Identifying semi-visible jets with darkCLR

Tanmoy Modak
6 November 2023
ML4Jets 2023
DESY and Universität Hamburg

Anomaly detection, Self-supervision

Anomaly detection

- Model-agnostic: No assumption on signal
- Density based
- The ability process high dimensional dataset
- Trained on background samples
- Not invariant under phase-space transformation

Self-supervision

- Self-supervision: model learns from the data itself. uses ‘pseudo-labels’ during training
- Control training such that the representation should have

  1) Invariance to certain transformations of the events/jets

  2) The discriminative power to anomalies

Contrastive Learning Representation

Contrastive learning representation (CLR)

- CLR: pseudo-labels are used network optimization via contrastive loss function
- Learns high-dimensional correlations in the data
- Learnt representations can be used for downstream tasks

- Positive-pair labels $\{(x_i, x'_i)\}$: each data point to an augmented version that does map to itself

- Negative-pair $\{(x_i, x_j) \cup (x_i, x'_j)\}$ for $i \neq j$: match each data point in the sample to every other that is not itself or an augmented/transformed version of itself

- *f* (typically a transformer encoder): $f : \mathcal{D} \to \mathcal{R}$

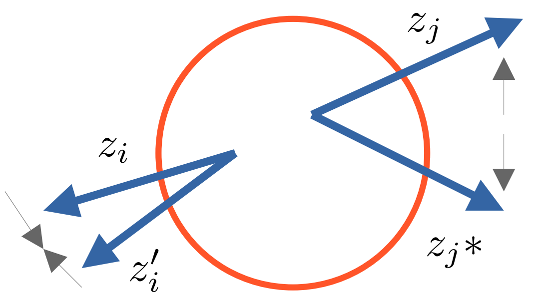
$$\mathcal{L}_{\mathrm{CLR}} = -\log \frac{e^{s(z_i, z'_i)/\tau}}{\sum_{j \neq i \in \mathrm{batch}} \left[e^{s(z_i, z_j)/\tau} + e^{s(z_i, z'_j)/\tau} \right]} \text{ with,}$$

$$z_i = f(x_i) \text{ and } z'_i = f(x'_i), s(z_i, z_j) = \frac{z_i \cdot z_j}{|z_i||z_j|} = \cos\theta_{ij} \ .$$

The AnomalyCLR

(B. M. Dillon, L. Favaro, F. Feiden, TM, T. Plehn; arXiv:2301.04660)

- Modified contrastive learning for anomaly detection
- Positive pairs (z_j,z_j'): physical augmentations, invariant
- Anomaly pairs (z_j,z_j*): anomaly augmentations
- AnomalyCLR: Anomaly augmentations



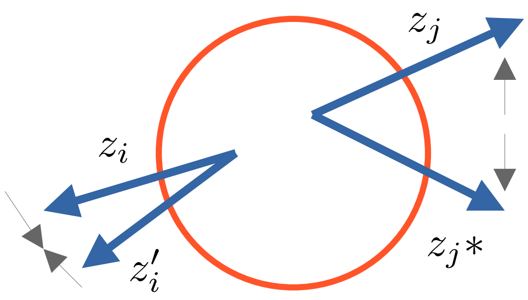
[anomalyCLR]

$$\mathcal{L}_{\text{AnomCLR}} = -\log \frac{e^{\left[s(z_i,z_i')-s(z_i,z_i^*)\right]/\tau}}{\sum_{j\neq i\in\text{batch}}\left[e^{s(z_i,z_j)/\tau}+e^{s(z_i,z_j')/\tau}\right]}$$

The AnomalyCLR

$$\mathcal{L}_{\text{AnomCLR}} = -\log \frac{e^{\left[s(z_i,z_i')-s(z_i,z_i^*)\right]/\tau}}{\sum_{j\neq i\in\text{batch}}\left[e^{s(z_i,z_j)/\tau}+e^{s(z_i,z_j')/\tau}\right]}$$

$$\mathcal{L}^{+}_{\text{AnomCLR}} = -\log e^{\left[s(z_i,z_i')-s(z_i,z_i^*)\right]/\tau} = \frac{s(z_i,z_i^*)-s(z_i,z_i')}{\tau}$$

