Challenges for unsupervised anomaly detection in particle physics

Katherine Fraser

Department of Physics
Harvard University
kfraser@g.harvard.edu

ML4Jets 2022

arXiv: 2110.06948, JHEP 03 (2022) 066

with S. Homiller, R. Mishra, B. Ostdiek, M. Schwartz

Outline

1. Introduction: Outlier Detection vs. Density Estimation
2. Two methods for Outlier Detection:
    A. Architectures
        i. Variational Autoencoders
        ii. Wasserstein Distances
    B. Results
3. Understanding Latent Space

Outline

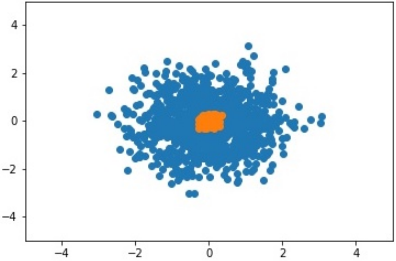
1. Introduction: Outlier Detection vs. Density Estimation
2. Two methods for Outlier Detection:
    1. Architectures
        1. Variational Autoencoders
        2. Wasserstein Distances
    A. Results
3. Understanding Latent Space

Why Anomaly Detection?

- The goal of unsupervised anomaly detection is to develop less model dependent methods.
- Try to develop methods that are trained only on background but can be used to find signals
- Can be divided into outlier detection and finding over densities

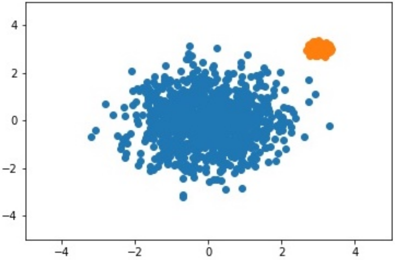
Two Types of Anomaly Detection

Finding Overdensities



[Collins et al: 1805.02664, D’Anglo + Wulzer: 1806.02350, Collins et al: 1902.02634, D’Anglo et al: 1912.12155, Nachman & Shih: 2001.04990, Stein et al: 2012.11638, Carron et al: 2106.10164, Hallin et al: 2109.00546, + many others]

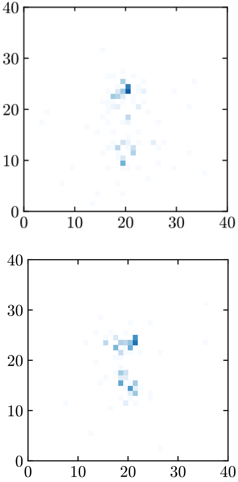
Outlier Detection



[Hajer et al: 1807.10261, Heimel et al: 1808.08979, Farina et al: 1808.08992, Cerri et al: 1811.10276, Roy + Vijay: 1903.02032, Atkinson et al: 2105.07988, Carron et al: 2106.10164, Ngairangbam et al: 2112.04958, + many others]

Simplifying the Problem

- Full event anomaly detection is hard
- Consider the simplified problem of detecting top and W jets in a QCD dijet background.
- Use jet images of simulated LHC jets, which have been preprocessed (flipped, rotated, discretized) and normalized by total pT.



Sample Images: QCD Jet (Above), Top Jet (Below)
[Fraser et al: 2110.06948]

Outline

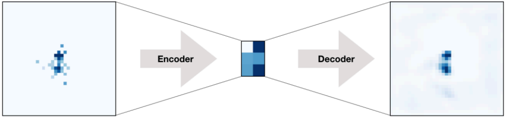
1. Introduction: Outlier Detection vs. Density Estimation
2. Two methods for Outlier Detection:
    A. Architectures
        i. Variational Autoencoders
        ii. Wasserstein Distances
    B. Results
3. Understanding Latent Space

