Faster and Robust anomaly detection w/ NuRD

Abhijith Gandrakota[1], Lily Zhang[2], Aahlad Puli[2], Nhan Tran[1], Jennifer Ngadiuba[1]

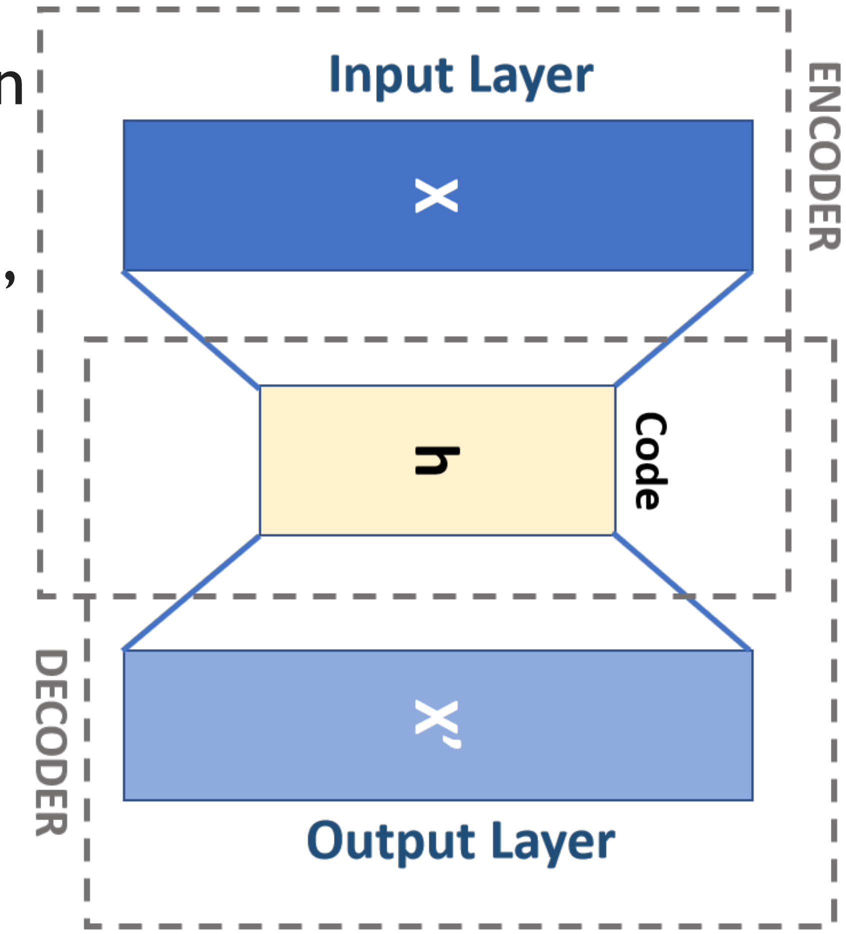
1: Fermilab
2: New York University

Fast Ml for Science '22, SMU

Arxiv: 2210.SOON

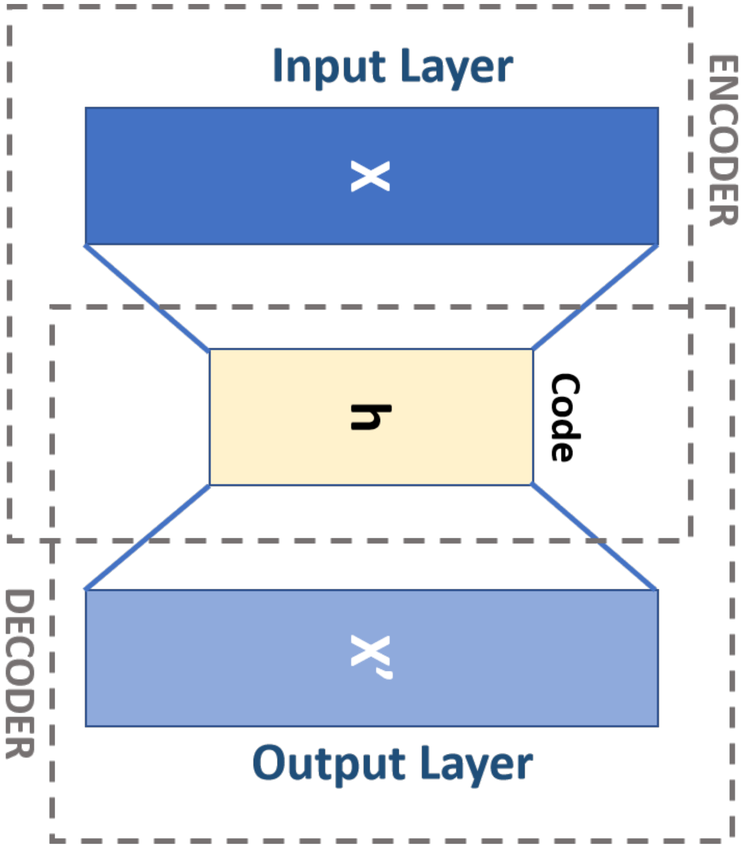
Introduction

- A standard approach for anomaly detection in High Energy Physics (@ LHC)
  - Look for "deviations" from expected (dominant) background physics
  - Encode the input information into a latent representation
  - Decode the representation back to initial representation, examine reconstruction loss (~MSE)
  - Use the reconstruction loss to find anomalies



Introduction

- A standard approach for anomaly detection in High Energy Physics (@ LHC)
  - Look for "deviations" for a expected (dominant) background physics
  - Encode the input information into a latent representation
  - Decode the representation back to initial representation, examine reconstruction loss (~MSE)
  - Use the reconstruction loss to find anomalies



- Primary concerns
  - Is the algorithm modeling the desired physics (e.g. semantics) correctly?
    - More importantly, is it learning anything we don't want it focus on ?
    - AEs model everything, even the unimportant features
- Different take in approaching this challenge using NuRD

Robust anomaly detection

- *More importantly, is it learning anything we don't want it to know ?*
- Objective: Distinguish between the animals ?

Our Training data:

Cows in a typical
Grass background

Robust anomaly detection

- *More importantly, is it learning anything we don't want it to know ?*
- Objective: Distinguish between the animals ?

Our Training data:

Cows in a grassland backdrop

Sure, we may detect penguins in snow

Robust anomaly detection

- *More importantly, is it learning anything we don't want it to know ?*
- Objective: Distinguish between the animals ?

Our Training data:

Cows in a grassland backdrop

Sure, we may detect penguins in show
Expected anomaly

This ?
Actual Anomaly

Robust anomaly detection

- *More importantly, is it learning anything we don't want it to know ?*
- Objective: Distinguish between the animals ?

Our Training data:

Cows in a grassland backdrop

Sure, we may detect penguins in snow
Expected anomaly

This ?
Actual Anomaly

How about this ?
Typical BKG in data

Robust anomaly detection

- *More importantly, is it learning anything we don't want it to know ?*
- Objective: Distinguish between the animals ?

Our Training data:

Cows in a grassland backdrop

Needs to learn this !

What if it learnt this ?