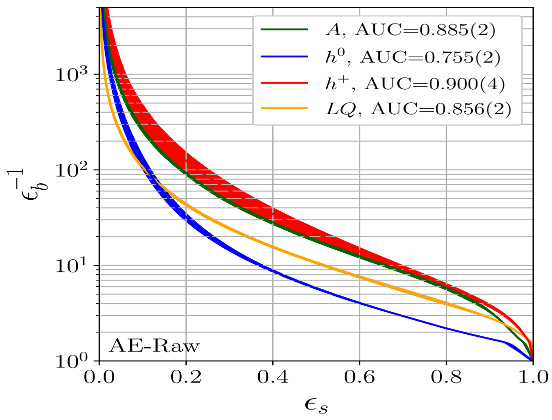


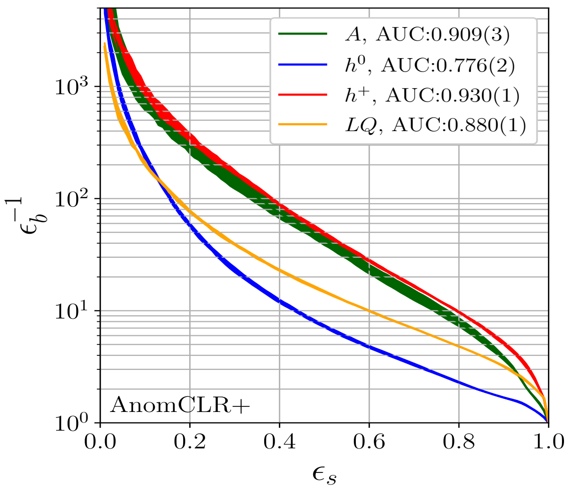
[anomalyCLR]

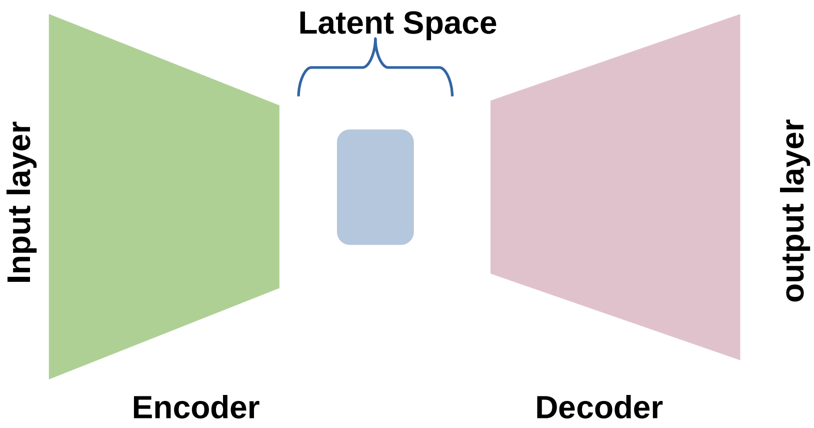
Anomaly augmentation

- Multiplicity
- Multiplicity shifts keeping total p_T and MET constant
- p_T and MET shifts

