Di Chai

Ph.D. Student of HKUST

Email: dchai@connect.ust.hk Homepage Google Scholar

Research Statement

Distributed computation is an essential topic and has many real-world applications. It could be broadly defined as the algorithms or systems that leverage the data and computational resources from multiple nodes to perform computations (e.g., model training). Many real-world research problems could be formulated as building effective distributed computation systems. For example, in urban computing, how to **intelligently** collaborate data from multiple nodes/regions to improve the overall learning performance; in medical/genomic data analysis, how to securely and efficiently utilize data from multiple institutions to enhance the accuracy and robustness of the analysis results; in generative model applications, how to utilize distributed computing resources to improve training and inference efficiency, etc. In the past few years, my research has focused on building **intelligent**, **secure**, and **high-performance** distributed computation systems.

Building such systems has the following challenges: 1) **Data collaboration**: Distributed data from multiple nodes are often closely related but also have distribution discrepancies. How to intelligently utilize data from multiple parties to improve overall learning performance is a challenge. 2) **Data protection**: Data from multiple nodes are often sensitive. How to ensure privacy and security during distributed computation is a challenge. 3) **High performance**: How to support large-scale data and efficiently utilize resources from multiple nodes is a challenge.

Intelligent Data Collaboration

In distributed computation, data from different nodes are closely related but also have distribution discrepancies. Designing intelligent algorithms to leverage the rich data from multiple nodes and mitigate the impact of distribution discrepancies is challenging. One representative application is the traffic prediction system in urban computing, where traffic data from different sites/regions are interrelated but also exhibit variations. To improve the effectiveness of data cooperation, we propose a distributed traffic prediction system based on graph neural networks (GNN). The main idea is to construct multiple types of graphs to capture the similarities and differences between nodes. For example, we build graphs based on traffic similarity, surrounding POI (Point of Interest) similarity, and geographical distance. We further propose a multi-graph fusion technique, such that these graphs, describing different types of similarities and discrepancies, can simultaneously contribute to distributed model training. And all the graphs are automatically fused during the learning process. Our solution significantly improves the effectiveness of distributed data cooperation. This work (C3) was published at SIGSPATIAL'18, the top conference in the field of geographic information systems (GIS), and achieved the **highest number of citations** (**300+ citations on Google Scholar**) among all papers at the conference that year. In the following study (J4), we analyzed the generality of different types of features and graphs from a system perspective and the paper was accepted by IEEE TKDE (2021).

Secure Data Protection

Another challenge in distributed computation is security. For example, in distributed medical and genetic data analysis, the data between multiple institutions is highly sensitive. It is essential to protect the private data during the distributed computation. Matrix decomposition (MF) is an essential primitive to support various real-world distributed applications, such as decomposing the genetic data from multiple parties for genome-wide association analysis (GWAS), decomposing the user profile data distributed in banks and online shopping companies for principal component analysis and linear regression modeling, decomposing the natural language processing data held by multiple parties for latent semantic analysis, etc. Our study is among the first to mathematically analyze the privacy leakage problem in distributed MF, proving that the gradients exchanged by multiple parties leak raw data. To solve this issue, we proposed a secure MF scheme based on homomorphic encryption. This work (J2) is published in IEEE Intelligent System (2020) and receives 300+ citations according to Google Scholar. The papers citing our work are widely published in top conferences and journals such as Nature Communications (as one of the main baselines), NeurIPS, WWW, ICML, etc. This study was also adopted by the FATE federated learning framework and deployed in the FedRec module, serving the secure distributed recommendation system. In the following studies, I continued to explore distributed MF resistant to inference attacks (J3), secure distributed singular value decomposition system over billion-scale data (C2), and decentralized distributed singular value decomposition system (C1), and these studies are published in top conferences/journals: ACM TIST'22, ACM SIGKDD'22, and USENIX ATC'24, respectively. By solving the privacy protection in distributed MF, we have provided secure solutions for most distributed applications based on MF.