Sure, we may detect penguins in show
Expected anomaly

This ?
Actual Anomaly

How about this ?
Typical BKG in data

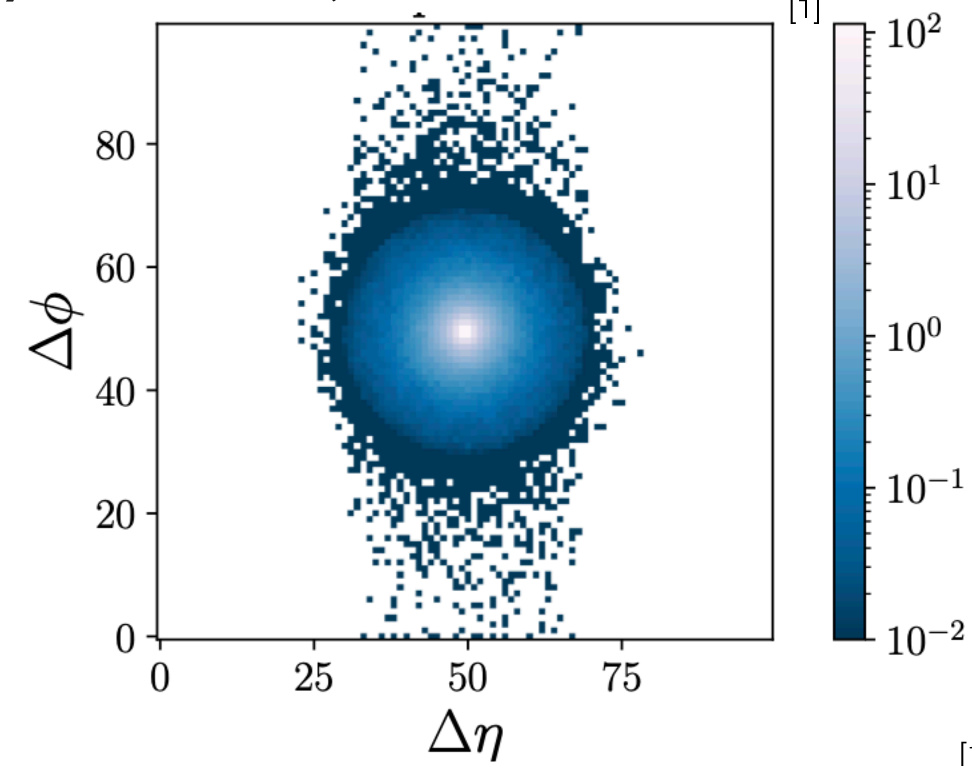
From inputs to representations

- Issue : Density estimation on the inputs models ***everything*** about the data
  - We want to model semantic features (*like jet structure*) while being decorrelated with nuisances (*like mass*)
- Solution: use different backgrounds to learn what is semantic
- Summary:
  - Use multiple known background labels (not just QCD).
  - Build representations to have maximum information with the labels.
  - Ensure representations do not vary w/ nuisances (Zhang et al. 2022, Puli et al. 2022).

The Inputs

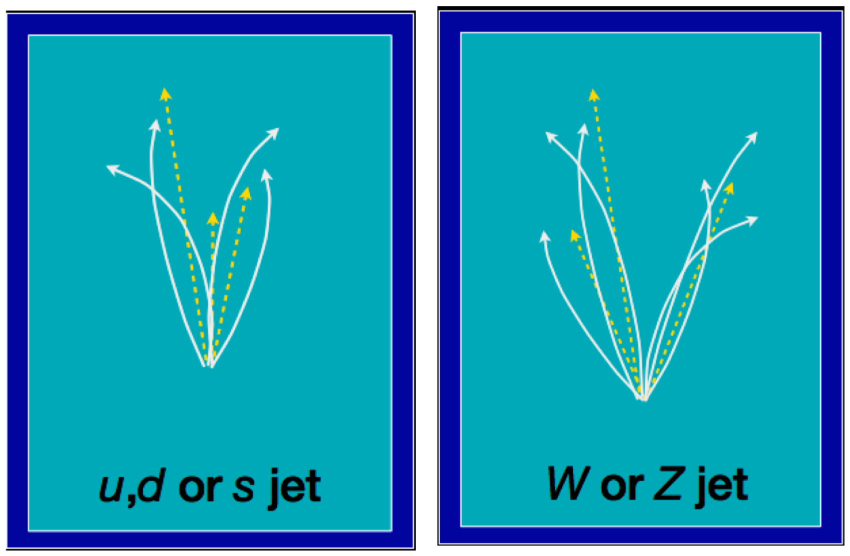
- For out dataset we have input features (X), labels for BKG types (Y), and Nuisance (Z)
- Objective is to learn particles decays at LHC, specifically hadronic jet shower

[1]



  - Input: Energy deposits in the detectors
    - Images ~ 50 X 50 pixels

[1]



- We have two background samples to learn semantics
  - We use QCD and WZ jets w/ labels
- We don't want the our representation to depend on the nuisance