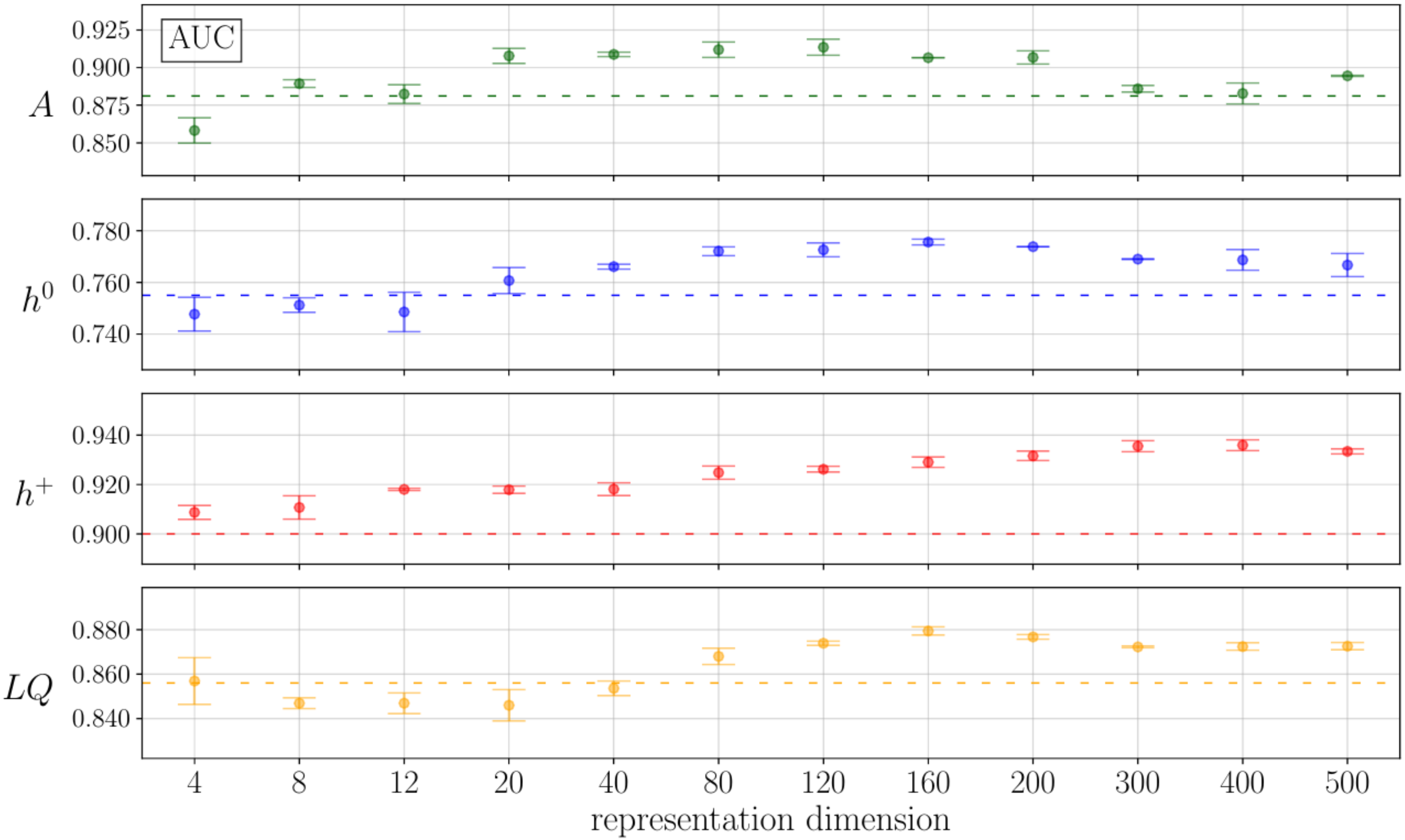
Anomaly score: anomalyCLR







(CMS anomaly detection data challenge, Govorkova et. al. 2107.02157)


AUC
A
0.925
0.900
0.875
0.850
h^0
0.780
0.760
0.740
h^+
0.940
0.920
0.900
LQ
0.880
0.860
0.840
4
8
12
20
40
80
120
160
200
300
400
500
representation dimension

DarkCLR

(Preliminary: B. M. Dillon, L. Favaro, TM, T. Plehn, J. Rüschkamp)

CLR represenation for semi-visible jets

Dataset

- Hidden-valley models
- 2 TeV heavy *Z'* resonance
- Dark quarks: q_d (500 MeV) charged under $SU(3)_d$
- Hadronizes to dark pion and ρ meson
- The fraction of constituents escaping detector $r_{\text{inv}} = 0.75$

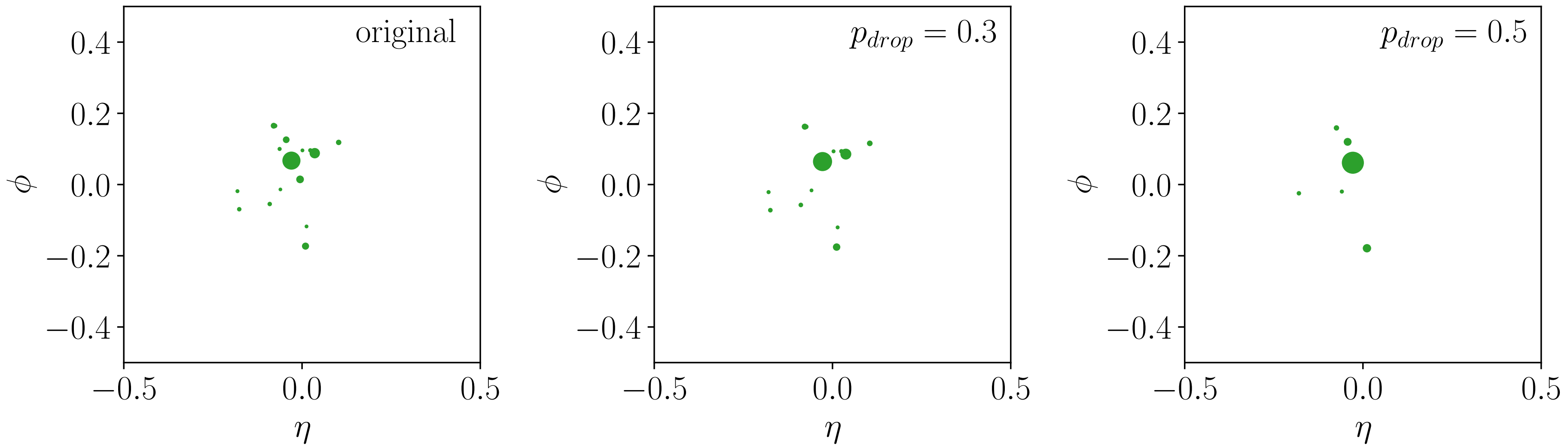
(Buss et. al. arXiv:2202.00686)

