Research Statement of Kanchan Chowdhury

1 OVERVIEW

My research interests are in the broad areas of machine learning systems, database systems, and geospatial data processing. The main objective of my research is to advance the utilization and interpretation of databases and geospatial datasets in the context of machine learning, driving new innovations in this field. This involves developing end-to-end systems and cutting-edge methods to reduce computational burdens while enhancing the accuracy and interactivity of approaches that integrate databases into AI and vice versa.

2 RESEARCH CONTRIBUTIONS

AI for DB: In the initial phase of my doctoral research, I was motivated by the prospect of employing machine learning technologies to automate tasks related to database management. For example, in an ongoing interaction session of a user with a database, given the historical query logs until the current timestamp, what if we can predict the next SQL query of the user? Predicting the next SQL query enhances database efficiency through adaptive indexing, speculative query optimization, and tuning. Speculative execution and result prefetching during user think time can bypass actual query execution, enabling faster retrieval from memory. Knowing upcoming query elements, even partially, is beneficial for proactive database management. In this study [5, 6], I collaborated on a project where we proposed a method to represent SQL queries as vector embeddings and compared the performance of sequence prediction models with several recommender system baselines. An empirical analysis with two real-world datasets leads to the conclusion that, out of all the evaluated algorithms, reinforcement learning combined with a numerical reward function performed the best.

Optimizing Spatial ML Training: In my quest to identify machine learning-based solutions for data integration and analytical challenges, I observed that geospatial datasets often fail to fully harness the potential of machine learning algorithms. This observation stands in stark contrast to the prevalent application of these algorithms to other data types, such as text, images, audio, and video. Consequently, this realization steered my research focus towards enhancing the efficacy of machine learning algorithms specifically tailored for spatial datasets. Spatial datasets are large in volume by nature, and training a spatial machine learning model with spatial datasets suffers from various problems, such as prolonged training time and excessive memory consumption. Spatial ML models usually represent the target dataset in terms of a grid by converting the geographical coverage into a spatial grid of fine-grained cells. As the number of cells increases, the size of the adjacency matrix also increases, and the training time and memory usage increase proportionally. As traditional approaches fail to preserve the spatial relationships essential for models, I mitigated this effect by developing a framework [1] that intelligently merges adjacent spatial cells into larger, more manageable units. This approach significantly reduces the size of the dataset without losing critical spatial information, adhering to a user-defined threshold for information loss. The proposed technique reduces the training time and memory requirements by up to 81% and 65% respectively, maintaining prediction or classification errors within a 5% margin.

Spatiotemporal DL Framework: Another limitation of deep learning with geospatial datasets is the lack of frameworks capable of effectively handling and processing raw spatiotemporal datasets for deep learning applications, especially in spatiotemporal vector and satellite imagery data analysis. Existing deep learning frameworks in the deep learning ecosystem suffer from several limitations when they are used for implementing raster and spatiotemporal deep learning models. Raw spatiotemporal datasets must undergo extensive preprocessing before they can be turned into trainable datasets. Owing to the large scale of these datasets, the use of non-scalable geospatial frameworks for preprocessing leads to slow processing times and memory issues. The need for specialized knowledge in distributed geographic data processing systems causes developers to depend solely on datasets that are pre-preprocessed and ready for utilization. In another pivotal piece of research, I addressed this gap by developing GeoTorchAI [2–4], a comprehensive framework that facilitates deep learning on raw spatiotemporal datasets. Distinguished by its ability to directly convert raw spatiotemporal datasets into trainable forms, GeoTorchAI also supports building, training and testing spatiotemporal models for both vector and raster imagery datasets. The data processing tasks run on Apache Spark in a distributed and scalable setting. Besides, it integrates seamlessly with existing PyTorch classes, in addition to offering a user-friendly interface for complex data processing tasks.

Geospatial Visualization: Exploring spatial datasets using visualization dashboards like Tableau and Apache Zeppelin typically involves multiple interactions between the dashboard and the data system. Existing geospatial visualization dashboards struggle with large datasets, leading to slow response times and hampering user interactivity. In response, my collaborative project, Tabula [7], introduced a middleware designed to sit between the data system and the dashboard. Tabula uses a sampling cube approach to

pre-materialize spatial samples, which speeds up data processing without sacrificing accuracy. Tabula enhances interactivity in data exploration, significantly reducing the data-to-visualization time.

