EPJ Web of Conferences **214**, 01007 (2019) https://doi.org/10.1051/epjconf/201921401007
CHEP 2018

Improving data quality monitoring via a partnership of technologies and resources between the CMS experiment at CERN and industry

Virginia Azzolini[1,*], *Michael* Andrews[2], *Gianluca* Cerminara[3], *Nabarun* Dev[4], *Colin* Jessop[4], *Nancy* Marinelli[4], *Tanmay* Mudholkar[2], *Maurizio* Pierini[3], *Adrian* Pol[5], and *Jean-Roch* Vlimant[6]for the CMS Collaboration

[1]Massachusetts Institute of Technology (US)
[2]Carnegie-Mellon University (US)
[3]CERN (CH)
[4]University of Notre Dame (US)
[5]Université Paris-Saclay (FR)
[6]California Institute of Technology (US)

Abstract. The Compact Muon Solenoid (CMS) experiment dedicates significant effort to assess the quality of its data, online and offline. A real-time data quality monitoring system is in place to spot and diagnose problems as promptly as possible to avoid data loss. The a posteriori evaluation of processed data is designed to categorize it in terms of their usability for physics analysis. These activities produce data quality metadata. The data quality evaluation relies on a visual inspection of the monitoring features. This practice has a cost in term of human resources and is naturally subject to human arbitration. Potential limitations are linked to the ability to spot a problem within the overwhelming number of quantities to monitor, or to the lack of understanding of detector evolving conditions. In view of Run 3, CMS aims at integrating deep learning technique in the online workflow to promptly recognize and identify anomalies and improve data quality metadata precision. The CMS experiment engaged in a partnership with IBM with the objective to support, through automatization, the online operations and to generate benchmarking technological results. The research goals, agreed within the CERN Openlab framework, how they matured in a demonstration application and how they are achieved, through a collaborative contribution of technologies and resources, are presented.

1 Introduction

Data Quality (DQ) assessment is an important aspect of every High Energy Physics (HEP) experiment. In the era of the Large Hadron Collider (LHC)[1], and of its highly sophisticated detectors, a prompt feedback on the quality of the recorded data is needed to maximize the effectiveness of data taking efforts, and offline quality verification is required to guarantee a good baseline for physics analysis. Following the CERN's [2] long standing tradition of collaboration with industry and research institutes, the CMS experiment[3] engaged into a

*e-mail: virginia.azzolini@cern.ch

partnership with IBM[4], within the CERN Openlab [5] framework, with the objective to improve operations, detector understanding and to generate benchmarking technological results. This document outlines, in 2, the particular physics requirements that govern the choice and design of the CMS-IBM project, while 3 describes in some detail the properties of the Machine Learning (ML) methods exploited in the two use cases CMS has studied. The preliminary results, presented at the conference and summarized in 4, illustrate the progresses achieved over the past few years and anticipate the way the future implementation will take place in everyday operations.

2 The project

The CMS-IBM partnership was born with the goal to improve data quality monitoring operations in two ways: facilitate the current monitoring of the quality of each data taken and look towards the automatic detection of anomalies in data.

The data collected by the CMS detector is monitored in real-time to guarantee its quality during the data taking, and to detect, as promptly as possible, deviations from normal behavior.

The real-time effort uses as tool a Graphical User Interface (GUI)[6]. This CMS typical graphical setup is composed of a set of histograms, produced for each component of the detector, in order to monitor key quantities and compare them to pre-defined reference distributions. Trained physicists monitor the produced histograms, mainly by visual inspection and through a set of statistical tests (χ2, Kolmogorov-Smirnov, etc) and report issues found. So far, the current system, used in LHC Run 1 (2010-2012) and Run 2 (2015-2018), has performed very well, but not without difficulties, that can be summarized as follow:

- it is expensive, in terms of *human resources*, about 250 people per year. The online active monitoring is in fact organized in pace of 8-hour shifts covering 24/7 operations, that happens smoothly only assuming shifters have received ample training and clear instructions are regularly maintained;
- its effectiveness is limited by the *volume budget*, about 50 elements (2D histograms, etc) that each shifter is expected to evaluate during a run. The amount of quantities, that a human can process in a finite time interval, can prevent an accurate real-time scrutiny;
- it makes *assumptions* on our level of understanding. To allow operations to be performed by informed, but not specifically expert physicists, it is necessary to introduce some a priori strategy, tailored towards some reference conditions and *failure modes*. Being this influenced by the present understanding of the detector conditions it could potentially imply a limitation in discovering problems.