[1] JEDI-Net, Eric A. Moreno et al

Nuisance Randomized Distillation

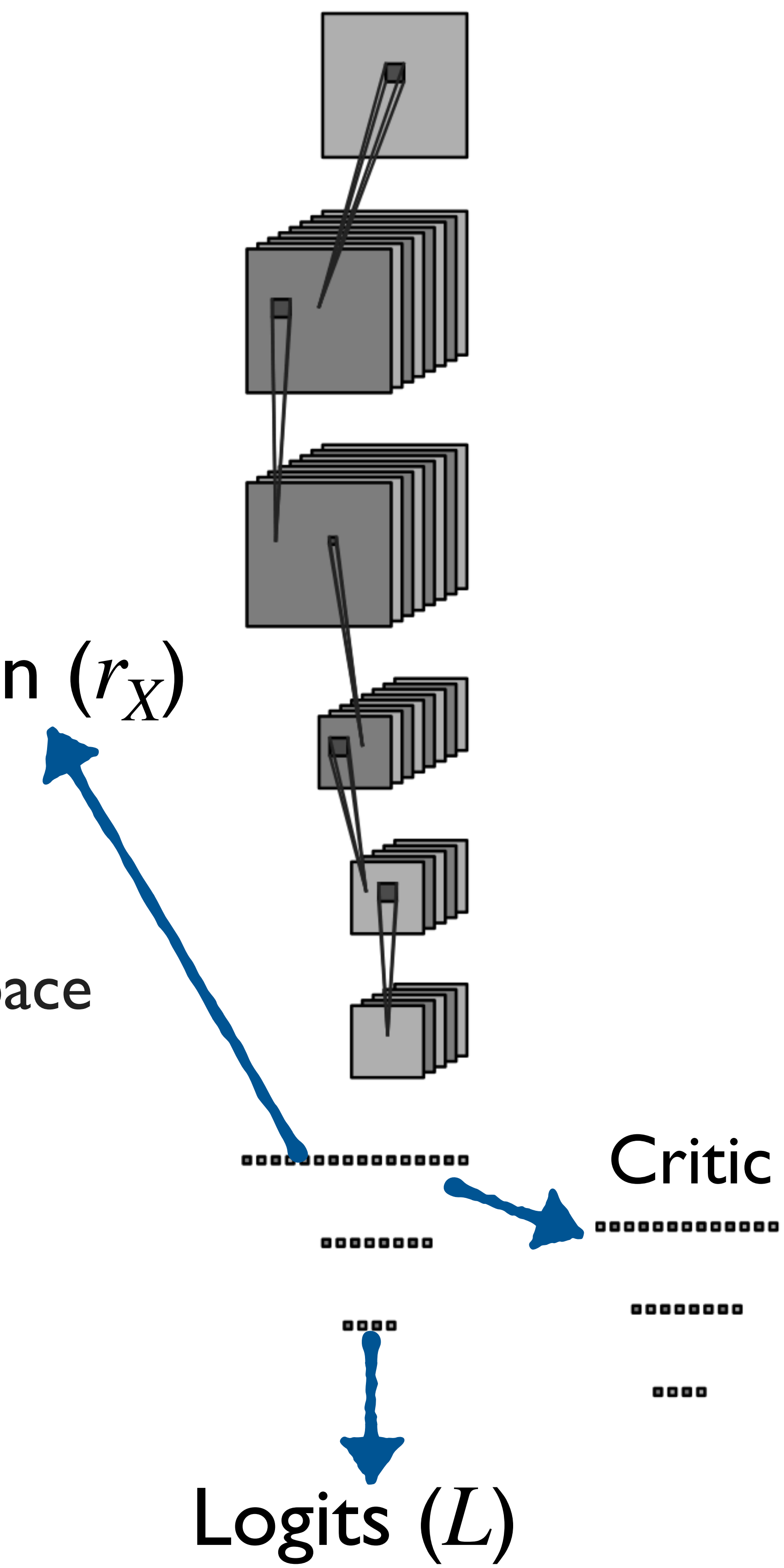
- For out dataset we have input features (X), labels for BKG types (Y), and Nuisance (Z)
- **Nu**isance **R**andomized **D**istillation::
  - I : Do not let model learn nuisance: break the dependence b/n label and nuisance.
    - Use importance weights w to break dependence.
  - II : Build informative representations that do not vary with the nuisance:
    - Intuitively, it shouldn't be possible to distinguish b/n
      - (r_X,Y,Z)
      - (r_X,Y, randomized nuisance($\hat{Z}$))
    - Can enforce this w/ critic model ϕ

$$\mathscr{L} = w\left(CE(Y_{pred}, Y_{true}) - \lambda \log \frac{p_\phi(r_X, Y, [Z, \hat{Z}])}{1 - p_\phi} \right)$$

