High dimensional active learning (edge-based challenge) for microscopy of nanoscale materials

Rama Vasudevan, Kyle Kelley, Stephen Jesse, Sergei Kalinin, Maxim Ziatdinov
Oak Ridge National Laboratory

Atomic force microscopy (AFM) is a premiere research tool utilized to explore material behavior at the nanoscale and has been a mainstay of nanoscience in the past three decades. It consists of a tip at the end of a cantilever that interacts with a sample to derive information on the sample properties and correlate the functional properties to microstructural features of the sample. Usually, in addition to regular raster-based scanning for high-resolution images of topography, the AFM enables individual point-based spectroscopy, where stimulus is applied to the tip or sample or environment and the response of the material is measured locally. Typically, the spectroscopy can be time consuming, as each pixel can take from ~0.1 to ~10s to acquire, and this has to be repeated across a grid of points to determine the response variability on spatially heterogeneous samples. One example is in measuring the relaxation of piezoelectric response in ferroelectric materials, as shown in our recent work (Kelley et al. npj Computational Materials **6**, 113 (2020)).

In this data, we acquired the piezoelectric response as a function of voltage, time and space in a 200nm thick ferroelectric film of $PbTiO_3$. Such data is critical to understanding the role of domain walls in enhancing the piezoelectric and dielectric properties of ferroelectric materials. The results are provided as a h5 file with tensors for the response and vectors for the applied voltage. These measurements are time consuming and difficult. One method to reduce the time is to instead explore active learning strategies, where only specific voltages and/or spatial locations are measured, and then the full response 'reconstructed' from this subset of measurements. Determining where to sample to optimize the reconstruction becomes an optimization problem which lies at the heart of this challenge. The dataset is of size (60,60,16,128) where the axes are (x,y,Voltage, time). The full details are available in the manuscript.

The data challenge questions revolve around developing and implementing a machine learning (ML) or statistical learning algorithm to best guide the instrument as to where to sample based on an existing subset to optimize the reconstruction, i.e., 'active learning'. Here, some subset of data is first captured, and then a set of new measurement conditions (e.g., certain spatial pixels) are given by the algorithm to sample next. The microscope captures that data, the new data is fed back into the algorithm to guide the next points, and so on until enough data is captured that is sufficient for a high quality reconstruction with sparse sampling.

Challenge Questions

The challenges are:

1. Perform reconstruction of the dataset based on different image reconstruction techniques (e.g., Gaussian process regression) in high dimensional spaces, to observe how redundant the information in the dataset is.
2. Develop an ML algorithm for optimized sequential sampling of the multidimensional dataset.
3. Implement the algorithm in a workflow, as if the microscope was actually taking datapoints, to showcase the method
4. Ensure that the method (a) reconstructs the true dataset to some tolerance (e.g., 90%), and that there is at least a 25% gain in efficiency (i.e., less number of spatial and or voltage/time points that need to be measured).

Notes on the challenge questions:

- The preference is for the code to be written in python, but other languages are not forbidden
- For (1), explore how much of the voltage, time and spatial data can be eliminated safely without dramatic loss in reconstruction accuracy
- Note that the efficiency gain needs to take into consideration algorithm running time on a dgx-2 machine for every iteration, because algorithms that take longer than several iterations of the actual full spectroscopy will not be useful to the instrument user.
- The successful algorithm will be deployed at the CNMS for a real experiment. Note that high use of GPU acceleration is desired given the availability of the dgx-2 system at the CNMS.