OJAS DESHPANDE

◇ osdeshpande@cs.stonybrook.edu ◇ +1(510)714-7690 ◇ LinkedIn ◇ Github ◇ Portfolio

EDUCATION

Master of Computer Science, Stony Brook University **May 2025**
Relevant Coursework: Distributed Systems, Network Security, Deep Learning. **CGPA: 3.8**

Bachelor of Computer Science, Manipal Institute of Technology **August 2022**
Relevant Coursework: Operating Systems, Object Oriented Programming, Data Structures. **CGPA: 8.76**

SKILLS

Languages	Python 3, SQL, C/C++, JavaScript, TypeScript, HTML5, CSS, Java, Golang.
Frameworks & Libraries	PySpark, Pandas, HDFS, TensorFlow, Pytorch, Numpy, scikit-learn, NLTK, Transformers.
Developer Tools	VS code, IntelliJ, Git, Docker, Jira, Shell scripting, Eclipse, Azure, AWS, Agile methods, Linux.

EXPERIENCE

Graduate Teaching Assitant September 2024-Present
Stony Brook University *Stony Brook, NY*

- Designed and graded assignments on **PyTorch**, transformers, and word embeddings (**Glove, Word2Vec**) for an NLP course.

Graduate Researcher January 2024 - Present
HLAB *Stony Brook, NY*

- Created **LLM** based embeddings using parallel training in **Pytorch** to efficiently process more than 1 billion tweets, which was tokenized and trained on multiple **GPUs** using **Distibuted Data Parallel** package.
- Trained large-scale language models using PEFT with a **Ridge regression** layer for predicting user-level depression scores based on PHQ-9 criteria. Aggregated embeddings from billions of **Tweets** to capture user behavior and linguistic patterns
- Designed and executed scalable cross-validation pipelines on **MySQL** tables using **SQLAlchemy**, optimizing predictive models with Ridge Regression and leveraging RoBERTa embeddings for feature engineering.
- Engineered spatio-temporal processing of 1 billion tweets using **Spark** on **YARN**, leveraging **SparkSQL** with partitioning and bucketing to execute heavy queries, improving data processing efficiency by 20%.
- Optimized data transfer between HDFS storage servers and GPU compute clusters using **Apache Arrow Flight**, **gRPC**, and **Linux** commands, reducing embedding generation times by 90%.

Associate Developer January 2022 - July 2023
SAP *Bengaluru, IN*

- Deployed ML models into production within SAP systems, integrating predictive analytics for insights on employee attrition and sales forecasting using **Python** and **XGBoost**.
- Evaluated existing AI systems by integrating **A/B testing** results with predictive models, using **Pandas** and **scikit-learn** to analyze KPIs like loading time and user engagement, assessed impact predictions for the new feature.
- Implemented parallel test execution in **Jenkins** using **Maven** and **Docker** containers for a **CI/CD** pipeline, reducing testing time from 45 to 15 minutes in build cycles, resulting in seamless build and deployment.

Research Intern March 2021 - September 2021
World Bank Research Group

- Developed an NLP pipeline using **Word2Vec** embeddings and **K-means clustering** to group company names from legal case **HTML** files. Extracted top cluster keywords and matched names to registered entities using **cosine similarity**.

PROJECTS

LLM-enabled Q/A system | *Hugging-face, transformers* February 2024 - April 2024

- Achieved significant improvements in question-answering accuracy by finetuning **OpenAI's GPT-2** and **LLaMA 13B** on the BoolQ dataset, leveraging the **LoRA** technique for efficient parameter tuning.
- Optimized model training and inference by deploying the system on **GCP**, implementing model parallelism on **GPU-based** instances with **CUDA** for efficient large-scale model training and parallel processing, significantly reducing training time.

Language Model with Attention Mechanism | *Python, TensorFlow, Pytorch* Feb 2024 - Mar 2024

- Implemented a language model combining **LSTM** networks with an **attention** mechanism to enhance sequence modeling, enabling dynamic input weighting for improved performance in **NLP** tasks like machine translation and text summarization.

Image Classifier | *Python, PyTorch, TensorFlow* February 2024 - April 2024

- Utilizing Residual Neural Network (ResNet) architecture, this project classifies CIFAR-10 dataset images. Implemented using **Python** libraries **PyTorch** and **TensorFlow**, used **Git** for version control