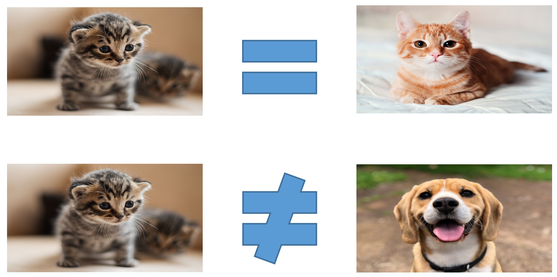
Self-Supervised Learning (SSL) for Jet Tagging

Zihan Zhao[1], Javier Duarte[1], Raghav Kansal[1], Farouk Mokhtar[1], Carlos Pareja[1]

[1]UC San Diego

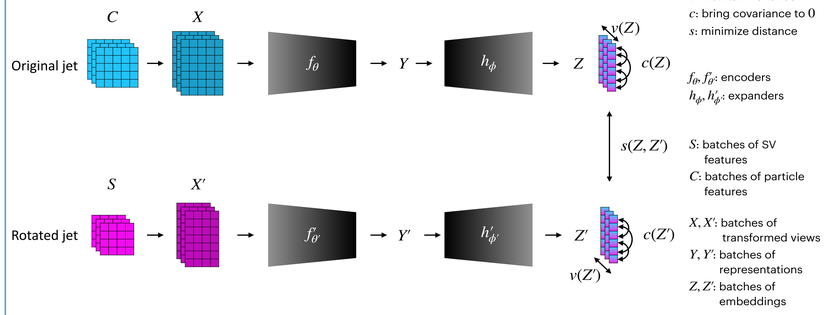
Jets and Self-Supervised Learning

- At the CERN_LHC, **jets** are showers of particles initiated by quarks and gluons.
- As opposed to supervised learning, which is limited by the availability of labeled data, **self-supervised** approaches can learn from vast unlabeled data.[1]
- One of the most powerful self-supervised learning approaches is **contrastive learning**: to learn the general features of a dataset without labels by teaching the model which data points are similar or different.



- To generate similar data points, we apply **augmentations** to jets. [2]
  - General: Translation, Rotations, cropping, etc.
  - Physics informed: soft splitting, colinear splitting, etc.

VICReg [3]



- The contrastive loss function has three terms:

$$\ell(Z, Z') = \underbrace{\lambda s(Z, Z')}_{\text{invariance}} + \mu[\underbrace{v(Z) + v(Z')}_{\text{variance}}] + \nu[\underbrace{c(Z) + c(Z')}_{\text{covariance}}]$$

- Invariance: the mean squared distance between embedding vectors, learns invariance to augmentations
- Variance: a hinge loss to maintain the standard deviation (over a batch) of each variable of the embedding above a given threshold. This term forces the embedding vectors of samples within a batch to be different.
- Covariance: a term that attracts the covariances (over a batch) between every pair of (centered) embedding variables towards zero. This term decorrelates the variables of each embedding and prevents an informational collapse in which the variables would vary together or be highly correlated.

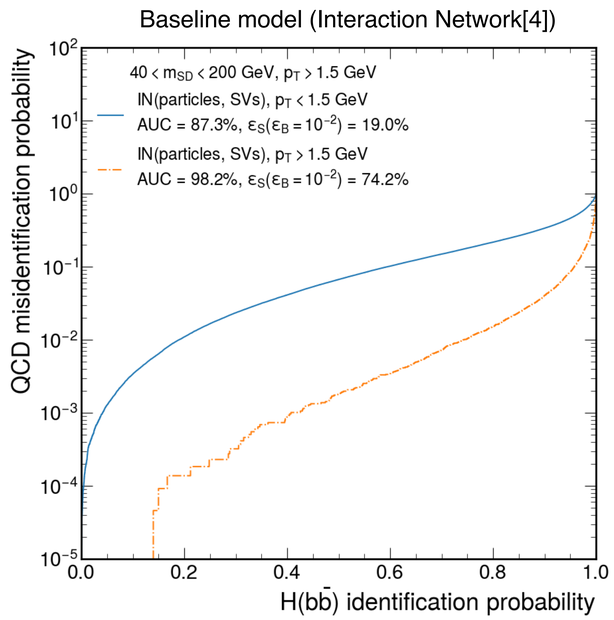
Datasets

- We plan on using three datasets:

