Deep Structured Analysis for Image Datasets from CFN and NSLS-II

Dantong Yu (dtyu@bnl.gov)
Kevin G. Yager (kyager@bnl.gov)
Masufumi Fukuto, Hanfei Yan, and Wei Xu
NSLS-II

- $912$M
- 791 m circumference
- 58 beam ports
- 3 GeV, 500 mA

- Each x-ray beam is $\sim 10^{13}$ ph/s
Modern scientific experiments generate massive amounts of data

Complex data analysis consumes scientists’ precious time, distracting from deep scientific questions

We can train machines to perform much of the workflow

Deep learning can extract meaningful insights and detect patterns from massive amount of data; well-suited to image-like datasets
Impact to Materials Science

- NSLS-II beamlines study materials from many perspectives:
  - Complex, multi-component, hierarchical materials
  - Diffraction, scattering, coherence experiments
  - Structure & dynamics across many scales

- If machine automation/learning become part of experimental workflow, scientist is liberated to focus on scientific discoveries

- Will shorten the latency between experiment to deep scientific insight, Impact for material design of battery components, solar PV, etc.

- Develop at CMS and CHX; and extend to other beamlines (SMI, LiX, FXI, HXN)

- To enable **automated materials discovery** across many synchrotron beamlines (Multimodal Analysis)
Objectives

• Low-level: identifying characteristic features in a diffraction image;
• Intermediate-level: detecting the occurrence of a physical process from a sequence of images;
• and 3) High-level: learning and predicting scientifically-meaningful trends.
• On-line Recognition and Prediction with Incremental Information
• The velocity of processing must be commensurate with that of data generation.
Preliminary Work

- Initial work has demonstrated the viability of applying machine-learning methods to synchrotron data.

- Applied machine-vision methods to tagging and classifying x-ray scattering images.

- Used advanced clustering methods to organize synchrotron data.

- Exploring machine-video methods to identify events in time-sequence scattering data.
  - Ongoing collaboration with M.H. Nguyen, Stony Brook University.
New Ideas

- Physical systems have natural hierarchies
- Deep-learning trains multiple levels of features/representations to extract meaning from data
- We will explore machine-learning hierarchies tuned to extract physics layers and meaning from scientific datasets
Technical Approach

- Synchrotron images analyzed using a combination of existing domain and image-analysis techniques, as well as new algorithms
- (Supervised/Unsupervised) Cluster and tag the data with physically-meaningful attributes
- Attributes/features used to extract higher-order trends, and to extract scientifically-relevant insights
- For example, this procedure could be mapped to a four-layer convolution neural network for trend analysis
On-Line Detection

- Off-line Training, On-line detection ➔ On-line Training, on-line detection
- Incremental Update to Existing Training Model
- On-line optimization
Co-Design Deep Learning Applications

Applications

<table>
<thead>
<tr>
<th>WATSON</th>
<th>TORCH</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM</td>
<td>Facebook</td>
</tr>
<tr>
<td>TENSORFLOW</td>
<td>CNTK</td>
</tr>
<tr>
<td>Google</td>
<td></td>
</tr>
<tr>
<td>THEANO</td>
<td>CAFFE</td>
</tr>
<tr>
<td>Université de Montréal</td>
<td></td>
</tr>
<tr>
<td>Berkeley</td>
<td></td>
</tr>
</tbody>
</table>

cuDNN

Tesla    TX-    GPUs    Titan

cuDNN is a library of primitives for deep learning
Future Machine Learning Aided Material Design

- X-ray scattering generates various ‘images’ that can be analyzed using machine-learning.

- Computer-directed beamline experiments would allow the instrument to explore physical parameter spaces, without human intervention.
Conclusion

- Machine-learning is a critical component of **automated materials discovery**; a new experimental mode that:
  - Liberates scientists to work on science
  - Enables computer-controlled ‘intelligent’ exploration of materials questions
  - Accelerate scientific discoveries

- Deep-learning is a crucial tool, allowing the computer to extract physically-relevant meaning from abstract datasets
CFN/NSLS-II Beamline: CMS

- CFN/X9 program has been extremely successful: premiere, highly-sought (>2:1) scattering instrument; highly productive (>25 publications/year)

- **Complex Materials Scattering** beamline will provide:
  - Sample environments for **in-situ** and stimuli-responsive studies of (non-equilibrium) nanomaterials
  - Automation and software for **intelligent exploration** of multidimensional parameter spaces
  - New paradigm for rapid **materials discovery**

**Constituents** + **Processing** → **Structure** → **Functionality**

- **Constituents**: small molecules, polymers, biomolecules, nanoparticles, colloids, porous materials, nanopatterned materials, ...
- **Processing**: compositions, concentration, temperature, pressure, stress, humidity, solvent annealing, ...
- **Structure**: unit cell, crystallinity, grain size, orientation, packing density, symmetry, ...
- **Functionality**: energy storage, catalysis, data storage, light management, electronics, surface-wetting, filtration, medicine, ...
CFN/NSLS-II Beamline: SMI

- **Soft Matter Interfaces** beamline: high-flux and high-resolution grazing-incidence scattering instrument
  - Wide energy range (2 to 24 keV) for resonant scattering on hybrid (soft/hard) materials, including edges relevant to soft matter (P, S, K, Ca)
  - Wide $q$-range for studies of hierarchical materials
  - Microbeams (~2 μm) for mapping of heterogeneous samples
  - High-flux and fast detectors for kinetic, in-situ, and in-operando experiments

![Diagram of SMI Beamline](image)