CMS collaboration hence imagined a project that, making use of the most modern ML techniques, could help reduce the operational flaws. We're guiding the development of the demo application to have three main properties:

- it will use **intelligent** deep learning technique to guarantee the present standards but at the cost of less burn for human shifters.
- it will add a **predictivity** layer to the operations, being not customized toward only few/specific scenarios.
- it will include time, as an extra dimension, being capable to **proactively** identify trending problems. For future development, the exploitation of sequence-based models, such as Long-Short Term Memory (LSTM)[7] to track the detector's time evolution and the inclusion of integrated detector metadata (like subsystems readiness to take data (Detector Control System LOW /HV voltage) are under investigation.

The current rule based system relies on two types of histograms to monitor physics and calibration quantities: Task and Client histograms. Task Histograms provide a purely statistical description/picture of key quantities (mean, RMS, etc.) without making a judgement about the quality of the data (e.g. 2D η/ϕ map of reconstructed and calibrated signal (commonly called *hits*); Client Histograms analyze the output of Task histograms and compare them against expected thresholds to provide an interpretation of the quality of the data.

With a ML-based quality monitoring, we aim to eliminate the need for hand-coded rules, with the simplest choice of a supervised approach, based on quality labels provided by experts, that would learn rules and statistics directly from raw data; while with semi-supervised learning of quality interpretation, we could potentially flag *unusual* looking features in Task-histograms and interpret them accordingly in Client-histograms, eliminating the need for hand-coded quality thresholds.

The success of the project will be to move from a rules-based assessment toward a supervised (and later unsupervised) interpretation of the data quality.

The data of two CMS sub detectors, Electromagnetic and Hadronic Calorimeter were used for this work. State of the art of application development is discussed in the following.

IBM company is the industry partner of CMS experiment in this project and we believe this partnership represents an evident benefit for all the developers and the analysis results. We established a two-fold collaboration based on activities of knowledge transfer and problems solving and provision to the project of an IBM Minsky Power 8+ cluster, hardware infrastructure, described in details in Table 1. The first built the substrate for a fast development of the project, the latter is absolutely critical now we're entering the consolidation phase of the project.

ML IBM cluster infrastructure		
login node	compute nodes	software
IBM S821LC (8001-12C)	IBM S822LC for HPC (8335-GTB)	CentOS 7.4 ppc64le
8 P8 cores at 2.32 GHz	16 P8 cores at 3.26/3.86 GHz	CernVM-FS
64 GB of DDR4 memory	4x NVIDIA P100	NVIDIA CUDA 9.1
	3584 CUDA cores, 16 GB HBM2	
8 TB HDD	4 TB HDD + 1.6 TB NVMe	IBM PowerAI 1.5
	IB EDR between nodes	slurm
1 GbE (/ 10 GbE)	1 GbE to the CERN network	IBM Spectrum Conductor
	1 Xilinx ADM-PCIE-8K5 (CAPI)	Deep Learning Impact

Table 1: ML IBM cluster infrastructure

3 Project Use cases and state of the Art

3.1 ECAL, a mature development

Physics at the LHC requires extremely high performance detectors. The CMS Electromagnetic Calorimeter (ECAL)[3], has been designed to facilitate the discovery of the Higgs boson and new physics in general, through the precise energy measurement of particles such as electrons and photons. The CMS ECAL is an hermetic homogeneous calorimeter, cylindrical shaped, made of ~61K lead tungstate ($PbWO_4$) crystals mounted in the central barrel part, closed by ~7K crystals in each of the two endcaps.

3.1.1 ECAL: supervised learning

We approached the DQ assesment problem in a simple intelligent way treating it as a typical binary classification problem, where good data are clearly different from all the expected

anomalies. In a preliminary phase, some time was spent to identify the most meaningful quantities to target and to plan the necessary preparation of the supporting data. We took as input the so-called rechit occupancy map, i.e. of the central part (*barrel*) of the CMS detector, that is a crucial metric to monitor in order to guarantee the correct measurement of the energy of electrons and photons. This distribution describes the energy deposition of rechit data which is expected to be approximately constant in regions of constant η, of the order of 0.1 GeV during collisions. A large irregular distribution or an unusually large energy range indicates potentially noisy trigger regions, bad data reconstruction, or issues with the calibration.

