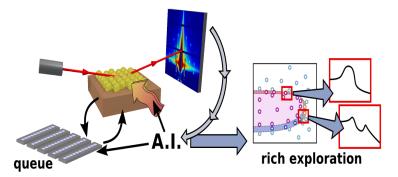# Deep learning for analyzing NSLS-II data stream

Kevin Yager[1], Dantong Yu[1], Minh Hoai Nguyen[2]
[1]Brookhaven National Laboratory, [2]Stony Brook University

Significance: The National Synchrotron Light Source II (NSLS-II) at Brookhaven National Laboratory (BNL) is now providing some of the world's brightest x-ray beams. This state-of-the-art scientific user facility offers enhanced capabilities and higher measurement throughputs, for a wide range of applications: biology and medicine, (nano) materials and chemical sciences, and environmental sciences. A series of imaging technologies and megapixel detectors with kilohertz frame-rates at the NSLS-IIs beamlines generate a variety of image streams in unprecedented velocities and volumes. A complete understanding of a complex material system often requires a suite of x-ray characterization tools that can reveal its elemental, structural, chemical and physical properties at different length-scales and time-scales. The flourish and continuing refinement of X-ray image techniques enable that the same sample may be studied with different perspectives and granularities, and at different time and locations, and thereby generate a big data problem of multiple image streams that outpaces any manual efforts and traditional data analysis practice. This data deluge can only be addressed using automated methods, ones that are capable of processing and anlyzing image streams in real-time and intelligent fashion. These methods would empower scientists by providing timely insights, allowing them to steer experiments efficiently during their precious x-ray beamtime allocation.

The x-ray scattering experiment is controlled by a A.I. software that runs a deep-learning algorithm and has access to a stream of samples. The software automatically analyzes the stream of x-ray scattering images, clustering them and organizing them so as to generate scientifically-meaningful results. The extracted trends are used to inform further automated measurements, leading to a rich exploration of parameter spaces.

Figure 1: **Illustration of a computer-directed beamline experiment.**

Goal: We envision be a transformative paradigm for synchrotron studies, where data acquisition and analysis are automated, and thereby liberating scientists to focus on deep scientific questions. Towards this goal, we propose to develop a set of intelligent automated methods, which will be the "brain" of a computed-directed beamline experiment, as illustrated in Figure 1. These methods will be used in real-time to extract hierarchical and physically-meaningful insights from scientific datasets collected at NSLS-II beamlines: 1) Low-level: identifying characteristic features in a diffraction image; 2) Intermediate-level: detecting the occurrence of a physical process from a sequence of images; and 3) High-level: learning and predicting scientifically-meaningful trends.

Approach and challenges: We propose to develop these methods based on the recent advances of Machine Learning. Machine Learning itself is undergoing a shift, with a re-thinking from traditional, naive, neural networks, towards deep learning models where the neural hierarchy is more rational, optimized, and informative. This has already led to clear advances in several fields including computer vision and speech recognition, and we aim to demonstrate similarly transformative gains with respect to scientific image streams. The core idea in deep learning is to design multiple levels of representations corresponding to a hierarchy of features, wherein the high-level concepts and knowledge are derived from the lower layers. This multi-level representation can capture the complex relationships hidden within rich datasets. This hierarchical representation motif closely matches how scientific knowledge is organized, and indeed how physical systems can be understood, as illustrated in Figure 2.

Our research objectives are to develop several deep learning algorithms for analyzing streams of x-ray images. However, the proposed algorithms are more than a naive application of existing techniques. Most current deep-learning algorithms focus on off-line processing, ignoring the tasks and challenges posed by real-time analysis of streaming data. First, most existing deep-learning algorithms were developed to address the recognition problem. They cannot be used for prediction and anticipation of the future. These requirements, however, are very crucial for real-time adaptive experimentation. Second, existing deep learning algorithms for image processing are computationally intensive, which cannot be used to analyze every single frame generated by a NSLS-II beamline. To address this issue, the proposed

Functionality

energy storage, catalysis, data storage, light management, electronics, surface-wetting, filtration, medicine, ...

