LG-254883-OLS, Virginia Polytechnic Institute and State University (Libraries)

Curating Very Large Biomedical Image Datasets For Librarian-In-The-Loop Deep Learning

Virginia Tech / Yale University

Project Justification

The Center for Digital Research & Scholarship at Virginia Tech Libraries, in partnership with the Focused Ion Beam Scanning Electron Microscopy Collaboration Core (F-SCC) at Yale University School of Medicine (YSM), requests $149,216 from IMLS for a 2-year National Leadership Planning Grant to prototype data curation pipelines for very large biomedical images. This project broadly addresses NLG program Goal 3 by collaborating with biomedical researchers on the effective use and broad dissemination of very large image data acquired from cutting-edge microscopy technologies. More specifically, we will address Objective 3.2 to explore innovative data curation approaches based on librarian-in-the-loop deep learning.

Recent breakthroughs in Focused Ion Beam Scanning Electron Microscopy (FIB-SEM) technology have made it possible to reliably acquire nanometer scale 3D imaging from sizable volumetric biomedical samples, each resulting in tens to hundreds of terabytes of raw image data. After generating landmark datasets for neuroscience and cell biology research, scientists at Yale are now bringing enhanced FIB-SEM to enable discoveries in translational and clinical research. Similar to many other data intensive science challenges, the bottleneck has now shifted from data collection and storage to data curation for the primary purpose of extracting insights and knowledge from data, albeit with more stringent requirements on efficiency, timeliness, replicability, and reusability.

Existing data curation models and frameworks are insufficient to address these challenges. In addition, the very large data volume has rendered comprehensive close reading and manual image annotation impractical. For example, it has been estimated that FIB-SEM images taken from a single cell may take up to 60 person-years to annotate manually. To make sense of these images, researchers increasingly resort to machine learning methods. Supervised deep learning has been applied to FIB-SEM images but its performance can be unreliable. Training a model for automatic image segmentation may take months on a GPU cluster and still result in overfitting. Thankfully, a recent study suggests that interventions from experienced and insightful domain experts and data curators may drastically speed up the training, although the performance gain originating from such human interventions has not been carefully benchmarked. Very large FIB-SEM datasets therefore present an archetypal test case on how to best orchestrate scientists, data curators, cyberinfrastructure, software, and deep learning algorithms to achieve best data-to-insight performance.

This project will draw insights from our prior IMLS funded project curating very large research datasets. Our past experience has shown that 1) data curators/librarians should be deployed in the big data pipeline as early as possible, even at the stage of physically acquiring data. Knowledge in data acquisition often affords pertinent opportunities to optimize the data pipeline. 2) Data curation should be driven primarily by data use and reuse, which closely aligns librarians/data curators with domain scientists. Long-term preservation activities are better performed as a side effect of data use and reuse. 3) The efficiency, cost, and performance of extracting insights from data are often the critical success factors for data curation and are closely associated with both the data format and the cyberinfrastructure options and choices. Experimenting and benchmarking are often the more effective way to achieve balanced results, therefore this prototyping project.

Project Work Plan

Task 1 (Lead: co-PI Pang): acquiring a new FIB-SEM dataset from scratch. The F-SCC at YSM will closely coordinate with the VT team in establishing the scientific goals for data acquisition and preparing data curators with sufficient background knowledge in biology, physics, and chemistry to understand the process of sample preparation, milling, and imaging.

Task 2 (Lead: PI Xie and co-PI Chen): setting up manual image annotation environment to establish ground truth. The potential target annotation environments may include VT Libraries own small GPU cluster and the Amazon cloud. We will attempt to run Paintera for manual annotation.

Task 3 (Lead: PI Xie and co-PI Chen): experimenting with the classic image segmentation with 3D UNet using VT Advanced Research Computing's Nvidia DGX A100 cluster and/or the Amazon cloud. Replicate the 2021 Nature paper using the newly acquired data.

Task 4 (Lead: PI Xie and co-PI Chen): experimenting with the human-in-the-loop enhancements. In particular, we will replicate various image pre-processing and transfer learning procedures described in the 2022 Biorxiv paper on the newly acquired data.

Task 5 (Lead: PI Xie and co-PI Chen): experiments and benchmarking. Using the same dataset, image format, and computing facility, we will parse, verify, and compare the previously published performance claims. It is important to repeat the experiments on randomized sample regions to eliminate the survivorship bias in the benchmarking. We will also evaluate the performance differences between different image labels (e.g., mitochondria vs. golgi vs. ER, etc.), cyberinfrastructure choices (public cloud vs. institutionally shared GPU cluster vs. designated small GPU cluster), image formats (tiff vs. H5 vs. Zarr vs. Parquet), and various curator intervention points. When benchmarking on the public cloud, we will gather cost information. We will also experiment with different communication patterns between data curators and domain scientists, e.g., constant contacts vs. feedback around results and changes vs. communicating only during deadlocks.

Results and Impact

By performing the above 5 tasks, we will be able to conduct fair benchmarking and gain insights between various factors impacting deep-learning performance of very large research datasets. The results will provide evidence based on which optimized curator-in-the-loop deep learning pipelines can be established for semi-automated curation of such datasets, which have become increasingly prevalent in STEM research.

Deliverables will include: a curated FIB-SEM image dataset with associated software code to perform and reproduce various benchmarking and performance tuning experiments; pipeline prototypes for curating very large image datasets; a white paper describing the state-of-the art of curating very large dataset and summarizing our findings, to be shared in venues such as JCDL, CNI, and/or IEEE Big Data. We will share all our documentation and software code on github as well as VT's institutional data repository (https://data.lib.vt.edu/).

Budget Summary

Our request to IMLS include the following direct cost breakdowns: $53,917 wages and fringe benefits (including $36,231 for 1.5 GRA-year counted as student support), $25,000 for acquiring one FIB-SEM dataset at F-SCC at YSM, and $22,948 tuition remission for the GRA which is also counted as student support. After applying Virginia Tech's federally negotiated indirect cost rate of 60%, our total IMLS request is $149,216.