AEs for Anomaly Detection

- In an autoencoder (AE), an encoder compresses inputs to a latent space, and then a decoder tries to map the latent space back to the original data by minimizing a reconstruction loss such as the mean power error:

$$d^{(\alpha)}_{MPE}(\mathcal{I}_1, \mathcal{I}_2) = \frac{1}{N_{pixels}} \sum_{i \in pixels} |\mathcal{I}_{1,i} - \mathcal{I}_{2,i}|^{\alpha}$$

- When the AE is trained on background, the reconstruction fidelity gives an anomaly score: background-like events should be reconstructed well while signal-like events should not [Heimel et al: 1808.08979, Farina et al: 1808.08992]



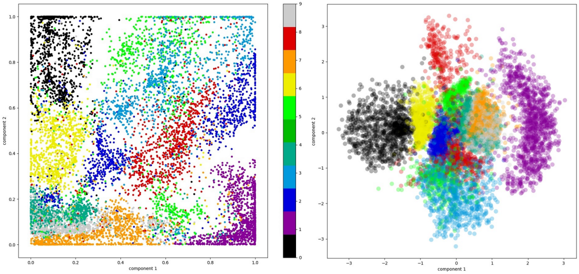
Schematic AE [Farina et al: 1808.08992]

Adapting Variational Autoencoders (VAEs)

- In a VAE, the latent space consists of multiple distributions (gaussians) that the decoder samples from, and a KL divergence is added to the loss to regularize training:

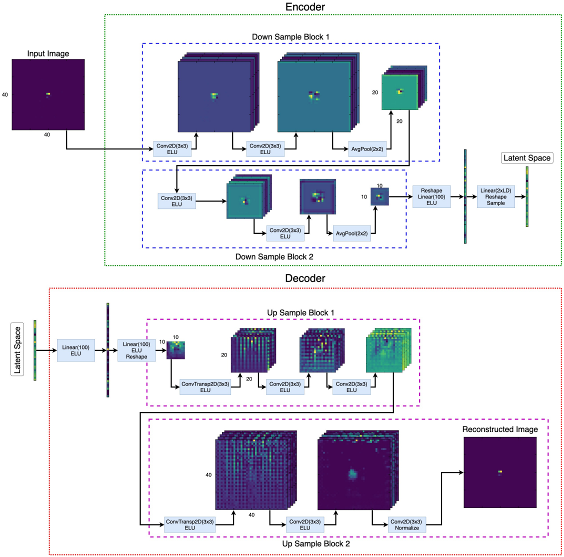
  $$\text{Loss} = (1 - \beta) \times \text{Reconstruction Loss} + \beta \times \text{KLD}$$

  This allows the VAE to be used for variational inference.
- This stochasticity gives distances in latent space meaning.



[PureAI]

Our Architecture



[Fraser et al: 2110.06948]

The VAE architecture contains:

- An encoder with downsampling blocks (each with convolutional layers, elu activations, and a pooling layer) and dense layers
- A decoder that mirrors the encoder.

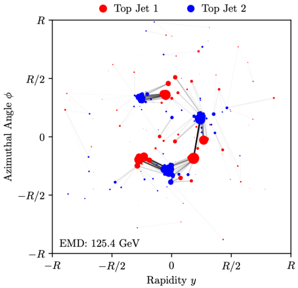
A More Physical Alternative

- Optimal transport (OT) is the minimum "effort" required to transform one event into another.
- The OT distance is

$$d_{OT} = min_f \sum_{i,j} f_{ij} c_{ij}$$

where f_{ij} is the transport plan (where and how to transport intensity) and c_{ij} is the cost function (how much work it takes to transport one unit of intensity).

- Optimal transport can be balanced or unbalanced. We normalize our images and restrict to balanced OT.