Anomaly augmentations

Invariance

- Rotation
- Translation
- permutation of the constituents

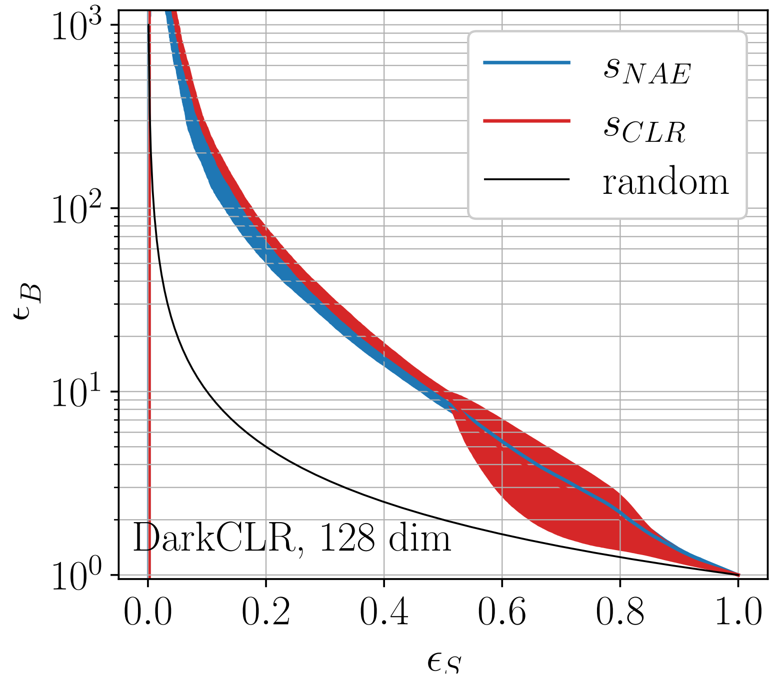
Dropping probability



Anomaly score

$$s_{CLR} = ||z||_{L_2}, \qquad z \in \mathbb{R}^{\mathbb{D}}$$

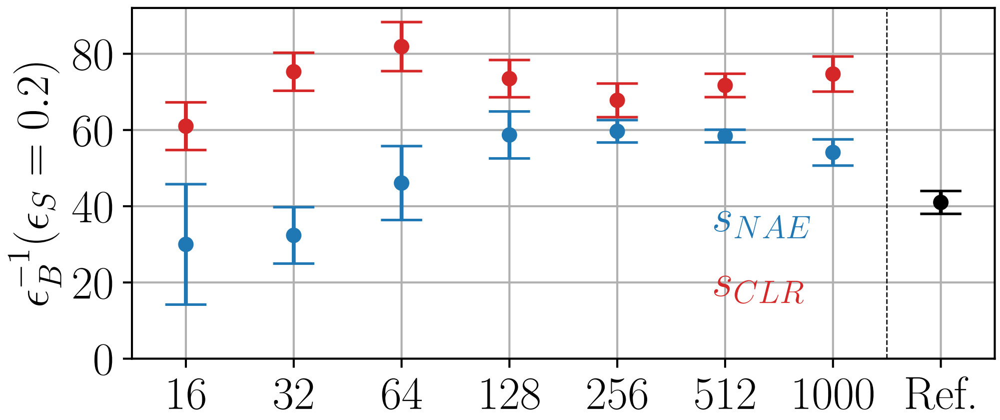
s_{NAE}: Reconstruction error of NAE

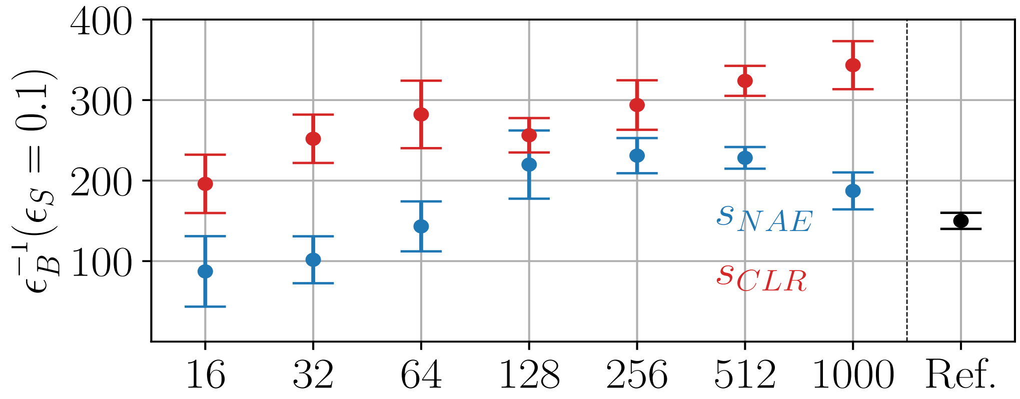


	DVAE	INN	NAE Jet images	DarkCLR
AUC	0.71	0.73	**0.76(1)**	**0.76 (1)**
$\epsilon_B^{-1}(\epsilon_S = 0.2)$	36	39	41 (1)	**73 (5)**

