Machine Learning in Calorimeter optimization

A. Boldyrev\textsuperscript{a}, D. Derkach\textsuperscript{a}, F. Ratnikov\textsuperscript{a,\textit{b}}, A. Shevelev\textsuperscript{a,1}

\textsuperscript{a}National Research University Higher School of Economics, Laboratory of Methods for Big Data Analysis, 11 Pokrovsky blvd., Moscow 109028, Russia
\textsuperscript{b}The Yandex School of Data Analysis, 11/2 Timura Frunze St., Moscow 119021, Russia

E-mail: andrew.shevelev@gmail.com

Abstract. The optimization of big industrial setups and the accompanying detailed simulations of internal physical processes require complex and time-consuming simulation calculations. We propose a versatile approach that can alleviate difficulties in solving this problem and show this using an example of electromagnetic calorimeter optimization at a Large Hadron Collider experiment. Our approach consists of a block representation of the calorimeter optimization process from setting sensitive characteristics of modules and their locations to obtaining a quality metric and applying machine learning methods. The main blocks are signal and background particles generation and their propagation to the calorimeter, the generation of electromagnetic showers of signal and noise in modules with a given technology, the combination of signal and noise with the simulation of different luminosities, the energy and spatial reconstruction of the signal and obtaining the final metric. This approach allows us to evaluate the operational characteristics of the calorimeter and find a more suitable configuration with the necessary quality without extensive use of time-consuming resources.

1. Introduction

One of the unsolved problems in physics is the baryon asymmetry also known as the matter asymmetry problem or observed imbalance in the baryonic and antibaryonic matter in the observable Universe. Neither the Standard Model of particle physics nor general cosmological theories provide an explanation of why this should be so. The Big Bang should have produced equal amounts of matter and antimatter based on the known accepted physical laws. However, why is there far more matter than antimatter in the observable Universe? The answer could lie in the domain of the $CP$ violation phenomena \cite{1}.

The dedicated tool to answer that and other questions is the LHCb detector that is one of the four main detectors at the Large Hadron Collider (LHC) at CERN (European Organization for Nuclear Research). The LHC is the world’s largest and highest-energy particle collider and the largest machine in the world \cite{2}.

The LHCb detector has a complex structure consisting of several subsystems. It is approximately 20 meters in length and 8 meters high. It consists of the following main parts: vertex locator, Cherenkov detectors, tracking system, magnet, electromagnetic calorimeter, hadron calorimeter, and muon system (see Figure 1).

The electromagnetic calorimeter (ECAL) is responsible for measuring the energy and the position of the incident particles. Its current configuration consists of three thousand and five

\textsuperscript{1} Corresponding author.
hundred modules. Each module is a measuring device with a distinctive structure and properties. The calorimeter data resembles a picture from an optical camera, but the principle of operation is different. The size of the ECAL is approximately 8 by 6 meters.

The current ECAL configuration has a rectangular shape and consists of three types of equal size modules. The first type of modules includes nine cells of size 4.04x4.04 cm$^2$ each. The second one comprises four cells of size 6.06x6.06 cm$^2$ each, and the third one consists of one cell of size 12.12x12.12 cm$^2$. The scheme of the current configuration and its modules is shown in Figure 2.

The collider is currently shut down for upgrade. The purpose of this break is to increase the luminosity of the collider to get a higher yield of rare events. The new conditions will be more difficult to operate with more background noise that clutter the signal detection. Thus, a modified reconstruction algorithm is required to handle a signal for each ECAL configuration. This article addresses the problem of fast generation of signal responses and a reconstruction algorithm to obtain a metric for each calorimeter configuration. In future, this result is foreseen to be used for a black-box optimization of the calorimeter configuration. In other words, we

---

2 LHCb - Large Hadron Collider beauty experiment cite: https://lhcb-public.web.cern.ch
answer the question of how to get the best measurement performance of the detector, given the resources in hand. The main goal of the article is to demonstrate that the simulation and reconstruction based on using machine learning techniques can provide a stable estimation for each configuration performance, the later can be used for the optimisation of calorimeters.

