The Journal of Open Source Software

giotto-deep: A Python Package for Topological Deep Learning

Matteo Caorsi [1¶], **Raphael Reinauer** [*2], **and Nicolas Berkouk** [2]

1 L2F SA, Rue du Centre 9, Saint-Sulpice, 1025, CH **2** Ecole Polytechnique Fédérale de Lausanne (EPFL), Laboratory for topology and neuroscience, Lausanne, 1015, CH ¶ Corresponding author

DOI: 10.21105/joss.04846

Software

- Review
- Repository
- Archive

Editor: Øystein Sørensen

Reviewers:

- @EduPH
- @leotrs
- @ismailguzel

Submitted: 04 October 2022
Published: 24 October 2022

Summary

Topological data analysis (TDA) has already provided many novel insights into machine learning (Carrière et al., 2020) due to its capabilities of synthesizing the shape information into a multiset of points in two dimensions: the persistence diagrams (Wasserman, 2016). Furthermore, many researchers in the field hope to give new insights into deep-learning models by applying TDA techniques to study the models' weights, the activation values in the different layers, and their evolution during the training phase (Naitzat et al., 2020). Orthogonally, TDA techniques have been used as feature engineering tools to extract novel information from the data, which are then used as standard features in a machine learning pipeline, with significant success in many fields (Hensel et al., 2021).

Statement of need

`giotto-deep` is a deep-learning Python package that seamlessly integrates topological data analysis models and data structures. Indeed, persistence diagrams (the core descriptors of topological data analysis) are intrinsically sets, and therefore require specific methods to be manipulated as tensors, and analyzed by neural networks. The library is founded on a PyTorch core due to the extensive use of the library in the machine learning community. The `giotto-deep` package was designed with usability in mind and provides a class-based interface to fast implementations of standard machine learning tasks, such as data preprocessing, model building, model training and validation, reporting and logging, as well as image classification, Q&A, translation, persistence diagram vectorization (via Persformer (Reinauer et al., 2021)). Additionally, giotto-deep supports more advanced tasks such as (distributed and multi-pod) hyperparameter optimization through its seamless integration with optuna (Akiba et al., 2019).

`giotto-deep` has been designed to be used by mathematics researchers and by machine learning engineers. The combination of speed, versatility, design, and support for TDA data structures in `giotto-deep` will enable exciting scientific explorations of the behavior of deep learning models, hopefully shedding new light on the generalisability and robustness of such complex and powerful models. In summary, `giotto-deep` is a powerful, easy-to-use tool that will help to incorporate topological data into machine learning models with little effort.

