Classification-based Anomalous Jet Tagging

Taoli Cheng

Joint work with Aaron Courville

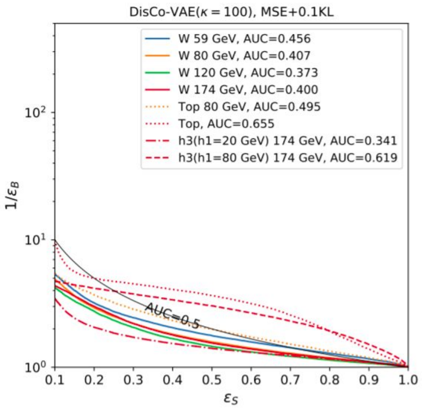
Jul. 06, 2021   ML4Jets

Mila

Université de Montréal

Motivation

- Generative models are not robust for Out-of-Distribution (OoD) detection in practice (slide)
- Supervised jet classifiers learn useful representations which could be generalized



[T. Cheng, et al. arXiv: 2007.01850]

Tasks	Base AUC	Transfer AUC
W/QCD → Top/QCD	0.926	**0.891**
g/q → Top/QCD	0.926	0.791
Top/QCD → W/QCD	0.957	**0.911**
q/g → W/QCD	0.957	0.822
W/QCD → q/g	0.861	0.763
Top/QCD → q/g	0.861	0.759

Table 2: Transferability results shown here. In *Base AUC*, the original trained AUC for the target task is shown, while in resulting in *Transfer AUC*, transferred embedding is used for training the classifier.

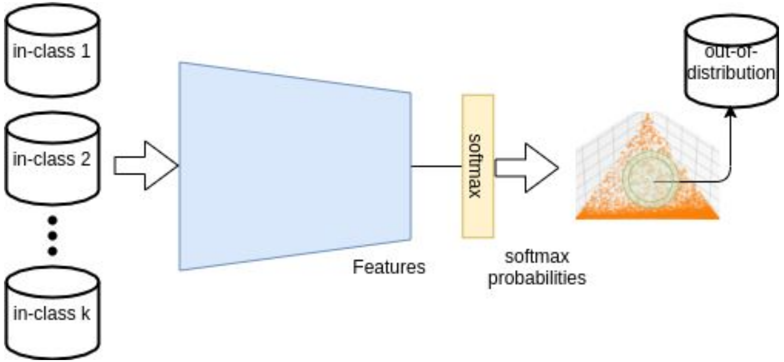
[T. Cheng, arXiv: 1911.01872]

- Opportunity to leverage sophisticated physics-inspired architectures: not only a jet classifier, but also a representation learning machine

Representation Learning + Interpretability + Anomaly Detection

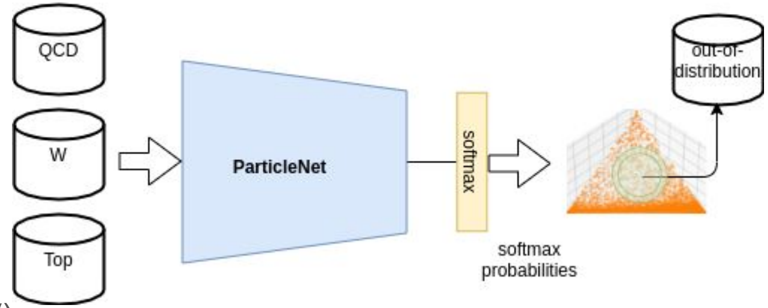
Classification based Anomaly Detection (CLF-AD) -- General Approach

- Basic assumption:
  - A well trained jet classifier will not be able to correctly classify out-of-distribution jets and thus give low confidence score
- Classifier architecture
- Anomaly Scores
  - Maximum Softmax probabilities / Confidence
  - Predictive Entropy / Uncertainty
  - ...
- Training procedure
  - Auxiliary tasks (outlier exposure)
- Predictive uncertainty (Ensembles)



