Available on CMS information server

CMS CR -2023/255

The Compact Muon Solenoid Experiment

Conference Report

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

03 November 2023 (v3, 04 December 2023)

Testing a Neural Network for Anomaly Detection in the CMS Level-1 Global Trigger test crate during Run 3

Noah Alexander Zipper for the CMS Collaboration

Abstract

We present the deployment and testing of an autoencoder trained for unbiased detection of new physics signatures in the CMS Global Trigger test crate during LHC Run 3. The GT test crate is a copy of the main GT system, receiving the same input data, but whose output is not used to trigger the readout of CMS, providing a platform for thorough testing of new trigger algorithms on live data, but without interrupting data taking. We describe the integration of the DNN into the GT test crate, and the monitoring, testing, and validation of the algorithm during proton collisions.

Presented at *TWEPP2023 Topical Workshop on Electronics for Particle Physics*

Prepared for submission to JINST

Testing a Neural Network for Anomaly Detection in the CMS Global Trigger Test Crate during Run 3

Noah Zipper[a,1] **on behalf of CMS collaboration**

[a] *University of Colorado Boulder,*
3100 Marine St, Boulder CO, USA

E-mail: noah.zipper@colorado.edu

Abstract: We present the deployment and testing of an autoencoder trained for unbiased detection of new physics signatures in the CMS Level-1 Global Trigger (GT) test crate during LHC Run 3. The GT test crate is a copy of the main GT system, receiving the same input data, but whose output is not used to trigger the readout of CMS, providing a platform for thorough testing of new trigger algorithms on live data, but without interrupting data taking. We describe the integration of the Neural Network into the GT test crate, and the monitoring, testing, and validation of the algorithm during proton collisions.

Keywords: Accelerator Subsystems and Technologies, Trigger algorithms, Trigger concepts and systems (hardware and software)

[1]Corresponding author.

Contents

1 Introduction

The CMS detector [1, 2] reads out far more data than can be processed, reconstructed, and analyzed. In order to use any of the terabytes per second being generated, a reduction of more than 99% is necessary. The job of the CMS Level-1 trigger (L1T), which it does in real-time on a chain of field programmable gate array (FPGA) hardware modules [3], is to perform this vital data reduction without missing the interesting physics needed for analyses. Operating on the clock of the LHC, where collisions occur every 25 nanoseconds, requires the entire system to adhere to microsecond latency constraints. Furthermore, stability is crucial for this system. Any error can lead to detector "dead time", where data is lost forever.

A potential problem of traditional trigger strategies is that they rely either on on a priori knowledge of signal or generic kinematic selections. This problem is addressed by triggering on how anomalous an event is. The use of a variational autoencoder (VAE) trained on real unbiased data to detect outliers in CMS data offers a solution that is both signal agnostic, as it is applicable to signatures that we have not had the foresight to target specifically, and highly sensitive, shown to effectively boost signal efficiency for multiple physics signatures [4, 5].

2 Anomaly Detection Trigger Algorithm

The VAE design uses an information bottleneck created by small-dimensional latent space, which enforces an efficient data encoding, and leads the model to learn what makes an event anomalous. For this Anomaly Detection implementation in the CMS Level-1 Global Trigger, called Anomaly eXtraction Online Level-1 Trigger aLgorithm (AXOL1TL), the inputs are taken from a set of standard L1T objects ($p_{\mathrm{T}}^{\mathrm{miss}}$, 4 e/$\gamma$, 4 μ, and 10 jets) as ($p_{\mathrm{T}}, \eta, \phi$) vectors. The missing transverse momentum vector $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ is computed as the negative vector sum of the transverse momenta of all the Particle Flow candidates in an event. Its magnitude is denoted as $p_{\mathrm{T}}^{\mathrm{miss}}$ [6] and it definitionally has an η value of 0. The design of the VAE, visualized in figure 1, was driven by the constraints of the L1T system and hardware. Multiple steps were taken to minimize latency and resource utilization,

