IRIS-HEP Fellows Program Project Proposal: Improving the Analysis Grand Challenge (AGC) Machine Learning Workflow

Con Muangkod
University of Colorado Boulder

Mentors: Elliott Kauffman[1], Alexander Held[2], Oksana Shadura[3]
[1]Princeton University, [2]University of Wisconsin-Madison, [3]University of Nebraska-Lincoln

May 24, 2024

Introduction

The Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP) is a center for software research and development in an attempt to prepare for the High-Luminosity LHC (HL-LHC) data that will be collected starting in 2029 [1]. The Analysis Grand Challenge (AGC) serves as an integration exercise that implements typical analysis workflow with developing techniques. The goals of the AGC project are to create a pipeline that is practical to the scope and scale of the analysis task as well as to improve user experience with potential interactive analysis [2].

The main developing workflow for the AGC is a ttbar cross-section measurement from CMS 2015 Open Data [3]. The machine learning (ML) technique, Boosted Decision Tree (BDT), has already been implemented for the jet classification which gives promising results. In this proposal, we aim to improve machine learning workflow by incorporating a more complex model by using a Graph Neural Network (GNN). We will, then, compare performance results, computational cost, and user experience from using a GNN with BDT and non-ML pipelines.

Project Description

The ttbar production is top quark-antiquark pair, dominantly produced at the LHC, that decay into four jets, a lepton, and a neutrino as final products, as shown in Figure A. However, we can never be certain which parents jets and a lepton belong to. The original non-ML method is to use the trijet system to reconstruct the top quark mass. This can be done by considering all combinations of trijet, and choosing one with the highest combined transverse momentum (pT). The reconstructed mass from this trijet is approximately the mass of the top quark. This implies that the chosen trijet is from a hadronic side. Then, we can calculate physics observables.

Another refined method is to use an ML Boosted Decision Tree (BDT). One can attempt assigning four labels to the jet system. This will give more information about the event rather than just reconstructed top quark mass. In ttbar production, there will be one lepton and one b-tagged jet (b_top_lep) on the side of the leptonic decay. The other is a hadronic side which consists of two indistinguishable W-jets and one b-tagged jet (b_top_had), shown in Figure A [4]. In each event, all permutations of jet assignments altogether with twenty features, for example, deltaR, combined mass, pT, btagCSVV2, and qgl, are put into the Boosted Decision Tree (BDT). The model, then, is trained to find the jet permutation with the highest score which gives the correct jet assignment. By knowing individual jet labels, the reconstructed top quark mass and other physics quantities can be calculated. This implementation of machine learning reduces the error of reconstructed top quark mass as shown in Figure B [4].

