6th Asia-Pacific Conference on Plasma Physics, 9-14 Oct, 2022, Remote e-conference

Machine Learning Analysis of Plasma-Science Data

Jong Youl Choi[1], Scott Klasky[1], Ralph Kube[2], Michael Churchill[2], CS Chang[2]
[1] Oak Ridge National Laboratory
[2] Princeton Plasma Physics Laboratory
e-mail (speaker): choij@ornl.gov

Data in plasma science is getting bigger as data generation and collection technologies are evolving faster than ever before. We observe big data in plasma science, ranging from fusion experiment facilities to simulations running on the world's largest supercomputers. One of the main challenges in large-scale plasma science data is how we process and analyze it in near-real time when the data size is big. Data compression is a well-known technique, but the preservation of the physics property simultaneously has been a research topic in the data-science community.

We have developed VAPOR, a deep-learning-based generative model based on Vector Quantized Variational Auto Encoder (VQ-VAE), focusing on compressing plasma data as well as preserving physics constraints [1]. Key features of VAPOR are three-fold; i) find a reduced representation of physics data, ii) reconstruct the data with a minimum loss, and iii) preserve physics information (e.g., mass, energy, moment conservation) (see Figs. 1).

For the experimental data side, we have developed DELTA [2,3], a system to address AI/ML data challenges in fusion science to support near-real-time streaming data analysis. We leverage ADIOS, an adaptable I/O library for data stream management. We construct deep-learning-based data analysis workflows for the fusion domain as an example. The capability demonstrated by this project is the basis for improving the state-of-the-art data federation for near-real-time data analysis amongst remote facilities.

We will discuss the challenges in performing AI/ML for plasma data, as an example, and data management. We present examples for deep-learning-based data analysis in the fusion domain, focusing on ongoing research work in i) decompressing and reorganizing data for analysis, ii) managing streaming data, iii) executing workflows for automated data analysis. Figure 1 and 2 show the outputs from our AI/ML workflows.

This research is supported by US DOE via the SciDAC Center for High-fidelity Boundary Plasma Simulation, and used resources of the Oak Ridge Leadership Computing Facility.

References

[1] Choi, Jong, et al. "Neural data compression for physics plasma simulation." *Neural Compression: From Information Theory to Applications--Workshop@ ICLR 2021*. 2021.

[2] Choi, Jong, et al. "Data federation challenges in remote near-real-time fusion experiment data processing." *Smoky Mountains Computational Sciences and Engineering Conference*. Springer, Cham, 2020.

[3] Kube, Ralph, et al. "Leading magnetic fusion energy science into the big-and-fast data lane." *Proceedings of the 19th Python in Science Conference*. 2020.

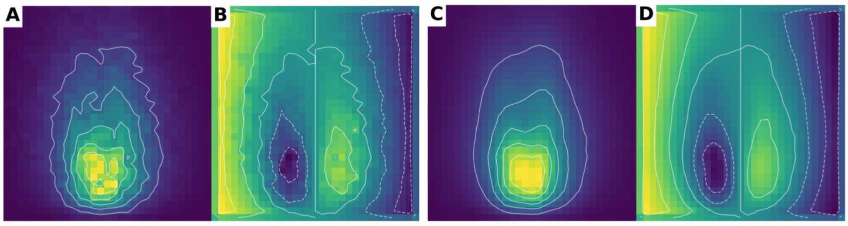

Figure 1. XGC 5D data reconstruction with VQ-VAE. A) Original data from XGC. B) corresponding energy density of (A). C) VAE-based reconstruction of (A), D) Energy density based on (C).

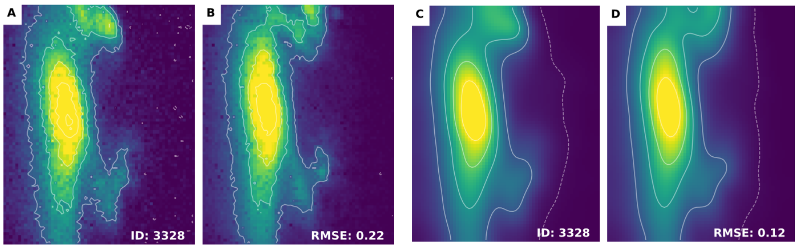

Figure 2: Demonstration of VQ-VAE reconstruction with NSTX GPI images. (A) original NSTX GPI image, (B) reconstructed image by VQ-VAE, (C) original image with Gaussian denoising treatment, and (D) reconstructed VQ-VAE image followed by Gaussian denoising. Frame numbers and Root Mean Square Error (RMSE) metrics are shown on the bottom-right corner.