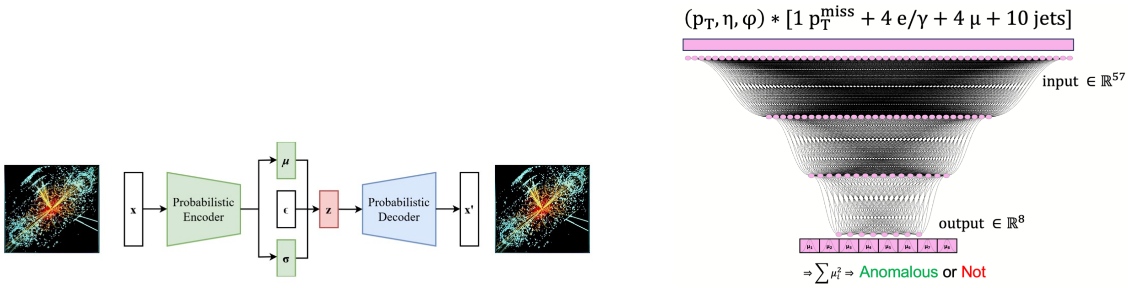

Figure 1. (Left) A typical design of a VAE, utilizing both an encoding and decoding network to reconstruct event information. (Right) The VAE model visualization for AXOL1TL, showing the layers, inputs, outputs, and the calculation of the anomaly score, our metric for triggering on interesting physics.

including the removal of the decoder network, and the simplification of the latent space loss term, shown in (2.1). The reconstruction term is computed from the difference between the input (x) and output ($\hat{x}$) of the VAE. The second, full regularization term, is the Kullback–Leibler divergence (KL-divergence) between the latent space distribution and a standard normal distribution with mean μ and standard deviation σ. The parameter β can be tuned to balance the reconstruction performance with more efficient latent space encoding. At inference time, the loss is approximated by the mean-squared term $\Sigma_i \mu^2$ of the KL-divergence for latency considerations. This approximation has no impact on performance.

$$\text{Loss} = (1-\beta)||x-\hat{x}||^2 + \beta\frac{1}{2}(\mu^2+\sigma^2-1-\log\sigma^2) \tag{2.1}$$

AXOL1TL is trained with unbiased data collected by the CMS Experiment during 2023 with a proton collisions at a centre of mass energy $\sqrt{s}$ = 13.6 TeV. From this dataset 10.5 million events were used: 50% for training and 50% for setting thresholds on the anomaly score. Quantization aware training, where inference-time quantization is emulated during training using the package QKeras [7], allowed for optimal performance in the final hardware implementation. Running anomaly inference on a set of events, and estimating the L1T rate for a given anomaly score threshold, a set of thresholds to demonstrate the range of performance possible with AXOL1TL were chosen, plotted in figure 2.

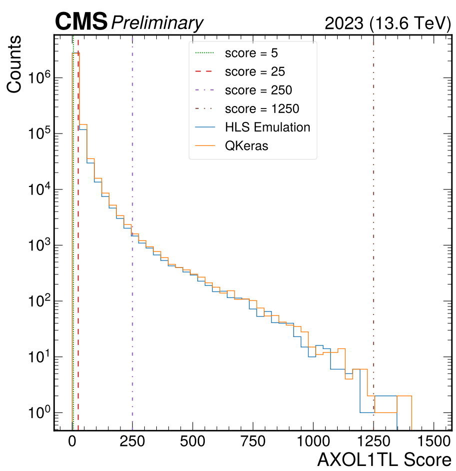

Figure 2. Anomaly score distributions for 2023 Ephemeral Zero-Bias events. Individual event scores/losses for the QKeras model in Python (orange) and standalone High-Level Synthesis (HLS) emulator (blue). Dotted lines represent scores that correspond to trigger paths in the μGT test crate.

For the 2023-trained model, an example of the significant performance improvement for Beyond the Standard Model (BSM) signals, measured in MC, by adding AXOL1TL to the 2023 trigger menu is shown in table 1.