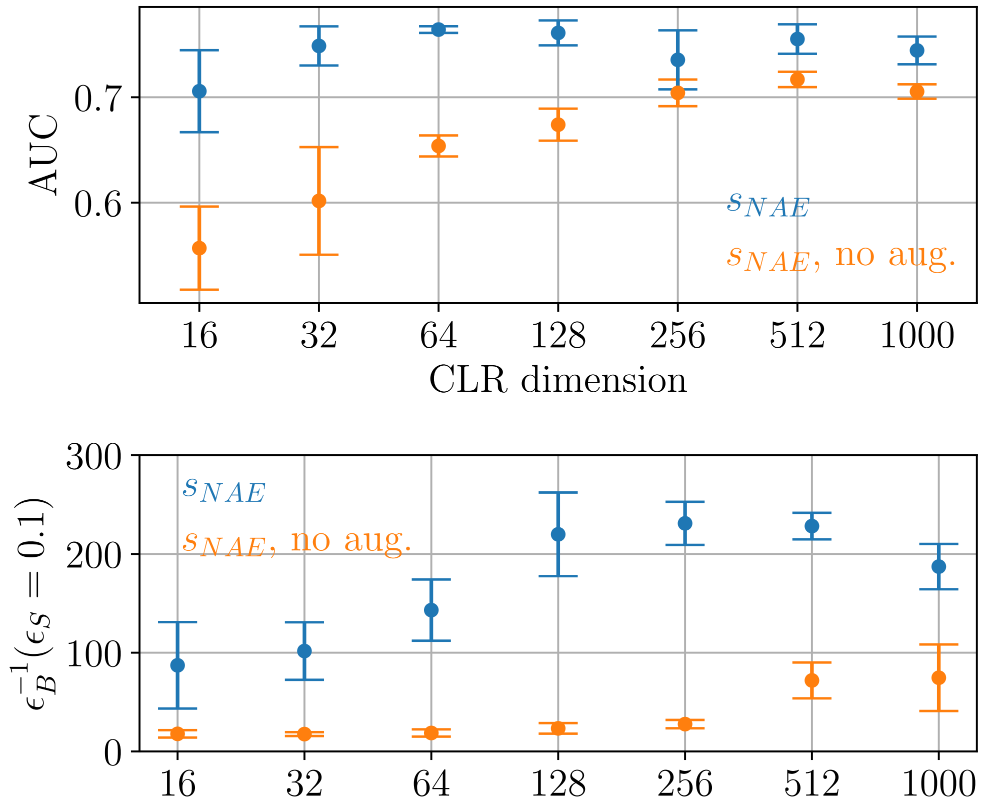
(DVAE, INN Buss et. al. arXiv:2202.00686; NAE, B. M. Dillon et. al. arXiv: 2206.14225)