Name	Size	Type	Features
Hbb vs QCD[4]	10 million jets	Real CMS simulation (stand in for “data”)	Tracks & SVs
JetClass[5]	100 million jets	Delphes simulation, less realistic (stand in for “simulation”)	Only particles
MultiJet primary dataset from Run of 2012[6]	10 million jets	Real CMS Open data	

Setup

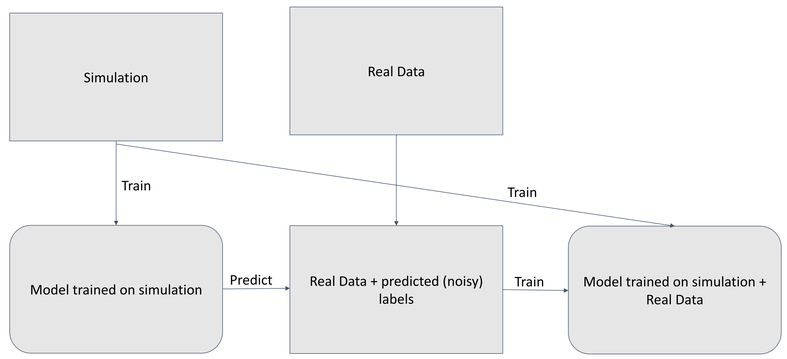
- To see whether SSL helps improve the performance of jet tagging, we devised two tests:
- **Domain adaptation test**: Does self-supervised pretaining in unlabeled data + finetuning in labeled simulation help the model adapt to the data domain?
  - Note: domain shift between unlabeled and labeled samples (e.g., differences in pileup distribution, etc.)



- **Sample size test**: Does self-supervised pretraining in (much larger) unlabeled data + finetuning in (much smaller) labeled simulation improve overall performance?
  - Note: no domain shift between unlabeled and labeled samples

Weakly Supervised Learning (WSL)

- The goal of WSL is to facilitate training with noisy labels.
- Although real data does not have labels, we can generate (noisy) labels for real data with ML models, and use these labels in a weakly supervised way to facilitate training



Summary

Limited by the lack of truth labels on real data, fully supervised ML algorithms are constrained to training only with simulated samples. With self-supervised learning, we can leverage vast amounts of unlabeled real data to facilitate training. We investigate the application of VICReg, a contrastive learning model, on a classification task: discriminating signal jets (e.g. H→bb̄ jets) from background jets (e.g. QCD jets). We also explore the use of jet augmentations in contrastive learning.

References

[1] Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., … Goldblum, M. (2023). A Cookbook of Self-Supervised Learning. Retrieved from https://arxiv.org/abs/2304.12210v1

[2] Dillon, B. M., Kasieczka, G., Olischläger, H., Plehn, T., Sorrenson, P., & Vogel, L. (2021). Symmetries, Safety, and Self-Supervision. *SciPost Physics*, *12*(6). https://doi.org/10.21468/SciPostPhys.12.6.188

[3] Bardes, A., Ponce, J., & LeCun, Y. (2021). VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. https://doi.org/10.48550/arxiv.2105.04906

[4] Moreno, E. A., Nguyen, T. Q., Vlimant, J. R., Cerri, O., Newman, H. B., Periwal, A., … Pierini, M. (2020). Interaction networks for the identification of boosted H →b b ¯ decays interaction networks for the identification of ... Moreno Eric A. *Physical Review D*, *102*(1). https://doi.org/10.1103/PhysRevD.102.012010

[5] Qu, H., Li, C., & Qian, S. (2022). Particle Transformer for Jet Tagging. https://doi.org/10.48550/arxiv.2202.03772

[6] CMS collaboration (2022). MultiJet primary dataset in AOD format from Run of 2012 (/MultiJet/Run2012A-22Jan2013-v1/AOD). CERN Open Data Portal. DOI:10.7483/OPENDATA.CMS.WING.7QKV

Contact

- Zihan Zhao: ziz078@ucsd.edu, z.zhao@cern.ch
- Javier Duarte: jduarte@physics.ucsd.edu

Acknowledgement

- This project is supported by the National Science Foundation under Cooperative Agreement" OAC-2117997