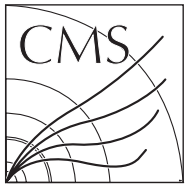
Available on CMS information server

CMS CR -2022/014



The Compact Muon Solenoid Experiment

Conference Report

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland

15 January 2022 (v4, 29 January 2022)

Machine Learning applications for Data Quality Monitoring and Data Certification within CMS

Vichayanun Wachirapusitanand for the CMS Collaboration

Abstract

The Compact Muon Solenoid (CMS) detector is getting ready for datataking in 2022, after a long shutdown period. LHC Run-3 is expected to deliver an ever-increasing amount of data. To ensure that the recorded data has the best quality possible, the CMS Collaboration has dedicated Data Quality Monitoring (DQM) and Data Certification (DC) working groups. These working groups are made of human shifters and experts who carefully watch and investigate histograms generated from different parts of the detector. However, the current workflow is not granular enough and prone to human errors. On the other hand, several techniques in Machine Learning (ML) can be designed to learn from large collections of data and make predictions for the data quality at an unprecedented speed and granularity. Hence, the data certification process can be considered as a perfect problem for ML techniques to tackle. With the help of ML, we can increase the granularity and speed of the DQM workflow and assist the human shifters and experts in detecting anomalies during data-taking. In this presentation, we present preliminary results from incorporating ML to highly granular DQM information for data certification.

Presented at *ACAT2021 20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research*

Machine Learning applications for Data Quality Monitoring and Data Certification within CMS

Vichayanun Wachirapusitanand on behalf of CMS collaboration

Chulalongkorn University, 254 Phyathai Rd, Wangmai, Pathumwan, Bangkok, Thailand, and Euproean Organization for Nuclear Research (CERN), Geneva Switzerland

E-mail: vichayanun.wachirapusitanand@cern.ch

Abstract. The Compact Muon Solenoid (CMS) detector is getting ready for datataking in 2022, after a long shutdown period. LHC Run-3 is expected to deliver an ever-increasing amount of data. To ensure that the recorded data has the best quality possible, the CMS Collaboration has dedicated Data Quality Monitoring (DQM) and Data Certification (DC) working groups. These working groups are made of human shifters and experts who carefully watch and investigate histograms generated from different parts of the detector. However, the current workflow is not granular enough and prone to human errors. On the other hand, several techniques in Machine Learning (ML) can be designed to learn from large collections of data and make predictions for the data quality at an unprecedented speed and granularity. Hence, the data certification process can be considered as a perfect problem for ML techniques to tackle. With the help of ML, we can increase the granularity and speed of the DQM workflow and assist the human shifters and experts in detecting anomalies during data-taking. In this presentation, we present preliminary results from incorporating ML to highly granular DQM information for data certification.

1. Introduction

Since the start of Large Hadron Collider (LHC) operations in 2009, the Compact Muon Solenoid (CMS) detector [1] has shown a stellar performance in recording particle collisions. During CMS Run 2 operations, which took place during 2015-2018, 98% of data recorded by the detector is classified as usable for physics analyses. This ratio illustrates the efficiency of the detector, along with the scientists and technicians who operate it, in recording data with such high quality.

In the next data-taking campaign, Run 3, starting in early 2022, CMS is expected to collect about twice the amount of pp collision data collected during Run 2 [2, 3, 4]. This translates to a recorded luminosity goal of 300 fb^{-1}, corresponding to 500 trillion collision events expected to occur in CMS. In this paper, we will discuss the motivation for using ML techniques to assist the existing human-centric data quality monitoring (DQM) and data certification (DC) process.

2. Current workflow of Data Quality Monitoring and Data Certification

To maintain a good amount of quality data for physics analyses, CMS Collaboration has DQM and DC workflows. Both workflows oversee live DQM during data-taking (also called online operations), and data certification after the data has been recorded from the detector and fully reconstructed (also called offline operations).

During collisions, the CMS detector sends a small portion of the data stream to the online DQM backend. The backend processes the data and sends histograms directly to the DQM GUI, a web application displaying histograms from the CMS detector for DQM shifters. The shifters then look at the histograms detailing several metrics from all subdetector components. If any anomalies appear in histograms, they must notify and consult with subsystem experts to determine whether they will affect the data quality.