A technical challenge has been to choose the data, apt for the exercise, and to format them to emulate a typical online streaming rhythm/conditions. Each of the inspected distributions are refreshed, in the online GUI application used for DQ, every 23 sec (commonly called *lumisection*, LS)[1]. We treated the single LS images as frozen frames of a streaming video (e.g. data taking) and approached the problem as an anomaly detection problem on images.

For the preliminary exercise we had access to a dataset of ~40K samples of data, collected by the CMS experiment in 2016/17, manually pre-scrutinized by a human certifier and declared of good quality. Because the anomalies encountered, in the considered data taking period, were so rare and limited in typology, we decided to artificially manufacture some ~8K samples of poor quality data, enriched of potential danger, as described in the following.

We analyzed them applying a supervised method, consisting of an Artificial Neural Network architecture with convolutional layers followed by dense layers, a rectifier is chosen as activation function (REctified Linear Units [8]), except for the last layer where a normalized exponential function (Softmax function)[9] is used. Further details can be found in Table 2.

We measure the classification performance quantifying the Area Under the Receiver Operating Characteristic Curve (also known as ROC AUC)[10], using recall (True Positive Rate) vs fall-out (True Negative Rate) as metrics; for a fall-out of 0.033 a recall of 0.9903 for training and validation set and 0.9944 for testing set respectively, is found. These excellent results convince us that the supervised application is ready to be implemented in the online ECAL monitoring tools already in the next months, allowing verification and validation of the specifications requested. The intended purpose to facilitate operations and to save human resources could be potentially be solved before the end of Run 2.

ECAL: supervised learning	
Activation	ReLU, but last layer that has softmax
Optimizer	ADAM
Performances	binary Cross Entropy loss
Net architecture	NN with convolutional layer followed by dense layers
Regularization	L2, dropout

Table 2: ECAL supervised approach: model details

3.1.2 ECAL: semi-supervised learning

While the supervised model seemed to generalize well as far as the classification has a pure binary solution, we desired to take our study further and to be able to detect anomalies, not foreseen earlier, in a way to understand better the ongoing behavior of the detector.
For this purpose a semi-supervised learning approach, based on Auto-Encoder(AE)[11], is what we tested on the data sample considered.

[1] A lumisection is in fact the minimum quanta of data time in LHC.

An AE Neural Network is a model for dimensionality reduction, composed by an encoding and a decoding section. The encoding section reduces the dimensionality of the data representation, while the decoding one expands it back to that of the input. The algorithm is trained to learn an approximation to the identity function, so as the reconstructed output is similar to the input. The identity function seems a particularly trivial function to be trying to learn; but, by placing constraints on the network (e.g. by limiting the number of hidden units), we can discover interesting structure about the data (e.g. some of the input features are correlated), and this is exactly one of the peculiarity interesting to us.

The model developed is an AE with convolutional layers, based on the Keras library framework (with TensorFlow[12] backend), using LeakyRelu[13] activation function in all the input-to-hidden and all hidden-to-hidden layers, and Linear activation in the hidden to output layer. We trained exclusively on normal instances, assuming imbalance of normal/anomalous instances; in this way the system learns to reconstruct the input minimizing the loss function (chosen here the *Minimum Squared Error*); for further details see Table 3.

ECAL: semi-supervised learning	
Preprocessing	data standardization, scale to 0 mean and standard deviation of 1
Activation	LeakyRelu activation for the input-to-hidden and hidden-to-hidden layers Linear activation for the hidden to output layer
Optimizer	AdaDelta
Loss function	MSE loss
Net Architecture	Auto-Encoder with convolutional layers, in framework Keras library (tensorflow backend)
Regularization	Dropout (10-15%)

Table 3: ECAL semi-supervised approach: model details

Two types of anomalies were analyzed to test the detection abilities of the proposed model: a hot tower[2] and a missing module[3], left columns (top and bottom raws respectively) of Figure 1. The artificially manufactured samples, previously described, were constructed by inserting, at random position, either a hot tower (red spot in the original image on the left) or a missing module (blue box in original image).

Looking at the image reconstructed by the AE, the architecture seems to perform well in case of hot tower events, the higher reconstruction loss can be associated with certitude to the spread around the hot area. The method does not work as effectively in event of missing module because the spread around the module is minimal with well defined edges.

Nevertheless the separation power of the ECAL semi-supervised method is evident in Figure 1 rightest column, where the Minimum Squared Error (MSE) loss is displayed as function of time (one entry per LS). The reconstruction error can be used to set a threshold between anomaly data and good ones; the separation power is absolutely indisputable in case of hot tower; less definitive, but still usable, in case of missing tower.

3.2 HCAL

Second use case of the CMS-IBM project is the Hadronic Calorimeter (HCAL)[3], chosen, as complementary to the ECAL study, for its capacity to identify and measure the energy of hadronic jets and to provide a good missing transverse energy determination. Because both of these quantities play a fundamental role in many analyzes of physics beyond the Standard Model, is crucial to have a trusty monitoring of the energy distribution maps. A dedicated

[2]Hot tower: large occupancy in a group of ∼5 x 5 crystal.
[3]Missing module: missing reading in a strip composed of 1700 crystals.

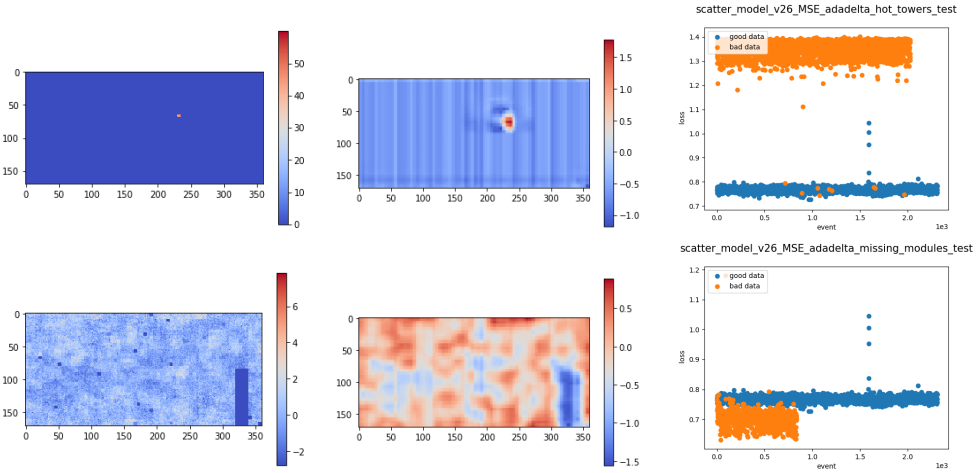

Figure 1: ECAL. From left to right: original image, reconstructed by AE image, MSE loss good data (blue) vs bad data (orange); top: Hot tower, bottom: Missing module

ML effort started only recently but it quickly reached the first faithful results following the lessons learnt from the ECAL study.

Also in this case the online running conditions were emulated, producing one set of images per LS, for all the 2017 data available, and artificially generating of 5x5 bad regions data (e.g. hot regions and dead regions principally), to accommodate the sparsity of anomalous real data.

At first a supervised method was adopted to verify the correctness of data pre-processing and ability over data duality classification.

Afterwards the step forward, towards the ability to predict anomalies, was taken via a semi-supervised approach, using a convolutional Auto-Encoder[11], the technical details of which can be found in Table 4. Fig. 2 reports the results of the experiments made on different samples of data: good data (top raw), hot channel (central raw) and dead channel (bottom raw). For each of these samples, columns, from left to right, present: the data type original image, the acquired image, as reconstructed by the AE, and the euclidean distance loss (related to its reconstruction error) respectively; the latter is precisely what we're interested in minimizing, because it represents how close our reconstruction is to the true input data.

Figure 3, the distribution of the max error made per reconstructed LS, confirms that good data (which reconstruction error is under 0.4) and anomalous data are very separable, even just applying a cut on the error value and we can classify them accordingly.

HCAL: semi-supervised learning	
Preprocessing	batch normalization
Activation	Relu activation for the input-to-hidden and hidden-to-hidden layers Sigmoid activation for the hidden to output layer
Optimizer	AdaDelta
Loss function	MSE loss
Net Architecture	Auto-Encoder with convolutional layers, in framework Keras library (tensorflow backend)
Regularization	maxPooling

