Nikhil Balwani

✉ nikhil.balwani@columbia.edu in nikhilbalwani nikhilbalwani nikhil.info ☎(646) 431-7195

Education

Columbia University **Aug 2022 – Dec 2023**

Master of Science in Computer Science; Machine Learning Track GPA: 4.12 / 4.33

- Relevant Courses: High-Performance ML, Natural Language Processing, Conversational AI, Distributed Systems

Ahmedabad University **Sep 2016 – May 2020**

Bachelor of Technology in Information and Communication Technology with Summa Cum Laude GPA: 3.60 / 4.00

- Relevant Courses: Data Structures and Algorithms, Software Engineering, Machine Learning, Artificial Intelligence
- Honors: Merit - Full Tuition Fee Scholarship (2019, 2020), Scholastic Distinction with “Highest Excellence” (Ranked 1/60 students)

Technical Skills

Languages: Python, Java, SQL, C, C++, R, Go, HTML/CSS, LaTeX, Bash.
Developer Tools: GCP, AWS, Google Cloud Shell, VS Code, Eclipse, Google Cloud Platform, Jupyter Notebooks, Git.
Databases: MySQL, PostgreSQL, GCP Spanner, MongoDB, Elasticsearch, BigQuery, RDBMS, HBase.
Technologies/Frameworks: Spark, TPU Training (GCP), DataFlow (GCP), Compute-Engine (GCP), Tensorflow, NumPy, Pytorch, scikit-learn, Flask, RESTful APIs, Apache Beam, MATLAB, CUDA.

Experience

Amazon **Seattle, WA**

Software Engineer Intern in ML Jun 2023 – Aug 2023

- Developed an LLM-based re-ranker for complementary recommendations using SentenceBERT and cosine similarity. Used FAISS cosine similarity index to speed up the search process by 10x.
- Architected a scalable solution to optimize an RNN-based inference API by 3x on AWS Batch Transform.
- Devised a real-time (latency $<$ 60 ms) use-case for RNN recommendations to enable interactivity on amazon.com.

Infocusp Innovations Private Limited **Ahmedabad, India**

Machine Learning Engineer, Confidential Project Jan 2021 – Jul 2022

- Managed the complete life cycle of 50+ Transformer LLMs - data preparation, model creation, and training on Google Cloud TPUs, and deployment of APIs on GCP Virtual Machines. Built an automated evaluation pipeline on GCP.
- Enhanced the accuracy of Transformer models by 30% points using a structure-aware attention mechanism for translation tasks. Exact match accuracy increased by 4%.
- Wrote scalable big data scripts in Apache Beam on Cloud Dataflow runner to process 100 million+ samples at scale.
- Created a document similarity engine API based on SimHash and Multi-Indexed Hashing that can search 21 million documents in under 150 ms.

Embibe - AI in Education **Bengaluru, India**

Data Scientist May 2020 – Jan 2021

- Implemented the organization’s first Knowledge Tracing model called Bayesian Knowledge Tracing (BKT) based on an HMM (Statistical Machine Learning) - the backbone of two different APIs.
- Architected an end-to-end data pipeline for BKT models to automatically update concept mastery scores using Spark.
- Trained an LSTM model for concept mastery - which led to an AUC performance gain of 0.21 on the validation set.
- Implemented an in-depth, scalable, and reproducible analysis of 10 million+ student attempts for Test-on-Test student performance and concept mastery improvement.
- Wrote an internal research article capturing 15 different Knowledge Tracing approaches in the literature - classified under Bayesian and Non-Bayesian techniques - which proved beneficial for new joiners for a comprehensive study.

Publications

- N. N. Balwani, D. K. Patel, B. Soni, and M. Lopez-Benitez **Long Short-Term memory-based spectrum sensing scheme for cognitive radio**, 2019 IEEE 30th Ann. Int. Sym. on PIMRC, Istanbul, Turkey, Sep. 2019.
  Used an LSTM-based approach to predict spectrum occupancy in Cognitive Radio Networks for 5G Wireless Systems.

Projects

Learning Universal Sentence Embeddings | *Python, Numpy, PyTorch, Deep Learning*

- Designed and implemented a teacher-student training strategy to achieve 0.67 Pearson Correlation score on the STS benchmark with just 4 million parameters.

MLify - Machine Learning from Scratch | *Python, Numpy, Machine Learning*

- Bare-metal implementations of some common supervised classifiers (Feed-Forward Neural Network, Decision Tree, Random Forest) and unsupervised clustering techniques (K-means, Gaussian Mixture Model using EM).

Achievements

- Won the first prize in Amazon F2/SL Hackathon for our project on “Real-Time Outfit Builder” during my internship.
- Achieved a perfect score of 600/600 and in the 99th percentile on CodeSignal’s Industry Coding Framework.