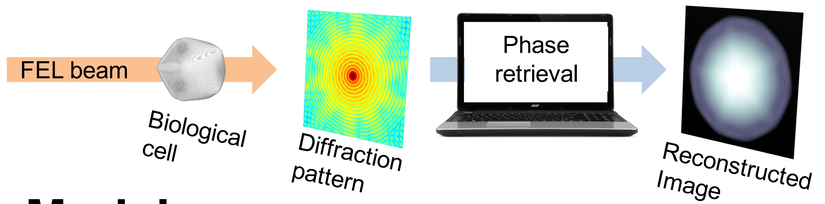
Data classification for diffraction images

NATIONAL ACCELERATOR LABORATORY

Po-Nan Li
Dept. of Electrical Engineering, Stanford & Biosciences Div., SLAC
liponan@stanford.edu

Abstract

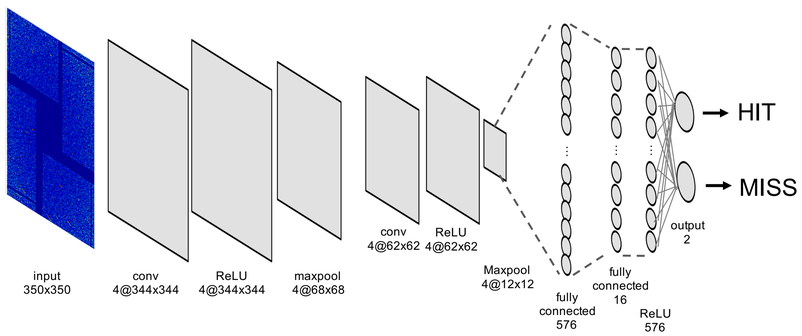
Nowadays the free-election laser such as LCLS has the capability of thousands of diffraction image in minutes, the transfer and storage of such large volume of data is however costly. While a large portion of images in the dataset are “miss,” currently crystallographers rely on hand-tuned peak finding algorithms to identify hits and misses [1]. Here we developed and trained a CNN-based data classifier for diffraction images, which achieves up to 95% accuracy.



Model

We trained a convolutional neural network (CNN) consisting of 2D convolution, ReLU, max-pooling and fully-connected layers and use it for binary classification, i.e. to predict a given image is a hit or miss.

- Loss function: binary cross-entropy
- Optimizer: stochastic gradient descent (learning rate: 0.002; momentum: 0.9)
- Batch size: 100; 3 epochs for each run

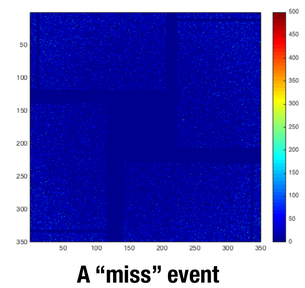


Data

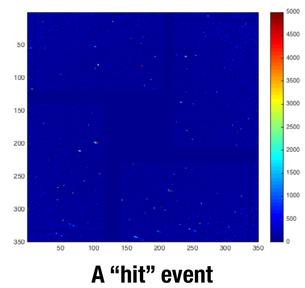
We used diffraction images from CXIC0415, a serial femtosecond crystallography (SFX) [2,3] experiment conducted at LCLS’s CXI beamline.

- 10,874 events from 3 runs for training
- 5,732 events from another 3 runs for validation.

Each event contains a 1750x1750 image, which represents diffraction intensities at different spatial frequencies. To speed up and reduce memory need, we cropped the image to its central 350x350 part, where Bragg peaks (if any) can be clearly observed.



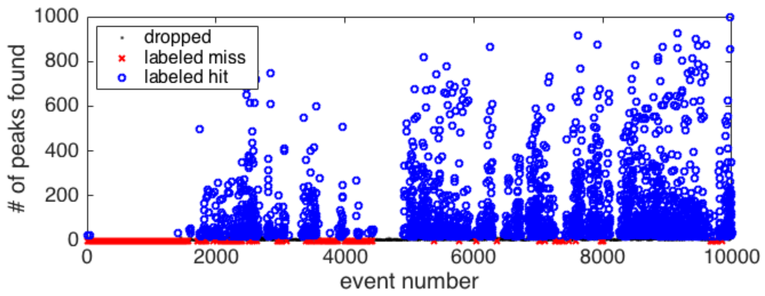
A “miss” event



A “hit” event

Feature and labels

A peak finding program was performed for each event, and we use the number of peaks found as criterion of hit/miss: events whose number of peaks are within first 25% percentile are labeled hit; last 25% are labeled miss.



Result

Prediction

Actual	Hit	Miss
Hit	0.49430	0.00570
Miss	0.00230	0.49770

Training data (50%+50%)
Sensitivity: 0.99537
Specificity: 0.98867

Prediction

Actual	Hit	Miss
Hit	0.44522	0.05478
Miss	0.00384	0.49616

Testing data (50%+50%)
Sensitivity: 0.99145
Specificity: 0.90057

Discussion

We reach very high sensitivity for hit events. Our CNN-based prediction is 100x faster than conventional method. False negative rate a bit high, probably due to weak hits.

Future work

- Train and test with full diffraction images
- Cross-experiment training and validation
- Extend to categorical classification for space groups, single/multiple-hit, etc.

Acknowledgements

The author is grateful to Drs. Chun Hong Yoon and Henry van den Bedem for their mentorship and providing access to experiment datasets.

References

[1] C.H. Yoon *et al.*, Sci. Rep. **6**, 24791 (2016).
[2] J. Tenboer *et al.*, Science **324**, 1246 (2014).
[3] M.S. Hunter *et al.*, Nature Comm. **7**, 13388 (2016).
[4] X. Duan *et al.*, Sci. Rep. **6**, 34406 (2016).