Properties

elastic modulus, refractive index, polarizability, wettability, ...

Structure

unit cell, crystallinity, grain size, orientation, packing density, symmetry, ...

Constituents

small molecules, polymers, nanoparticles, colloids, ...

CAR PERSON ANIMAL

Output (object identity)

3rd hidden layer (object parts)

2nd hidden layer (corners and contours)

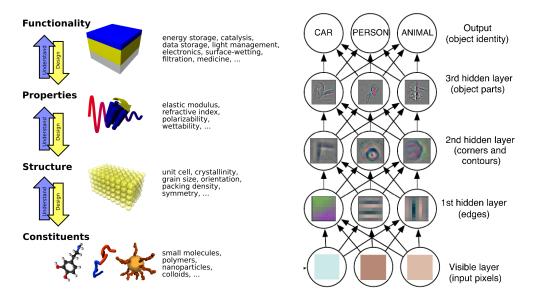1st hidden layer (edges)

Visible layer (input pixels)

Figure 2: **Comparison of hierarchies underlying physical systems and deep learning models.** Constituents organize into well-defined structures, which give rise to emergent properties, which in turn dictate functional response. By developing a machine-learning hierarchy closely aligned with this physically-relevant hierarchy, we will enable meaningful insights to be automatically extracted from scientific data at multiple levels. Similarly, deep learning recognizes raw images represented as a collection of pixel values to an object identify by breaking the complicated mapping into a series of simple mappings recursively, which creates a hierarchy. The input pixels are fed to the bottom visible layer, then a series of hidden layers abstract complex features from the lower layer. Both of them demonstrate similarity in feature extractions.

algorithms must leverage the most advanced computer architectures such as multiple GPUs and FPGAs. More importantly, the calculation for each frame must be simplified, removing the analyses that have no or little benefits for the tasks being considered. Aiming to maximize the overall performance, the proposed algorithms must determine the optimal trade-offs between spatial analysis and temporal consideration (i.e, how much time to spend analyzing a frame versus how many frames we can analyze per second).

Focus area – nanomaterial scientific case: Modern materials research investigates complex motifs, including multi-component assemblies and composites, self-assembly of nanomaterials, and the influence of process history. These trends all yield a corresponding explosion of the size of the associated experimental datasets. Multi-component materials inherently require exploring multi-dimensional phase-spaces; responsive materials require also in-situ and operando studies across a myriad of environments (temperature, pressure, solvent, humidity, etc.) and applied fields (electric, shear, temperature-gradients, etc.). The process history of materials assembly is also crucial. Self-assembly is usually described as an energy-minimization process—that is, driven by equilibrium physics—whereas it is now being appreciated that assembly is often strongly pathway-dependent and kinetically-trapped. All of these trends give rise to large, complex, multi-dimensional parameter spaces. It is no longer feasible to explore these complex physical systems manually. Instead, automation at both the hardware and software levels is crucially required. More specifically, understanding these multi-component and multi-scalar materials requires a hierarchy of scientific insights: at the level of the constituents, at the level of their organizations, at the level of the emergent properties, and finally with respect to their functional response. As a concrete example: photo-voltaic devices involve a delicate layering of different materials, where each layer is typically multi-component, and exhibiting properties that depend strongly on the nanostructure and microstructure. Only through precise control of material composition and processing history, can optimal solar energy collection performance be achieved.

This white paper will address the complexity of hierarchical self-assembling nano-materials by developing a corresponding hierarchy of analysis methods. Importantly, these analysis results will be organized into a computational hierarchy, where low-level analysis methods feed into higher-level insights. In this way, we will measure physical properties of material systems at all levels in the hierarchy (from constituents to properties), allowing the facile extraction of high-level scientific insights.