Anshita Saxena

anshita333saxena@gmail.com +1 (514) 569-5243 Anshita Saxena

Summary: I am a self-motivated result-oriented person and diligent person which is demonstrated by several awards I earned at companies such as IBM and Ericsson in machine learning, data science, and big data domains. I have knowledge and experience from Big Data to Machine learning to deep learning and over 6 years of working experience. I am open to relocate to anywhere across Canada.

Work Experience

Ericsson Global AI Accelerator Lab Montreal, Canada
Machine Learning Engineer (Internship) (Jan 2024 - Apr 2024)

- Developed end-to-end RAG based LLM application which got accepted as part of Ericsson Developer Conference 2024. This application uses langchain and ray for scalability, GPT-NeoX-20b model as LLM, fast API for model serving, streamlit for frontend and huggingface for quantization, LoRA, and QLoRA. Awarded as best implementation with technical insights presenter at Ericsson Global AI Lab Montreal.
- Executed models on the distributed multi-node cluster using a Distributed Sampler for data distribution. Implemented a novel Distributed decentralized sparse model using Resnet18 as the base model.

Hydroquebec Research Institute (IREQ) Montreal, Canada
Applied Deep Learning Researcher (Internship) (May 2023 - Dec 2023)

- Conducted a Research Project focused on developing a virtual sensor for estimating displacements in high-frequency turbine labyrinth seals. This involved leveraging Deep Learning techniques to extrapolate high-frequency characteristics from low-frequency measurements on time-series dataset.
- Proposed and developed a novel and successful way to employ the WaveNet model, which is generally used for Speech Generation, for generating signals from a rotating machinery (time-series dataset) to address a complex problem statement characterized by diverse information sources operating at various levels. Ran on a super-computer (cluster of GPU nodes). Improvements done on BiLSTM.
- Used MAE as a loss function, initial training loss was 7.5 which was reduced to 0.7. Generated signals overlapped the target signals for about 89% of test data. All models were implemented in PyTorch, JAX, and Keras. Total of 4 million datapoints for high-frequency data in comparison to 10k datapoints for low-frequency data.

IBM Bengaluru, India
Data Scientist/Big Data Engineer (Full Time) (Apr 2018 - Aug 2022)

- Achieved IBM Excellence and Eminence Award 2019, IBM Significant Business Impact Award 2020, IBM Service Excellence Award 2021, and IBM Growth Award 2022.
- End-to-end pipeline for a loan prediction machine learning model on our IBM Cloud Pak for Data using IBM Watson, showcasing a seamless process from ingesting raw data into S3 buckets to model construction. Leveraging PySpark to develop and build containerized models, including logistic regression, XGBoost, and Ensemble methods (Decision Trees, LGBM, Random Forests). Used Shapley and Lime to analyze the feature importance of the features. Visualization dashboards for clients using Palantir for IBM Cloud Pak for Data provided comprehensive insights. The implementation covered data from 8 different bank clients with 1.2k customers, where we built the Canonical Model to refine data into one format, aiding in identifying potential loan defaulter customers using Random Forest and XG-Boost models. Additionally, a White Paper for the Practice Department was authored, responding to a request from senior leadership. Led the project responsibilities with senior data scientists and data architects.
- Worked on a time-series forecasting model using LSTM to predict future revenue for the products. Batch inference predictions were served daily on the dashboard.
- Developed microservice RESTful APIs for Kafka ingestion and NoSQL database retention on a multi-tenant highly available distributed architecture. Processed 2600 messages/min at 500 concurrency level in a multi-datacenter mode. Innovatively developed a microservice API on Kubernetes, facilitating the transmission of error and exception messages from model microservices to a Slack channel. This not only streamlined early alerts and remediation but also significantly reduced the monitoring workload by 60 hours.

- Contributed significantly to cost savings by developing automation scripts for data extraction, manipulation, and reposting into Kafka using Python. The reposting script, integrated into the Production system, played a pivotal role in timely package tracking, resulting in a cost saving of 2 million dollars. Use Spark (cassandra-spark connectors) for data extraction, data manipulation, and data transferring for large datasets.
- Successfully managed the migration of data from one ScyllaDB Cluster to another using Spark. Employed PySpark for data transformation and Apache Spark (datastax spark-cassandra-connector Scala framework) for data extraction. The processed data, amounting to 21 TBs, was handled on a highly available distributed cluster with 2-factor node replication, showcasing a robust and scalable approach. Data Ingestion of 7TB from different local clients to Hadoop cluster, and ran spark jobs and python scripts for processing.

Relevant Projects/Proof of Concept/Work-Experience as Side Projects

End-to-End Large Language Model (LLM)-based Scalable Retrieval Augmented Generation (RAG)-powered Question-Answering (QA) App

- Application integrates a semantic embedding model to represent queries as vectors and utilizes a vector database for retrieving top-k relevant contexts.
- Employing a Eleuther AI's GPT-Neo 20B LLM, the project demonstrates fine-tuning in a multi-GPU cluster environment, showcasing data ingestion, context loading, fine-tuning, embedding, and indexing processes.
- App takes a new query, conducts vector similarity search, retrieves relevant contexts, and passes them, along with the query, to the LLM, which generates context-aware answers.
- Libraries and Techniques Used: Streamlit, Parameter-Efficient Fine-Tuning, Ray (for distributed LLM Fine-Tuning), Deta (To access Deta Vector Database), LangChain, PyTorch, FastAPI (To serve production-ready LLM App)

