augmentation strategies by generating synthetic data, enabling effective fine-tuning of LLAMA2 language model.
- Fine-tuned LLAMA2 using Parameter-Efficient Fine-Tuning (PEFT), LORA, and QLORA significantly reducing compute costs.
- Deployed the model as a chatbot in an end-to-end application using Docker and FastAPI.

Machine Translation Research Assistant - ***Center for Language and Speech Processing*** Jan 2023 – Jun 2023

- Initiated and led a project focused on improving machine translation of medical terminologies, thus democratizing access to crucial healthcare information in low-resource languages.
- Designed and deployed automated web scraping tools, amassing a database of over 15,000 medical terms.
- Engineered a robust translation pipeline capable of handling thousands of terms across 300 languages, while implementing advanced compound-splitting algorithms to increase translation accuracy by 25%.

Business Technology Analyst - ***Deloitte USI*** Jul 2021 – Jun 2022

- Served as an SQL Developer in the database team, playing a pivotal role in database management, Visualisation,bug resolution, and script development for API data delivery.
- Optimized SQL-based stored procedures, achieving a 20% reduction in processing times for tax data management tasks.
- Cultivated partnerships across departments to execute and maintain enterprise-level data management solutions, contributing to a 30% increase in client retention rates.

Technical Projects

CodeTalk - ***Code Editing via Natural Language Instructions***

- Compiled a comprehensive dataset for training language models aiming to assist in code editing tasks. Utilized Dijkstra's algorithm to identify closely related codes based on edit distance, improving the quality of training data to Instruction fine-tune CodeLlama.

AcademiSearch - ***Research Paper Search Engine for CLSP***

- Developed a full-stack web application to query a database of over 3,000 research papers published by CLSP.
- Enhanced user experience and understanding by integrated ChatGPT for generating concise summaries of each paper.
- Achieved efficient search and retrieval capabilities by transforming paper summaries into vectors using the BERT model and storing them in a vector database.
- Implemented article clustering based on similarity metrics and nearest neighbors, categorizing using topic modeling techniques.
- The project serves as a comprehensive solution for researchers at CLSP to quickly find relevant articles, and showcases skills relevant to software engineering, data engineering, data science, and machine learning roles.

ImageClassify-VT - ***Unsupervised Image Classifier with Vision Transformers***

- Crafted an innovative image classification model using Meta's DINOv2, attaining a remarkable 86% ImageNet accuracy.