Aniket Saxena

✉ aniketrec.123@gmail.com ☎ +1 (587) 938-9321 Aniket Saxena 🏠 Aniket Saxena

Education

Aug 2014 – Jul 2018

University: Dr. A.P.J. Abdul Kalam Technical University, AKTU
College: Meerut Institute of Technology, MIET Group, Meerut.
Degree: B.Tech in Computer Science and Engineering
Division: First Division with Honours

Sep 2023 – Present

University: University of Montreal, Canada
(MILA- Montreal Institute of Learning Algorithms)
*D*egree: Masters in Computer Science (Machine Learning Specialization)
Courses: Fundamentals of Machine Learning, Data Science, Representation Learning, Natural Language Processing (NLP)

Achievements and Extracurricular Activities

- Submitted two Graph Neural Networks (***GNNs***)-based Patent applications to the US Patent Office, including one **first-inventor Patent application** and hence felicitated by **2023 IBM Research Invention Achievement Award**.
- Awarded **Academic Excellence Award** for exceptional performance during Academic Year 2015-2016.
- **Service Excellence Award** by IBM Systems Development Lab for the year 2022.
- **Diversity Award (Tuition Fees Scholarship/Exemption)** by University of Montreal.

Publications

2021 *V. BK, B. Ganesan, **A. Saxena**, D. Sharma, A. Agarwal, "Towards Automated Evaluation of Explanations in Graph Neural Networks," 2021 International Conference on Machine Learning (ICML) Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI (XAI). **(Link)***

2019 *__A. Saxena__ and A. Saxena, "Optimal Partition Search," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1-8, doi: 10.1109/ICECCT.2019.8869459. **(Link)***

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

Work Experience

IBM Global Business Services, India Bangalore, India
Data Scientist (Sept 2018 - July 2020)

- Built a machine learning platform from scratch for monitoring and tracking the traversed path of 120k vessels.
- Wrangled 1.5TB of vessel trading data stored in the Synapse database to train a Machine Learning regressive prediction algorithm and implemented an offline batch inference method for vessel trajectory prediction (in terms of longitudes and latitudes), which reduced the overall vessel traversal logistics time by 27%.
- Led a cost-saving initiative by devising an automated ML approach to predict the Time Charter (TC) Rate to identify the best vessel at the lowest price based on the availability of the vessels, leading to an annual cost reduction exceeding $\$70,000$ in logistics expenditures.
- Optimized the developed ML platform by implementing an explainability solution to identify influential features in the data for the prediction algorithm, resulting in improved trained ML model accuracy from 65% to 86% for both TC Rate and vessel trajectory prediction tasks.
- Integrated a data visualization framework, built using the *Dash-Plotly* library, into the platform for creating intuitive as well as user-interactive visualizations and launched the developed machine learning platform for production use, allowing 2k+ end clients to discover, on the fly, daily patterns in vessels movement and track the regular impact of vessel trades on liquidity.

Technical University of Dortmund, Germany Dortmund, Germany
Research Intern (Dec 2020 - June 2021)

- Implemented a unique post-hoc model-agnostic GNN explainability framework that doesn't necessitate retraining for generalizing explanations to new instances for node and graph classification, and link prediction tasks. It takes a trained

GNN model and some graph meta-data as input, computes layer-wise edge importance weights, and outputs an explanatory subgraph. The implemented framework is available on PyTorch Geometric. ***(link)***

IBM Research, India Bangalore, India
Research Associate (July 2020 - Dec 2022)

- Developed a framework to convert domain-specific ontologies into large-scale enterprise heterogeneous knowledge graphs. The framework processes ontologies as property graphs, models relations in the knowledge graph with a multi-layered ***(RGCN)*** model, and outputs predicted label for each domain-level entity in the property graph.
- To make the overall approach interpretable, introduced an explainability component in the framework to produce human-intelligible explanations for every single prediction being made for the input knowledge graph.
- Proposed a novel automated approach to evaluate GNN explanations to boost the reliability of predictions made by complex ML models applied to the Graph-structured data domain. Work was published in the ***Explainable AI Workshop at ICML 2021***.
- Deployed the developed framework as a production-ready API that was also employed to create a novel large-scale graph representation learning dataset called the *TACRED people* dataset, which is being used by other teams in the lab to accelerate graph ML research.
- Created an efficient, unsupervised framework for clustering heterogeneous graphs within Master Data entities. The framework partitions each graph to produce minimal sub-structures, clustering nodes with significant similarities.

IBM Research, Zurich Rüschlikon, Switzerland
Research Associate (Feb 2021 - Dec 2021)

- Developed a novel unsupervised contrastive learning-based approach for the detection of similar bugs on 3TB of unstructured bug text corpus to facilitate faster log analysis.
- The approach extracts documents from the corpus of bug text and then implements Text Variational Autoencoders (T-VAEs)-based unsupervised anomaly detection technique to detect anomalies in the bugs corpus. Subsequently, learns a latent representation using a fine-trained BERT language model to form a document embedding for each bug document, and then measures the degree of similarity between these learned document embeddings of all extracted documents using novel Word Centroids (WC) and Word Mover Distance (WMD) algorithms.
- Introduced quantization techniques to optimize this transformer-based similarity detection method, achieving a 38% smaller and 1.8× faster model and attaining 3× inference speedup and 5% improvement on both CPU and GPU.
- Upon obtaining the semantic-similarity scores, the approach clusters the similar bugs together based on semantic textual similarity indexing, which is governed by these computed scores.
- Included main measures to evaluate the approach (e.g. novel and repeated). Experiments conducted on the bugs corpus strongly recommended the 100% elimination of manual detection of similar bugs. This transition not only resulted in an impressive 85% reduction in response time to address client-reported bugs but also empowered the support team to focus solely on handling new and unique bug instances.
- Rolled out the devised method as a real-time IBM Storage Insights product feature functionality, accessible to a broad customer spectrum.