- Use the representations to detect anomalies.

Model and the OOD Score

- Building out representation:
  - Main model: CNNs w/ final dense layers output to logits
    (Similar to the CNN Encoder architecture used in QCD AE)
    - Representation is the output from N-1 layer
  - Critic: Simple MLP, output to Logits
  - OOD Dataset: Top quarks



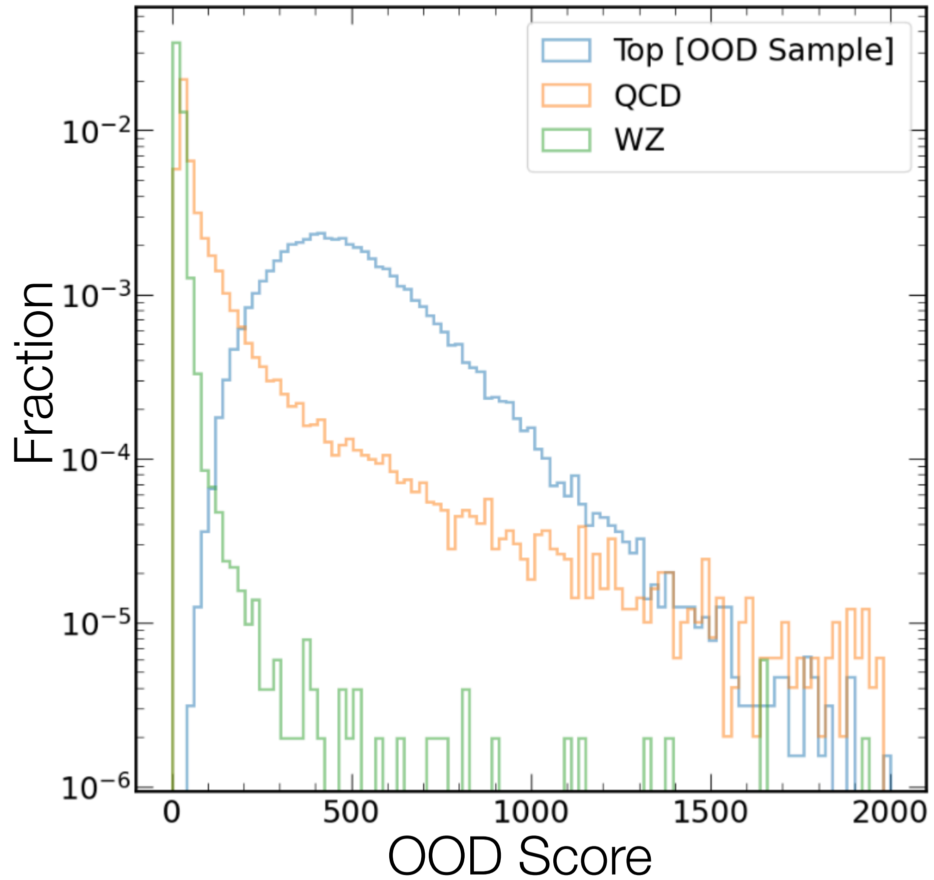
- OOD Score:
  - Calculate the distance from samples in representation space