2. Simulation and Machine Learning Method
To optimize ECAL configuration, one needs a function taking module configuration as input and an algorithm to measure the quality of the configuration as output. To obtain physical performance of ECAL subsystem, we have to complete a dozen of steps: from proton bunch interaction and parton formation to particle passage through matter simulation and getting signal from the calorimeter.

We estimate the quality of the particular calorimeter configuration by evaluating the width of the reconstructed invariant mass peak of the neutral pion.

The quality of the calorimeter configuration depends on the module type and signal reconstruction algorithms. This means that we not only have to carry out the simulation but also to tune reconstruction algorithms for each configuration. The main problems are:

- Detailed simulation is CPU consuming (about two minutes per one event [3], while an estimation of metric value for specific configuration requires about one million events).
- Current reconstruction algorithm has to be manually fine-tuned for each ECAL configuration that requires human manual efforts and resources.

To solve these problems, first of all, we modify standard simulation approach and secondly we use machine learning technique to reconstruct energy and spatial position of particles.

Simulation of physical events is performed using **Pythia** 8 [4], **GEANT4** [5], **Gaudi** [6] is used to propagate the particles through matter. For the signal sample we have 1.2 million photons from \( B_s^0 \rightarrow J/\psi(\rightarrow \mu^+\mu^-)\pi^0(\rightarrow \gamma\gamma) \) decays. For the background we take 30 million particles from the LHCb MC Minimum Bias sample for upgraded geometry at 14 TeV including \( \gamma, \pi^+, \pi^-, e^+, e^-, n, \) and \( p \) particles. Next, we propagate each particles to the front surface of the calorimeter by Gaudi.

We produce electromagnetic clusters for each incident particle listed above in 2.02x2.02 cm\(^2\) cells in order to obtain bigger cell clusters by convolution-like operation to avoid spending time to resimulate full processes. In this way, we obtain the response for cells of all needed size. In addition, we produce different levels of noise by mixing signals with more background particles for different background density (a number of Primary Vertices = nPV). Examples of generated energy deposits with cell size 2.02, 4.04, 6.06, and 12.12 cm\(^2\) by vertical scale and 0, 10, 20, and 30 nPV by horizontal scale can be seen in Figure 3.

A distinguished task in the pipeline is to obtain an algorithm for each configuration that solves an inverse problem: based on the calorimeter image, reconstruct the energy and position of the incident particle. This part is of crucial importance, since it directly affects the results of the metric evaluation for each configuration [7, 8]. To solve the problem of energy and spatial reconstruction algorithm tuning for each ECAL modules configuration we use machine learning methods based on extreme gradient boosting on decision trees [9, 10]. These methods have shown stability and reliably in automatic mode. For feature input parameters, we consider 24 cells of energy deposits around the most energetic cell (the brightest spot in Figure 3) in the cluster and cell itself. Cluster barycenter, reference time information, and various sums and squared sums of energy deposits are also used. We develop distinct algorithms for spatial and energy reconstruction. That is there are three models: for x coordinate, y coordinate, and energy. To train the models, we split the data into train and test samples as 50/50 percent. As a preventative measure against overfitting, we use the k-fold cross-validation method.
As a final metric of the calorimeter quality, we consider the reconstructed invariant mass peak width of the neutral pion decayed on two photons. The invariant mass formula includes reconstructed particle energy $E$, and reconstructed momentum $p$ each of two photons: $M^2 = (E_1 + E_2)^2 - ||p_1 + p_2||$.

3. Results

In the presented approach, the cluster generation stage (the results of which can be reused for the given technologies) takes 0.9 second per event, and 0.2 seconds per event is taken by both the reconstruction stage and obtaining the metric. This result is almost two orders of magnitude faster than the standard procedure. Moreover, with this approach there is no need to manually optimize the reconstruction algorithm for every separate configuration, that saves a lot of expert person-hours.

Figure 4 shows the results of the measurement of $\pi^0$ mass peak width with the application of machine learning methods for both energy and spatial reconstruction for different ECAL modules configurations and different luminosity conditions. We conclude that the machine learning reconstruction (marked as blue and red dots) works stably in autonomous mode in those very different conditions. The asterisk marks the current configuration of the calorimeter used in LHCb before the current LHC long shutdown [11]. The green crosses are an approximation of the current reconstruction and ECAL configuration in different luminosity conditions.

