Bachelor’s Programme in Science and Technology

Data lineage and observability for ensemble Machine Learning serving in the edge

Dung Nguyen Anh

Aalto University

BACHELOR’S THESIS

Aalto University
BACHELOR'S THESIS 2023

Data lineage and observability for ensemble Machine Learning serving in the edge

Dung Nguyen Anh

Thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Technology.
Otaniemi, 29 Jan 2023

Supervisor: professor Maarit Käpylä
Advisor: Minh-Tri Nguyen

Aalto University
School of Science
Bachelor's Programme in Science and Technology

Aalto University
Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Abstract

Author
Dung Nguyen Anh

Title
Data lineage and observability for ensemble Machine Learning serving in the edge

School School of Science

Degree programme Bachelor's Programme in Science and Technology

Major Data Science **Code** SCI

Supervisor professor Lauri Malmi

Advisor Minh-Tri Nguyen

Level Bachelor's thesis **Date** 27 Nov 2017 **Pages** 70 **Language** English

Abstract

Abtract

Keywords keyword

urn https://aaltodoc.aalto.fi

Contents

Abstract **ii**

Contents **iii**

1. Introduction **iv**

2. Background **vi**

2.1 Data quality and Data dimensions vi

2.2 Data lineage vi

2.3 Related work vii

3. Data lineage for ensemble ML serving **viii**

4. Experiment and Results **ix**

5. Conclusion **x**

Bibliography **xi**

1. Introduction

Recently, Machine Learning (ML) has been applied extensively to many problems, including customer journey optimization in marketing [19], particle identification in Physics [8], and mental health problems prediction in healthcare [18]. Then, with the increasing number of ML applications and their proven performance, several organizations are seeking third-party ML service providers to improve their operations, creating a new business model — ML as a Service (MLaaS). As with other kinds of service, MLaaS requires a contract which is usually called a Service Level Agreement (SLA) that sets the expectations and describes the delivered service to preserve the benefits of both stakeholders. However, the current SLA has yet to fulfill its preservation job because ML-specific attributes, such as data quality, inference accuracy, and explainability, have created new challenges. For instance, in Internet of Things (IoT) applications, many issues of the ML models have their root causes in the quality of data, which is impacted by many factors such as erroneous measurement, environmental noise, and discrete observations [15]. Moreover, because of the relationship between data quality and inference accuracy, ML models can produce false predictions when data quality problems exist, causing ML providers to be penalized based on the SLA. Another issue is that in MLaaS, the customers only submit the data and does not cooperate with the ML provider in the prediction task. Thus, from their perspective, ML models are considered like a black box, which is insufficient in human life-related applications such as e-health and autonomous vehicles where uncertain decisions cannot be tolerated.

One approach to resolve these two challenges is increasing model explainability by explaining the inference result to the customer [12]. However, explainability aspects of ML models have been researched mostly in the training task [15], making it unclear about the appropriate utilization and implementation of ML-specific attributes and constraints. As an approach to support ML-specific service contracts, Linh et al., proposed QoA4ML framework which outlines the essential components of such contracts, the definition of ML attributes and constraints, and the guidance on how these elements should be monitored and assessed [20]. Although QoA4ML has created a foundation for robust ML monitoring, current implementation focuses on a general ML model such as the final confidence of the ensemble model instead of the base models' confidence. This can be inefficient for some sophisticated applications such as digital assistants where a complex chain of models are used for inference task. Additionally, with the increasing number of research in autoscaler for inference serving system [16] and automated ensemble [2], it requires monitoring all underlying base models to ensure dynamic and robust inference capabilities of ML solutions.

In this research, I perform an analysis of data lineage monitoring and its implementation in edge environment. From that, I propose an approach to improve QoA4ML framework with data lineage, which can be combined with other techniques to improve the current MLaaS. I provide a prototype with QoA4ML as the foundation so that ML developer and provider can integrate it with their deployed service.

The rest of this paper is organized as follows. Section 2 discusses the background of the research, and section 3 explains data lineage implementation in the QoA4ML framework. Section 4 describes the experiment and its result, while section 5 concludes the research.

2. Background

This section presents the background on data quality in the edge and discusses the data lineage definition, characteristic and its relationship with data quality. We then identify the challenges of monitoring data lineage for ML serving.

2.1 Data quality and Data dimensions

In [13], Liu et al. defines data quality as the suitability of the data for the application and data dimension as metrics that describe the level of data quality. As the definitions are pretty general, data dimensions depend highly on the application, and different organizations value data dimensions differently [5]. For example, although completeness is valuable when working with tabular data, it is not as relevant in computer vision. Instead, attributes such as sharpness, noise, or dynamic range, are more significant and can affect the inference accuracy heavily. In edge computing and IoT applications, data characteristics are uncertain, erroneous, and noisy [13], so the requirements will focus on recognizing methods for capturing data quality and the causal relationship between data quality and the performance of machine learning services for dynamic inferences. [15].