AXOL1TL Rate	1 kHz	5 kHz	10 kHz
$H \to aa \to 4b$ Signal Efficiency Gain	46%	100%	133%

Table 1. Efficiency improvement of AXOL1TL trigger bits to 2023 L1 Menu with respect to the BSM signal of a Higgs decaying to two (pseudo)scalars of mass 15 GeV, where (pseudo)scalars decay to bottom quark pairs. The model used is trained on Run 3 ZeroBias events (Run 367883). Efficiency gains from AXOL1TL at various triggering rates are compared.

3 CMS Global Trigger Firmware

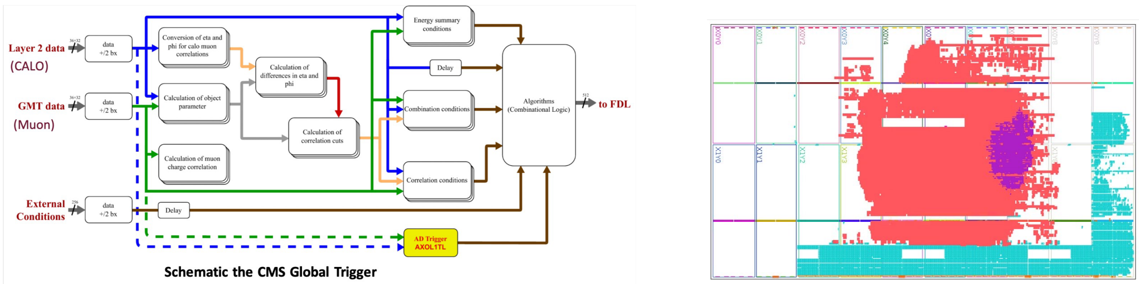

Figure 3. (Left) Schematic of the μGT board, showing AXOL1TL score calculation in yellow. (Right) Floorplan of the Xilinx Virtex-7 FPGA chip [8] on the μGT board. MP7 firmware payload (coral), MP7 infrastructure (cyan), and the AXOL1TL network (purple) resource utilization are shown.

The firmware for the Anomaly Detection algorithms had to be integrated into the existing CMS Global Trigger (μGT) board structure [9], shown in in figure 3. To meet timing, the score is calculated in concert with other global trigger quantities, fed into the trigger combination logic, and its corresponding trigger decisions are read out via the standard links.

High-level synthesis (HLS) was used to synthesize the hardware code (VHDL) for the MP7 μGT FPGA, utilizing the hls4ml package to efficiently synthesize the neural network [10]. To incorporate the trigger information into the common CMS software tools, an external CERN Gitlab repository for HLS dependencies was made. This code generates the bitfiles that configure the MP7 boards.

Vivado co-simulation tests, shown in table 2, demonstrate the the synthesized VHDL code meets the L1T latency requirement of 50 ns and takes up only a small fraction of the resources available on the Xilinx Virtex-7 chip [8].

Further validation performed to ensure trigger decision agreement for given thresholds. A ModelSim emulator, the standard CMS tool for L1T menu [11] validation was used. Test vector files with the Level-1 objects, detector conditions, and a separately emulated reference decision were read into the environment, and table 3 shows the perfect trigger bit agreement observed.

	Latency	LUTs	FFs	DSPs	BRAMs
AXOL1TL	2 ticks (50 ns)	2.1%	~ 0%	0%	0%

Table 2. Vivado timing and resource utilization report for Anomaly Detection trigger on Xilinx Virtex-7 FPGA. Results show that firmware build meets L1 latency requirements, fitting within 2 clock cycles @ 40 MHz. The resource utilization of the module is also relatively small. Columns refer to different FPGA components: look-up tables (LUTs), flip-flops (FFs), digital signal processing (DSPs) slices, and block random access memory (BRAMs)

L1 Menu Index	L1 Menu Algorithm Name	Test Vector Count	Hardware Emulation Count	Agreement
94	L1_ADT_20000	0	0	✓
95	L1_ADT_4000	29	29	✓
103	L1_ADT_400	2618	2618	✓
108	L1_ADT_80	3331	3331	✓

Table 3. Test Crate firmware validation. The table shows trigger bits for the L1 menu including 4 Anomaly Detection thresholds: scores >1250, >250, >25, and >5 from top to bottom (factor of 16 difference between physical anomaly score and hardware integer). Test vector column is generated from standalone emulator inference results and hardware count comes from standard global trigger firmware simulation workflow using ModelSim, using the same events from Run 368566. Perfect bit agreement is observed.

