Machine learning in data lakes

Rihan Hai

TU Delft
r.hai@tudelft.nl

Machine learning (ML) training data is often scattered across disparate collections of datasets, called *data silos*. This fragmentation poses a major challenge for data-intensive ML applications: integrating and transforming data residing in different sources demand a lot of manual work and computational resources. With data privacy and security constraints, data often cannot leave the premises of data silos, hence model training should proceed in a decentralized manner.

In this talk, I will first briefly give an overview of data lakes [2], and our recent works on model zoos [4, 3]. I will focus on our vision of how to bridge the traditional data integration (DI) techniques with the requirements of modern machine learning [1]. We explore the possibilities of utilizing metadata obtained from data integration processes for improving the effectiveness and efficiency of ML models. Towards this direction, we have analyzed two common use cases over data silos, feature augmentation and federated learning. Bringing data integration and machine learning together, we highlight new research opportunities from the aspects of systems, representations, factorized learning and federated learning. Finally, I will introduce our recent progress in building novel data lakes that support both data management and machine learning tasks [5].

References

[1] R. Hai, C. Koutras, A. Ionescu, Z. Li, W. Sun, J. Van Schijndel, Y. Kang, and A. Katsifodimos. Amalur: Data integration meets machine learning. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*, pages 3729–3739. IEEE, 2023.

[2] R. Hai, C. Koutras, C. Quix, and M. Jarke. Data lakes: A survey of functions and systems. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[3] Z. Li, H. Kant, R. Hai, A. Katsifodimos, and A. Bozzon. Macaroni: Crawling and enriching metadata from public model zoos. In *International Conference on Web Engineering*, pages 376–380. Springer, 2023.

[4] Z. Li, M. Schönfeld, W. Sun, M. Fragkoulis, R. Hai, A. Bozzon, and A. Katsifodimos. Optimizing ml inference queries under constraints. In *International Conference on Web Engineering*, pages 51–66. Springer, 2023.

[5] W. Sun, A. Katsifodimos, and R. Hai. Accelerating machine learning queries with linear algebra query processing. In *Proceedings of the 35th International Conference on Scientific and Statistical Database Management*, SSDBM '23, New York, NY, USA, 2023. Association for Computing Machinery.