2.2 Data lineage

Data lineage or provenance describes the origin of data, its derivation and changes over time [11]. Based on which questions it can answer, Ikeda and Widom [11] classify data lineage into two types: *where*-lineage and *how*-lineage. Then for each class, there are two *granularities*:

- Schema-level (coarse-grained) answers data lineage questions at the general level. It can be which datasets our current data originate from for where-lineage and which transformations have been applied for how-lineage

- Instance-level (fine-grained), on the other hand, clarifies the origin of a specific data point and how different data points are combined to achieve such results.

The answers that data lineage can provide will shine in edge computing because the complicated data pipeline makes it very time-consuming and costly for ML providers to trace down the root cause of the problems manually. However, with data lineage, they can immediately identify the data input's origin and derivation and then confirm the data quality based on them, helping reduce the required work [17]. Although data lineage is precious in MLaaS, there are some challenges requires considering in this research:

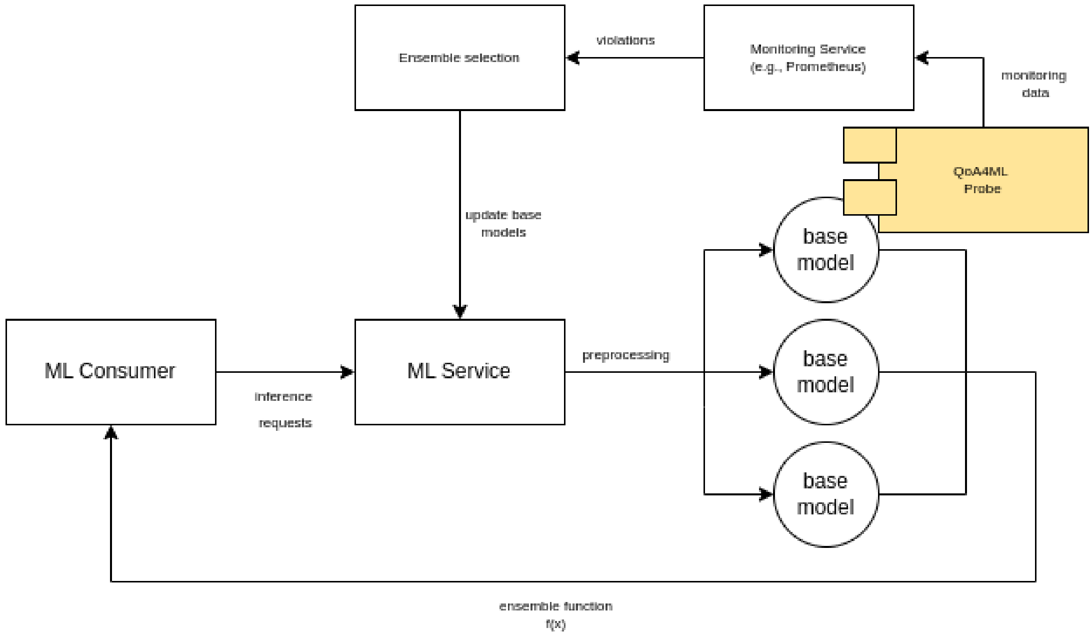
- Storing and querying lineage can be expensive when the number of IoT devices and the complexity of the pipeline increase[11]. Additionally, a resource-constrained environment like the edge demands an even more efficient processes.

- Current ML solution adopts many open source application and library that can be considered as a black box, so how can we accurately capture data provenance for black box operations and many others [17]

2.3 Related work

Data lineage/provenance: Data provenance research has been active for three decades with the first paper dated back to 1990 [21]. Since then, data lineage has been applied in scientific database[22] [3] [1] [10], datawarehouse [6] [7], and recently big data platform [14] [9] and IoT. Although data lineage can be used for ML interpretability as their goals are clearly similar [4], the number of papers about this application is moderate. In this research, I analyzes the application of data lineage in increasing data interpretability which can be utilized in optimizing ensemble ML model for dynamic inference.

3. Data lineage for ensemble ML serving



4. Experiment and Results

5. Conclusion

Bibliography

[1]

[2] Yang Bai, Lixing Chen, Mohamed Abdel-Mottaleb, and Jie Xu. Automated Ensemble for Deep Learning Inference on Edge Computing Platforms. *IEEE Internet of Things Journal*, 9(6):4202–4213, March 2022.