Data recorded from the CMS detector is stored as a *run*, or a period where CMS had stable detector conditions. A run contains several *lumi sections* (LS), where each LS is defined as a subsection of one run for 2^{18} LHC orbits, or approximately 23 seconds. Due to the large amount of data, DQM shifters and subsystem experts certify the data on a run-by-run basis, both during online and offline operations. This poses two problems in the current workflow:

- The certification process currently implemented does not have enough granularity. Anomalous LS may occur inside a run, and certifying data on a run-by-run basis may miss them, even with information available from the Detector Control System explaining which subsystem is currently active during each LS.
- Even with a run-by-run certification, there are a large number of runs that have to be certified by shifters, and for each run, shifters are required to inspect hundreds of histograms derived from subdetector information. This can cause fatigue after inspecting a large number of histograms, resulting in human errors in judging the quality of the data.

As stated earlier, Run 3 data taking is starting in early 2022 and will have a much larger amount of data to be certified, which can cause additional problems for shifters. On the other hand, ML techniques can efficiently handle a large amount of data and make predictions at a speed unmatched by humans. Furthermore, once the High-Luminosity LHC upgrade [5] is in place in the late 2020s, the amount of data due to an increased luminosity will be growing still. Therefore, ML has the potential to substantially aid shifters to detect anomalies.

3. Current unsupervised ML techniques being explored

The CMS Collaboration is currently exploring the possibility of ML techniques to classify anomalous data and find ways to implement them into the DQM and DC workflow. In this current stage, several techniques that are being explored by the CMS Tracker team will be discussed below. The following work is only applied to the CMS Tracker system [6] as a proof-of-concept, but can be applied to other subsystems as well. More information regarding the work presented in this section can be found in [7].

3.1. Non-negative Matrix Factorization (NMF)

Non-negative Matrix Factorization (NMF), as described in [8] is a matrix decomposition technique which decomposes one large matrix V of dimension $m \times n$ into two smaller matrices H with dimensions $m \times r$ and W with dimensions $r \times n$ such that $V = HW$.

For anomaly detection, we can represent histograms from several good LS in a form of a column matrix of $m \times 1$ dimensions, where one histogram contains m bins, find a matrix H of $m \times n$ dimensions (where $n < m$), and rewrite the histograms in a form of $n \times 1$ dimension column matrix. Effectively, we can expect that the histogram from a good LS can be written as a linear combination of n *components*. If a good LS is deconstructed with this method, we can reconstruct it perfectly. On the other hand, if a bad LS is deconstructed with this method, its histogram will be reconstructed badly with a lot of errors.

In this work, the NMF technique is applied to histograms representing cluster charge data from Pixel barrel layer 2, recorded in 2017. As seen from Figure 1, bad LS can be detected based on unusually high reconstruction errors, as shown in Figure 1 (right), which is expected. Hence, this method of anomaly detection is promising in this use case.

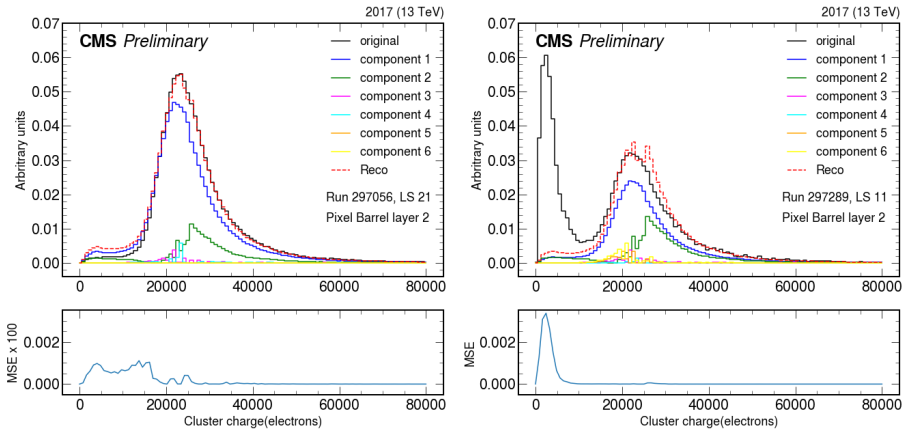


Figure 1. Example of histogram reconstruction with NMF technique from a good LS (left) and a bad LS (right). [7] The training dataset, recorded in 2017, is decomposed into 6 components. A reconstructed histogram (in red) is then calculated based on each component. For a good LS, the reconstruction is almost perfect, while for a bad LS, the reconstruction shows a high error depicted in the bottom plot.

