Contribution ID: **34**

Type: **Talk**

Enhancing Transparency and Reproducibility in AI Model Training through Provenance-Enabled Data Preprocessing and Workflow Documentation

Wednesday 11 October 2023 11:10 (20 minutes)

As the scale and complexity of AI models continue to grow, the demand for vast amounts of data, including unlabelled and uncurated datasets, has become increasingly prevalent. To address this challenge, the role of data preprocessing, filtering, augmentation, and curation using automated methods has gained increasing significance in ensuring optimal model performance. However, while AI models themselves are increasingly documented, the transparency surrounding the preprocessing techniques applied often remains incomplete, impeding reproducibility.

This paper proposes a novel approach aimed at improving transparency and reproducibility in training scenarios by capturing concrete provenance information throughout the data preprocessing workflow. By recording the sequence of transformations applied to the data, researchers and developers gain the ability to recreate workflows and analyze sample provenance, improving informed decision-making. Simultaneously, this approach offers a pathway for gathering metadata, which serves debugging, monitoring, and development purposes, exposing developers to valuable insights.

The proposed solution is realized through the introduction of a lightweight data pipeline library designed for seamless chainable stream operations. Focused on this common use case, this library enables the systematic capture of structured provenance information, reducing the overhead of subsequent analysis. Demonstrated in the context of a compact computer vision use case, the methodology not only exposes useful training metrics but also comprehensively documents the workflow with negligible performance implications.

Further work involves expanding the application of this approach to encompass larger and more complex use cases. Additionally, the integration of intermediate dataset versioning, facilitated by a dedicated DVC plugin, allows for including intermediate data versioning and traceability, thereby improving on the overall reproducibility and transparency of AI model training workflows.

Please assign your contribution to one of the following topics

Metadata annotation and management close to the research process

Please specify ”other” (stakeholder)

In addition please add keywords.

Provenance, Metadata collection, transparent AI, ML Training workflows

Please assign yourself (presenting author) to one of the stakeholders.

Researchers

Primary author: HOFFMANN, Nils

Presenter: HOFFMANN, Nils

Session Classification: Parallel Track 2