4 Test Crate Implementation

Once validated, the μGT firmware was implemented on the CMS Global Trigger Test Crate; The TC is a set of identical MP7 boards that are used as backup for the production system, as well as for testing new trigger strategies. In this configuration, the TC is set to read in the same inputs, but without actually triggering on events and saving data.

The TC is connected to the CMS Prometheus monitoring server that can query trigger metrics in real-time. This gives us the ability to monitor real time trigger rates for the anomaly trigger paths while data is being taken, shown in figure 4.

For particular runs, events that were triggered and saved by the production system contain TC information, showing which trigger bits were fired. This allows for a final validation of the anomaly score performance with respect to emulation. Figure 5 and table 4 show that for such events we see minimal mismatches and reasonable agreement between hardware and emulation.

5 Summary

In this article, we have shown a signal-agnostic trigger model sensitive to interesting physics. A firmware implementation for this trigger algorithm was successfully integrated into the CMS Level-1 trigger architecture. Using the CMS Global Trigger test crate, we showed an active hardware trigger that performed consistently during 2023 collisions. Finally, validation was performed for all steps using HLS emulation. We plan to update this algorithm and prepare downstream trigger logic to implement the Anomaly eXtraction Online Level-1 Trigger aLgorithm (AXOL1TL) in the Level-1 trigger for 2024 data-taking.

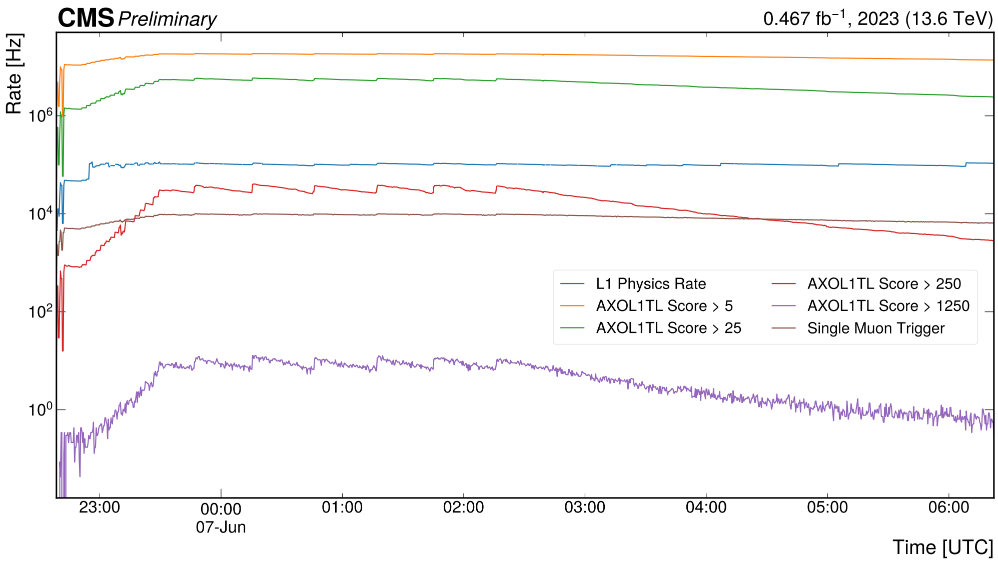

Figure 4. Test crate rate monitoring time series. L1 trigger rates shown for 4 Anomaly Detection threshold triggers, overall L1 physics rate, and the L1_SingleMu22 un-prescaled single muon reference trigger. Time-averaged rates are read from μGT test crate monitoring software via Prometheus server at a ~20s buffer rate during good data-taking conditions in 2023. AXOL1TL model is trained on 2018 ZeroBias data and thresholds are chosen to test possible range of accessible trigger rate. Thresholds are not meant to model realistic trigger rates. Consistent performance is shown over the course of partial fill-cycle. The turn-on corresponds to the beginning of an LHC fill and the sawtooth pattern corresponds to luminosity levelling.

