Improving machine-learning classification of x-ray scattering images

Ronald Lashley [e] , Nicole Meister [a, b], Ziqiao Guan [b], Bo Sun[f], and Dantong Yu [c, d]
[A] Stony Brook University, [b] Centennial High School, [c] Brookhaven National Laboratory,
[D] New Jersey Institute of Technology, [e] Lincoln University (PA), [f] Rowan University

ABSTRACT

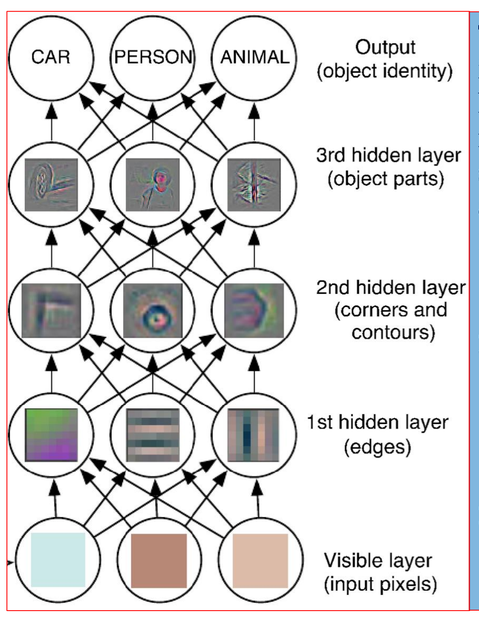
In technology, image recognition software uses various methods to extract information from an image. X-ray scattering images contain visual features such as rings, spots, and halos Using Machine learning, we are proposed to increase the accuracy of image classification by increasing the amount of data. Previously, synthetic images that have been generated have a size of 256 x 256 pixels. Now, we have increased the resolution of the synthetic images by generating synthetic images with a size of 1,000 x 1,000 pixels. Increasing the resolution of the image provides us with more information of the attributes, increasing our ability to recognize and classify the images by tagging them. Here at the Computer Science Initiative, we have changed a few parameters in the previous python code that was used for the original dataset size to increase the size of the generated images to 1,000 x 1,000 pixels. We used the generated synthetic images to train and test our machine learning system that classifies and tags the x-ray scattering images. The real dataset that needs to be classified and tagged are from the National Synchrotron Light Source II (NSLS-II) at Brookhaven National Laboratory (BNL). Once the training and testing is verified to be efficient enough we will apply the machine learning with the real dataset.

INTRODUCTION

X-ray scattering is a powerful technique for probing the physical structure of materials at the molecular and nanoscale, where strong X-ray beams are shined through a material to learn about its structure at the molecular level. This can be used in a wide variety of applications, from determining protein structure to observing structural changes in materials. Modern x-ray detectors can generate 50,000 to 1,000,000 images/ day, thus it's crucial to automate the workflow as much as possible.