Model dependence

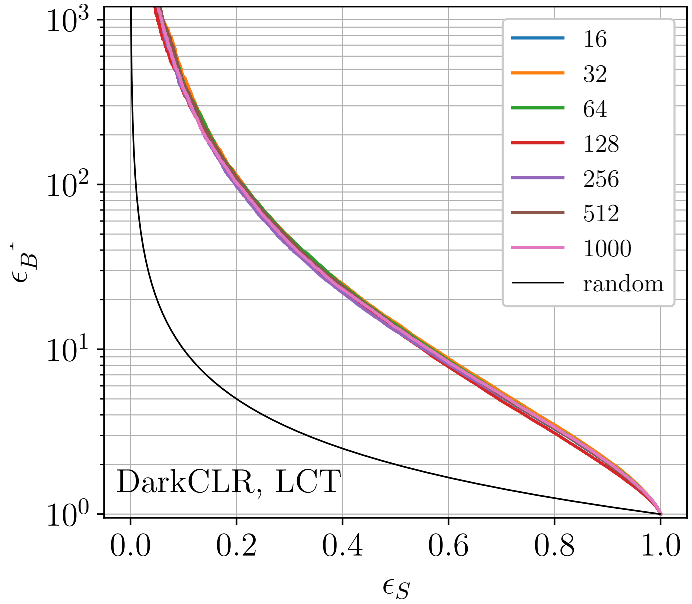


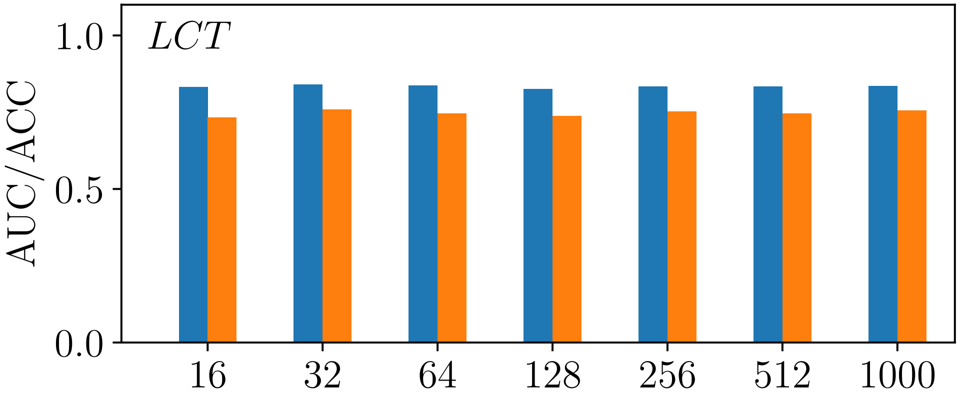


Model dependence



- simple linear classifier test
- Take representations before the head
- Same ROC regardless of the embedding
- AUC of **0.83 (0.78,** raw data**)** from the LCT test





Summary and Outlook

- Self-supervision: Offers unique way to identify anomalous objects in data
- Model agnostic.
- AnomalyCLR: Anomaly detection for events
- DarkCLR: CLR for semi-visible jet detection

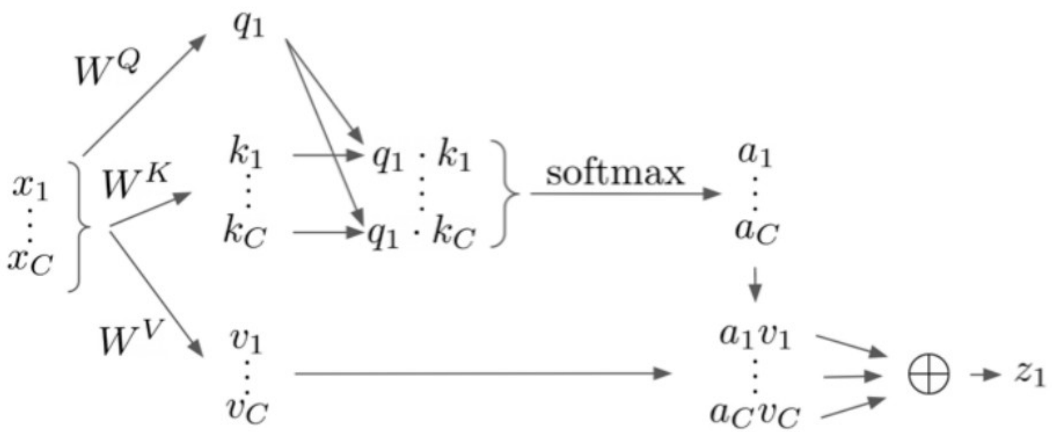
  1) Apply preprocessing via invariances to transformations

  2) Downstream task: Anomaly detection

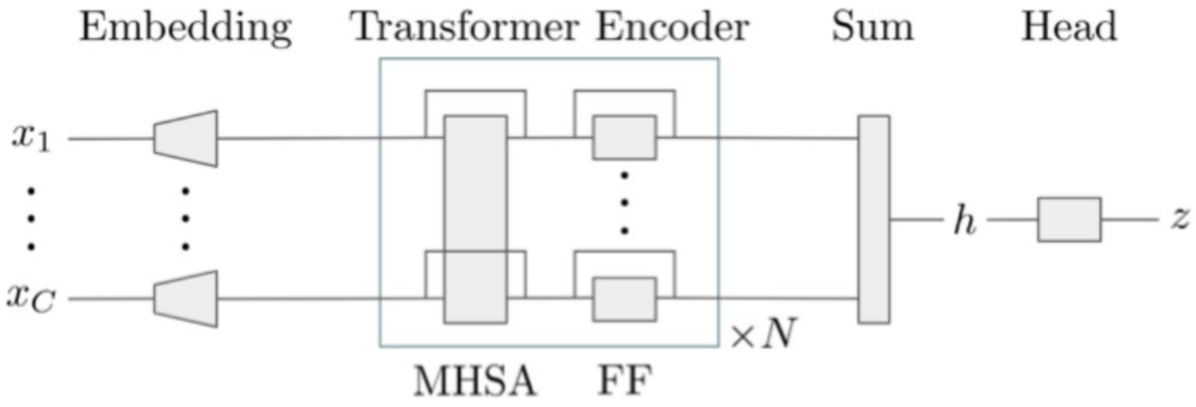
Additional Slides

Transformer Encoder

Self-Attention:



Network:



Contrastive loss function

$$\mathcal{L}_{\mathrm{CLR}} = -\log \frac{e^{s(z_i, z_i')/\tau}}{\sum_{j \neq i \in \mathrm{batch}} \left[e^{s(z_i, z_j)/\tau} + e^{s(z_i, z_j')/\tau} \right]}$$

x_i :Data point e.g. a event

x_i' : Augmented version of the augmented data point

$\{(x_i, x_i')\}$: Positive pairs

$\{(x_i, x_j)\} \cup \{(x_i, x_j')\}$: Negative pairs

$$z_i = f(x_i) \text{ and } z_i' = f(x_i')$$

$$s(z_i, z_j) = \frac{z_i \cdot z_j}{|z_i||z_j|} = \cos\theta_{ij}$$

The representation

- Transformer: projects each object to a larger vector of the embedding dimension
- embeddings passed through the transformer, with a feed-forward network.
- Output transformer: ($n\times$ model dimension). n is number of objects in an event
- Output: sum over the n vectors, enforces the permutation invariance
- The output of this head network is what is passed to the loss function
- For AD: representation output of the transformer network

(For details of transformer: B. M. Dillon, G. Kasieczka, H. Olischlager, T. Plehn, P. Sorrenson and L. Vogel, SciPost Phys. 12(6), 188 (2022))

CLR for anomaly detection

- Contrastive learning: Allows the function f to encode the nontrivial features of background data since it is optimized on background data.
- This means representation learnt by f only focuses on background features
- Anomalous data is not out-of-distribution
- In this form CLR will not achieve competitive performance in anomaly detection

NAE Loss

$$p_\theta(x) = \frac{e^{-E_\theta(x)}}{Z_\theta} \qquad \text{with} \qquad Z_\theta = \int_x dx\, e^{-E_\theta(x)} \,,$$

$$\mathcal{L}(x) = -\log p_\theta(x) = E_\theta(x) + \log Z_\theta \qquad \Rightarrow \qquad \mathcal{L} = \left\langle E_\theta(x) + \log Z_\theta \right\rangle_{x\sim p_{\text{data}}}$$

Single headed self-attention mechanism

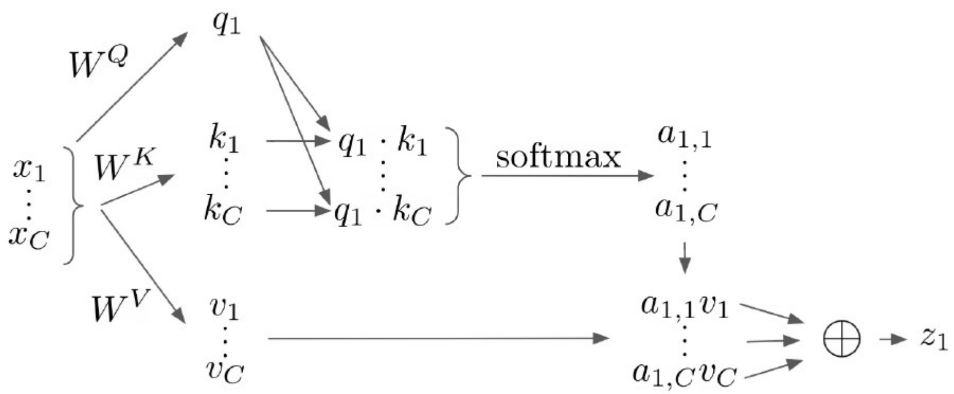
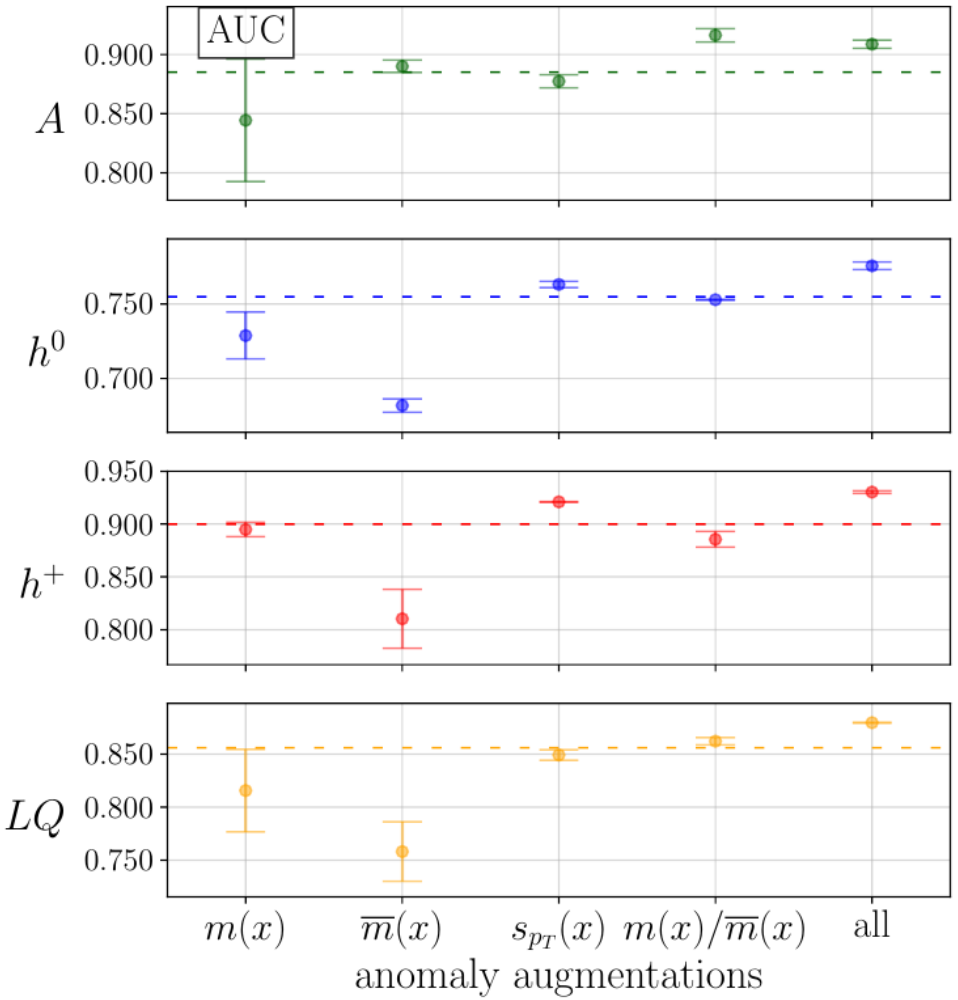