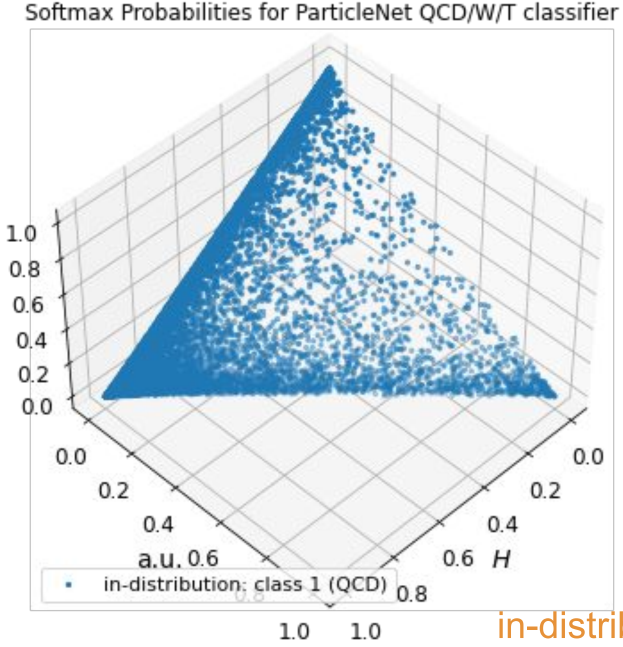
Workflow for Anomalous Jet Tagging

- In-distribution training data
  - Simulated large-cone QCD/W/Top jets with pT~600 GeV
  - Low-level jet constituent 4-vectors (or variants)
- Model (a decent baseline: ParticleNet) *[Huilin Qu, Loukas Gouskos. arXiv: 1902.08570]*
- Training
  - One-vs-All binary classification
  - All-vs-All multiclass classification
- Post-processing: anomaly score
- Out-of-distribution test sets
  - OoD class 1: H (174 GeV) → hh (h→ jj) with h (20 GeV)
  - OoD class 2: H (174 GeV) → hh (h→ jj) with h (80 GeV)
  - OoD class 3: “Top” (174 GeV) with W (20 GeV)

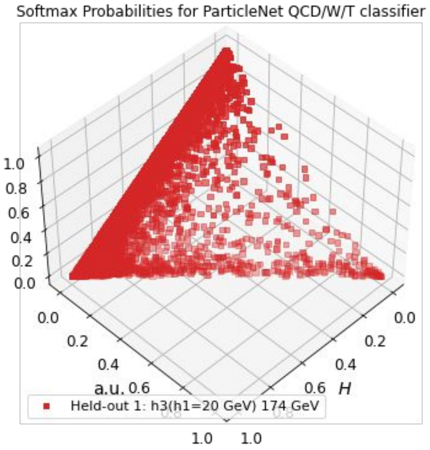


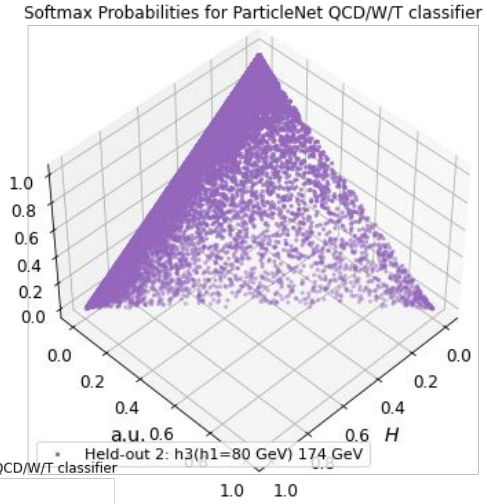
Softmax Probability Simplex

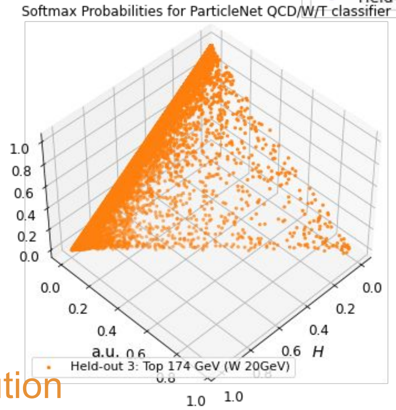
- Test set type affects the simplex distributions $\{p_i(x)\}$