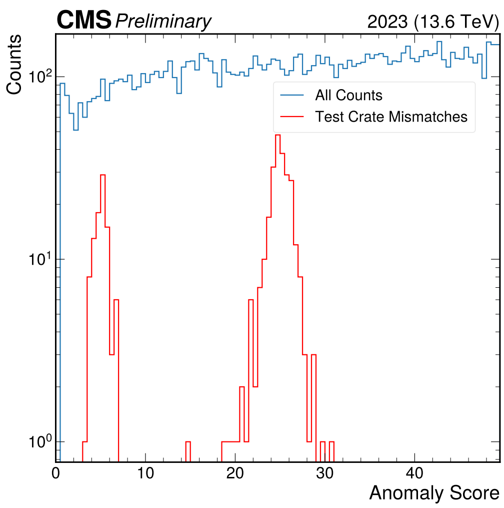

Figure 5. Anomaly Detection hardware and emulator trigger counts. Events from promptly reconstructed 2023 Ephemeral ZeroBias data where hardware bits are recorded from configured μGT test crate. Red segments represent mismatches between hardware and emulation. Clustering near decision boundaries implies issue is due to precision/rounding problem. Minimal mismatches in hardware vs. emulation (<1%) are observed.

L1 Menu Algorithm Name	Test Crate Count	Standalone Emulation Count	Mismatches
L1_ADT_20000	1	1	0
L1_ADT_4000	742	741	19
L1_ADT_400	21236	21229	253
L1_ADT_80	25468	25481	93

Table 4. Anomaly Detection hardware and emulator trigger counts. Events from promptly reconstructed 2023 Ephemeral ZeroBias data where hardware bits are recorded from configured μGT test crate. Test Crate Count shows events triggered in hardware and read out into data and Standalone Emulator Count is evaluated via offline inference with L1 objects.

References

[1] CMS Collaboration, *The CMS experiment at the CERN LHC*, *Journal of Instrumentation* **3** (2008) S08004.

[2] CMS Collaboration, *Development of the CMS detector for the CERN LHC Run 3*, Tech. Rep. CMS-PRF-21-001, CERN-EP-2023-136, CMS-PRF-21-001-003, CERN, Geneva (2023).

[3] CMS Collaboration, *The Phase-2 Upgrade of the CMS Level-1 Trigger*, Tech. Rep. CERN-LHCC-2020-004, CMS-TDR-021, CERN, Geneva (2020).

[4] E. Govorkova et al., *Autoencoders on field-programmable gate arrays for real-time, unsupervised new physics detection at 40 MHz at the Large Hadron Collider*, *Nature Mach. Intell.* **4** (2022) 154 [2108.03986].

[5] J. Duarte, S. Han, P. Harris, S. Jindariani, E. Kreinar, B. Kreis et al., *Fast inference of deep neural networks in FPGAs for particle physics*, *Journal of Instrumentation* **13** (2018) P07027 [1804.06913].

[6] CMS Collaboration, *Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector*, *JINST* **14** (2019) P07004 [1903.06078].

[7] C.N. Coelho, A. Kuusela, S. Li, H. Zhuang, T. Aarrestad, V. Loncar et al., *Automatic heterogeneous quantization of deep neural networks for low-latency inference on the edge for particle detectors*, *Nature Mach. Intell.* **3** (2021) 675 [2006.10159].

[8] AMD Xilinx, "Virtex 7 FPGA." https://www.xilinx.com/products/silicon-devices/fpga/virtex-7.html.

[9] M. Jeitler et al., *The level-1 global trigger for the CMS experiment at LHC*, *JINST* **2** (2007) P01006.

[10] FastML Team, *hls4ml*, May, 2023. 10.5281/zenodo.7933047.

[11] B. Herbert, "l1menus." https://github.com/herbberg/l1menus, 2020.

[12] CMS Collaboration, *Anomaly Detection in the CMS Global Trigger Test Crate for Run 3*, Tech. Rep. CMS-DP-2023-079, CERN, Geneva (2023).