Anomaly Detection in the CMS Global Trigger Test Crate for Run 3

Chang Sun
on behalf of the CMS Collaboration
Why Anomaly Detection?

• At the CMS experiment, we need to discard ~99.8% of data in real time within micro seconds.
  • See Thea's talk this morning

• We have signals that are hard to trigger on with the current trigger system.

• The signature for new physics is unknown.
Our Solution

• We propose a new ML-based L1 global trigger at the CMS experiment for new physics detection, AXOL1TL
Our Solution

- Use a Variational Autoencoder (VAE) based trigger algorithm, AXOL1TL, to select anomalous events in real-time
  - Sensitive to a wide variety of signals
  - VAE is a common method for anomaly detection
- VAE trained on ZeroBias data to detect data outliers
  - ZeroBias data: minimally biased, unfiltered data from the detector
- Firmware generation with hls4ml+vivado toolchain
- Standalone HLS emulation of Anomaly Detection output for validating performance
Workflow

2023 Zero Bias Collisions Dataset

Train QKeras

Model.h5

Evaluate QKeras

Plot Thresholds

Fast ML 2023

Workflow

Development

Conversion

Implementation / Validation

uGT HLS (C++)

Model HLS (C++)

Standalone Emulator (Python)

CMSSW Emulator

cms-hls4ml

L1 NTTuples (w/ Test Crate bits)

RAW Data File

Collisions

Online Monitoring

Prometheus

L1 keys

uGT bitfiles

Model VHDL

HDL Simulation

Test Vectors

L1 Menu xml

uGT VHDL fwk

uGT HLS fwk

Caltech PMA

Fast ML 2023
Anomaly metric

• The loss for a VAE is

\[
\text{Loss} = (1 - \beta) \| x - \hat{x} \|^2 + \beta \cdot \frac{1}{2}(\mu^2 + \sigma^2 - 1 - \log \sigma^2)
\]

Reconstruction term

KL regularization term

• Observe the KL term is a good anomaly metric, skip the decoder:

\[
2 \cdot \text{KL} = \mu^2 + \sigma^2 - 1 - \log \sigma^2
\]

• As the \( \sigma \) term is insignificant compare to the \( \mu \) term, skip the \( \sigma \) term:

\[
2 \cdot \text{KL} = \mu^2 + \sigma^2 - 1 - \log \sigma^2
\]

• These modifications reduce resource and latency of the network on FPGAs
  • Remove the decoder
  • Skip the slow exponential/logarithm operation
Dataset

- The model is trained with ZeroBias data collected by the CMS Experiment during 2023 with $\sqrt{s}=13.6$ TeV
  - Using the first 4 e/gamma, 4 muons, 10 jets, and missing ET objects as inputs
  - Each object has $p_t$, $\eta$, $\phi$ in hardware integers
  - 10.5 million events are used
    - ~50% for training
    - other ~50% for setting anomaly thresholds based on the score distribution
- Preprocessed by scaling with bit-shift and clipping extreme values.
- Signal samples are Monte-Carlo generated.
Architecture

The anomaly score
Anomaly Score and Thresholds

- Anomaly score distributions for 2023 ZeroBias events.
- Event scores for the QKeras model and standalone HLS emulator are consistent.
- Dotted lines represent the score thresholds implemented in the μGT trigger test crate.
Physics Performance

• We observed significant performance improvement on various SM and BSM signals by adding AXOL1TL to the 2023 trigger menu
  
  \[ \text{Improvement} = \frac{\text{L1 Efficiency w/ AXOL1TL@freq}}{\text{L1 Efficiency w/o AXOL1TL}} - 1 \]

• Shown in the table is the performance improvement for a Higgs decaying to 2 (pseudo-)scalars to bottom quarks.
  