in-distribution







out-of-distribution

B. Lakshminarayanan, et al. arXiv: 1612.01474
Gianni Franchi, et al. arXiv: 2006.00954

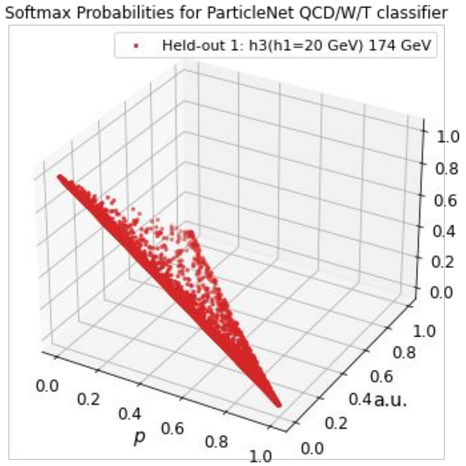
Improving Uncertainty Estimate

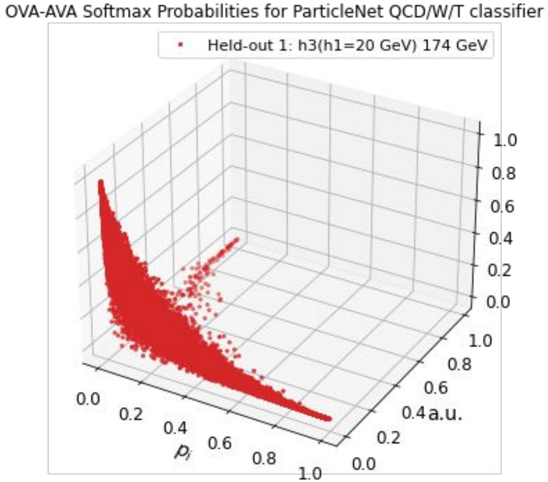
- Deep ensembles: training M models and averaging over the predictions (alternative uncertainty estimation approach w.r.t. Bayesian Neural Networks)
- One-vs-All (OvA) classification combined with All-vs-All (AvA) classification: brings sharper decision boundary

$$p_i^{\mathrm{OVA-AVA}}(x) = p^{\mathrm{i-OVA}}(x) \times p_i^{\mathrm{AVA}}(x)$$

One-vs-All (OvA) combined with All-vs-All (AvA)

- Combining OvA and AvA softmax probabilities
  - AvA classification pulls OoD samples to the center
  - OvA classification pulls OoD samples away from the closed-world simplex





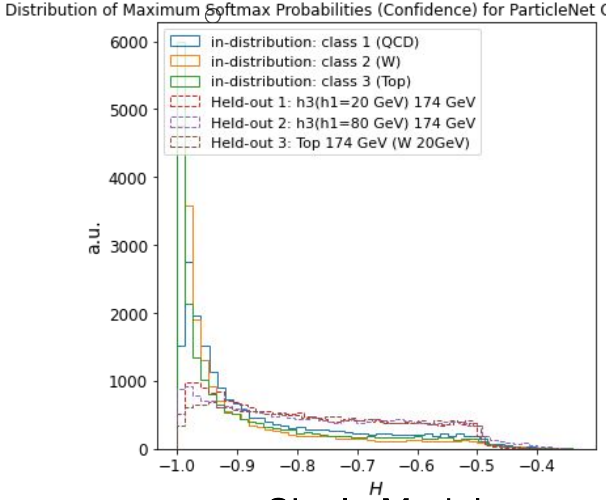
Anomaly Scores

Softmax probabilities vs Logits vs Representation Layer (Final Features)

