doi:10.1017/S1431927622011412
Microsc. Microanal. 28 (Suppl 1), 2022

Lossless Image Compression for 4D-STEM Datasets

Jacob Hinkle[1], Debangshu Mukherjee[1], Kevin Roccapriore[3], Alexander Rakowski[2], Christopher Nelson[3], Ondrej Dyck[3], Stephen Jesse[3], Nageswara Rao[1], Colin Ophus[2], Andrew Lupini[3] and Olga Ovchinnikova[1]

[1.] Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA.
[2.] National Center for Electron Microscopy, Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.
[3.] Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN, USA.
* Corresponding author: hinklejd@ornl.gov

Recent progress in scanning transmission electron microscopy has enabled the collection of four-dimensional datasets (4D-STEM) that are incredibly rich in information at atomistic levels of detail. These datasets typically occupy 20-50 GiB on disk. Montaging and related automated acquisition methods promise to increase these data sizes many-fold in the near term. As data collection has advanced, so has our statistical modeling and computational capacity. Electron ptychography via iterative optimization is a computationally intensive method that uses 4D-STEM datasets to reconstruct 2D and 3D images with sufficient resolution to resolve individual atoms. Modern approaches to ptychography use high-performance computing with distributed computation to compute these image reconstructions in an acceptable timeframe. This necessitates transferring of tens to hundreds of gigabytes of data between microscope control computers and data centers which are typically housed in separate buildings or even separate facilities. Still, most 4D-STEM data collection methods write data in formats that are not optimized for network transfers or space efficiency. In this work, we have explored an array of existing and novel compression methods and found that custom approaches provide the highest compression ratios for this task.

It is now standard practice to incorporate deep learning (DL) into many phases of microscopy analysis pipelines. Within the past few years, the DL community has made great strides in developing models that can learn compression codecs directly from image datasets [1]; see for example the learned lossless compression (L3C) model used in our present work [2]. We investigated the use of DL as well as other compression methodologies that exploit the unique structure in 4D-STEM datasets to drastically reduce bandwidth and storage requirements. We have developed novel techniques for compressing 4D-STEM datasets which leverage the unique features of this type of data. Using the fact that important image contrast is contained in the center of the image while the periphery is dominated by background noise, we have employed entropy coding using pixel-specific histograms. The results show that this simple approach can significantly reduce the size of 4D-STEM data, up to nearly nine-fold compared to the current uncompressed output provided by the Nion Swift software. We have validated this approach using real observed 4D-STEM datasets and are extending these results to simulations covering a diverse range of materials and experimental conditions. Our results suggest that up to ten-fold lossless compression of 4D-STEM data may be achievable in practice. In future work, we plan to explore much larger compression ratios using lossy compression methods that preserve reconstructed image quality [3].

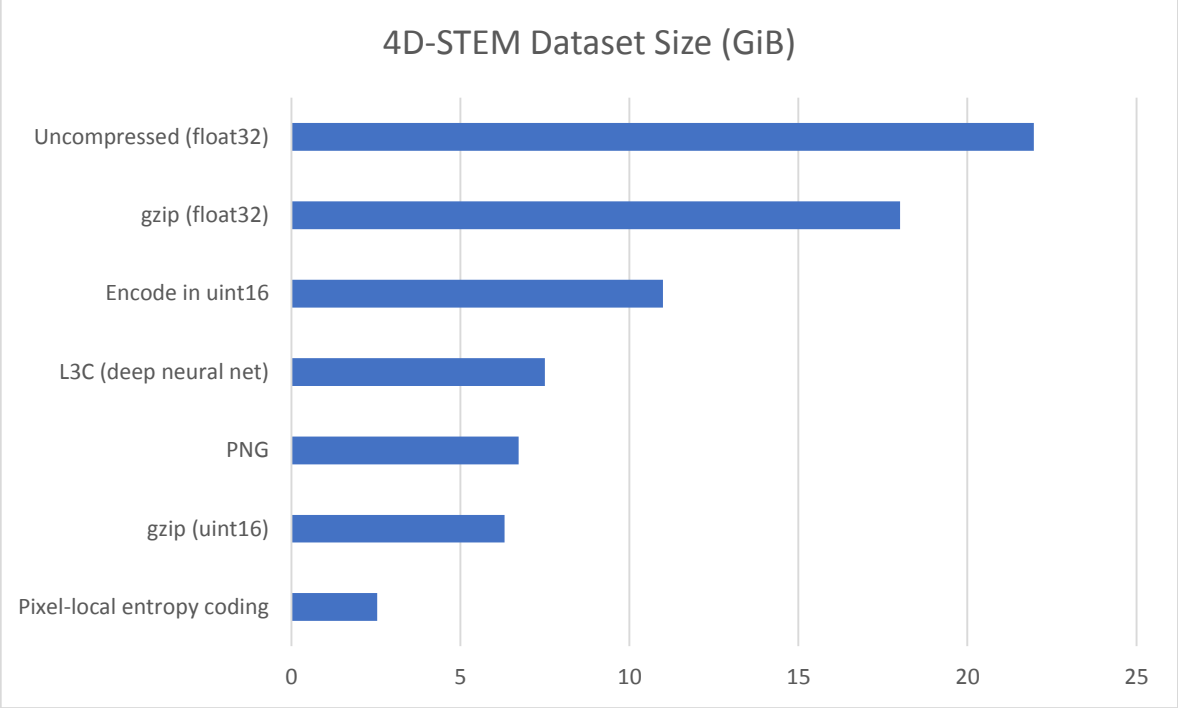

Figure 1. Size of a 4D-STEM dataset compressed using various lossless compression methods. Pixel-local histograms with arithmetic coding performs best, achieving an 8.6x compression ratio, or 3.70 bits per pixel. Our method surpasses off-the-shelf methods like gzip and PNG by at least two-fold. Notably, the deep learning approach we tried (L3C) showed lackluster performance, most likely because its convolutional architecture is unable to leverage location-specific differences in pixel distributions found in 4D-STEM data.

References:

[1] Mentzer et al. Proc. CVPR (2019).
[2] Oord et al. NIPS (2016).
[3] This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan)

https://doi.org/10.1017/S1431927622011412 Published online by Cambridge University Press