[3] Peter Buneman, Adriane Chapman, and James Cheney. Provenance management in curated databases. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, page 539–550, New York, NY, USA, 2006. Association for Computing Machinery.

[4] Peter Buneman and Wang-Chiew Tan. Data provenance: What next? *SIGMOD Rec.*, 47(3):5–16, feb 2019.

[5] Corinna Cichy and Stefan Rass. An Overview of Data Quality Frameworks. *IEEE Access*, 7:24634–24648, 2019.

[6] Yingwei Cui, Jennifer Widom, and Janet L. Wiener. Tracing the lineage of view data in a warehousing environment. *ACM Trans. Database Syst.*, 25(2):179–227, jun 2000.

[7] YW Cui and J. Widom. Lineage tracing for general data warehouse transformations. *The VLDB Journal — The International Journal on Very Large Data Bases*, 12:41–58, 09 2001.

[8] Denis Derkach, Mikhail Hushchyn, Tatiana Likhomanenko, Alex Rogozhnikov, Nikita Kazeev, Victoria Chekalina, Radoslav Neychev, Stanislav Kirillov, Fedor Ratnikov, and on behalf of the LHCb collaboration. Machine-learning-based global particle-identification algorithms at the lhcb experiment. *Journal of Physics: Conference Series*, 1085(4):042038, sep 2018.

[9] Boris Glavic. *Big Data Provenance: Challenges and Implications for Benchmarking*, volume 8163, pages 72–80. 01 2014.

[10] Thomas Heinis and Gustavo Alonso. Efficient lineage tracking for scientific workflows. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1007–1018, New York, NY, USA, 2008. Association for Computing Machinery.

[11] Robert Ikeda and Jennifer Widom. Data lineage: A survey. Technical report, Stanford InfoLab, 2009.

[12] Senthil Kumar Jagatheesaperumal, Quoc-Viet Pham, Rukhsana Ruby, Zhaohui Yang, Chunmei Xu, and Zhaoyang Zhang. Explainable AI Over the Internet of Things

(IoT): Overview, State-of-the-Art and Future Directions. *IEEE Open Journal of the Communications Society*, 3:2106–2136, 2022. Conference Name: IEEE Open Journal of the Communications Society.

[13] Aimad Karkouch, Hajar Mousannif, Hassan Al Moatassime, and Thomas Noel. Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications*, 73:57–81, September 2016.

[14] Dionysios Logothetis, Soumyarupa De, and Kenneth Yocum. Scalable lineage capture for debugging disc analytics. In *Proceedings of the 4th Annual Symposium on Cloud Computing*, SOCC '13, New York, NY, USA, 2013. Association for Computing Machinery.

[15] My-Linh Nguyen, Thao Phung, Duong-Hai Ly, and Hong-Linh Truong. Holistic Explainability Requirements for End-to-End Machine Learning in IoT Cloud Systems. In *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*, pages 188–194, Notre Dame, IN, USA, September 2021. IEEE.

[16] Kamran Razavi, Manisha Luthra, Boris Koldehofe, Max Mühlhäuser, and Lin Wang. FA2: Fast, Accurate Autoscaling for Serving Deep Learning Inference with SLA Guarantees. In *2022 IEEE 28th Real-Time and Embedded Technology and Applications Symposium (RTAS)*, pages 146–159, May 2022. ISSN: 2642-7346.

[17] Mingjie Tang, Saisai Shao, Weiqing Yang, Yanbo Liang, Yongyang Yu, Bikas Saha, and Dongjoon Hyun. SAC: A System for Big Data Lineage Tracking. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1964–1967, April 2019. ISSN: 2375-026X.

[18] Ashley E. Tate, Ryan C. McCabe, Henrik Larsson, Sebastian Lundström, Paul Lichtenstein, and Ralf Kuja-Halkola. Predicting mental health problems in adolescence using machine learning techniques. *PLOS ONE*, 15(4):1–13, 04 2020.

[19] Alessandro Terragni and Marwan Hassani. Optimizing customer journey using process mining and sequence-aware recommendation. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 57–65, New York, NY, USA, 2019. Association for Computing Machinery.

[20] Linh Truong and Tri Nguyen. QoA4ML – A Framework for Supporting Contracts in Machine Learning Services. 2021 IEEE International Conference on Web Services (ICWS);, page 11, 2021.

[21] Y Richard Wang and Stuart E Madnick. A Polygen Model for Heterogeneous Database Systems: The Source Tagging Perspective.

[22] A. Woodruff and M. Stonebraker. Supporting fine-grained data lineage in a database visualization environment. In *Proceedings 13th International Conference on Data Engineering*, pages 91–102, 1997.