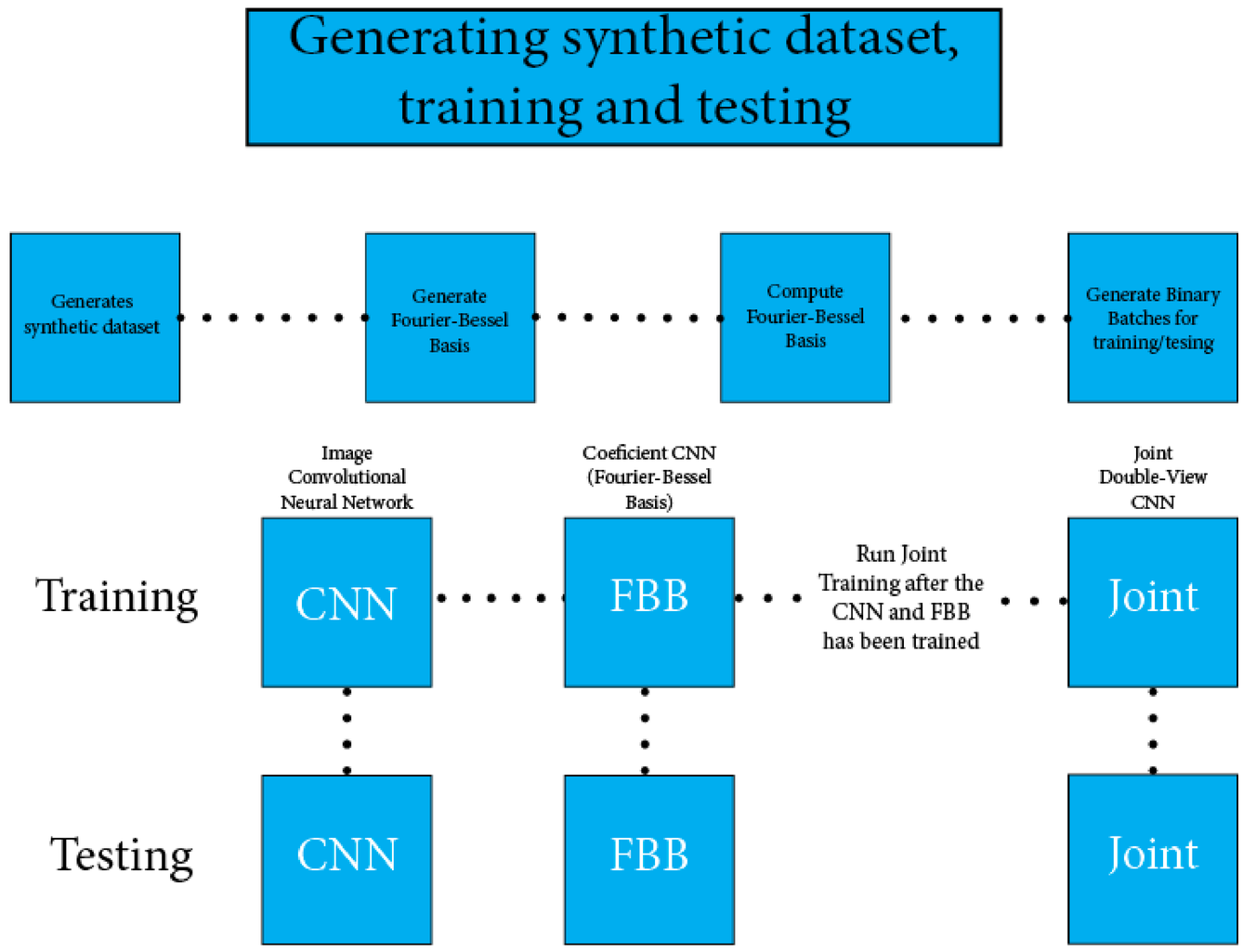
Machine Learning itself is undergoing a shift, with a re-thinking from traditional, naive, neural networks, towards deep learning models where the neural hierarchy is more rational, optimized, and informative. This has already led to clear advances in several fields including computer vision and speech 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.

For machine-learning, there are two types of datasets that are used. The first is real dataset, which is collected by shining powerful x-rays through a particular material and the attribute is labeled by material experts. The second type dataset is synthetic scattering dataset, where the data is generated by simulation software. The simulation software is able to synthetic scattered images based on physics laws. Previously, the synthetic images that were generated had a size of 256 x 256 pixels. Now, we believe if we increase the size of the synthetic images that we generate, we will be able to obtain more information/ data to improve the classification and tagging of the images. If this is seen to be true and efficient then the scientist at the National Synchrotron Light Source II (NSLS-II) can build an x-ray machine or calibrate it to get larger images. The size of synthetic data that we have increased to are 1,000 x 1,000 pixels.



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.



