University of Oxford

Computer Science Department

Project Dissertation

Machine Learning in High Energy Physics
Quark and Gluon Jet Discrimination

Author:
Maxence DRAGUET

Supervisors:
Dr. Anisoara CALINESCU
Prof. Alan BARR

Trinity Term
2019-2020

Abstract

This is a multidisciplinary project, implementing modern data analysis techniques and advanced machine learning in the field of particle physics. The objective is to derive a new algorithm to perform quark versus gluon jet discrimination with the ATLAS detector, one of the main experiment on the Large Hadron Collider at CERN. Different architectures targeting generated level simulations (not accounting for detector effects) have recently been explored in the literature with promising results. Among them, the paradigm of recurrent-learning has been proven to lead to state-of-the-art gluon versus quark jet discrimination. The approach learns a probabilistic model scaffolded on a leading-order model of physics (by the use of jet factorisation algorithm to produce trees) by recurrent learning and MLP mapping. The model thus learnt has the added advantage of being interpretable. The project aims to implement this method into an efficient discriminator for the ATLAS collaboration, using the particular conditions of its detector and appropriated simulations that are comparable to the data collected (the reconstructed level).

CHAPTER 1

Introduction

1 Introduction:

General introduction to the context: particle physics, quantum "revolution" with models of increasing reach. Current status: Standard Model incredibly precise though lacunar. Need to explore in-depth the particle world to find inconsistencies. This multidisciplinary project targets a tool to analyse data produced in large experiments. Improving jet tagging/discrimination to increase the precision of physics searches.

Structure of the text:

- Chapter II: the context of the problem. Describe very generally the building blocks of the project such as: the Standard Model components and "mechanisms", the experimental setup, and the interest for jets (why tagging them, why discriminating quark from gluon jets and why low-level inputs are interesting: trigger). Mention that the field is extremely dependent on state-of-the-art data analysis and computational techniques, making it a perfect context for applying advanced ML (a lot of high-quality data with a powerful and well-understood underlying model). This should motivate the "why" of the project. The second part of this chapter should explain how the task was carried out in the past and serve as a sort of “state of the art”. It should describe the several ways to see a jet: an image, a physical object with its properties (engineered variable) or a recurrent-pattern. Motivate why using a more advanced model like JUNIPR makes sense in this context (avoiding engineered features → lower-level hence closer to the detector level and realistic for online operation, information about the jet is not restricted to choosing variables but is in the structure of the model itself thanks to C/A tree,

- Chapter III: the methods explored to solve the problem in this project. It will follow the pipeline of the project:
  - from data collection (use of Monte Carlo simulations, the reconstructed level, collecting the information from ATLAS files, the ATLAS software to process DAOD files into ROOT files with the need for quality cuts and so on),
  - to data processing (the need to further cut the data (upRoot) for consistency and extract relevant information),
  - the modelling of a jet: BDT, NN, and introduce thoroughly the idea behind JUNIPR. Explain why the Graph NN and the CNN where not pursued in the end (first one by lack of time, second one because it is known to offer lower performance, as should have been shown in chapter II from the literature).
  - A final part described the set-up here (anti-kT algorithm, C/A algorithm, pplx-int access, submission on Condor, the need for multiprocessing when possible and possibly the ARC) as well as the tool used (FastJet through PyJet, UpRoot, Root, PyTorch, ...). Quick words on some of the computational challenge met (particularly those related to performance and how JUNIPR can be speeded up with multiloader batches, batch-oriented losses, efficient used of padding, ...).
- Chapter IV: this is the first part of the actual work carried out. It describes the challenges met in collecting the data (large volume, need to implement a complex algorithm that typically takes a few months for a student to get used to, the interest of the GRID submission and why in the end it was not necessary, mentioning the job failure due to ATLAS team updating the distributions). It should present the

distributions of our two signals (different variables), trying to expose some structural elements of importance. Particularly, the different energy distribution problem has to be addressed. The necessity to reweight the samples so that they match in distributions should clearly appear (it would make things far too easy for the BDT and does not correspond to the tool intended: a discriminator of jets based on their structure, not their energy distribution due to some process and slice bias). Finally, this section should explore the processing for JUNIPR jets (why we collect this information) and describe why the jet re-clustering (the second one) sometimes fails to return a single object (show some eta-phi maps and the histogram). A word on the 1 GeV constituent cut also has to be addressed (necessity to maintain clean nodes in the tree to not introduce noise in the probability distribution). The need to scale data appropriately for JUNIPR should also be described.

- Chapter V: introduce the benchmarking models: BDT/NN. A quick word on their implementations (nothing special to mention) and show the ROC curves that clearly indicate these models saturate in performance (modifying the NN leads to no important change in efficiency). I will try to produce something indicating the importance of the different features for the BDT and reconnect this with the literature (gluon jets tend to be wider, ...).

- Chapter VI: chapter in two steps. First on unary JUNIPR. Explain how they are trained and the architecture used (a quick word on the improve in performance when going from RNN cell to LSTM cell and the daughter branch being split for each variable). What does the model learn? Show the probability distribution obtained that proves there is a good agreement. Show some jet trees displaying in what way the model is interpretable. The second part is on binary JUNIPR. Explain the freedom on the binary objective (maximising the likelihood ratio test based on the truth label or the BCE with the truth label). Show the ROC curves (hopefully it beats the NN/BDT, It already performs as well but with one of the corrupted model!). Display some jets with the likelihood ratio test at nodes and discuss this from a "physical point of view" (it might be challenging to do this given I have not spent much time working on the physical aspect of the question but more on the algorithmic tools).

Should I show some learning curves? They bring little information.

- Conclusion: what was the objective vs what is the result. What can be improved upon this such as training JUNIPR on pT slices (increase precision) and implementing it for a trigger: when calorimeter gathers something that could be a jet, summing the pT collected, it calls the appropriate version of unary JUNIPR's during the clustering process in real-time so that, by the end of the clustering, a tag can be given. Also, discuss some technical aspects of the model: how could it be made more powerful (changing the structure, adding other sorts of information such as another MLP layer between the two models that would take the probability at each node to give a tag from finer information than the global jet probability). Finally, discuss how incredibly general JUNIPR is (it can: discriminate, generate, re-weight and analyse jets) and how this could be implemented in different tasks such as quark jet discrimination (my possible DPhil project), multi-jet analysis (a tree made of several jet trees and the probability of the event is that of the trees given a process) and even complex event topology (such as the VBF production mechanism: how could use the two outside jets and some elements of the internal structure to derive a tagging tool for these events).

CHAPTER 2

Jet-Tagging in High Energy Physics

CHAPTER 3

Methodology

CHAPTER 4

Collecting and Processing the Data

CHAPTER 5

Benchmarking Models

CHAPTER 6

JUNIPR implementation

CHAPTER 7

Conclusion

Bibliography

[1] The Standard Model (of Physics) at 50. https://blogs.scientificamerican.com/observations/the-standard-model-of-physics-at-50/. Accessed: 2020-04-18. (Not cited)

[2] Wikipedia: The Standard Model (of Physics). https://simple.wikipedia.org/wiki/Standard_Model. Accessed: 2020-04-18. (Not cited)

[3] The ATLAS Collaboration website. http://www.atlas.ch/. Accessed: 2020-04-18. (Not cited)

[4] The ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation 3*, S08003, 2008. (Not cited)

[5] D. Guest, K. Cranmer, and D. Whiteson. Deep Learning and its Application to LHC Physics. *Ann. Rev. Nucl. Part. Sci.*, 68:161–181, 2018. (Not cited)

[6] P. Baldi, P. Sadowski, and D. Whiteson. Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nature Commun.*, 5:4308, 2014. (Not cited)

[7] G. Kasieczka, Plehn N. Kiefer, Tilman, and J.M. Thompson. Quark-Gluon Tagging: Machine Learning vs Detector. *SciPost Phys.*, 2019. (Not cited)

[8] The ATLAS Collaboration. Quark versus Gluon Jet Tagging Using Charged-Particle Constituent Multiplicity with the ATLAS Detector. *ATL-PHYS-PUB-2017-009*, 2017. (Not cited)

[9] P. T. Komiske, E. M. Metodiev, and J. Thaler. An Operational Definition of Quark and Gluon Jets. *JHEP*, 11, 2018. (Not cited)

[10] J. Hernandez-Gonzalez, I. Inza, and J.A. Lozano. Weak Supervision and Other Non-Standard Classification Problems: a Taxonomy. *Patt. Recog. Lett.*, 69:49–55, 2016. (Not cited)

[11] L.M. Dery, B. Nachman, F. Rubbo, and A. Schwartzman. Weakly Supervised Classification in High Energy Physics. *JHEP 05*, 145, 2017. (Not cited)

[12] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz. Learning to Classify from Impure Samples with High-Dimensional Data. *Phys. Rev. D*, 98, 2018. (Not cited)

[13] E.M. Metodiev, B. Nachman, and J. Thaler. Classification without labels: Learning from mixed samples in high energy physics. *JHEP 10*, 174, 2017. (Not cited)

[14] The ATLAS Collaboration. Quark versus gluon jet tagging using jet images with the ATLAS detector. *ATL-PHYS-PUB-2017-017*, 2017. (Not cited)

[15] A. Andreassen, I. Feige, C. Frye, and M.D. Schwartz. JUNIPR: a Framework for Unsupervised Machine Learning in Particle Physics. *Eur. Phys. J.*, C79, 2019. (Not cited)

[16] A. Andreassen, C. Frye I. Feige, and M.D. Schwartz. Binary JUNIPR: an interpretable probabilistic model for discrimination. *Phys. Rev. Lett.*, 123, 2019. (Not cited)

[17] T. Cheng. Recursive Neural Networks in Quark/Gluon Tagging. *Computing and Software for Big Science 2.1*, 2018. (Not cited)

[18] H. Qu and L. Gouskos. ParticleNet: Jet Tagging via Particle Clouds. 2019. (Not cited)

[19] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P.S. Yu. A Comprehensive Survey on Graph Neural Networks. 2019. (Not cited)

[20] P.T. Komiske, E.M. Metodiev, and J. Thaler. Energy flow polynomials: A complete linear basis for jet substructure. 2017. (Not cited)

[21] The ATLAS Collaboration. Measurement of the charged-particle multiplicity inside jets from $\sqrt{s} = 8$ TeV pp collisions with the ATLAS detector. *The European Physical Journal C*, 76, 2016. (Not cited)

[22] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands. An Introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015. (Not cited)