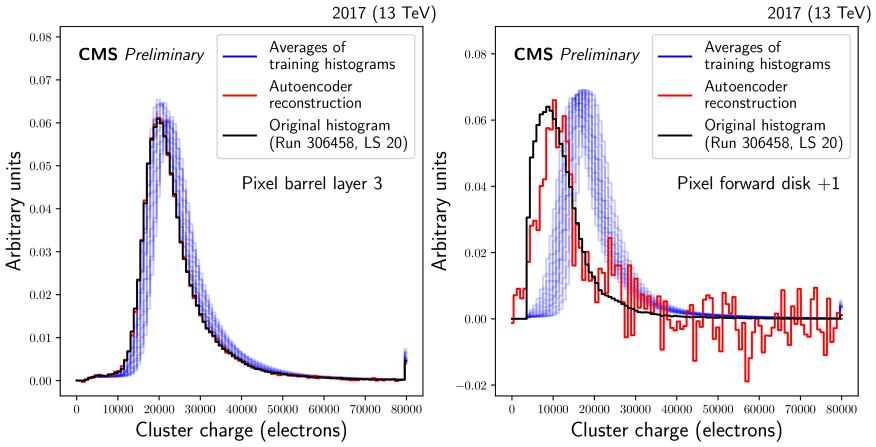


Figure 2. Reconstruction example of a one-dimensional autoencoder trained with data recorded in 2017 [7]. Left: histogram from a good LS; and right: histogram from an anomalous LS. Each plot shows different types of histograms or monitoring elements. The blue histograms represent histograms that are used to train an autoencoder. The red histogram is an anomalous reconstruction of the autoencoder when compared to the original histogram in black.

3.2. One-dimensional autoencoders

Traditional autoencoders can also be used as a method to detect anomalous histograms. They can be trained with a collection of known good histograms to reconstruct an input histogram with good accuracy. With a trained autoencoder, we can expect that if a histogram is from a good LS, there will be a low reconstruction error. On the other hand, we can expect to see a high reconstruction error if an input histogram is from a bad LS.

As shown in Figure 2, the autoencoder is trained with histogram data representing cluster charge monitoring histograms for Pixel barrel layers and forward disks, recorded in 2017. An anomalous histogram can be detected via a bad reconstruction of an autoencoder trained with a collection of good histograms.

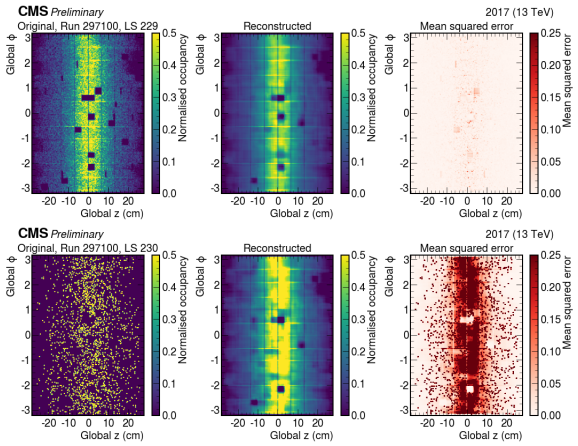


Figure 3. Reconstruction example of ResNets trained with data recorded in 2017. [7] Each row represents one LS. The left column represents original histograms, the middle column represents histograms reconstructed from ResNets, and the right column represents reconstruction error. As seen in the bottom row, there is a discrepancy in an original histogram and a reconstructed histogram, and its reconstruction error is shown as dark red indicating that there is a high reconstruction error for that LS.

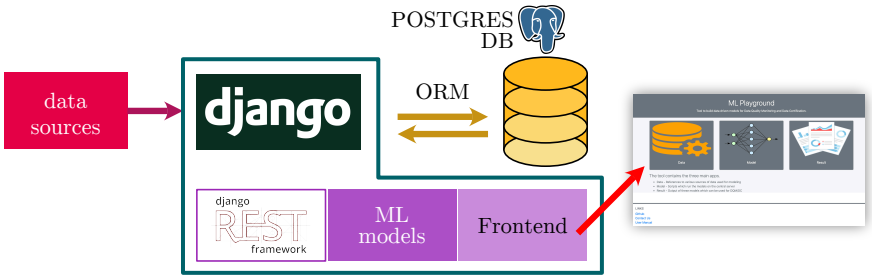


Figure 4. Rough schematics of ML playground, based on Django web framework.

