JetNet: A Python package for accessing open datasets and benchmarking machine learning methods in high energy physics

**Raghav Kansal** [1,2]¶**, Carlos Pareja** [1]**, Zichun Hao** [3]**, and Javier Duarte** [1]

1 UC San Diego, USA **2** Fermilab, USA **3** California Institute of Technology, USA ¶ Corresponding author

Summary

JetNet is a Python package that aims to increase accessibility and reproducibility for machine learning (ML) research in high energy physics (HEP), primarily related to particle jets. Based on the popular PyTorch ML framework, it provides easy-to-access and standardized interfaces for multiple heterogeneous HEP datasets and implementations of evaluation metrics, loss functions, and more general utilities relevant to HEP.

Statement of need

It is essential in scientific research to maintain standardized benchmark datasets following the findable, accessible, interoperable, and reproducible (FAIR) data principles (Chen & others, 2022), practices for using the data, and methods for evaluating and comparing different algorithms. This can often be difficult in high energy physics (HEP) because of the broad set of formats in which data is released and the expert knowledge required to parse the relevant information. The JetNet Python package aims to facilitate this by providing a standard interface and format for HEP datasets, integrated with PyTorch (Paszke et al., 2019), to improve accessibility for both HEP experts and new or interdisciplinary researchers looking to do ML. Furthermore, by providing standard formats and implementations for evaluation metrics, results are more easily reproducible, and models are more easily assessed and benchmarked. JetNet is complementary to existing efforts for improving HEP dataset accessibility, notably the EnergyFlow library (P. T. Komiske et al., 2020), with a unique focus to ML applications and integration with PyTorch.

Content

JetNet currently provides easy-to-access and standardized interfaces for the JetNet (Kansal et al., 2022), top quark tagging (Butter & others, 2019; Kasieczka et al., 2019), and quark-gluon tagging (P. Komiske et al., 2019) reference datasets, all hosted on Zenodo (European Organization For Nuclear Research & OpenAIRE, 2013). It also provides standard implementations of generative evaluation metrics (Kansal et al., 2021, 2023), including Fréchet physics distance (FPD), kernel physics distance (KPD), 1-Wasserstein distance (W1), Fréchet ParticleNet distance (FPND), coverage, and minimum matching distance (MMD). Finally, JetNet implements custom loss functions like a differentiable version of the energy mover's distance (P. T. Komiske et al., 2019) and more general jet utilities.

Impact

The impact of JetNet is demonstrated by the surge in ML and HEP research facilitated by the package, including in the areas of generative adversarial networks (Kansal et al., 2021), transformers (Kach et al., 2022; Kach & Melzer-Pellmann, 2023; Kansal et al., 2023), diffusion models (Leigh et al., 2023; Mikuni et al., 2023), and equivariant networks (Buhmann et al., 2023; Hao et al., 2023), all accessing datasets, metrics, and more through JetNet.

Future Work

Future work will expand the package to additional dataset loaders, including detector-level data, and different machine learning backends such as JAX (Bradbury et al., 2018). Improvements to the performance, such as optional lazy loading of large datasets, are also planned, as well as community challenges to benchmark algorithms facilitated by JetNet.

Acknowledgements

References

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., & Zhang, Q. (2018). *JAX: Composable transformations of Python+NumPy programs* (Version 0.3.13). http://github.com/google/jax

Buhmann, E., Kasieczka, G., & Thaler, J. (2023). EPiC-GAN: Equivariant Point Cloud Generation for Particle Jets. *SciPost Phys.*, *15*, 130. https://doi.org/10.21468/SciPostPhys.15.4.130

Butter, A., & others. (2019). The Machine Learning landscape of top taggers. *SciPost Phys.*, *7*, 014. https://doi.org/10.21468/SciPostPhys.7.1.014

Chen, Y., & others. (2022). A FAIR and AI-ready Higgs boson decay dataset. *Sci. Data*, *9*, 31. https://doi.org/10.1038/s41597-021-01109-0

European Organization For Nuclear Research, & OpenAIRE. (2013). *Zenodo*. CERN. https://doi.org/10.25495/7GXK-RD71

Hao, Z., Kansal, R., Duarte, J., & Chernyavskaya, N. (2023). Lorentz group equivariant autoencoders. *Eur. Phys. J. C*, *83*(6), 485. https://doi.org/10.1140/epjc/s10052-023-11633-5

Kach, B., Krücker, D., & Melzer-Pellmann, I. (2022). *Point Cloud Generation using Transformer Encoders and Normalising Flows*. https://arxiv.org/abs/2211.13623

Kach, B., & Melzer-Pellmann, I. (2023). *Attention to Mean-Fields for Particle Cloud Generation*. https://arxiv.org/abs/2305.15254

Kansal, R., Duarte, J., Su, H., Orzari, B., Tomei, T., Pierini, M., Touranakou, M., Vlimant, J.-R., & Gunopulos, D. (2021). Particle cloud generation with message passing generative ad-

---

versarial networks. *Advances in Neural Information Processing Systems*, *34*. https://papers.neurips.cc/paper_files/paper/2021/file/c8512d142a2d849725f31a9a7a361ab9-Paper.pdf

Kansal, R., Duarte, J., Su, H., Orzari, B., Tomei, T., Pierini, M., Touranakou, M., Vlimant, J.-R., & Gunopulos, D. (2022). *JetNet* (Version 2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6975118

Kansal, R., Li, A., Duarte, J., Chernyavskaya, N., Pierini, M., Orzari, B., & Tomei, T. (2023). Evaluating generative models in high energy physics. *Phys. Rev. D*, *107*(7), 076017. https://doi.org/10.1103/PhysRevD.107.076017

Kasieczka, G., Plehn, T., Thompson, J., & Russel, M. (2019). *Top quark tagging reference dataset* (v0 (2018_03_27)) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.2603256

Komiske, P. T., Mastandrea, R., Metodiev, E. M., Naik, P., & Thaler, J. (2020). Exploring the Space of Jets with CMS Open Data. *Phys. Rev. D*, *101*(3), 034009. https://doi.org/10.1103/PhysRevD.101.034009

Komiske, P. T., Metodiev, E. M., & Thaler, J. (2019). Metric Space of Collider Events. *Phys. Rev. Lett.*, *123*(4), 041801. https://doi.org/10.1103/PhysRevLett.123.041801

Komiske, P., Metodiev, E., & Thaler, J. (2019). *Pythia8 quark and gluon jets for energy flow* (Version v1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3164691

Leigh, M., Sengupta, D., Quétant, G., Raine, J. A., Zoch, K., & Golling, T. (2023). *PC-JeDi: Diffusion for Particle Cloud Generation in High Energy Physics*. https://arxiv.org/abs/2303.05376

Mikuni, V., Nachman, B., & Pettee, M. (2023). Fast point cloud generation with diffusion models in high energy physics. *Phys. Rev. D*, *108*(3), 036025. https://doi.org/10.1103/PhysRevD.108.036025

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, p. 8024). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf