Anomalies detection with autoencoders

Piotr W. Nowak for the ALICE Collaboration,
Warsaw University of Technology, Faculty of Physics
piotr.wladyslaw.nowak@cern.ch

Abstract

Data Quality Assurance plays a crucial role in all high-energy physics experiments. However, methods currently employed in the ALICE experiment at the Large Hadron Collider rely on traditional data analysis techniques based on statistical analysis of the data. The machine and deep learning techniques could handle higher-dimensional and more complex problems than the traditional ones. In this work we show a new method for data quality assessment in the ALICE experiment which leverages from the autoencoder processing. We present and compare several architectures of deep autoencoders and variational autoencoders. Considering a specific physics channel, using limited statistics of real data from the ALICE experiment, we show that our approach outperforms current methods. In the considered analysis we gain better separation of anomalies saving up to 35% of data examples set as anomaly and get a continuous score instead of binary cut.

1 Introduction

Data quality assurance at the CERN Large Hadron Collider [1] is preventing the storage and usage of data containing anomalies for analysis. ALICE [2], one of four big experiments at CERN, is focused on the study of heavy-ion collisions. In our work we focus on the ALICE offline quality assurance [2] of the Time Projection Chamber (TPC) [2]. TPC is the main tracking detector of ALICE and its performance is crucial for the vast majority of the physics analyses. ALICE TPC is the largest time projection chamber ever built and it is very sensitive to the properties of the gas - proper quality control is of great importance in the case of this detector.

For the data quality assurance the current approach is an automatic system that monitors around 40 parameters defined by detector experts. Some parameters are related to low-level features of the reconstructed tracks. Quality assurance in ALICE is performed run-by-run, where a run covers the period of data taking corresponding more or less to a single fill of the LHC (around 12 hours). After a couple weeks of collecting of the data and gathering values for many runs, the system calculates the mean and standard deviation for each parameter and applies outlier label for values deviating by more than 3 σ from the mean. Finally, all flags are combined into a single flag that classifies a given run to be right for further analyses.

We propose to use autoencoders [3] for this anomaly detection problem. Because of the generalisation enforced by the bottleneck of a standard autoencoder, we can treat observations with the highest reconstruction error as anomalies. As observed in [4], abnormal examples are usually underrepresented in the latent space of properly trained autoencoder and this should result in high reconstruction error. We compare several architectures of deep autoencoders and variational autoencoders. To ensure extreme generalisation we evaluate models with up to two neurons in the latent space.

We analyse the results of the proposed technique considering a specific physics analysis and using real data from the ALICE experiment. Our tests indicate, that our solution does not only discovers the same data anomalies as the standard quality assessment technique, but it also discards only the most abnormal examples – true data anomalies. Our physics studies performed on the selected subset

Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019), Vancouver, Canada.

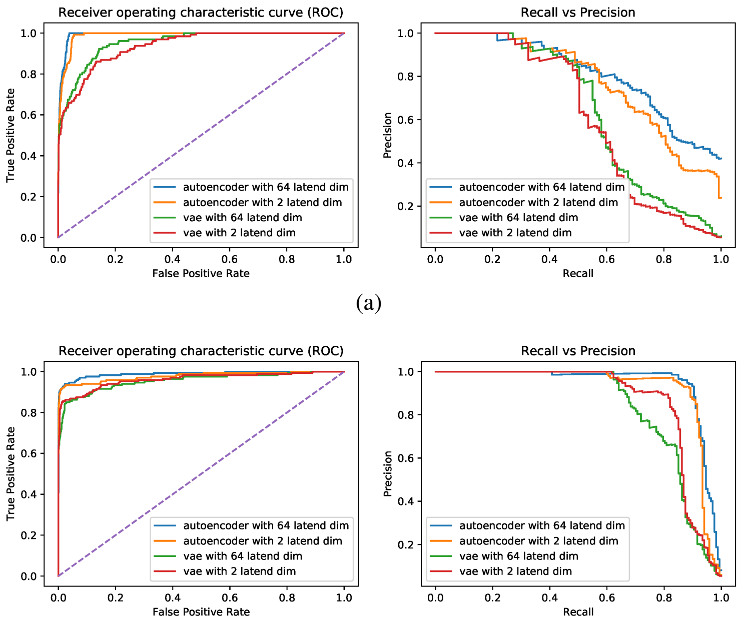
of data revealed that, while using autoencoders, we can save, in the considered analysis, up to 35% of data examples set as outliers while retaining the same purity of the sample.

Related Works
Researchers from CMS proposed similar solution [5] [6] for CMS offline quality assurance. Studies reviled that autoencoder architecture performed the best in terms of anomalies discovery reaching AUC = 0.905. This approach was valuable for quality assurance experts with easy to interpret result.

2 Proposed solution

For the purpose of studies presented in this work, the set of tracked parameters was extended to over 200 quantities related to working conditions inside the detector. Second major difference is the usage of data taking periods of much finer granularity, corresponding to around 10-15 min of data acquisition. We gathered the data from five periods, two from lead-lead collisions (LHC18q,LHC18r), and three from proton-proton collisions (LHC18f,LHC18o,LHC18p). Because of the vast difference between collecting data from lead-lead and proton-proton collisions we separated these into two datasets. We then performed a standard data assurance procedure by assigning to each data chunk an automatic quality label. In total this processing resulted in 8116 data samples out of which 296 were identified as outliers.
To find unseen relations and correlations between parameters we use the deep learning model. After reducing the number of input parameters from 200 to 97, by using only parameters that are physical attributes, we built an autoencoder model with 2 hidden layers encoder with 512 and 128 neurons, a latent space with 64 neurons and a symmetric decoder with 2 hidden layers with 128 and 512 neurons. We transform features by scaling each to the range [0,1]. We use Leaky ReLu activation and Sigmoid activation function[3] for the last layer. For our loss function, we take a mean squared error loss. The neural network was implemented in Python using Keras library with TensorFlow backend [7]. To ensure that our deep model will not develop a way to encode anomaly data into latent space we use the automatic quality label to use data of outliers only for testing. We have built a variational autoencoder using similar architecture. For both models we check their results obtained with a latent space changed to only 2 neurons to ensure extreme generalisation.

3 Results



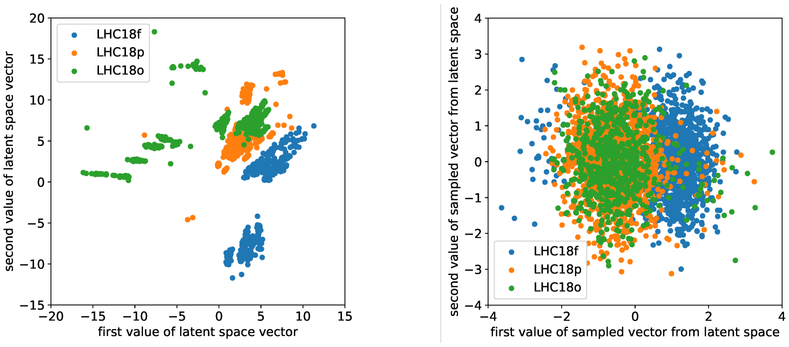
(a)

(b)

Figure 1: ROC curve and precision-recall plots for semi-supervised anomalies detections with autoencoder and variational autoencoder for **(a)** lead-lead collisions **(b)** proton-proton collisions.

To analyse our unsupervised learning methods we confronted the results with standard quality assurance procedures. As is visible in the ROC plot (Fig. 1), with fine tuned training, we are able to recreate almost the same classification as with the standard procedure. On the other hand, we can tune an autoencoder to provide softer cuts on the data and therefore increase the efficiency of data selection. ROC plot (Fig. 1) shows that 2 neurons in the latent space is enough to find outliers with surprisingly excellent accuracy.

Visualization of a two-dimensional latent space (Fig. 2) for proton-proton collisions shows that there are general differences between examples within different data-taking periods. Additionally, visualization reveals that autoencoder groups examples by periods, while variational autoencoder sampling of these is also done in groups but with much smoother transitions between them.

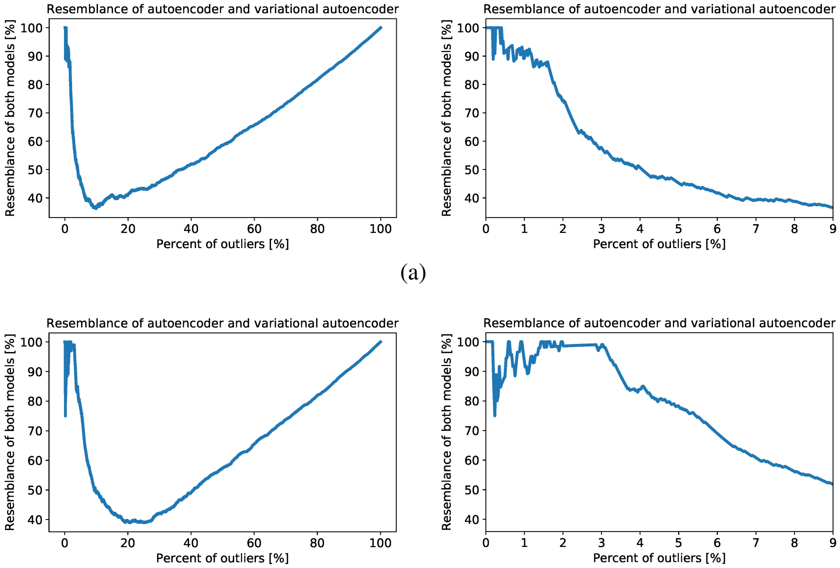


(a)

(b)

Figure 2: Visualization of 2-D latent space in **(a)** autoencoder and **(b)** variational autoencoder sampling for proton-proton collisions with marked different data-taking periods.

