AI Data Quality Monitoring with Hydra

Thomas Britton, David Lawrence, Torri Jeske, Kishansingh Rajput
Thomas Jefferson National Accelerator Laboratory, Newport News, VA 23606, USA

Abstract. Hydra is an extensible framework for training and managing AI for near real time monitoring that aims to replace the tedious and repetitive data quality monitoring activities the shift crew and online monitoring coordinator typically perform. It continuously scans incoming data in the form of monitoring plots for signs of problems, flagging them for human review. A web app was developed such that experts can efficiently label images for training. Labels are stored in a database for use in training and model validation. Backed up by a comprehensive database, it utilizes an additional web based front-end for viewing the current monitoring status from anywhere in the world. The system has been in production use for the GlueX experiment at Jefferson Lab for more than 2 years with new features still under active development.

1. Introduction

Online monitoring is a critical component to successful data taking during nuclear physics experiments. Given the large expense of experiments and the quantity of data generated, it is essential to identify anomalous or faulty detector behavior as soon as possible. This is difficult when most experiments rely on shift takers with varying levels of expertise and a lack of standardized monitoring intervals.

At Jefferson Lab, the four experimental halls (A, B, C, and D) generate monitoring histograms throughout their respective data taking sessions, usually referred to as "runs". For each run, a unique "run number" will contain a set of histograms generated for each detector system. In Hall B, experiments using the CLAS12 spectrometer [1] generate over 400 monitoring histograms per run number which are analyzed by the shift crew for potential problems. The sheer volume of histograms to look at often leads to missed problems. In Hall D, an Online Monitoring Coordinator is charged with an additional review of all plots from the preceding 24 hours and providing summaries as to the status of the detectors. This amounts to reviewing hundreds of plots each day, a time consuming and labor intensive task. This is made even more difficult when special tests are conducted and not well documented. These tests can often look like problems to the non-expert when in fact they represent a specific special configuration and should not be treated as problematic.

2. Hydra

With numerous histograms to analyze and a natural variation among the observers with which the analysis is done, there is a need to standardize and automate online monitoring. Hydra is an AI based system that aids shift takers with online monitoring during experiments. Hydra provides a scalable framework for managing the training, deployment, and running of a multitude of machine learning models. Hydra models are trained to classify monitoring images from various detector systems in near real time during data taking. For the GlueX experiment in Hall D [2],

this amounts to classifying each monitoring histogram approximately once per minute (a time period necessary to gather sufficient statistics). Inception V3 [3], created by Google, was chosen as the default model type in order to reduce the overhead spent developing a bespoke model topology to complete image classification tasks for a specific plot.

2.1. Data Set

Hydra initially utilized the already saved and collated monitoring images from the different experiments in each Hall at Jefferson Lab. To make labeling historical data sets easier, a web site was developed that displays available labels and each monitoring histogram, sorted by detector and time-ordered by the unique run number. These labels and references to the corresponding images are saved into a database to be used for training. A snapshot of the web page is shown in Fig. 1. Detector experts are typically labeling a subset of images throughout the duration of the experiments. This unbiased set of saved images is configurable within Hydra. In addition, Hydra is configured to save specific images it thinks it needs for future training and validation, namely images indicating problematic behavior and images for which the model is less confident in. These images automatically appear on the labeling website.

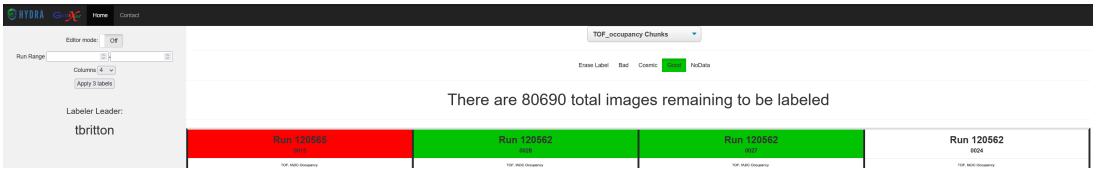

Figure 1. The Data Labeler web page. Monitoring histograms are sorted by detector and run number. Each image can then be quickly labeled utilizing the available labels that appear above the set of images. The images and labels are stored in a database to be used for training.

