Detect New Physics with Deep Learning Trigger at the LHC

Zhenbin Wu (UIC)
Thong Nguyen (Caltech), Maurizio Pierini (CERN)
--on behalf of the CMS collaboration

CPAD Instrumentation Frontier Workshop
December 8-10, 2019
The LHC Big Data Problem

Data Flow

L1 Trigger
- 40M bunch crossing per second
- Logging rate: ~100 kHz
- Non-zero suppressed RAW data rate ~1 PB/s
- Zero suppressed data rate is ~20 TB/s
- Coarse local reconstruction implemented on FPGA/hardware.

HLT Farm
- Logging rate ~1 kHz
- Data rate ~1 GB/s distributed over dozens of primary datasets
- Simplified global reconstructions implemented on CPUs.

Offline Computing
- Roughly 1 GB/s data rate
- Global reconstruction fully optimized for accuracy with software implemented on CPUs.

Data Analysis
- User-written code, plots, theses, talks, etc.
- ~100 papers of 10 MB each year, less than 1 kB/s
Trigger at LHC

- The interested physics productions are much smaller comparing to inelastic production
- Trigger in LHC: finding a needle in a haystack scenario (anomaly)
- Event not trigger will be lost forever!
Trigger at HL-LHC

• The High-Luminosity of LHC:
  • Higgs, Flavour, Gauge Hierarchy, Supersymmetry, Dark Matter
  • O(100) GeV mass scales $\rightarrow$ O(50) GeV endpoints $\rightarrow$ O(20) GeV thresholds

• Weak-scale physics $\rightarrow$ Large statistics $\rightarrow$ High luminosity $\rightarrow$ Harsh environment!

• Great effort on upgrading Phase 2 Trigger system at HL-LHC

• Science potential of HL-LHC determined by datasets it collects
Workflow of Searches

- Pick a new model
- Design Trigger
- Collect Data
- Analysis
- No deviation
Workflow of Searches
The LHC Big Data Problem

A Drastic Data Reduction

Could new physics have been discarded somewhere in this process?
Model-Independent Searches in HEP

- Traditional new physics search relies on hypothesis testing with specific alternative models.
- Motivated multiple attempts for model-independent searches in high-energy physics over the years.
An Alternative Approach

• General approach by model-independent searches:
  • Look for discrepancy from the kinematic distribution of data versus expectation from Monte Carlo, taking into account of detector’s effects.
  • Look-elsewhere effect dilutes the discovery power with large number of bins.
• ATLAS’ proposal: use the analysis to identify an excess, but establish the significance with a traditional method (supervised) on an independent dataset.

Same spirit we have in mind for what follows…
Autoencoders in a Nutshell

• Compression-decompression algorithm that learns to describe the a given dataset in terms of point in lower-dimension *latent space*, from which it reconstructs the original data.

• **Unsupervised learning**, used for data compression, generation, clustering, etc.

• Anomaly: any event whose decompressed output is “far” from the input, in some metric of the autoencoder loss.

\[
\text{Loss} = f(\text{input} - \text{output})
\]
Autoencoders @ Level-1 Trigger

Standard Triggers

AE Triggers

- A Model-Agnostic Trigger for anomaly events with autoencoder (AE) model
- Deployment at Level-1 trigger to avoid any bias from upstream
- But limited by the resource and latency requirement on the Level-1 trigger system
CMS Phase 2 Level-1 Trigger

- Sketch of upgraded CMS Phase 2 Level-1 Trigger system
- Produce Particle Flow particles, combining Calo/Muon/Tracker information
- Produce PUPPI weight of each particles for pileup mitigation
- Outputs of each trigger systems send to Global Trigger for Level-1 decision
Example AE Model

- Train with simulated ZeroBias event at 200 pileup
- Use simulated Puppi Jet/MET/MHT inputs (18 inputs) with preprocessing

- Activation function: ReLU
- Loss function: L1Loss
- Training - validation ratio : 0.8
- Number of epochs: 100-200 epochs
- Number of layers: 8 layers

- Model is designed with simplicity for firmware implementation and resource/latency requirement
AE Performance

- Model was trained and validated with simulated Zerobias events, no knowledge of signal during training.
- Use the reconstruction loss of AE inputs and outputs as discriminator.
- Inference with signal samples show the separation power.

![Graph showing signal efficiency against rate in kHz for CMS Phase-2 Simulation at 14TeV, 200PU. Charts are labeled VBF H → Inv., VBF H → bb, and HH → 4b. Text on the graph indicates “Work in progress.”]
AE Implementation

• Use the hls4ml package to implement the AE model into FPGA firmware

• With additional logic for L1Loss function calculation

• Fully unroll AE with minimal latency, well within the Phase 2 Global Trigger latency budget

• With Xilinx Virtex UltraScale+ (VU9P) FPGA, the AE consumes ~10% of DSP resource, ~1% of Flip Flop and LUT

• To be included in the upcoming CMS Phase 2 Level-1 Trigger TDR
How to use the stream?

• Not to claim a discovery!

• Use as a resource to guide new physics searches in subsequent data takings, with some extra ingredients:
  • Data mining & visual inspection,
  • BSM-agnostic hypothesis testing.
Data Mining & Visual Inspection

- Macroscopic and microscopic views of the saved data stream.
- Learn any repeated patterns of events.
- Select a set of anomalies for visual inspection.
Learning New Physics from a Machine

Agnolo & Wulzer, arXiv:1806.02350

• Use SM MC as null hypothesis, run hypothesis testing without specifying alternative hypothesis.

• Allow for isolation of anomalous events by looking at their contribution to the likelihood ratio.
Conclusions

• The LHC has an enormous potential of discovering physics beyond the Standard Model, given the unprecedented collision energy and the large variety of production mechanisms that proton-proton collisions can probe.

• We propose a model-independent anomaly detection technique, based on deep autoencoders, to identify new physics events.

• Simple AE model can be implemented at the Level-1 trigger level.

• More advanced AE model can be designed for HLT or 40MHz scouting system (arXiv:1811.10276).

• Stay tune for the CMS Phase 2 Level-1 Trigger TDR.
BACKUP