Search for new physics using unsupervised machine learning for anomaly detection with ATLAS

Rui Zhang
on behalf of the ATLAS collaboration
University of Wisconsin-Madison, Wisconsin
Large Hadron Collider Physics Conference 2023
22–26 May 2023, Belgrade, Serbia
Introduction

- Model independent search provides broader and more efficient searches for new physics
  - Try not to be model specific, namely loose event selection criteria
  - Usually lead to high background

- Anomaly detection (AD) can help identify rare events that differ significantly from majorities

Today: Using anomaly detection to look for new phenomena in two body invariant masses

[Diagram]

- Apply kinematics selection
- Background shape modelling
- Look for excess "BumpHunter"
- No excess? Set limits on cross section

Apply "AD" selection

Expect AD to give high background suppression and low signal efficiency loss

ATLAS-CONF-2023-022
Analysis strategy

Trigger (one lepton) and pre-selection
\( p_T^l > 60 \text{ GeV}, p_T^{\text{jet}} > 30 \text{ GeV} \)

Reconstruct Rapidity Mass Matrix for each event

Train autoencoder using 1% ATLAS Run2 data

Define signal region using reconstruction loss from autoencoder

Fit invariant mass spectrum, statistical analysis, look for bumps
Advantage of using Anomaly Detection

- Not relying on specific signal hypothesis — model independent search.

- Unsupervised anomaly detection trained on data — no MC modelling dependence

- Using event topologies on the standard reconstructed objects (jet, b-jet, e, μ, γ, met) — object type & multiplicity, 4-momenta, two-body system information, all will contribute to the anomaly score.
Event representation: Rapidity Mass Matrix

Multi-jet QCD process

Higgs process

- Expected to have different characteristics for different processes

Source: 1810.06669

Analog to a QR code
Train autoencoder

- Using randomly selected 1% collision data
  - Sufficient statistics to train and well represent the full collision dataset
  - Split to 7:3 for training and validation, monitor validation loss for early stopping
- Tried other architectures such as Variational (convolutional) autoencoder and various sizes of the autoencoder. The selected one gives better performance

\[ \text{loss} = \frac{1}{n} \sum_{i=1}^{1287} (x_i - \hat{x}_i)^2 \]
Anomaly signal test

Can model detect anomalous?

Make up anomaly events by shuffling objects and re-calculating RMM

- Anomaly 1: jets beyond 2 are set to photon
- Anomaly 2: jets beyond 2 are set to b-jets
- Anomaly 3: keep only 2 jets and 1 lepton
  - Less anomaly even than the original events

They all show up with large loss, in addition:

- More anomalous is seen when b-jet or photon multiplicity increases - expected
- Less anomalous is seen when multiplicity is low, anomalous may come from large pT, Etmiss, etc - expected
Anomaly region definitions

- Anomaly region should enhance BSM signal and suppress SM bkg
  - Need enough bkg for modelling

- Select events corresponding to 10pb, 1pb, 0.1pb ($\times 140$ fb$^{-1}$) as 3 anomaly regions to cover different sensitivities

- 10pb AR is the main one to study in this analysis
Background modelling

- The mass spectra of background (SM) are expected to be smoothing falling.
- Using SM MC and loose-lepton control regions to establish background function form (five parameters, hence “p5”)

\[
f(x) = p_1 (1 - x)^{p_2} x^{p_3} + p_4 \ln x + p_5 \ln^2 x
\]

- Also used an alternative function form to estimate systematics
  - Replace the highest order term with a different from that will affect the tail most

\[
f(x)^{\text{alt}} = p_1 (1 - x)^{p_2} x^{p_3} + p_4 \ln x + p_5 / \sqrt{x}
\]
Results
BumpHunter results for 10 pb WP

- Agree with p5 fit
  - Tests of normality on pulls passed for all masses passed
  - Background shape uncertainty shown in yellow

- Largest deviation reported by BumpHunter is $m_{j\mu}$ at ~4.8 TeV
Demonstration of sensitivity to BSM signals

- Sensitivity improvement quantified by $\Delta Z$

$$\Delta Z = \left( \frac{Z_{AE}}{Z} - 1 \right) \times 100\%$$

$$Z = \sqrt{2 \left( (s + b) \ln \left( 1 + \frac{s}{b} \right) - s \right)}$$

$$\Delta Z > 0 \Rightarrow \text{improvement}$$
Limit setting using generic Gaussian signals

- Uncertainties include:
  - Luminosity
  - Experimental uncertainties in signal derived from SSM W'/Z' MC simulation and applied to Gaussian signal
    - JES, JER, Lepton energy scale, etc
  - Alternative bkg functional shape in bkg modelling
    \[ f(x) = p_1 (1 - x) p_2 x^{p_3 + p_4 \ln x + p_5 \ln^2 x} \] \( \text{VS} \) \( f(x)^{alt} = p_1 (1 - x) p_2 x^{p_3 + p_4 \ln x + p_5 / \sqrt{x}} \)
  - Spurious signal uncertainty
Upper limits of Gaussian signals

- Signal width of $\sigma=0$ and $\sigma/m=15\%$ are shown
  - Narrow signals have better limits as expected
- Error band is from $\sigma=0$
- Waves are similar, $\sigma=0$ is subject to local fluctuations
- Local $2.9 \sigma @ m_{j\mu} = 4.8\text{TeV}$, $2.8 \sigma @ m_{j\mu} = 1.2\text{TeV}$
Conclusions

- An successful application of unsupervised machine learning for anomaly detection using event level information
- Searched for new phenomena in 9 invariant masses for jet+X (b-jet+X) from 3 outlier regions
- Largest deviation for j+μ near 4.8 TeV is consistent with a statistical fluctuation
- Analysis method shows improvement of sensitivity up to ~300%
- Model-independent limits (gaussian signal) is presented
An event passing 10 pb Anomaly Region with $m_{j\mu} = 4.72$ TeV.

Left grey cones: $p_T^j = 1376$ GeV, $\eta = -1.79, \phi = 3.059$

Red: $p_T^\mu = 430$ GeV

Green: $E_T^{\text{miss}} = 256$ GeV

Thank you!
Backup
Loss distributions for BSM models with high mass
Fit results without applying Anomaly Region cut