High Performance and Future Research Directions

The third challenge in distributed computation pertains to achieving high performance. Supporting large-scale data and efficiently utilizing the computational resources of multiple nodes present a significant challenge. In our series of studies on distributed MF, we have delved deeply into achieving high performance. Specifically, in our work on distributed MF over billion-scale data (KDD'22), we propose to reduce computational complexity using sparse block matrix multiplications and reduce the storage overhead by delicately analyzing data access patterns. This billion-scale distributed MF system is adopted by BGI Genomics in decomposing large-scale genome data across multiple institutions for genome-wide association studies. In our work of efficient decentralized MF (to appear in ATC'24), we performed a quantitative analysis of the communication complexity of the design space and selected the path with the lowest communication overhead. Furthermore, we managed to reduce 66% of communication rounds through overlapping the system pipelines.

In the designing of high-performance distributed MF, I have gained a wealth of experience in optimizing storage, computation, and communication for distributed matrix computations. In future research, I am planning to explore high-performance distributed training and inference systems for generative AI, where the core steps are also distributed matrix computations. I will explore using an algorithm and system co-design to solve the issues of high computational overhead, high computational resource requirements, high data quality requirements, and high data privacy issues in generative models.

Conference Paper

C1. Efficient Decentralized Federated Singular Vector Decomposition.
Di Chai, Junxue Zhang, Liu Yang, Yilun Jin, Leye Wang, Kai Chen, and Qiang Yang.
USENIX ATC'24 Accepted.

C2. Practical Lossless Federated Singular Vector Decomposition Over Billion-Scale Data.
Di Chai, Leye Wang, Junxue Zhang, Liu Yang, Shuowei Cai, Kai Chen, and Qiang Yang.
ACM SIGKDD'22.

C3. Bike Flow Prediction with Multi-Graph Convolutional Networks.
Di Chai, Leye Wang, and Qiang Yang.
SIGSPATIAL/GIS'18. [Google Scholar 300+ Citations]

C4. Sphinx: Enabling Privacy-preserving Online Learning over the Cloud.
Han Tian, Chaoliang Zeng, Zhenghang Ren, Di Chai, Junxue Zhang, Kai Chen, and Qiang Yang.
IEEE S&P'22.

Journal Paper

J1. A Survey for Federated Learning Evaluations: Goals and Measures.
Di Chai*, Leye Wang*, Liu Yang, Junxue Zhang, Kai Chen, and Qiang Yang. (*Co-first Authors)
IEEE TKDE Accepted (2024).

J2. Secure Federated Matrix Factorization.

Di Chai, Leye Wang, Kai Chen, and Qiang Yang.
IEEE Intelligent Systems, 36(5): 11-20 (2021). [Google Scholar 300+ Citations]

J3. Efficient Federated Matrix Factorization against Inference Attacks.
Di Chai, Leye Wang, Kai Chen, and Qiang Yang.
ACM TIST, 2022, 13(4): 1-20.

J4. Exploring the Generalizability of Spatio-Temporal Traffic Prediction: Meta-Modeling and an Analytic Framework.
Leye Wang, Di Chai, Xuanzhe Liu, Liyue Chen, and Kai Chen.
IEEE TKDE, 2021, 35(4): 3870-3884.

Workshop Paper

W1. Aegis: A Trusted, Automatic and Accurate Verification Framework for Vertical Federated Learning.
Cengguang Zhang, Junxue Zhang, Di Chai, and Kai Chen.
IJCAI FL-Workshop (2021). [Best Application Award]

W2. Practical and Secure Federated Recommendation with Personalized Mask.
Liu Yang, Junxue Zhang, Di Chai, Leye Wang, Kun Guo, Kai Chen, and Qiang Yang.
International Workshop on Trustworthy Federated Learning (2022).

W3. Secure Forward Aggregation for Vertical Federated Neural Networks.
Shuowei Cai, Di Chai, Liu Yang, Junxue Zhang, Yilun Jin, Leye Wang, Kun Guo, and Kai Chen.
International Workshop on Trustworthy Federated Learning (2022).