  • $H \rightarrow a a [15 \text{ GeV}] \rightarrow 4b$

<table>
<thead>
<tr>
<th>AXOL1TL Rate</th>
<th>1 kHz</th>
<th>5 kHz</th>
<th>10 kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal Efficiency Gain</td>
<td>46%</td>
<td>100%</td>
<td>133%</td>
</tr>
</tbody>
</table>
Firmware Validation

- Xilinx Virtex-7 XCVU9P is used
  - 1728000 LUTs, 3456000 FFs, 12288 DSPs,
- No timing violation
- Resource and latency requirement met
  - 2 ticks@40MHz w/ initialization interval=1
  - LUT usage of O(1%)
Test Crate Monitoring

• An earlier version of AXOL1TL is installed in the test crate
  • Trained on 2018 ZeroBias data
Test Crate Monitoring

• An earlier version of AXOL1TL is installed in the test crate
  • Trained on 2018 ZeroBias data
• The rates shown are
  • AXOL1TL with 4 different thresholds, selected for testing purpose only. **They are not meant to be model realistic trigger rates.**
  • Single muon trigger
  • Total L1 physics rate
• Consistent performance observed through the partial fill cycle.
  • Initial rise-up corresponds to the beginning of a LHC fill
  • sawtooth pattern corresponds to luminosity leveling
Test Crate Validation

<table>
<thead>
<tr>
<th>L1 Menu Algorithm Name</th>
<th>Test Crate Count</th>
<th>Standalone Emulator Count</th>
<th>Mismatches</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1_ADT_20000</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>L1_ADT_4000</td>
<td>742</td>
<td>741</td>
<td>19</td>
</tr>
<tr>
<td>L1_ADT_400</td>
<td>21236</td>
<td>21229</td>
<td>253</td>
</tr>
<tr>
<td>L1_ADT_80</td>
<td>25468</td>
<td>25481</td>
<td>93</td>
</tr>
</tbody>
</table>

- Test Crate Count
  - from configured μGT test crate.

- Standalone Emulator Count
  - From offline inference with reconstructed L1 objects

- Minimal mismatches present (≤ 1%)

- Events are from promptly reconstructed 2023 ZeroBias data where trigger bits are recorded

- On the figure, red bins are mismatched events. As mismatches are concentrated around decision boundaries, rounding/precision issues are suggested.
Event Display

- Shown is the event with highest anomaly score that is not triggered by the normal L1 menu.
  - Ephemeral Zero Bias 2023 Run 367883.

- This is a very busy event with respect to the given the run condition

  - At L1, this event has 12 jets (of which 11 has $E_T > 20$ GeV), and 1 muon of 3 GeV.
  - Offline construction identifies 7 jets (PUPPI algorithm) with $p_T > 15$ GeV, and 1 muon.
  - Has 75 reconstructed vertices
    - Given the pile-up profile of the data taken in Run 2 and Run 3, this is highly unlikely.
Conclusion

• We demonstrated that our VAE-based model 🐸AXOL1TL is performant in anomaly detection as an L1 physics trigger.

• Our model meets the timing and resource requirements imposed by the L1 trigger system. A working copy of the model is installed in the test crate at the CMS Experiment.

• Further integration with the L1 trigger system is underway.
Backup: Test Crate Firmware Validation

- Number of trigger issued with four thresholds
  - (1250, 250, 25, 5)
- Test Vector Count
  - By standalone emulator
- HW Count
  - By standard global trigger firmware simulation workflow with ModelSim

<table>
<thead>
<tr>
<th>Idx</th>
<th>L1 Menu Algorithm Name</th>
<th>Standalone Emulator Count</th>
<th>HW Count</th>
<th>Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>94</td>
<td>L1_ADT_20000</td>
<td>0</td>
<td>0</td>
<td>✓</td>
</tr>
<tr>
<td>95</td>
<td>L1_ADT_4000</td>
<td>29</td>
<td>29</td>
<td>✓</td>
</tr>
<tr>
<td>103</td>
<td>L1_ADT_400</td>
<td>2618</td>
<td>2618</td>
<td>✓</td>
</tr>
<tr>
<td>108</td>
<td>L1_ADT_80</td>
<td>3331</td>
<td>3331</td>
<td>✓</td>
</tr>
</tbody>
</table>

Test vectors generated from Run 3 data
Backup: References


• FastML Team. hls4ml (Version v0.7.1) [Computer software]. https://doi.org/10.5281/zenodo.1201549.