[*] Co-first author

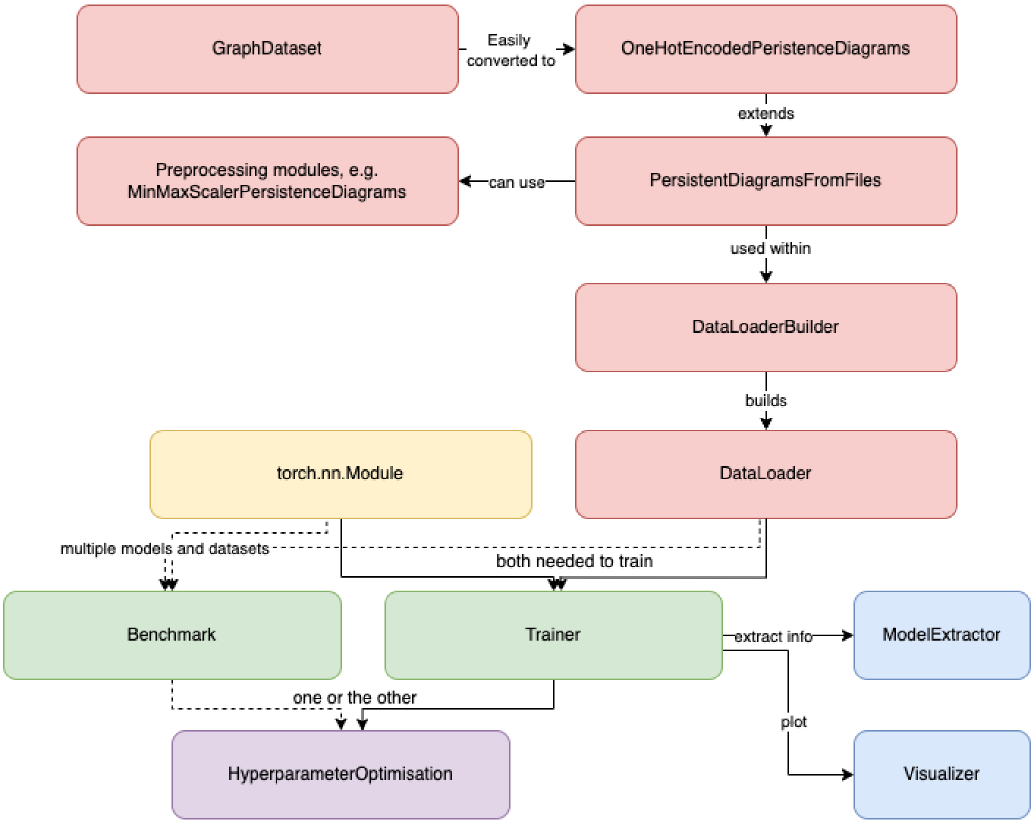

Figure 1: Simplified architecture diagram.

The `giotto-deep` architecture is schematized in figure Figure 1.

The hyperparameter searches (HPO) can also be distributed on a `kubernetes` cluster using python RQ, speeding up the computation: this is an essential aspect when dealing with large models and large hyperparameter searches. Many topological computations in `giotto-deep` are performed by `giotto-tda` (Tauzin et al., 2021).

`giotto-deep` handles the whole pipeline: from data preprocessing up to the hyperparameter search, the `k-fold` cross-validation, and the deployment of the models. We provide various preprocessing and training pipelines already implemented, but we invite users to extend and improve them. The eventual goal is to create a readable code base that is easy to learn and simple to implement so that the contribution of new features would be naturally encouraged. Additionally, we provide classical TDA datasets in a dedicated dataset cloud; any user can access and download the dataset from the cloud fully automated from the `giotto-deep`.

Research projects using `giotto-deep`

The current most relevant scientific application of this software is the Persformer: a novel algorithm to automatize the persistence diagrams vectorization (Reinauer et al., 2021). We also showcased in Jupyter notebook how to apply the library to get the results of (Perez & Reinauer, 2022) to extract topological information from the attention scores of the BERT-Transformer model.

Acknowledgments

The authors would like to acknowledge the financial support of the Swiss federation: Innosuisse project 41665.1 IP-ICT.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.

Carrière, M., Chazal, F., Ike, Y., Lacombe, T., Royer, M., & Umeda, Y. (2020). Perslay: A neural network layer for persistence diagrams and new graph topological signatures. *International Conference on Artificial Intelligence and Statistics*, 2786–2796.

Hensel, F., Moor, M., & Rieck, B. (2021). A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, *4*, 681108.

Naitzat, G., Zhitnikov, A., & Lim, L.-H. (2020). Topology of deep neural networks. *J. Mach. Learn. Res.*, *21*(184), 1–40.

Perez, I., & Reinauer, R. (2022). The topological BERT: Transforming attention into topology for natural language processing. *arXiv Preprint arXiv:2206.15195*. https://doi.org/10.48550/arXiv.2206.15195

Reinauer, R., Caorsi, M., & Berkouk, N. (2021). Persformer: A transformer architecture for topological machine learning. *arXiv Preprint arXiv:2112.15210*.

Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Medina-Mardones, A. M., Dassatti, A., & Hess, K. (2021). Giotto-tda:: A topological data analysis toolkit for machine learning and data exploration. *J. Mach. Learn. Res.*, *22*(39), 1–6.

Wasserman, L. (2016). Topological data analysis. *arXiv Preprint arXiv:1609.08227*. https://doi.org/10.48550/arXiv.1609.08227