2.2. Training

A model is trained for each detector system that has the associated monitoring histograms labeled. Most of the time, the detectors work as designed and this results in many more 'Good' images compared to few examples of 'Bad' images. In order to balance the training, we can strategically reduce (i.e. under sample) the number of "Good" images. An example "Normal" and "Under" sampling are shown in Fig. 2. Essentially, we use all examples of "Bad" images, which tend to have much higher variance than nominal operation, and under sample the "Good" images, which tend to be much lower in variation. This has the effect of training an equally accurate model in about one fifth of the time (for detectors with a sufficiently large number of labeled plots). If there are no examples of a specific label type, that label is automatically dropped from the training. Details from the training, including number of epochs, loss, model name, and the path to saved model location is stored in a database.

2.3. Testing

Each model is validated before being put into production. After training, inference is run on each image associated with that model. If there are discrepancies between the label given by the expert and Hydra's classification, then those images, labels, and Hydra's confidence are tabulated for further review by the associated expert. Typically, a model will be between 90% and 95% accurate after training. Further review and correction of human errors in labeling raises that value to about 98% without retraining. If there is a "large" number of these discrepancies, as compared to the total data set, the model is retrained. An "enhanced" confusion matrix (Fig. 3) is formed that displays counts in each matrix element, as is typical for confusion matrices,

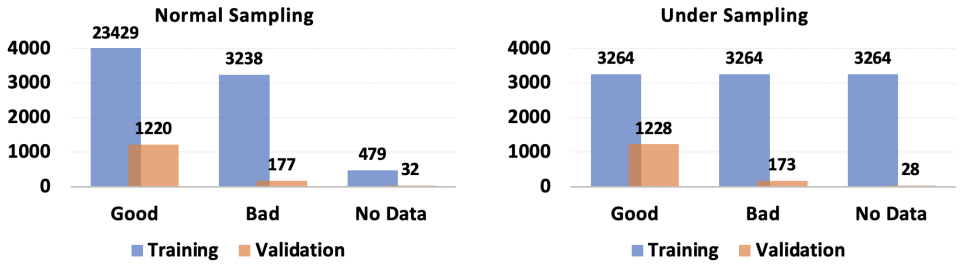

Figure 2. 'Normal' and 'Under' sampling schemes. Note the counts are different between the two schemes due to changes in random selection. Given enough statistics the 'Under' sampling scheme produces models just as accurate as other sampling methods in around one-fifth the time.

along with the associated confidence levels of the model's predictions. This adds a level of interpretability and allows users to answer questions such as: "When the model is wrong, how confident is it?". This can help distinguish between modalities, which may need more work/training to resolve and those that are borderline and may not be imminently correctable.

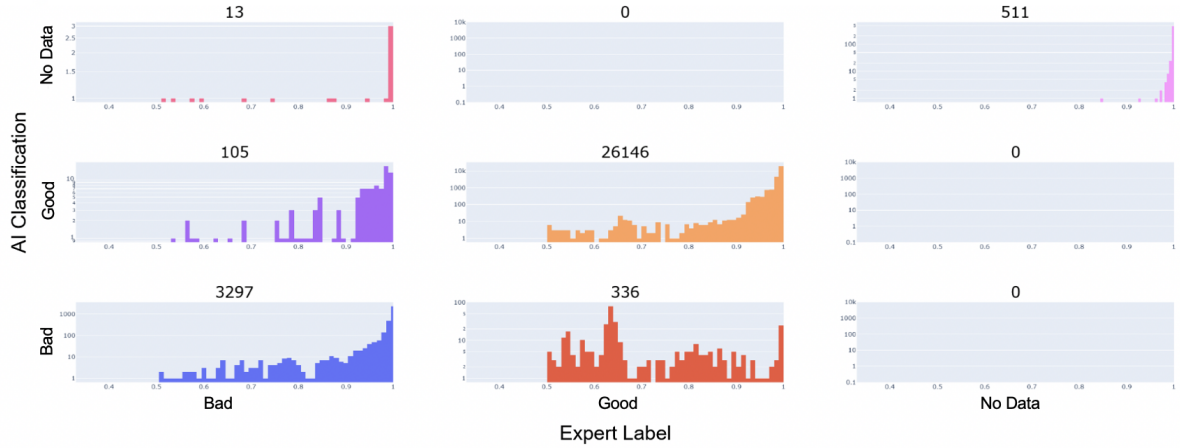

Figure 3. An example 'enhanced' confusion matrix. In this matrix all combinations of AI prediction and Expert Labels are plotted; the number above the distribution represents the count in each element and when taken alone is the standard confusion matrix. The distribution represents the confidence (between 0.0 and 1.0) the model had in its classification.