The performance metric does vary for different configurations, thus we conclude that the optimal layout of ECAL modules can be based based on using the presented approach.
Figure 4. $\pi^0$ peak width of different ECAL configurations with different luminosity conditions.

4. Conclusions
The main goal of this work is to develop a pipeline with fast energy deposit cluster simulation in different luminosity conditions, and for different module types. This pipeline is based on the stable ML-based automatically trained reconstruction model that is tuned for every given ECAL modules configurations.

We developed a tool for a faster estimation of the calorimeter performance. We demonstrate that machine learning methods work successfully for spatial and energy reconstruction methods, thus may be used to evaluate physics properties of the calorimeter.

Modular pipeline structure allows us to easily change its constituent blocks, such as simulations or metrics block, that can be conveniently used in further applications.

The pipeline can be also used for the automatic tuning of the calorimeter setup in given run conditions.

Acknowledgments
The research leading to these results has received funding from Russian Science Foundation under grant agreement n° 17-72-20127.

This research was supported in part through computational resources of HPC facilities at NRU HSE.

References
[1] A.D. Sakharov (1967). "Violation of CP invariance, C asymmetry, and baryon asymmetry of the universe”. Journal of Experimental and Theoretical Physics Letters. 5: 24–27. and in Russian, A. D. Sakharov (1967). "Violation of CP invariance, C asymmetry, and baryon asymmetry of the universe”. ZhETF Pis’ma. 5: 32–35. republished as A. D. Sakharov (1991). "Violation of CP invariance, C asymmetry, and baryon asymmetry of the universe”. Soviet Physics Uspekhi (in Russian and English). 34 (5): 392–393. Bibcode:1991SvPhU..34..392S. doi:10.1070/PU1991v034n05ABEH002497.
[2] Joel Achenbach (March 2008). "The God Particle”. National Geographic Magazine. Retrieved 25 February 2008
[3] LHCb Collaboration Computing Model of the Upgrade LHCb experiment. Technical report LHCb-TDR-018, CERN, Geneva
[4] T. Sjostrand, S. Mrenna, P. Skands A Brief Introduction to PYTHIA 8.1 Comput.Phys.Commun. 178 (2008) 852-867 e-Print: 0710.3820 doi: 10.1016/j.cpc.2008.01.036

[5] S. Agostinelli et al. GEANT4: A Simulation toolkit. Nucl. Instrum. Meth., A506: 250–303, 2003

[6] G. Barrand, I. Belyaev, P. Binko, M. Cattaneo, R. Chytracek, G. Corti, M. Frank, G. Gracia, J. Harvey, E. van Herwijnen, P. Maley, P. Mato, S. Probst and F. Ranjard, GAUDI - A software architecture and framework for building HEP data processing applications, Comput. Phys. Commun. 140 (2001), 45-55 doi:10.1016/S0010-4655(01)00254-5

[7] F. Ratnikov, D. Derkach, A. Boldyrev, A. Shevelev, P. Fakanov, L. Matyushin Using machine learning to speed up new and upgrade detector studies: a calorimeter case https://arxiv.org/abs/2003.05118

[8] F. Ratnikov Using machine learning to speed up and improve calorimeter R&D. Journal of Instrumentation Volume 15, May 2020 JINST 15 C05032 doi:10.1088/1748-0221/15/05/C05032

[9] A. Boldyrev, D. Derkach, F. Ratnikov, A. Shevelev ML-assisted versatile approach to Calorimeter R&D Journal of Instrumentation, Volume 15, September 2020 JINST 15 C09030 doi:10.1088/1748-0221/15/09/C09030

[10] T. Chen, C. Guestrin XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754 [cs.LG], March 2016

[11] O. Deschamps, F.P. Machefert, M.H. Schune, G. Pakhlova, I. Belyaev Photon and neutral pion reconstruction. CERN. LHCB-2003-091. Sep 2003.