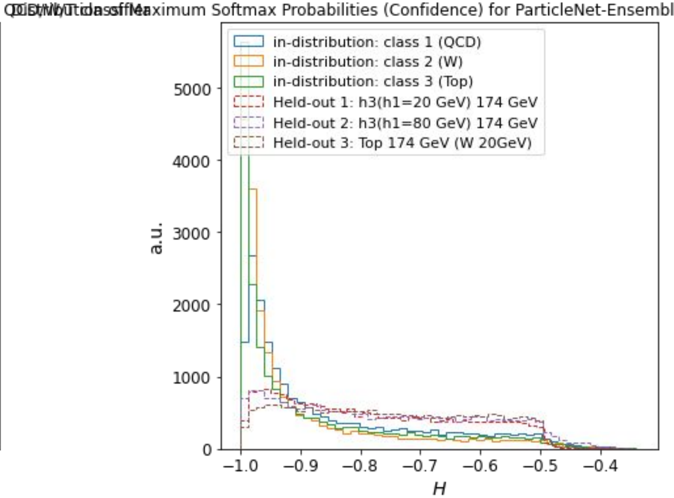
- Softmax probabilities based scores
  - Maximum Softmax Probability: $-\max\{p_1, p_2, p_3\}$
  - Softmax Probabilistic Entropy: $-\sum_{i=1}^{k} p_i log(p_i)$
- Logits based scores
- Representation based scores
  - Distance in feature space
  - Distance-based logits: Replacing logits with feature distance for softmax

Confidence Distribution -- Maximum Softmax Probability

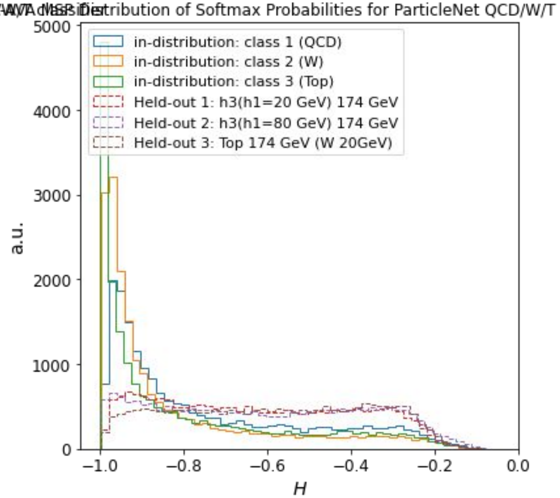
- Taking the (negative) maximum softmax probability $-\max\{p_1, p_2, p_3\}$ as the OoD score
  - In-distribution samples -> close to - 1.0
  - Out-of-distribution samples -> extreme case - 0.33 (for classical softmax outputs)