3. Operation

Hydra provides near real-time monitoring in the form of a web page, referred to as the Hydra Run page. A snapshot of the page for Hall B at Jefferson Lab is shown in Fig. 4. The page displays all of the latest monitoring histograms and the associated run number, date and time, and Hydra's prediction and confidence. In the event Hydra is confident that an image is "Bad", the square is highlighted in red to notify the shift crew. Hydra can also be connected to the alarm system and provide an audible alarm to the shift crew. A separate page, referred to as Hydra Log, contains histograms that Hydra either labeled as "Bad" or was not confident in. This page aids experts in reviewing the status of the detectors by focusing on images associated with potential issues.

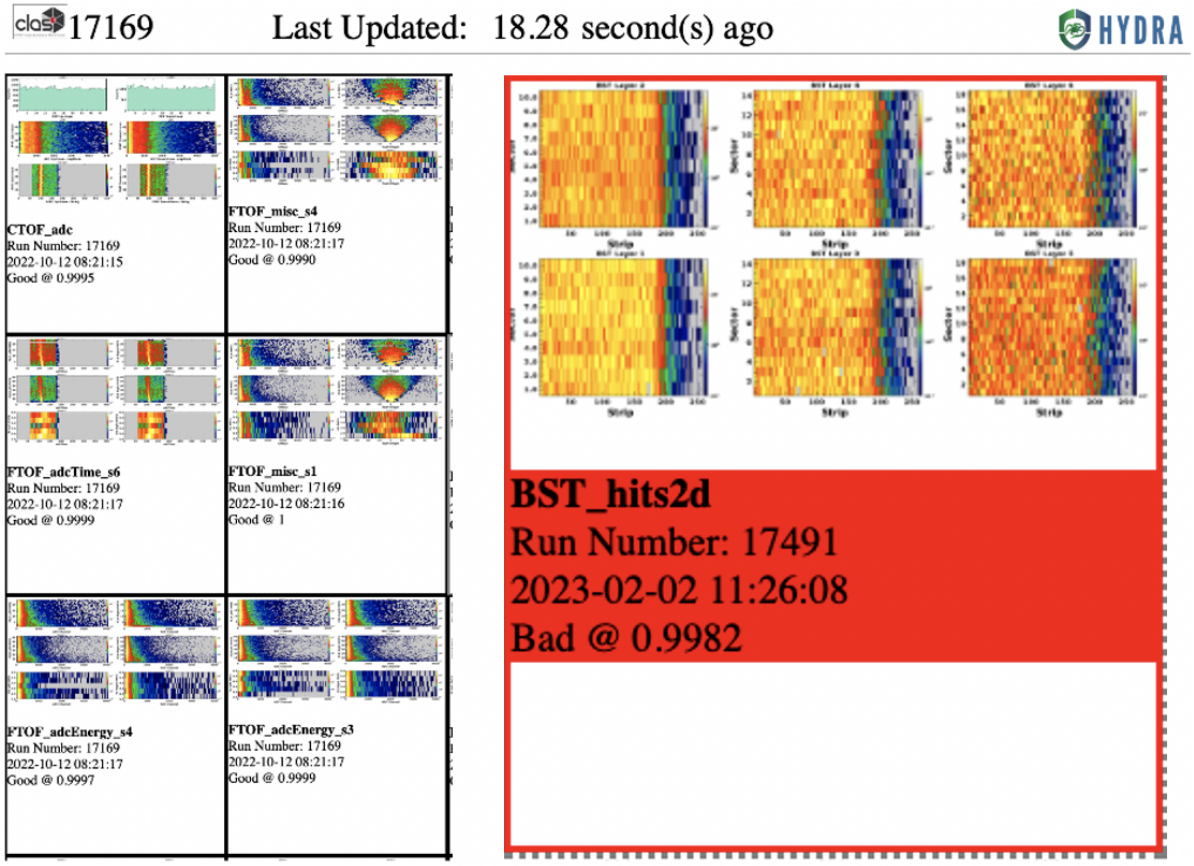

Figure 4. The Hydra Run page as seen by the shift crew and detector experts. Each block contains the image name, run number, date and time, and Hydra's classification and confidence. If Hydra is confident that an image is 'Bad', the block is highlighted in red and, if configured, an audible alarm will alert the shift crew to the problem.