Figure 2: Illustration of scaled dot-product single-headed self-attention. All elements are defined in the text.

$$z_1 = \sum_i \operatorname{softmax}\left(\frac{q_1 \cdot k_i}{\sqrt{d}}\right) v_i = \sum_i \operatorname{softmax}\left(\frac{(W^Q x_1) \cdot (W^K x_i)}{\sqrt{d}}\right) W^V x_i$$


AUC
A
h^0
h^+
LQ
0.900
0.850
0.800
0.750
0.700
0.950
$m(x)$
$\overline{m}(x)$
$s_{p_T}(x)$
$m(x)/\overline{m}(x)$
all
anomaly augmentations

DarkCLR

Hyper-parameter	Value
Model (embedding) dimension	128
Feed-forward hidden dimension	512
Output dimension	512
# self-attention heads	4
# transformer layers (N)	4
# head architecture layers	2
Dropout rate	0.1
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	5×10^{-5}
Batch size	256
# constituents (α)	50
# jets	100k
# epochs	150

More on anomalyCLR

CLR (Ting Chen, Simon Kornblith, M. Norouzi, G. Hinton,arXiv:2002.05709)

hyper-parameter	
model (embedding) dimension	160
feed-forward hidden dimension	160
output dimension	160
# self-attention heads	4
# transformer layers (N)	4
# layers	2
dropout rate	0.1

hyper-parameter	
optimiser	Adam(β_1=0.9, β_2=0.999)
learning rate	5×10^{-5}
batch size	128
# epochs	500

$$-\log e^{\left[s(z_i,z_i')-s(z_i,z_i^*)\right]/\tau}, \;\; s(z_i,z_j) = \frac{z_i \cdot z_j}{|z_i||z_j|} = \cos\theta_{ij}$$

AE trained with anomalyCLR

Encoder: $e : \mathbb{R}^D \to \mathbb{R}^B$
Decoder: $d : \mathbb{R}^B \to \mathbb{R}^D$
AutoEncoder: $h = e \circ d : \mathbb{R}^D \to \mathbb{R}^D$

Minimise the mean-squared-error (MSE) $\mathcal{L}(\vec{x}, \theta) = (\vec{x} - \vec{x}')^2$, where $\vec{x}$ is input data, $\vec{x'}$ is output data and θ is learnable parameters of the AE.

- Encoder with five feedforward layers with dimension: 256, 128, 64, 32, 16
- Bottleneck: 5
- Decoder: 16, 32, 64, 128, 256
- Batch size: 4096
- Epoch: 100
- Learning rate: 0.001
- Adam optimizer