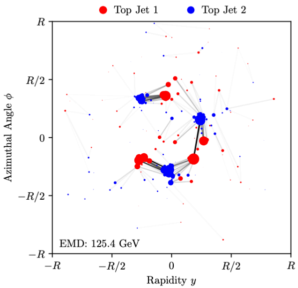
Example OT Plan
[Komiske et al: 1902.02346]

A More Physical Alternative

- Examples of OT metrics include the Energy Movers Distance [Komiske et al: 1902.02346, 2004.04159] and more general Wasserstein distances

$$d^{(p)}_{Wass} = \left(min_f \sum_{i,j} f_{ij} (c_{ij})^p \right)^{1/p}$$

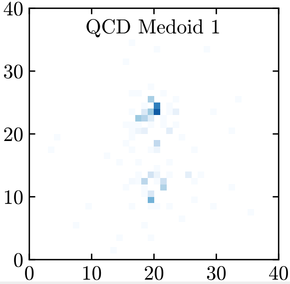
where c_{ij} is the Euclidean distance in the (η, ϕ) plane.

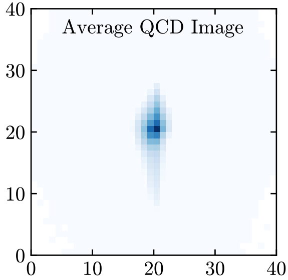


Example OT Plan
[Komiske et al: 1902.02346]

Using Optimal Transport Distances

- OT gives the distance between events. How can we use it to get a score for the “distance” to a distribution?
- Pick reference samples and use OT distances to the references as an anomaly score.
- Test average QCD image and k-medoids of the QCD jets as the reference, with k chosen by the elbow method. Medoids perform better.





Outline

1. Introduction: Outlier Detection vs. Density Estimation
2. Two methods for Outlier Detection:
    A. Architectures
        i. Variational Autoencoders
        ii. Wasserstein Distances
    B. Results
3. Understanding Latent Space

Key Questions

- How do the VAE and Event-to-Ensemble OT compare?
- How robust are the VAE and Event-to-Ensemble distance?
  - Do results depend on reconstruction loss/distance choice? (Ex. MAE, MSE, Wasserstein distance - implemented with the Sinkhorn approximation through the GeomLoss package)
  - How model independent are the best choice of reconstruction loss and other hyperparameters? (Ex. β, number of downsampling blocks)

VAE Results

Signal Training Metric	Down Samplings	Anomaly Metric	Top jet AUC	Top jet $\epsilon_S(\epsilon_B = 0.1)$	W jet AUC	W jet $\epsilon_S(\epsilon_B = 0.1)$
Supervised	-	-	**0.94**	**0.81**	**0.96**	**0.91**
MSE	2 ($\beta = 10^{-7}$)	Loss	0.83	0.48	**0.65**	0.14
	3 ($\beta = 10^{-8}$)	Loss	**0.84**	0.49	0.65	0.12
		MSE	0.84	0.48	0.65	0.12
		MAE	0.81	0.39	0.53	0.04
		Wass(1)	0.84	0.51	0.52	0.05
		Wass(2)	0.82	**0.51**	0.54	0.08
Wass(1)	2 ($\beta = 10^{-8}$)	Loss	0.79	0.37	0.46	0.04
		MSE	0.76	0.33	0.61	**0.15**
		Wass(1)	0.79	0.37	0.46	0.04

- The VAE performs best with MSE loss and 2-3 downsampling layers.
- Wasserstein loss doesn’t perform as well for most benchmarks

OT Results

Metric	Number of medoids	Method	Top jet AUC	Top jet $\epsilon_S(\epsilon_B = 0.1)$	W jet AUC	W jet $\epsilon_S(\epsilon_B = 0.1)$
Wass(1)	-	Avg	0.81	0.33	0.62	0.02
	1	Medoid	0.83	0.28	0.63	0.02
	3 (elbow)	Medoids (min)	0.85	0.43	0.67	0.04
	5	Medoids (min)	**0.87**	**0.54**	0.60	0.05
	7	Medoids (min)	0.87	0.54	0.61	0.05
Wass(5)	4 (elbow)	Medoids (min)	0.67	0.22	0.41	0.04
MAE	1	Medoid	0.82	0.40	**0.71**	0.07
	3 (elbow)	Medoids (min)	0.82	0.49	0.61	**0.08**

- Best Top vs. QCD: 1-Wasserstein metric
- MAE does well for QCD vs. W; correlated with Wass(1) here
- Find worse performance for larger p: small pixel differences become comparatively less important, agrees with [Finke et al: 2104.09051] for AEs.