Comparison of our models is done with standard methods set as ground truth. We know that the automatic system is not ideal and that it can classify some samples wrongly. Comparing autoencoder and variational autoencoder models (Fig. 3) show that both models find same most abnormal samples.



(a)

(b)

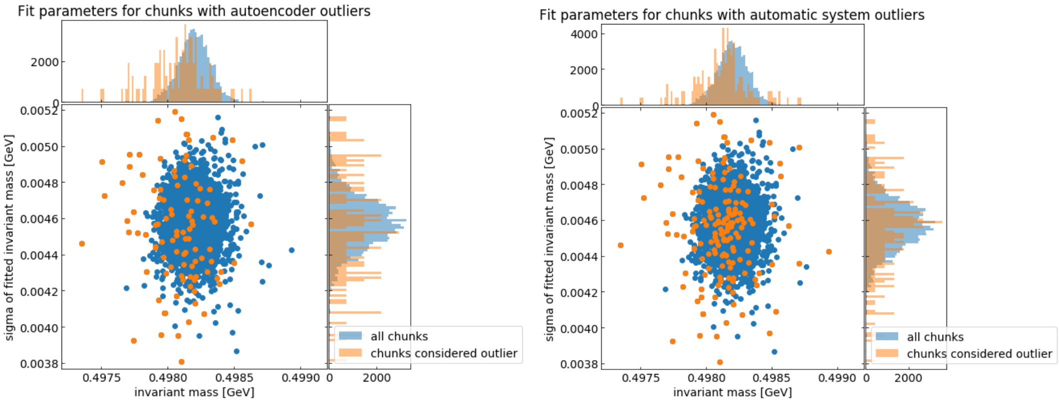
Figure 3: Resemblance of autoencoder and variational autoencoder models for **(a)** lead-lead collisions **(b)** proton-proton collisions.

Choosing the number of outliers that both models found with 90% resemblance provides us with a threshold that we compare with the number of outliers found with the standard method (Tab. 1).

Table 1: Comparision of outliers found with standard methods and autoencoders

Periods	Standard methods	Autoencoders
Lead-lead collisions	2.7%	1.9%
Proton-proton collisions	4.9%	3.1%

To validate our observations we performed additional studies, by running a simple physics analysis on the whole collected data samples. For each data chunk we fit a peak of the invariant mass of K_S^0. As presented in (Fig. 4), our approach with autoencoders tends to tag as outliers mostly data chunks for which the invariant mass lies on the far edges of the global distribution. On the contrary, the standard anomalies detection method selects also data samples with central mass selection. This might suggest that standard methods tag too many data as anomalies.



(a)

(b)

Figure 4: Mean and σ of K_S^0 invariant mass distributions for outlier selection with autoencoder **(a)**, and standard procedure **(b)**.

4 Summary

Using unsupervised machine learning methods in the ALICE offline quality assurance of the Time Projection Chamber can not only recreate standard methods results but also improve separation of anomalies. Autoencoders and variational autoencoders are not only good algorithms to find outliers but are also great tools to visualize and analyze results. Even using only a 2 dimensional latent space our models show excellent accuracy in finding anomalies. Comparing our models we found only most abnormal examples – true data anomalies. With this information we save 35% examples tagged with standard methods as outliers in the considered statistics. We also confirm our conclusions by testing our methods with the analysis of the invariant mass distributions of K_S^0.

5 Acknowledgements

The authors acknowledge the support from the Polish National Science Centre grants no. UMO-2016/21/D/ST6/01946, UMO-2016/22/M/ST2/00176, UMO-2018/31/N/ST6/02374 and the Ministry of Science and Higher Education grant no. DIR/WK/2016/2018/17-1.

References

[1] Lyndon Evans and Philip Bryant. “LHC Machine”. In: *JINST* 3 (2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[2] K. Aamodt et al. “The ALICE experiment at the CERN LHC”. In: *JINST* 3 (2008), S08002. DOI: `10.1088/1748-0221/3/08/S08002`.

[3] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[4] Mayu Sakurada and Takehisa Yairi. “Anomaly detection using autoencoders with nonlinear dimensionality reduction”. In: *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. ACM. 2014, p. 4.

[5] Adrian Alan Pol. *Anomaly detection using Deep Autoencoders for the assessment of the quality of the data acquired by the CMS experiment*. Tech. rep. 2018.

[6] Adrian Alan Pol et al. “Detector monitoring with artificial neural networks at the CMS experiment at the CERN Large Hadron Collider”. In: *Computing and Software for Big Science* 3.1 (2019), p. 3.

[7] François Chollet et al. *Keras*. https://github.com/fchollet/keras. 2015.