Table 4: HCAL semi-supervised approach: model details

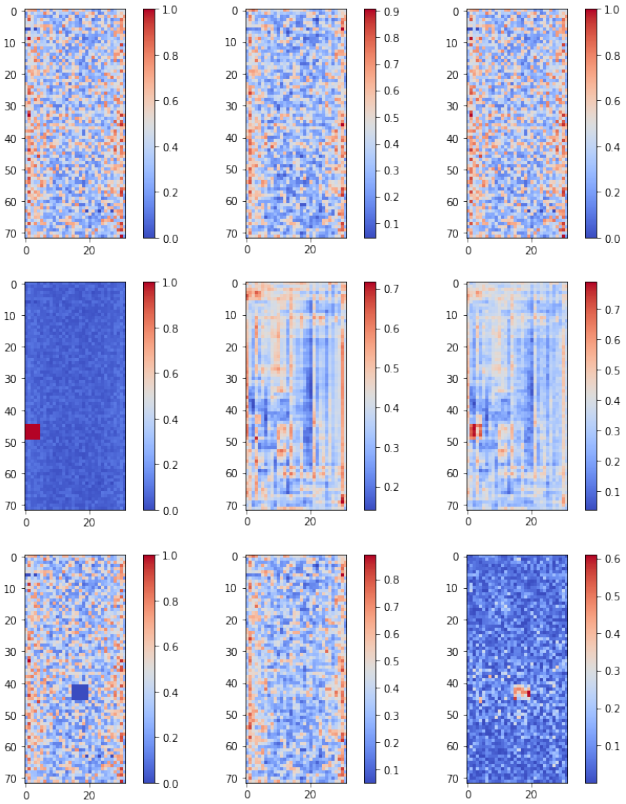

Figure 2: HCAL. From left to right: original image, reconstructed by AE image, reconstruction error; from top to bottom: Good data, Hot channel, Dead module

Figure 3: Separation power HCAL. Max error per reconstructed LS: Good data(green), Hot channel(red), Dead module(blue)

4 Summary and Future plans

The CMS experiment engaged in a partnership with the IBM company, with CERN Openlab as facilitator, to investigate the benefits of a future ML-based quality monitoring. Two types of models were developed:

- a Supervised model, for learning quality rules and statistics directly from raw data, and eliminate the need of the present hand-coded rules. The first validated prototypes proved, with a ROC AUC ~0.99, to be able to provide a robust binary classification of known incidents in the data taking. Next months will be spent in pinning down computing details, as data processing speed in the ML approach and the amount of computing resources used from the Minsky system and in cross referencing the results, comparing the human certification and the intelligent one, once the latter will be implemented in online monitoring tools.
- a convolutional Auto-Encoder method, tested on known cases, manifested a satisfying ability of differentiating good and bad event. CMS would rely on its semi-supervised learning technique, for flagging any unforeseen unusual looking features. The strength of the predictive method will be completed by the ongoing tests of more sophisticated models and extension to the whole detector.

The final benchmarking results, we believe, will be reached once time will enter in the matrix of features, via a time recurrent algorithm, and we will be able to include detector metadata to proactively foresee and alarm the collaboration about potential future hardware failures.

Acknowledgement

Authors would like to thank the CMS collaboration for believing in this project, for dedicating some woman/manpower to it and for allowing the use of the data analyzed. We acknowledge the support of the CERN openlab project in creating the best conditions of communications and in providing an exclusive technical support infrastructure. The participation to the CHEP 2018 conference has been possible thanks to the funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement n. 772369), authors are grateful for the endorsement.

References

[1] LHC web page: https://home.cern/topics/large-hadron-collider, accessed: 05/03/19.
[2] CERN web page: https://home.cern/, accessed: 05/03/19.
[3] CMS Collaboration, JINST **3** S08004 (2008).
[4] IBM Switzerland Ltd (IBM).
[5] A.Di Meglio at al., CERN openlab whitepaper, doi: 10.5281/zenodo.998694.
[6] F. De Guio, The CMS data quality monitoring software. JPCS, **513** 032024 (2014).
[7] S. Hochreiter, J. Schmidhuber, LSTM memory, Neural Comput., **9** no. 8, 1735 (1997).
[8] V. Nair, G. Hinton, Rectified Linear Units Improve Restricted Boltzmann Machines, *ICML '10*, 807, (2010).
[9] C. Bishop, *Pattern Recognition and Machine Learning*, (Springer 2006).
[10] C. E. Metz, Sem. Nucl. Med. **8**, 283 (1978).
[11] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, (MIT Press 2016).
[12] M. Abadi et al. http://tensorflow.org/ (2015).
[13] Maas et al. Rectifier nonlinearities improve neural network acoustic models. *ICML*, **30** (2013).