Machine Learning Conference for X-Ray and Neutron-Based Experiments, Munich 2024

Contribution ID: **60**

Type: **Invited talk**

MLExchange: A Machine Learning Platform for On-the-fly Data Analysis at Scientific User Facilities

Monday 8 April 2024 16:50 (30 minutes)

With the continuous enhancement of experimental capabilities at scientific user facilities, the demand for computational tools that seamlessly guide users through their data lifecycle grows exponentially. These tools play an important role in facilitating the application of machine learning (ML) techniques to accelerate materials discovery. In light of this, MLExchange introduces a collaborative web-based platform to democratize diverse workflows for on-the-fly data visualization, rapid ML-based data analysis, automated experiments, and other applications. Currently, MLExchange offers a selection of web-based graphical user interfaces (GUI) for image segmentation, latent space exploration, data labeling, and classification [1].

In particular, its data labeling pipeline, *Label Maker*, aims to accelerate the demanding and time-consuming process of labeling scientific data sets through similarity-based querying, clustering, and classification approaches. To achieve this, its architecture connects four independent GUIs: (1) *Data Clinic* for latent space extraction, (2) *MLCoach* for data classification, (3) *Latent space explorer* for dimension reduction, latent space visualization, and clustering, and (4) *Label Maker* for data visualization and label assignment. Across this pipeline, the web applications make use of an assortment of ML-based techniques, including principal component analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) for dimension reduction, Density-based spatial clustering (DBSCAN) and Mini Batch K-means for data clustering, and tunable deep learning algorithms for latent space extraction and data classification. *Label Maker* has shown potential applications for cross-facility learning by using *Tiled* for data access, which has enabled the visualization of Resonant Soft X-ray Scattering data collected at the National Synchrotron Light Source II. Furthermore, we have successfully demonstrated its effectiveness in enhancing the fine-tuning process of foundational models with human feedback.

Overall, the MLExchange platform offers a collaborative ecosystem to easily deploy ML-based algorithms for scientific data analysis. Among these efforts, MLExchange aims to enhance its capabilities to handle complex workflows, such as mitigating training biases with foundational models and enabling cross-facility model training.

[1] Z. Zhao, T. Chavez, E. A. Holman, G. Hao, A. Green, H. Krishnan, D. McReynolds, R. J. Pandolfi, E. J. Roberts, P. H. Zwart, H. Yanxon, N. Schwarz, S. Sankaranarayanan, S. V. Kalinin, A. Mehta, S. I. Campbell, and A. Hexemer, "MLExchange: A web-based platform enabling exchangeable machine learning workflows for scientific studies," in *2022 4th Annual Workshop on Extreme-scale Experiment-in-the-Loop Computing (XLOOP)*, 2022, pp. 10–15. doi: 10.1109/XLOOP56614.2022.00007

Primary author: CHAVEZ, Tanny (Lawrence Berkeley National Laboratory)

Co-authors: ZHAO, Zhuowen (Lawrence Berkeley National Laboratory); JIANG, Runbo (Lawrence Berkeley National Laboratory); HOLMAN, Elizabeth A. (Lawrence Berkeley National Laboratory); GREEN, Adam (Lawrence Berkeley National Laboratory); KRISHNAN, Harinarayan (Lawrence Berkeley National Laboratory, Center for Advanced Mathematics in Energy Research Applications); KOEPP, Wiebke (Lawrence Berkeley National Lab); MCREYNOLDS, Dylan (Lawrence Berkeley National Lab); PANDOLFI, Ronald (Lawrence Berkeley National

Laboratory, Center for Advanced Mathematics in Energy Research Applications); ROBERTS, Eric J. (Lawrence Berkeley National Laboratory, Center for Advanced Mathematics in Energy Research Applications); ZWART, Petrus H. (Lawrence Berkeley National Laboratory, Center for Advanced Mathematics in Energy Research Applications); HAO, Guanhua (Lawrence Berkeley National Laboratory); YANXON, Howard (Argonne National Laboratory); SCHWARZ, Nicholas (Argonne National Laboratory); GANN, Eliot H. (Brookhaven National Laboratory); ALLAN, Daniel B. (Brookhaven National Laboratory); USHIZIMA, Daniela (Lawrence Berkeley National Laboratory, Center for Advanced Mathematics in Energy Research Applications); BARNARD, Edward (Lawrence Berkeley National Laboratory); MEHTA, Apurva (SLAC National Accelerator Laboratory); SANKARANARAYANAN, Subramanian (Argonne National Laboratory, University of Illinois Chicago); HEXEMER, Alexander (Lawrence Berkeley National Lab)

Presenter: CHAVEZ, Tanny (Lawrence Berkeley National Laboratory)

Session Classification: Session 4

Track Classification: MLC