Comparison

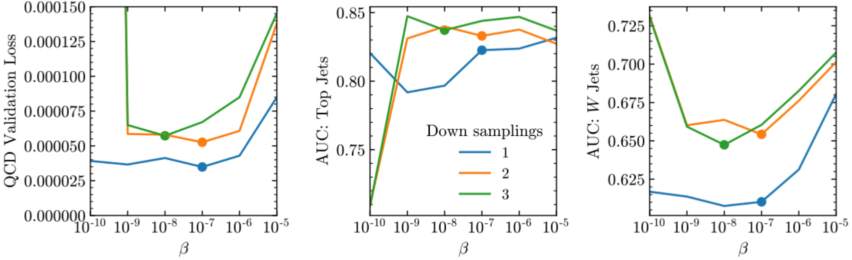
Signal Training Metric	Down Samplings	Anomaly Metric	Top jet AUC	Top jet $\epsilon_S(\epsilon_B = 0.1)$	W jet AUC	W jet $\epsilon_S(\epsilon_B = 0.1)$
Supervised	-	-	**0.94**	**0.81**	**0.96**	**0.91**
MSE	2 ($\beta = 10^{-7}$)	Loss	0.83	0.48	**0.65**	0.14
	3 ($\beta = 10^{-8}$)	Loss	**0.84**	0.49	0.65	0.12
		MSE	0.84	0.48	0.65	0.12
		MAE	0.81	0.39	0.53	0.04
		Wass(1)	0.84	0.51	0.52	0.05
		Wass(2)	0.82	**0.51**	0.54	0.08
Wass(1)	2 ($\beta = 10^{-8}$)	Loss	0.79	0.37	0.46	0.04
		MSE	0.76	0.33	0.61	**0.15**
		Wass(1)	0.79	0.37	0.46	0.04

- Reference samples slightly outperform the VAE with most benchmarks

- Best hyperparameters are signal dependent

Metric	Number of medoids	Method	Top jet AUC	Top jet $\epsilon_S(\epsilon_B = 0.1)$	W jet AUC	W jet $\epsilon_S(\epsilon_B = 0.1)$
Wass(1)	-	Avg	0.81	0.33	0.62	0.02
	1	Medoid	0.83	0.28	0.63	0.02
	3 (elbow)	Medoids (min)	0.85	0.43	0.67	0.04
	5	Medoids (min)	**0.87**	**0.54**	0.60	0.05
	7	Medoids (min)	0.87	0.54	0.61	0.05
Wass(5)	4 (elbow)	Medoids (min)	0.67	0.22	0.41	0.04
MAE	1	Medoid	0.82	0.40	**0.71**	0.07
	3 (elbow)	Medoids (min)	0.82	0.49	0.61	**0.08**

Hyperparameter Dependence



- There is no signal independent way of choosing hyperparameters.
- Choices that best represent the background are often not best for signal detection: β with the lowest loss on the validation samples is NOT best for QCD vs. W classification
- Also applies to choosing metric/number of medoids for reference samples

Semi-Supervised Results with OT

Reference Sample	Metric	Number of medoids	Method	Top jet AUC	Top jet $\epsilon_S(\epsilon_B = 0.1)$	W jet AUC	W jet $\epsilon_S(\epsilon_B = 0.1)$
Supervised	-	-	-	**0.94**	**0.81**	**0.96**	**0.91**
QCD Ref Best	MAE/Wass(1)	Various	Medoids	**0.87**	**0.54**	**0.71**	**0.08**
Top Reference	Wass(1)	3 (elbow)	Medoids (min)	0.32	0.07	0.79	0.53
		5	Medoids (min)	0.45	0.12	**0.84**	**0.62**
	Wass(5)	2 (elbow)	Medoids (min)	0.72	**0.32**	0.70	0.06
		3	Medoids (min)	0.66	0.20	0.61	0.04
		5	Medoids (sum)	**0.73**	0.30	0.58	0.02