Question Answering

- Implemented complete pipeline having data cleaning, tokenization, initializing document store, retriever, reader, and evaluation. Model used: Minilm-uncased-squad2, MiniLM-L12-H384-uncased.
- Compared BM25 and Dense Passage Retriever for Reader based on Recall Evaluation Metrics for top-[1, 3, 5, 10, 20] reader and reader. Results showed that Dense Passage Retriever boosted performance by 0.1 for top-[10,20], and saturated for top-[3, 5] retrievers and readers.
- Implemented domain adoptation and fine tune on SQUAD, SubjQA, SQUAD+SUBJQA datasets to compare the performance. Results showed 25% improvement on using the SQUAD+SUBJQA fine-tuned dataset.
- Evaluation and Exact Match/F1 score for Retriever Evaluation.
- Used Huggingface, Pytorch, and Haystack for transformers model building and evaluation.

Real-Fake News Detection

- Used embeddings such as DistilBert-base-uncased, Bert-base-uncased, roberta-base, all-MiniLM-L6-v2 sentence transformer along with cosine similarity to group several categories into true and false news. Total classes: 391, which were grouped into true/false categories.
- Used SMOTE Analysis with and without attention-masks to balance the classes.
- Accuracy was improved from 76% to 84.6%.
- Used Huggingface and Pytorch for the whole implementation.

Hockey NHL Project

- Downloaded the live hosted data using Python scripts and API. Cleaned the data according to the project requirements. Created the visualizations using contours, the intention is to showcase the shot generation intensity.
- Applied logistic regression, XGBoost, Neural Networks, Ensemble methods (Decision Trees, LGBM, Random Forests), Log models in Comet ML. Used Shapley and Lime to analyze the feature importance of the features.
- Developed flask API for prediction and download model from the registry. Created first docker containers for passing the data to generate prediction service (using Flask API) and second docker container for WEB UI through Streamlit. Establish the docker communication network.

Education

Sep 2022 – Apr 2024 | **University: University of Montreal, Canada**
(MILA- Montreal Institute of Learning Algorithms)
*D*egree: Masters in Computer Science (Machine Learning), **Grade**: A, **GPA**: 3.98, **Category**: Excellent
*C*ourses: Machine Learning, Data Science, Deep Learning, Natural Language Processing, Geometric Data Analysis

Aug 2013 – Jul 2017 | **University: Dr. A.P.J. Abdul Kalam Technical University (Formerly known Uttar Pradesh Technical University)**
College: Meerut Institute of Technology, MIET Group, Meerut.
Degree: B.Tech in Computer Science and Engineering, **Percentage**: 82.26/100 [Honours (With Merit)]

Selected Honors And Awards

2024 | **Celebrating Excellence Award**: *Organization: Ericsson, Montreal, Canada*

2022, 2023 | **Diversity Award (Tuition Fees Scholarship/Exemption)**: *Organization: University of Montreal*

2019, 2020, 2021, 2022 | **Awards: IBM Eminence and Excellence Award, IBM Impact on the Business and Significant Achievement, IBM Service Excellence Award, IBM Growth Award, ScyllaDB Innovator Award**
Organizations: IBM India Pvt. Ltd, ScyllaDB (NoSQL Product Company)

2014, 2015, 2016 | **Awards: Academic Excellence Award, Codezilla Coding Award**
Organizations: Meerut Institute of Technology, Meerut Institute of Engineering & Technology, MIET Group

Skills: Soft and Technical Skills, Certifications, and Courses - Online Learning

Soft Skills | Leadership, Teamwork, Adaptability, Positivity, Interpersonal Skills, Creative thinking

Technical Skills | Tech- LLMs, Natural Language Processing, Machine Learning System Design, VS Code, IBM Watson, IBM Cloud, RHEL, Git, Docker, Kubernetes, Transformers, Nltk, Flask, Spark, Kafka, Microservices, Jupyter Notebook, Snowflake, object-oriented principles, Grafana, Prometheus, Kibana; Libraries - PyTorch, pandas, numpy, NLTK, Scikit-Learn, Streamlit, HuggingFace, CometML, Optuna, Ray, LangChain, Tensorflow
Programming Languages- Python; Databases- NoSQL, MySQL, PostgreSQL, Elastic-search

Certifications and Online Courses | Microsoft Azure Fundamentals: 2021, IBM Certified Big Data Engineer: 2020, Enterprise Design Thinking Practitioner, IBM Watson Knowledge Catalog Essentials, Cognitive Practitioner (IBM), Python for Data Science (IBM), Deep learning specialization (Coursera)

Publications

2019 | **Optimal Partition Search**
A. Saxena and A. Saxena, "Optimal Partition Search," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-8, doi: 10.1109/ICECCT.2019.8869459. Link: https://ieeexplore.ieee.org/abstract/document/8869459

2017 | **DeepCoder: An Approach to Write Programs**
A. Saxena, A. Saxena, J. Patel, "DeepCoder: An Approach to Write Programs," 2017 International Conference on Advanced Research and Innovation in Engineering (ICARIE), 2017 International Journal of Engineering and Manufacturing Science (IJEMS), 2017, pp. 9-13, Vol. 7, No. 1, Research India Publications. Link: https://www.ripublication.com/ijems_spl/ijemsv7n1_02.pdf