Single Model



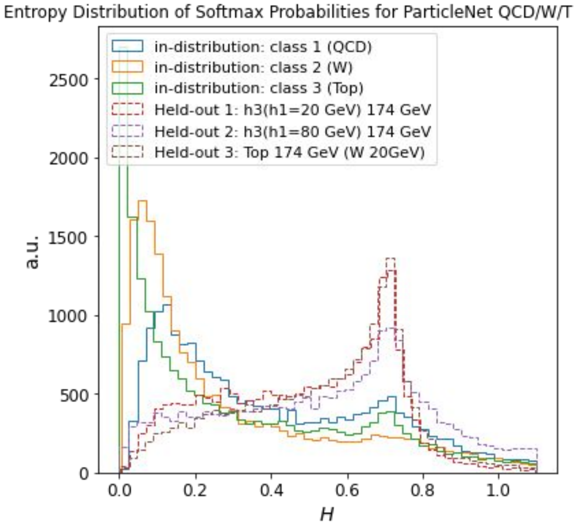
Ensemble-10



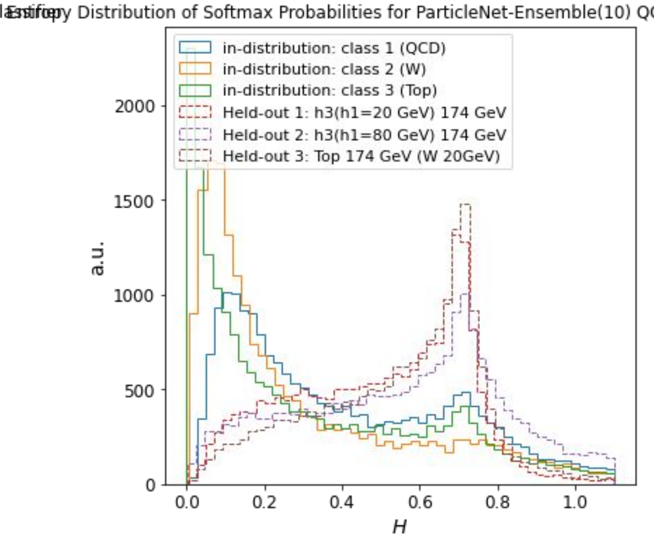
OVA-AVA

Softmax Probabilistic Entropy Distribution

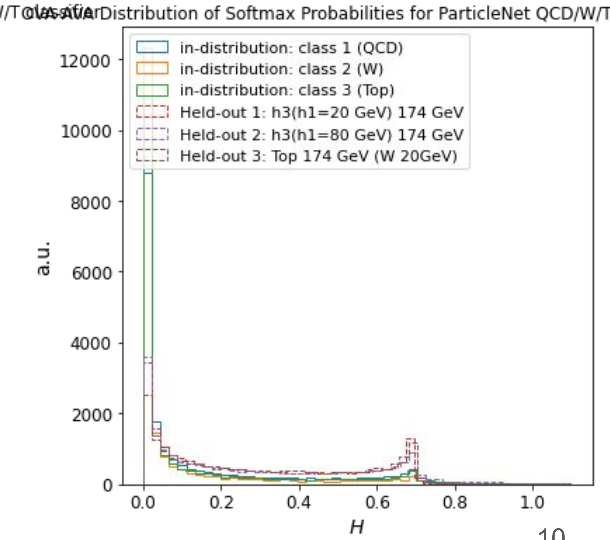
- Taking softmax entropy of (p1, p2, p3) $-\sum_{i=1}^{k} p_i log(p_i)$ as the OoD score
    - In-distribution samples -> close to 0
    - Out-of-distribution samples -> peaks at ~0.7 (entropy of ~(0.5, 0.5))



Single Model



Ensemble-10



OVA-AVA

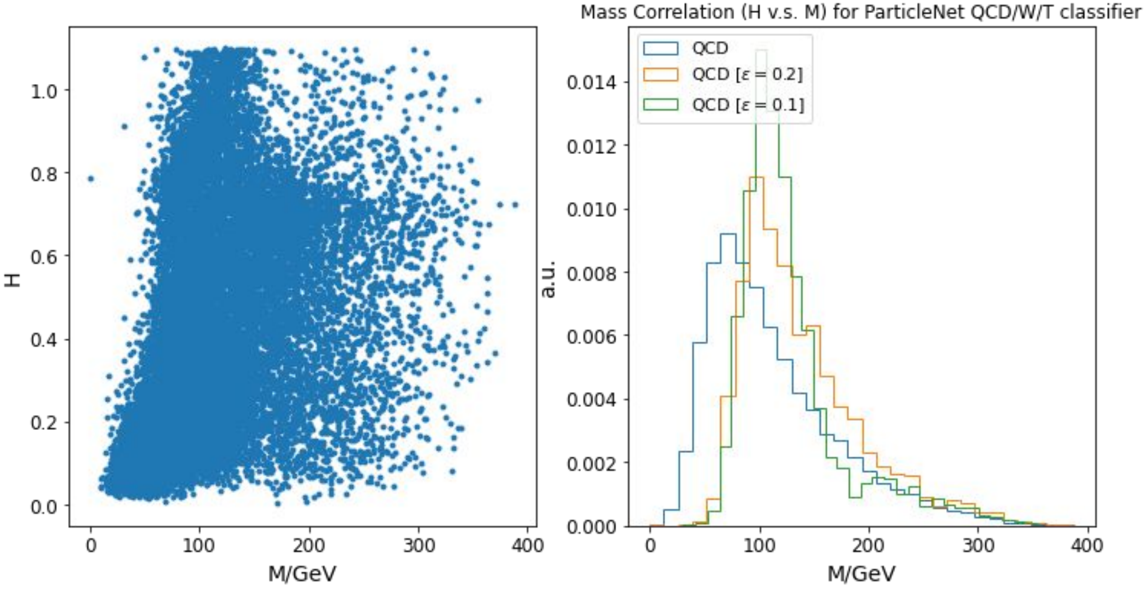
Scenario Comparison -- Area Under ROC Curve

- Discriminating QCD (in-class 1) and OoD classes
- Better uncertainty estimate → better OoD detection
- OVA-AVA with anomaly score $-\sum_{i=1}^{k} p_i log(p_i)$ performs best in all the OoD test sets.