Optimizing DB ML Pipeline: This project concentrates on refining the data preprocessing and machine learning inference pipeline, particularly focusing on scenarios where the data preprocessing segment involves computationally intensive join queries. In a wide array of practical applications, features for machine learning inference are collected from various independent datasets or data repositories. For instance, the classification of traffic patterns for urban planning purposes necessitates the joining of data collected from diverse sources, including sensor-collected data, road network details, and public transport statistics. It is widely recognized that a substantial number of data processing tasks experience bottlenecks due to join operations involving inputs or outputs with high cardinality. To alleviate the impact of this bottleneck, I worked on co-optimizing the overall pipeline of data preprocessing and model inference. My proposed algorithm analyzes the data flow graph of the inference query and detects opportunities for decomposing the input layer of the model into multiple components and pushing the join operations down the decomposed model. Preliminary results on two datasets show that the proposed optimized pipeline can reduce the overall pipeline runtime by up to 17 times.

3 FUTURE RESEARCH DIRECTION & CONCLUSION

Broadly speaking, my research impacts the fields of geospatial database systems and machine learning in profound ways, paving the way for more efficient, accurate, and user-friendly approaches to handling large-scale spatial datasets.

DB & AI: Looking forward, I am committed to exploring deeper into the realms of databases and AI, particularly focusing on developing technologies that can either improve the involvement of databases in ML systems optimizations or optimize database systems utilizing the automation power of ML, enhancing the user experience in the intersection of databases and AI. My future research will continue to align with the evolving needs in this area, addressing the emerging challenges and harnessing the potential of new technologies, such as large language models, generative AI, and federated learning, in the database domain. I want to explore research problems such as serving DL models from relational databases by utilizing the massive parallelism power of databases and optimizing queries with complex ML-based filters where cardinality estimation is a nontrivial task.

Future Geospatial AI: Research in the intersection of databases and AI can also benefit from my skills and experiences in geospatial database technology. As an example, currently, researchers in the Database/AI community are working actively on optimizing database engines by harnessing the power of ML. However, these algorithms do not consider the exceptional features related to geospatial databases. For example, partitioning of large-scale geospatial datasets needs to be done based on their spatial proximity, and selectivity estimation of spatial operators does not follow the rules of regular equality and range filters. I can contribute to this area by utilizing my experience working with geospatial databases. Another problem that I would like to explore involves generating geospatial database queries from natural language questions. Although there are lots of efforts going on by the NLP community to synthesize regular database queries, none of these methods focus on the geospatial queries for spatial databases. Existing works on this research problem make wide use of word-to-vector models and large language models such as BERT and GPTs. However, these models lack the context of the geospatial domain, operators, and keywords used in geospatial query languages. My goal is to enable the automatic synthesis of spatial database queries from natural language questions by embedding geospatial context and keywords into large language models.

Research Grants: Future smart cities, powered by AI, will introduce novel research directions at the intersection of geospatial data and AI. Agencies such as NSF, the US Geological Survey, the US Department of Agriculture, and NASA, along with companies like Google and Esri, are actively investing in grants for this domain, while the emerging in-DB AI research is attracting grants from NSF and big companies. I am confident in securing funding from these agencies. Besides, due to handling sensitive user data, both in-DB AI and geospatial AI systems are vulnerable to inference attacks, extending opportunities for collaboration with cybersecurity research, along with the potential for more grants.

REFERENCES

[1] Kanchan Chowdhury, Venkata Vamsikrishna Meduri, and Mohamed Sarwat. 2022. A Machine Learning-Aware Data Re-partitioning Framework for Spatial Datasets. In *IEEE 38th ICDE*. 2426–2439.

[2] Kanchan Chowdhury and Mohamed Sarwat. 2022. GeoTorch: A Spatiotemporal Deep Learning Framework *(SIGSPATIAL '22)*. Article 100.

[3] Kanchan Chowdhury and Mohamed Sarwat. 2023. A Demonstration of GeoTorchAI: A Spatiotemporal Deep Learning Framework *(SIGMOD)*.

[4] Kanchan Chowdhury and Mohamed Sarwat. 2024. Deep Learning with Spatiotemporal Data: A Deep Dive into GeotorchAI. In *IEEE 40th ICDE*.

[5] Vamsi Meduri, Kanchan Chowdhury, and Mohamed Sarwat. 2019. Recurrent neural networks for dynamic user intent prediction in human-database interaction. *Advances in Database Technology - EDBT* (2019).

[6] Venkata Vamsikrishna Meduri, Kanchan Chowdhury, and Mohamed Sarwat. 2021. Evaluation of Machine Learning Algorithms in Predicting the Next SQL Query from the Future. *ACM Trans. Database Syst.* (2021).

[7] Jia Yu, Kanchan Chowdhury, and Mohamed Sarwat. 2020. Tabula in Action: A Sampling Middleware for Interactive Geospatial Visualization Dashboards. *Proc. VLDB Endow.* 13, 12 (aug 2020), 2925–2928.