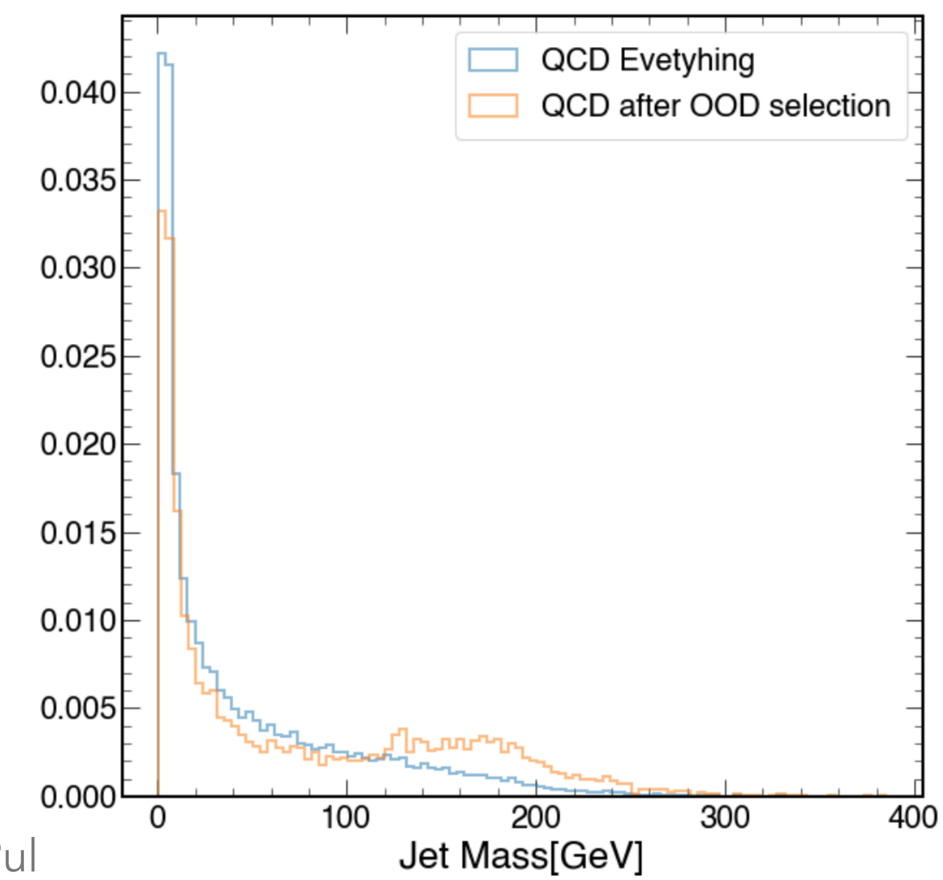
$$d_A = (r_X - < r_A >)\ \Sigma(r_A)^{-1}\ (r_X - < r_A >)^T$$

  - Get the distance of from both backgrounds, $[d_{QCD}, d_{WZ}]$
- Detect out of distribution using this information

Experiments and Results

- Trained on QCD and WZ labeled data to build out the representation space
  - Representation space is has a dimension of 20
  - The critic model :3 layers w/ 256, 128, 68 neurons
- Enforced Joint Independence with two weights
  - Lambda values of: 0.001,   1
  - Leading to AUC of: 0.94,   0.83
    (Baseline: AUC w/ plain AE : 0.88)
  - Corr. of OOD Score and Mass (QCD) : 0.012
- Representation w/ Joint independence gives us robustness:
  - Performance guarantees across different BKG-distributions





Why is this relevant for Fast ML ?

- We can shrink the model's size !
  - Can Reduce by roughly by half compared to Autoencoders like density estimation methods
  - Leads to Algorithms w/ smaller footprint and faster inference times
  - We can do this while making it ROBUST !
  - Best of all, the detection procedure is FPGA friendly.
- But what the catch ?
  - The Critic is re-trained for every batch !
    - Dramatically increases the *training* time, scales like ~ n^2
  - We need labeled backgrounds to build the representation space.

Summary

- In HEP (often many other fields) we have multiple backgrounds. We should use information contained in all of them.
- This is a new take on building a representation space to detect anomalies:
  - Training w/ background labels gives us good performance.
  - NuRD, via joint independence, helps
    - Maximize physics while decorrelating nuisances
- This technique although takes longer to train, results in smaller models
  - A primary benefit of increased robustness.
- Paper will be out on Arxiv soon with code.

Thank you