Model-Score / AUC	OoD class 1 $H^{h=20GeV}_{174GeV}$	OoD class 2 $H^{h=80GeV}_{174GeV}$	OoD class 3 $Top^{W=20GeV}_{174GeV}$
SingleModel-Entropy	0.624	0.654	0.664
SingleModel-MSP	0.654	0.666	0.692
Emsemble10-Entropy	0.633	0.665	0.675
Emsemble10-MSP	0.665	0.677	0.706
OVA-AVA-Entropy	**0.677**	**0.681**	**0.708**
OVA-AVA-MSP	0.668	0.670	0.705
SingleModel-EnergyScore	0.552	0.675	0.599
OE-VAE (previous works)	0.736	0.624	0.721

Mass Correlation -- CLF-AD

- Not strongly mass-correlated compared with generative models
- Picking average mass of in-distribution classes



Results -- Discussion

- Classifier architectures (MLP, ParticleNet, etc.)
  - Better classification performance → better OoD detection
- Anomaly scores
- Increased uncertainty estimate helps with OoD detection
- Different mass correlation
  - Depends on in-distribution classes
  - Carefully choosing in-distribution classes helps in this case

Discriminative vs Generative

- Representation-driven approach
- Extra freedom of in-distribution classes
- Mass correlation depends on in-distribution classes
- Sensitive to jet types

- Likelihood-driven approach
- Sensitive to dominant correlations (in cases without further learning guidance)
- Strong mass correlation
- Possibility of assigning high likelihood to OoD samples (observed in both computer vision and jet physics)

Summary

- We introduce an alternative supervised discriminative approach for anomalous jet tagging
- QCD/W/Top as in-distribution classes; tested on held-out jet types
  - Better classification accuracy → Better OoD detection
  - Better uncertainty estimation → Better OoD detection
- Combining One-vs-All and All-vs-All classification to improve OoD detection
- Focuses on reporting softmax-probability-based anomaly scores → other options
- Only reporting on limited test OoD types → to further expand the test spectrum

Thanks!

Backup

Anomaly Detection can Fail

Slide from T. Cheng, ML4Jets 2020

back

- Outliers can be assigned higher probability sometimes, this happens in a general scope of anomaly detection using generative models
- Quick example: MSE based anomaly metric has intrinsic mass dependence → naive VAE assigns higher probability to lower mass jets

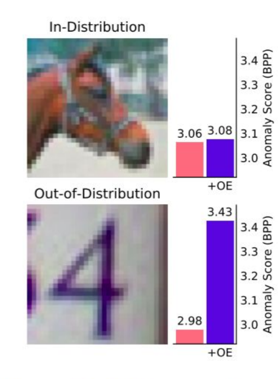
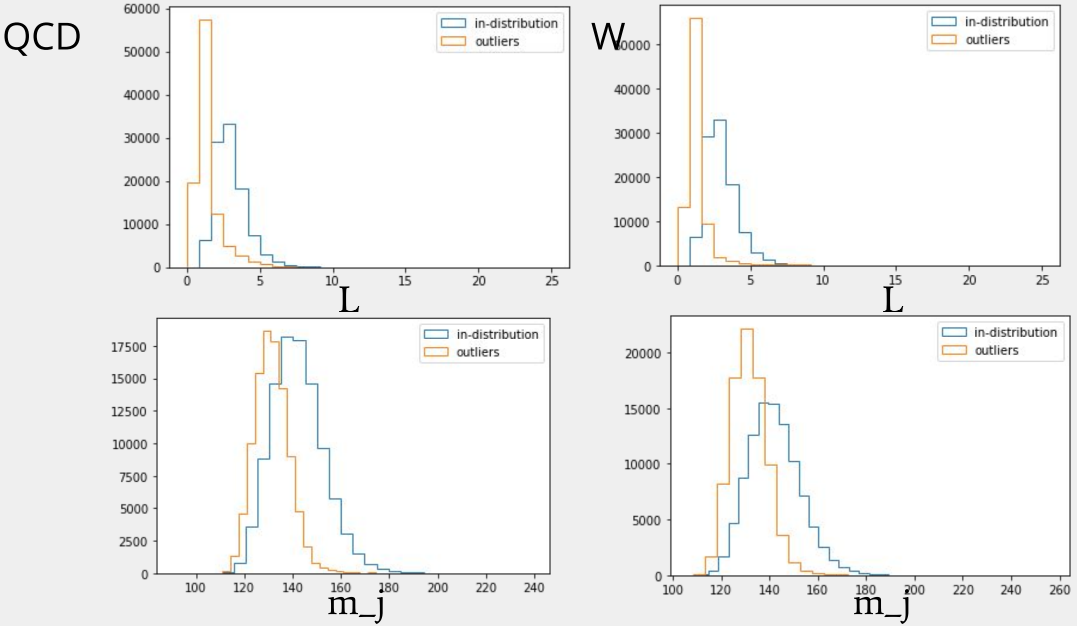


