Contribution ID: **785** Type: **Oral Presentation**

Data Quality Assurance Metrics for Federated Machine Learning

Wednesday, 15 May 2024 11:40 (15 minutes)

Machine/deep learning (ML/DL) have emerged as powerful tools for driving science innovation. These methods tend to be data hungry, with large volumes that are not likely to be collocated. Further, industry or professional companies typically have a vested interest in keeping their data private. Recent work has shown the viability of federated learning (FL), a ML /DL framework where multiple groups (called clients) collaboratively train a single model without exchanging locally stored, private data. We apply this framework to a problem common in reservoir engineering and digital rocks physics: we use a convolutional neural network with the specific task of predicting the flow velocity field through a porous sample using its segmented x-ray microtomography image as input. The results are nevertheless applicable to any task with spatial data stored across multiple locations participating in training the same model.

As with any typical DL workflow, the quality of the training data plays a key role in the prediction accuracy and the generalizability of a network model. Additionally, the interest in enabling the reusability of these models is guided by the findable, accessible, interoperable, and reusable (FAIR) principles. In the context of FL, the assurance that each client provides quality and representative training data that adhere to the FAIR principles has proven challenging because the data and its metadata remain private.

Here, we first explore how differences in local data can affect prediction behavior once individual models are aggregated. We then propose several preprocessing metrics that can be quickly computed and shared with the central server to assure the quality of the training set but without ability to infer the entire dataset from them. For our specific task of predicting a flow field, we employ common techniques of quantifying pore geometries based on their images. These include, but are not limited to, pore space morphological and topological descriptors, heterogeneity characterization, representativeness quantification, and resolution checks using morphological drainage.

We validate our trained models to test that the proposed metrics properly capture the representativeness of the new training data while still providing sufficient diversity to further the model's learning. The scope of this work is to show that a set of quality assurance metrics enables model reuse under FAIR principles while maintaining clients'data privacy during FL. In a broader context, these metrics are also useful for preprocessing classical machine learning training data and in developing constitutive relationships between complex pore geometries and transport properties. Last but not the least, we discuss how we apply this framework with two industry partners.

Acceptance of the Terms & Conditions

Click here to agree

Student Awards

Country

United States

Porous Media & Biology Focused Abstracts

References

Conference Proceedings

I am interested in having my paper published in the proceedings.

Primary author: CHANG, Bernard (The University of Texas at Austin)

Co-authors: Mr TURHAN, Cinar (The University of Texas at Austin); Mr MOHAMED, Ali (The University of Texas at Austin); Dr ESTEVA, Maria (The University of Texas at Austin); PRODANOVIC, Masa (The University of Texas at Austin)

Presenter: CHANG, Bernard (The University of Texas at Austin)

Session Classification: MS15

Track Classification: (MS15) Machine Learning and Big Data in Porous Media