PROCEEDINGS OF SCIENCE

A pipeline to test Graph Neural Network algorithms for flavour tagging

Greta Brianti[a,b,c,*]

[a] *University of Trento,*
Via Sommarive 14, 38123 Povo, Trento (TN), Italy
[b] *Trento Institute of Fundamental Physics and Applications INFN-TIFPA,*
Via Sommarive, 14, 38123 Povo, Trento (TN), Italy
[c] *Fondazione Bruno Kessler,*
Via Sommarive 18, 38123 Povo, Trento (TN), Italy

E-mail: greta.brianti@unitn.it

The flavour tagging, i.e. the identification of jets originating from heavy flavour quarks, is an essential task for the Standard Model and Beyond the Standard Model research at colliders. Machine Learning-based algorithms have been playing a key role since long time in this task. Graph Neural Networks (GNNs) are a type of machine learning tool where input datasets are represented and processed as graphs. In the context of flavour tagging, GNNs can be particularly useful as they can represent and exploit the internal structure of jets for the identification of the original parton by utilizing the tracks associated with the jets. In this article, we present AUTOGRAPH (Automatic Unified Training and Optimization for Graph Recognition and Analysis with Pipeline Handling), a fully automated and totally customizable pipeline based on GNNs dedicated to flavour tagging.

The European Physical Society Conference on High Energy Physics (EPS-HEP2023)
21-25 August 2023
Hamburg, Germany

[*] Speaker

1. Introduction

Since the middle of Run 2 [1], due to the increase in the expected integrated luminosity delivered by the Large Hadron Collider (LHC) [2], a growing emphasis has been put on algorithmic improvement of Machine Learning techniques for online tagging of heavy flavours [3]. Nowadays, the CMS [4] and ATLAS [5] collaborations, LHC general-purpose experiments, utilize Graph Neural Networks (GNNs) for flavour tagging. Moreover, GNN algorithms are applied to offline analysis for background-signal classification tasks [6]. For these reasons, a pipeline that allows easy access to this state-of-the-art technology could be advantageous for many analyses.

2. The AUTOGRAPH pipeline

The pipeline architecture is divided into two main components - the user interface and the automated steps, illustrated in Figure 1. The interface comprises a single configuration file that enables the user to access and manage the jet-graph structure, network architecture, and training hyperparameters. The automated steps are the underlying structure of the pipeline, consisting of sub-programs written in Python language managed by the user interface. The automated steps are executed by launching a single Python script.

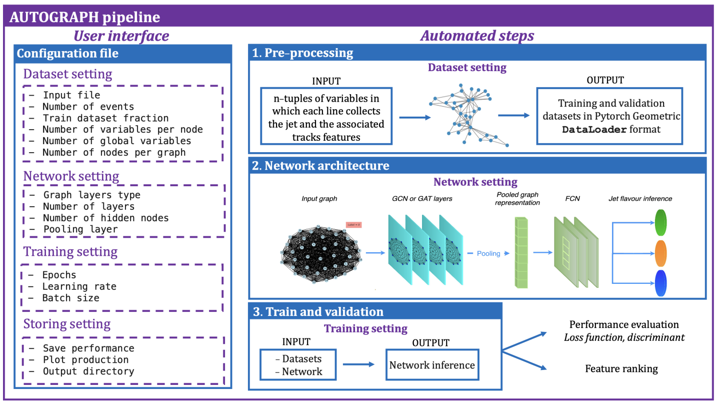

Figure 1: Pictorial pipeline representation. The *user interface* section represents the configuration file wherein the user can choose the setting for the *automated steps*. The latters are divided in three main processes illustrated schematically.

2.1 Dataset handling

To feed the GNN, the array-structured data collected in High Energy Physics collisions are recast to graph description. The graph is a jet representation where the tracks associated with the jet constitute the fully connected neurons. Through the configuration file, the user can customize the graph architecture, selecting the number of tracks per jet, the features associated with each track, and the global-jet features. Finally, the resulting graph list is converted into Pytorch Geometric DataLoader format [7]. To train supervised machine-learning networks, it is necessary to access the truth information level included exclusively in the Monte Carlo simulation (MC). For this reason, the pipeline is conceived to work on a simulated dataset.

2.2 Network architecture

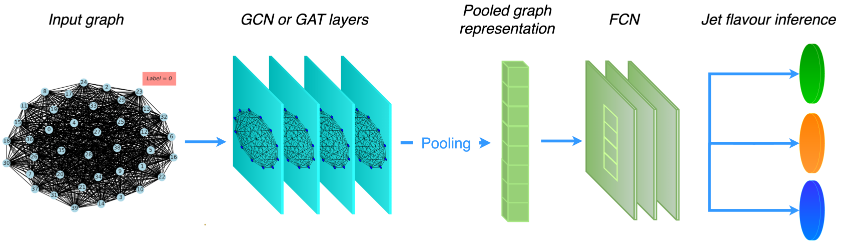

Figure 2: Representation of the network architecture. The input graph, labeled by the MC truth-level information, is classified from the Fully Connected Network in three classes corresponding to l-, c- or b-jet.

The default structure of the network, shown in Figure 2, consists of a series of Graph Layers followed by a pooling function. A three-classes Fully Connected Network (FCN) executes the final classification. The FCN outputs p_b, p_c and p_l are the network output probabilities representing the probability that the input jet is generated, respectively, from a bottom quark, a charm quark, or a light quark. The user is given a choice between two Message-passing-based layers: the Graph Convolutional Layers [8] and the Graph Convolutional Attention Layers [9]. The configuration file allows the selection of the graph layer number, the hidden nodes per layer, the pooling function and the FCN architecture. In conclusion, the network architecture is fully customizable and can include the attention mechanism [10].

2.3 Training and performance evaluation

After the dataset preparation and the network architecture setting, the user can train the selected model with a fully customizable set of hyperparameters. Among the latter are the number of epochs for which the model is trained, the batch size, the learning rate and the optimization algorithm.

3. Application of the pipeline on simulated datasets

Two simulated datasets were used to evaluate the performance of the pipeline. The first dataset was a next-to-leading order $t\bar{t}$ simulation, while the second dataset was a leading order $Z'H$ simulation. In the second dataset, H refers to the Higgs boson and Z' is a dark matter mediator candidate that is restricted to decay to hadrons. Three Monte Carlo simulation frameworks interfaced with

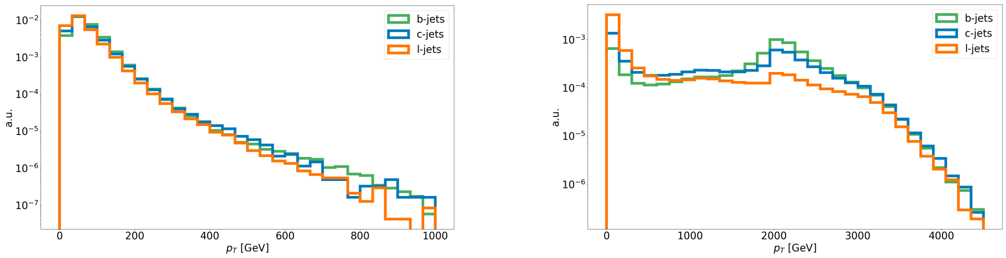

Figure 3: Jet transverse momentum distribution for the $t\bar{t}$ at next-to-leading order dataset (left) and the $Z'H$ at leading order dataset with $m_{Z'} = 2$ TeV (right). The distributions are divided for jet truth flavours.

each other have been exploited to obtain the datasets. Firstly, MadGraph_aMC@NLO [11] was used to generate parton-level hard processes, followed by Pythia 8.3 [12] which provides the parton showering and hadronization. Finally, Delphes 3.5.0 [13] covers the detector response simulation. The $Z'H$ dataset with $m_{Z'} = 2$ TeV shows an extended transverse momentum range of the jets, as can be seen in Figure 3. The tracks have been associated to the jet with the $\Delta R = \sqrt{(\eta_{jet} - \eta_{trk})^2 + (\phi_{jet} - \phi_{trk})^2}$ criteria: for jet $p_T \leq 150$ GeV is required $\Delta R \leq 0.45$, while for jet $p_T > 150$ GeV $\Delta R \leq 0.26$ [14]. Furthermore, the pipeline can provide the distribution

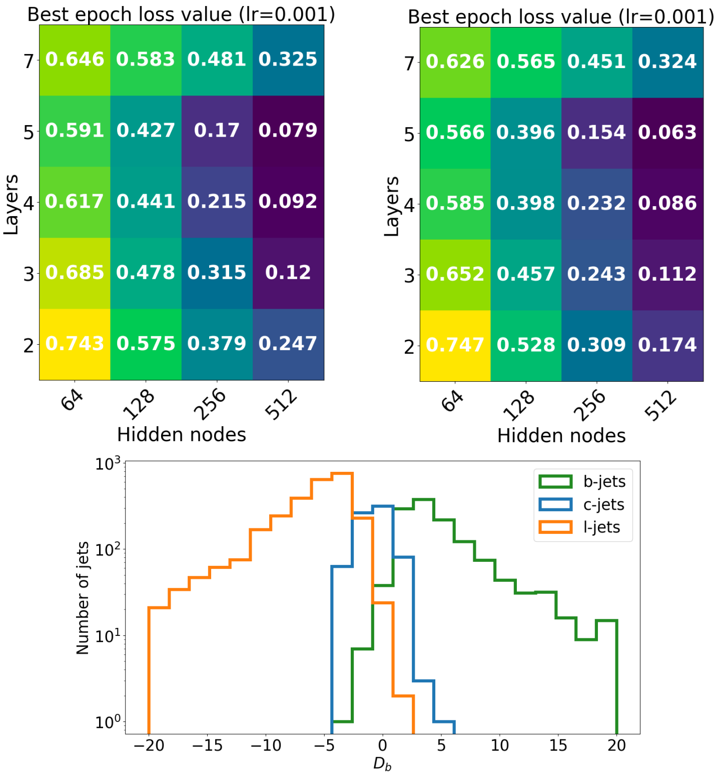

Figure 4: Results from the grid search. Starting from the left: (a) Heat map with best epoch loss value for the grid search on $t\bar{t}$ at the next-to-leading order dataset, (b) Heat map with best epoch loss value for the grid search on $Z'H$ at the leading order dataset, (c) Discriminant distribution for the best architecture with the $Z'H$ at leading order dataset divided per flavour.

of the discriminant D_b, defined as $D_b = log\left(\frac{p_b}{p_c f_c + (1-f_c) p_l}\right)$, where p_b, p_c and p_l are the network output probabilities, introduced in the subsection 2.2, and f_c is the fraction of c-jets in the dataset. This variable represents the network's capability to distinguish between jet flavours. The better the separation of flavour contributions, the better the network's performance. The discriminant distribution for the architecture of 5 graph layers and 512 nodes is shown in Figure 4 (c).

4. Conclusion

The AUTOGRAPH pipeline is designed to provide straightforward access to state-of-the-art flavour tagging algorithms based on GNNs. This paper presents the pipeline along with an application case on two simulated datasets with different kinematic characteristics.

References

[1] J. T. Boyd. LHC Run-2 and Future Prospects. 2020. arXiv: 2001.04370 [hep-ex].

[2] O. Brüning, H. Burkhardt, S. Myers, The large hadron collider, Progress in Particle and Nuclear Physics, Volume 67, Issue 3, 2012, Pages 705-734, ISSN 0146-6410, https://doi.org/10.1016/j.ppnp.2012.03.001.(https://www.sciencedirect.com/science/article/pii/S0146641012000

[3] Mitrevski J 2015 Journal of Physics: Conference Series 664 072034

[4] CMS Collaboration et al. "The CMS experiment at the CERN LHC". In: (2008).

[5] The ATLAS Collaboration and G Aad. "The ATLAS Experiment at the CERN Large Hadron Collider". In: Journal of Instrumentation 3.08 (Aug. 2008), S08003–S08003. DOI: 10.1088 /1748-0221/3/08/s08003. URL: https://doi.org/10.1088/1748-0221/3/08/s08003.

[6] DeZoort, G., Battaglia, P.W., Biscarat, C. et al. Graph neural networks at the Large Hadron Collider. Nat Rev Phys 5, 281–303 (2023). https://doi.org/10.1038/s42254-023-00569-0

[7] "PyTorch Geometric: A Library for Geometric Deep Learning on Graphs and Manifolds, https://pytorch-geometric.readthedocs.io/en/latest/

[8] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. 2017. arXiv: 1609.02907 [cs.LG].

[9] Petar Veličković et al. Graph Attention Networks. 2018. arXiv: 1710.10903 [stat.ML].

[10] Ashish Vaswani et al. Attention Is All You Need. 2023. arXiv: 1706.03762 [cs.CL].

[11] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: Journal of High Energy Physics 2014.7 (July 2014). doi: 10.1007/jhep07(2014)079. url: https://doi.org/10.1007%2Fjhep07%282014%29079

[12] Christian Bierlich et al. A comprehensive guide to the physics and usage of PYTHIA 8.3.2022. arXiv: 2203.11601 [hep-ph].

[13] J. de Favereau et al. "DELPHES 3: a modular framework for fast simulation of a generic collider experiment". In: Journal of High Energy Physics 2014.2 (Feb. 2014). doi: 10.1007/jhep02(2014)057. url: https://doi.org/10.1007%2Fjhep02%282014%29057

[14] Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run. Tech. rep. All figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYS-PUB-2017-013. Geneva: CERN, 2017. url: https://cds.cern.ch/record/2273281