Figure 2: OOD scores from PixelCNN++ on images from CIFAR-10 and SVHN.

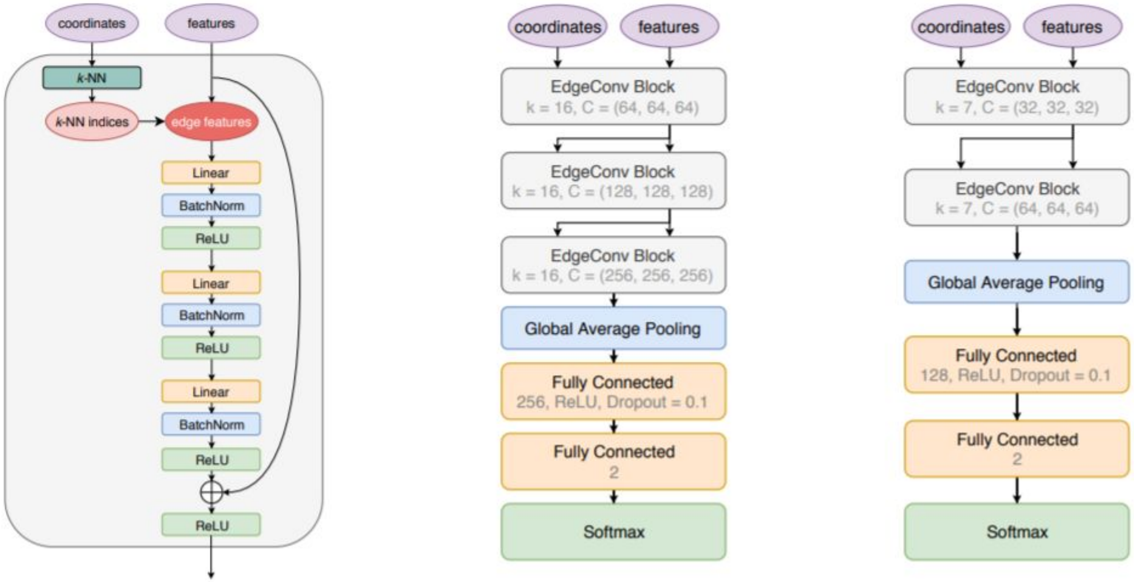


- D. Hendrycks, M. Mazeika, T. Dietterich.
  Deep Anomaly Detection with Outlier Exposure.
  arXiv: 1812.04606

Slide from Huilin Qu, ML4Jets 2020

THE PARTICLENET ARCHITECTURE

- Based on EdgeConv and DGCNN, we developed PARTICLENET, a customized architecture for jet tagging on particle clouds



EdgeConv block *ParticleNet architecture* *ParticleNet-Lite*

[arXiv:1902.08570](https://arxiv.org/abs/1902.08570)

Mass Correlation -- VAE