Data versioning in machine-learning architecture

Peter Bartoš, Stanislav Krištof

Faculty of Informatics and Information Technologies
Slovak University of Technology in Bratislava

September 29, 2024

Abstract

Data versioning plays a crucial role in modern machine learning architecture, ensuring that the complex and ever-evolving datasets that provide the basis for models can be tracked, compared, and managed efficiently. At its core, data versioning refers to the practice of creating unique references for different states of a dataset over time, allowing us to trace changes, restore previous versions, and debug issues. This is vital in machine learning workflows, where even small changes in data can significantly impact model performance.

In this domain, data versioning supports reproducibility by maintaining a consistent link between datasets and the models trained on them. Without version control, it becomes challenging to recreate experiments, leading to inconsistencies in predictions and hindering model audits. Versioning also simplifies collaboration across teams, enabling multiple stakeholders to work on the same data without overwriting each other's progress.

Basic approaches to data versioning include full duplication of datasets, where copies are saved with each change, and metadata-based versioning, where timestamps indicate the validity of each record. Advanced solutions (like lakeFS and DVC) deal with versioning as a core component of machine learning architecture. They enable storage-efficient data commits, branching, and comparison, similar to how Git handles version control in software development.

Overall, data versioning enhances productivity, reduces errors, and fosters an engineering-driven approach to handling data in machine learning pipelines, ultimately enabling smoother transitions between development stages and more robust model deployment. The objective of the project is to go over these systems and provide detailed overviews of how data versioning has such a crucial role in machine learning architecture.