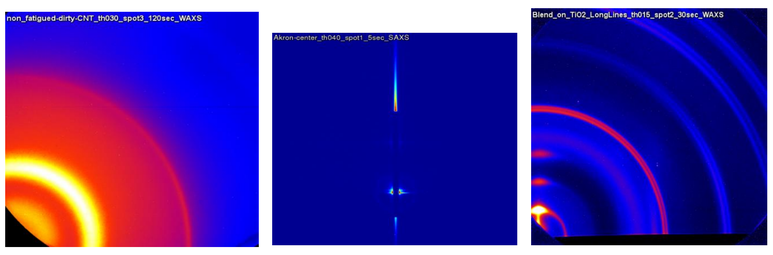
Flow Chart on how the synthetic dataset is generated, trained and tested

METHODS

First we will generate synthetic images to use for testing and training. The synthetic images will be generated using various experiments that have certain variables placed to develop x-ray scattering images. Some of these variables are constant and will be the same across an experiment(s). Previously, the image size that was generated were 256 x 256-pixels. Now we have changed a few parameters so that the image that is generated is 1,000 x 1,000-pixels. The benefit of generating a larger image with a higher resolution is that it provides us with more information of the attributes, increasing our ability to classify the images.

After running the simulation code which generates the synthetic images we generate and then compute the Fourier-Bessel Basis. When it came to computing the Fourier-Bessel Basis, we had to make some changes in the code. Instead of using the Multiprocessing pooling python module like we did with the 256 x 256 dataset, we had to use a loop for parsing. This change had to be made because of the pooling module not being able to handle the larger size dataset. Following that we generated Binary Batches for training and testing. There are 10 batches that are generated which each contain data from the images.

Once the batches were completed we entered training. The 3 types of training are Image Convolutional Neural Network, Coefficient CNN, and Joint Double View CNN (combination of the two). For the 256 x 256 dataset, Image CNN can achieve mean average precision of ~0.5-0.6 with training. The Coefficient CNN can outperform Image CNN by ~0.1. Joint double-view CNN should be slightly better than both. From this we predict that the results will be proportionally the same with the Joint Double-view CNN being better than both the Image CNN and Coefficient CNN individually. For each type of training we used 9 of the 10 batches and saved the 10th batch for testing. Image CNN and Coefficient CNN must be trained before we trained the Joint Double-view CNN but we were able to train the Image CNN and Coefficient CNN simultaneously.



Example of false color images with the "Ring" tag. Tags can include a diverse selection of images, which makes classification of x-ray scattering images difficult

CONCLUSIONS

After training and testing was complete and the accuracy error rate was satisfactory, we can now use the real dataset from NSLS-II to test the system. Once that is done and is satisfactory, we will look to increase the data size again from 1,000 x 1,000 pixel size to 2,000 x 2,000 pixel size. We are looking to gain even more data to improve machine learning classification of x-ray scattering images. With the dataset now being larger, we need to convert from using Alexnet to using VGGnet. VGGnet is suitable for larger data sizes while Alexnet is suitable for smaller data sizes. The process of using Alexnet was successful for the 256 x 256 pixel images. By switching to VGGnet, we expect the training and testing to be more efficient.

ACKNOWLEDGEMENTS

This project was supported in part by the National Science Foundation, Louis Stokes Alliances for Minority Participation (LSAMP) at Lincoln University of Pennsylvania under the LSAMP Internship Program at Brookhaven National Laboratory. I wish to thank my host, Dr. Dantong Yu, for his professionalism and generosity during the NSF program. I would also like the thank my Visiting Faculty Professor and Mentor, Dr. Bo Sun, for her assistance, guidance, and for selecting me as the student to join her team. Further, I would express my gratitude to the Office of Science Education Program, and all who continue to so willingly assist interns in that branch. I very much appreciate the efforts of the National Science Foundation, LSAMP at Lincoln University of Pennsylvania with regard to their support. Finally, I wish to acknowledge the hospitality and kindness of Brookhaven National Laboratory and the Department of Energy.