3.3. Two-dimensional residual networks

Residual networks, or ResNets [9], is a class of neural networks. In traditional neural networks, one hidden layer is connected directly to the next layer in the network and so on. ResNets, on the other hand, consists of skipping connections where one hidden layer can have its input added to the output of another layer in the network. ResNets can also be constructed in the same fashion as autoencoders, where the network must reconstruct input histograms.

In this work, ResNets are trained with occupancy figures from pixel barrel layer 1 in global z and phi coordinates, recorded in 2017. This type of figure represents the occupancy of particle tracks that go through the CMS Tracker subsystem, which is crucial to the detection of particles produced from collisions. As shown in Figure 3, an anomalous histogram can be detected via a high reconstruction error (calculated as the mean squared error of each pixel in a histogram). In this work, ResNets are used to detect anomalies in two-dimensional histograms, but they can also be used to detect one-dimensional histograms as well.

3.4. ML playground

In addition to ML techniques that are being actively explored, a Django-based framework [10] called ML playground is being developed in parallel. This framework aims to consolidate training dataset information, automate ML training based on selected datasets, and generate training reports based on the ML model performance.

4. Challenges to developing ML based anomaly detectors and integration to the human-centric certification process

The current certification process relies heavily on human shifters. Implementing ML-based anomaly detectors in this human-centric process poses several challenges, both due to the

development of the detector itself and the integration to the process as follows:

- As with all anomaly detection tasks, it has a class imbalance problem. The training data is usually obtained from recorded data which has most anomalies removed by human experts during data taking over a long period. Thus, anomalous histograms that can be used as testing datasets are scarce compared to good histograms.
- Relying on human-based labels seems to be a reasonable approach during training. However, human-based labels may be incorrect, both due to run-by-run certification in the past, human error and fatigue. The final decision for one LS to be included in physics analyses is also derived by subdetectors under the condition that all subsystems must be good for the data to be labelled as good.
- Although ML-based anomaly detection algorithms can detect anomalies with very high accuracy, they may miss some anomalies that have not been seen before. As such, this ML system does not intend to replace human shifters and experts completely, but rather assist humans to focus on problematic histograms and skip histograms that look good.
- With an upcoming High Luminosity LHC upgrade [5], expected to start operations in 2025, the data taking rate is expected to increase significantly, making the use of ML even more compelling. The detector conditions during data-taking, such as calibration, alignments, and changing detectors, are expected to change over time, and ML techniques will have to adapt to those changes as well. Thus, there is ongoing work to assess the amount of data needed for the optimal training of ML algorithms.

5. Summary

Throughout this paper, we have discussed several unsupervised ML techniques that can aid humans in DQM/DC workflow. These techniques are a good starting point into developing a ML-based system that can be implemented during Run 3 data taking or the High Lumiosity LHC era in the future.

Acknowledgments

The author would like to acknowledge Chulalongkorn University and the Chulalongkorn Academic into Its 2nd Century Project Advancement Project (Thailand) for financial support.

References

[1] CMS Collaboration, 2008, *JINST* **3** S08004
[2] CMS Collaboration, 2021, *EPJ* C **81** 800
[3] CMS Collaboration, 2017, CMS-PAS-LUM-17-004 https://cds.cern.ch/record/2621960
[4] CMS Collaboration, 2018, CMS-PAS-LUM-18-002 https://cds.cern.ch/record/2676164
[5] O. Aberle et al., 2020, *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report* (Geneva: CERN)
[6] CMS Collaboration, 2017, *The Phase-2 Upgrade of the CMS Tracker* CERN-LHCC-2017-009 (Geneva: CERN)
[7] CMS Collaboration, 2021, CERN-CMS-DP-2021-034 https://cds.cern.ch/record/2799472
[8] D.D. Lee and H.S. Seung, 2000, *Adv. Neural Inf. Process. Syst.* **13** 556-562
[9] K. He, X. Zhang, S. Ren, and J. Sun, 2015, Deep Residual Learning for Image Recognition *Preprint* 1512.03385
[10] Django Software Foundation, 2022, Django (Version 4.0.1) https://www.djangoproject.com/