IBM Systems Development Lab, India Pune, India
Data Scientist (Apr 2022 - June 2023)

- Created a multivariate time series-based neural learning framework that utilizes generic machine learning algorithms (LSTM, GRU, and RNN) to impeccably predict run-out memory times of 300k+ storage systems in the resources pool, which resulted in the reduction of resource exhaustion time by 85%.
- Developed a novel approach to segment storage systems based on their respective predicted behaviors in terms of average growth in usable capacity over a specific forecasting window period, which in turn helped fill the missing data present in the time series.
- Designed and implemented pipelines for various stages of development, starting from extracting high-dimensional data in the form of multivariate time series (with underlying non-continuous dynamic) from a storage data lake to training ML time-series algorithms and predicting the storage systems' capacities.
- Created and optimized online inference approach to quickly serve the incoming requests for storage systems' capacity predictions at scale, leading to a 70% decrease in the framework's time to respond to user traffic requests.
- Introduced the entire framework as a fully operational machine learning-driven IBM Storage Insights product feature with low latency and high throughput, serving numerous customers' requests in real-time.

Projects

- **Contrastive Learning-based Few-shot Image Classification** ***(Link)***
  Libraries/Framework: Numpy, Matplotlib, PyTorch, and TorchVision
  - Defined a custom contrastive loss and trained a few-shot version of Siamese Networks to do n-way k-shot image classification by mapping the image similarity task into a fully-supervised classification learning task.
- **Graph Representation Learning for Classification** ***(GitHub)*** ***(Deployed App)*** ***(App Demo Video)***

Libraries/Framework: PyTorch, PyTorch Geometric, DVC, MLFlow, Optuna, and Streamlit

- Developed and deployed an end-to-end Node and Graph classification app, enabling real-time node or graph classification on different datasets (*Cora* and *ENZYMES*).
- Utilized an abstract variant of *Binomial Distribution* to compute feature and edge importance weights that result in providing explained feature mask and explanation subgraph.

- **LLM-RAG-powered-QA-App** ***(GitHub)*** ***(App Demo Video)***
  Libraries/Framework: Transformers, Ray, LangChain, FastAPI, and PyTorch
  - Fine-tuned a 20B parameters Large Language Model (LLM) in a multi-GPU cluster environment by leveraging the distributed training paradigm.
  - Developed a production-ready, scalable Retrieval Augmented Generation (RAG)-based context-aware Question Answering (QA) App that first finds contexts relevant to the incoming query by implementing fast vector similarity search within the pre-defined embedding space and then sends these contexts alongside the query to the fine-tuned LLM model to generate the answer.
  - Implemented scalable major ML workloads for contexts (load, embed, and index the contexts in the vector database) across multiple workers with different compute resources and served the LLM App in a highly robust and scalable manner.
- **Molecule Graph Generation** ***(Link)***
  Libraries/Framework: PyTorch, PyTorch Geometric, Numpy, and NetworkX
  - Implemented a minimal version of Variational Graph AutoEncoders (VGAE) to generate new molecular graphs that have a similar statistical distribution as that of the learned distribution of the training dataset.
  - Experiments were performed using the ZINC dataset consisting of 12k molecular graphs with 38 heavy atoms.

Open-Source Machine Learning Contributions

Contributed to **PyTorch Geometric** to add new features and operators such as Relational Graph Attention Networks, Clustering for Graph-structured Data using Graph Neural Networks, and Model-agnostic Explainability Framework for Graph Neural Networks.

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

Soft Skills

Leadership, Teamwork, Adaptability, Positivity, Interpersonal Skills, Problem-Solving Skills, Creative thinking

Technical Skills

Interests- Generative Modelling (Generative AI), Natural Language Processing (NLP), Time Series, Machine Vision, Learning on Graphs, MLOps

Databases- MySQL (SQL), PostgreSQL (extended SQL). **Programming Languages-** Python, LaTeX. **Framework/Libraries-** Tensorflow, Keras, PyTorch, PyTorch Geometric, DGL (Deep Graph Library), Scikit-learn, Numpy, Pandas, Dash-Plotly, Matplotlib, NetworkX, HuggingFace (*transformers*), FastAPI, Flask, MLFlow, Optuna, LangChain, Ray (distributed machine learning framework)

Other Tools- Git, Docker, Streamlit, MS Azure, Google Cloud Platform (GCP), GitHub Actions, DVC (for data, model, and code version control)

Certifications and Skill Courses

Machine Learning A-Z with Python Course (Udemy), Machine Learning (Coursera), Machine Learning Crash Course (Google), Columbia University's AI Specialization, Python for Data Science (IBM), Deep learning specialization (Coursera)