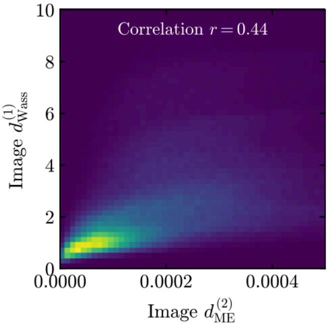
- OT is easy to apply to other reference samples: also use top jets as a reference and try to detect QCD vs. Top jets or QCD vs. W jets
- Comparing to a Top reference sample is better than comparing to a QCD sample for QCD vs. W classification but not Top vs. W Classification.
- For Top vs. QCD classification with top reference samples, higher p is better.

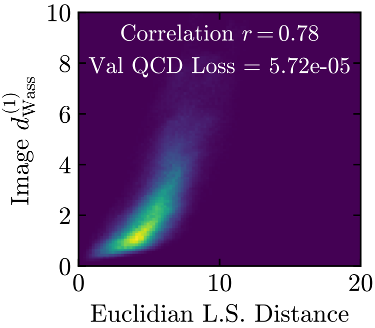
Outline

1. Introduction: Outlier Detection vs. Density Estimation
2. Two methods for Outlier Detection:
    A. Architectures
        i. Variational Autoencoders
        ii. Wasserstein Distances
    B. Results
3. Understanding Latent Space

Understanding the Latent Space

- Can we use the latent space to understand what the VAE is learning?
- Distances between events in the VAE latent space are correlated with Wasserstein OT distances between the same pairs. Downsampling helps generate these correlations.





Summary

1. There are general challenges with outlier detection. For both VAEs and OT with reference samples, choices that best represent the background are often not best for signal detection. Outlier detection is inherently signal dependent and hard to optimize.
2. The event-to-ensemble Wasserstein distances do as well or better than the VAE because Wasserstein OT distances and VAE latent space distances are correlated. This suggests there could be cases where they can be interchanged.

Back Up Slides

Variational Inference with VAEs

- Data x, Latent space elements z
- Let where $q_\phi(z|x)$ is the VAE encoder. Then p(x) =

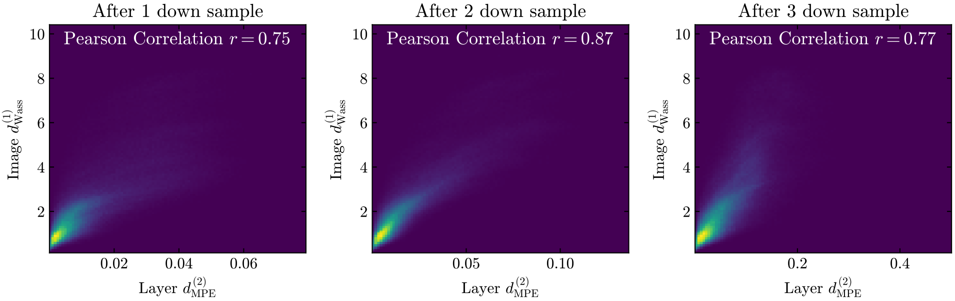
$$\mathbb{E}_{p(z)}[p(x|z)] = \int p(x|z)p(z)dz$$

$$= \int q_\phi(z|x)\frac{p(x|z)}{q_\phi(z|x)}p(z)dz = \mathbb{E}_{q_\phi(z|x)}\left[\frac{p(x|z)p(z)}{q_\phi(z|x)}\right]$$

- $\Rightarrow \log p(x) = \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p(x|z)p(z)}{q_\phi(z|x)}\right]$

$$\geq \mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p(x|z)p(z)}{q_\phi(z|x)}\right)\right] = \mathbb{E}_{q_\phi(z|x)}\left[\log p(x|z) - \log\left(\frac{q_\phi(z|x)}{p(z)}\right)\right]$$

Downsampling vs. Layers



The Elbow Method