4. Future Work

Hydra has been used in Hall D at Jefferson Lab since 2019. Deployments in all of the other Halls at the lab are ongoing. In addition, Gradient Class Activation Maps [4] (GradCAM) are being implemented for the purpose of understanding why Hydra classified an image as "Good" or "Bad". An example use case of GradCAM is shown in Fig. 5. This is a sample online monitoring image from Hall B. The detector expert labeled the image as "Bad", and Hydra correctly classified the image. In order to understand Hydra's classification, we can utilize GradCAM to identify regions of the image that are important for the classification. For this particular image, Hydra is looking at the low occupancy histogram in the bottom right panel. This is confirmed by the detector expert.

5. Summary

Hydra is an AI based system that performs online monitoring in a standardized, efficient manner. It uses the same plots developed for human shift takers to monitor data quality during an experiment. Hydra is able to classify images in near real time and can alert the shift crew of significant problems as they occur. Hydra has a user friendly interface via a web browser that allows the shift crew and detector experts to monitor experiments from anywhere in the world.

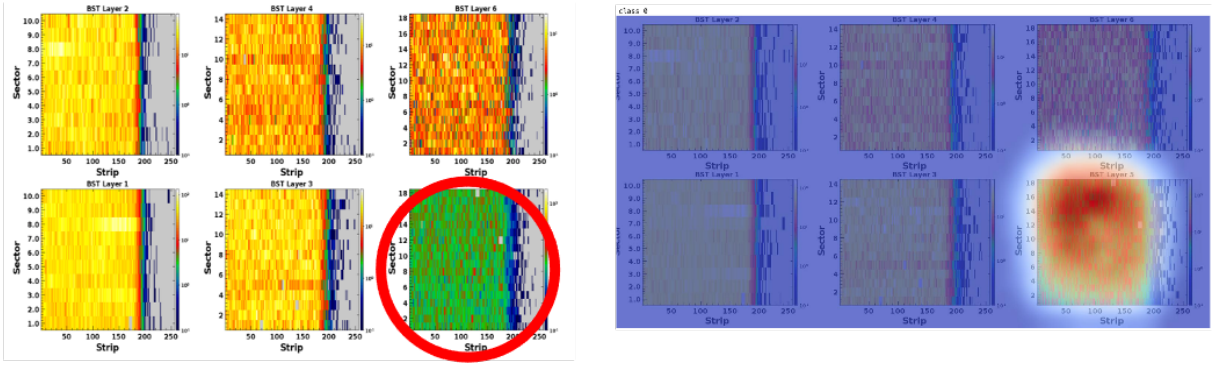

Figure 5. Left: A typical online monitoring histogram from Hall B. The red circle indicates a low occupancy, as identified by the detector expert. Right: The original image with a GradCAM heat map overlayed, identifying the regions of the image that are important for the specific classification.

Initially deployed in Hall D at Jefferson Lab, further deployments in each of the remaining 3 halls at Jefferson Lab is in progress. In addition, advances in computer vision related to interpretability are being implemented to help identify particular areas of the images that are important for its classification. Such techniques could be used to identify specific types of detector behavior, such as a dead channel.

6. Acknowledgements

Jefferson Science Associates, LLC operated Thomas Jefferson National Accelerator Facility for the United States Department of Energy under U.S. DOE Contract No. DE-AC05-06OR23177.

References

[1] V.D. Burkert et al. The clas12 spectrometer at jefferson laboratory. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 959:163419, 2020. URL: https://www.sciencedirect.com/science/article/pii/S0168900220300243, https://doi.org/https://doi.org/10.1016/j.nima.2020.163419 doi:https://doi.org/10.1016/j.nima.2020.163419.

[2] S. Adhikari et al. The GlueX beamline and detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 987:164807, jan 2021. https://doi.org/10.1016/j.nima.2020.164807 doi:10.1016/j.nima.2020.164807.

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL: http://arxiv.org/abs/1512.00567, http://arxiv.org/abs/1512.00567 arXiv:1512.00567.

[4] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. https://doi.org/10.1109/CVPR.2016.319 doi:10.1109/CVPR.2016.319.