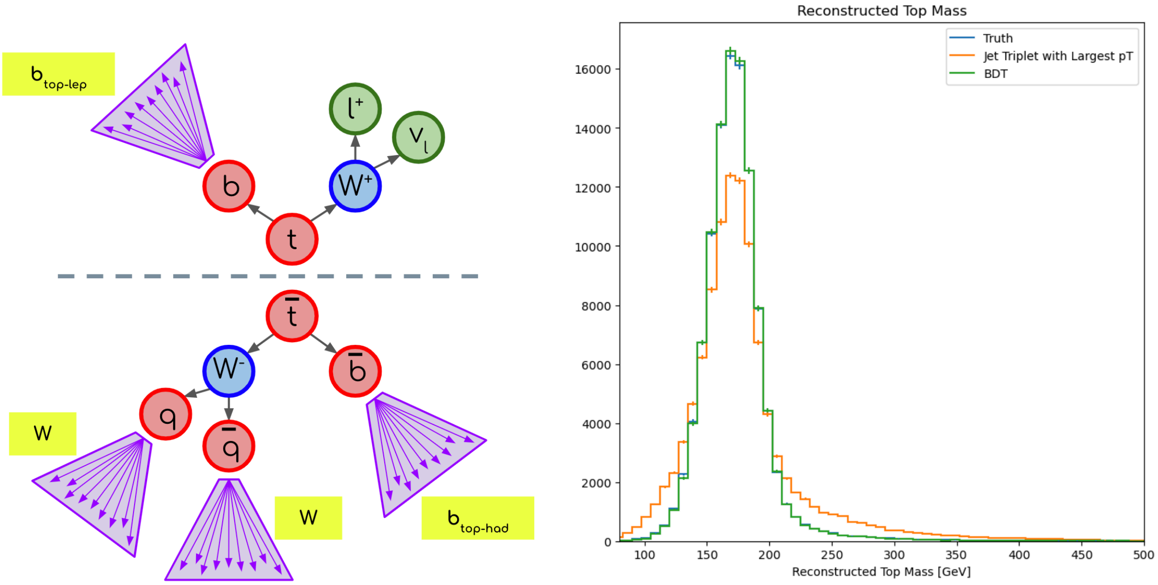

Figure: (A) The illustration of ttbar production separated into two regions; leptonic (top) and hadronic (bottom) sides. (B) The comparison of reconstructed top quark mass between the highest pT trijet method and implemented BDT method.

The existing jet assignment method using a BDT has improved the performance of jet identification. However, with the upcoming HL-LHC, there will be more data collected. In other words, we will have better insight into more complex events. In order to extract meaningful events from the background events, we need a more sophisticated model that can reflect the user's need for the analyses. We can then compare the new model with the existing one. Thus, this project will take a step further to implement a more complex machine learning algorithm using Graph Neural Network (GNN).

Data in particle physics is often heterogeneous and sparse in space which does not suit grid-like data structure or image representation [5]. As a result, a GNN is a good alternative method to handle this heterogeneous data. It is also possible that imposing order, used in BDT, will impact the learning performance, hence prediction.

To construct the GNN graph, we can represent jets as unordered particles embedded in two-dimensional eta-phi space. These will be the nodes in the graph with node features pT, btagCSVV2, and qgl. The edges, connecting two nodes, will represent the relationship between final state particles with edge features deltaR, combined mass, and combined pT [6]. We can, then, apply learning functions or message passing layers as a form of training. Finally, we can make predictions on graph-level or node-level, depending on the physics task, with classification scores. Similar to BDT, the jet permutation with the highest score gives correct jet assignment.

The output performance, however, depends on the GNN architecture. In this pipeline, there are different functions for node transformation, aggregation, and update. The most important is message-passing layers which can be, for instance, Graph Convolutional Networks (GCN), Multi-Layer-Perceptron (MLP), Graph Attention Networks (GAN), Gated Graph Neural Networks (GGNN), etc. Thus, this project will design a GNN architecture such that it will impose the correct model of nature and interaction in ttbar production.

Preliminary Timeline:

Week 1-2:

- Understand theoretical physics of ttbar production
- Familiarize myself with terminologies as well as software packages; ServiceX, coffea, dask, awkward, pyhf, mlflow, pyhf, Hist, cabinetry, NVIDIA Triton, pyTorch, etc.

Week 3-4:

- Run the analysis using the first workflow without machine learning implementation. Find the trijet system with highest pT and determine reconstructed top quark mass.
- Document analysis results and user's experience in each step of the workflow.

Week 5-6:

- Run the analysis using the existing machine learning BDT workflow. Use permutation of jets assignment and related features as inputs to BDT. Find the highest BDT score and determine reconstructed top quark mass.
- Document analysis results and user's experience in each step of the workflow.

Week 7-8:

- Run the analysis using the implemented GNN workflow. Construct graph using nodes as jet permutations with corresponding node-features and edge-features. Make graph-level predictions that give classification scores. Use the highest score to calculate reconstructed top quark mass.
- Document analysis results and user's experience in each step of the workflow.

Week 9-10:

- Compare the performance of reconstructed top quark mass from each workflow; non-ml, BDT, and GNN.
- Summarize analysis experience, computational cost and user experience, between non-ml, BDT, and GNN.

Week 11-12:

- Document results from each workflow.
- Suggest potential improvement and implementation as well as possible usage to the other analysis.
- Address any unexpected issues or interesting findings during the project.

Reference

[1] P. Elmer, M. Neubauer, M.D. Sokoloff, *Strategic Plan for a Scientific Software Innovation Institute (S2I2) for High Energy Physics* (2017), arXiv:1712.06592 [physics.comp-ph]

[2] A.Held, E. Kauffman, O. Shadura, A. Wightman, *Physics analysis for the HL-LHC: concepts and pipelines practice with the Analysis Grand Challenge* (2024), arXiv:2401.02766v1 [hep.ex]

[3] CMS Data preservation and open access group, Getting Started with CMS 2015 Open Data, https://opendata.cern.ch/docs/cms-getting-started-2015 (2022)

[4] E. Kauffman, A. Held, O. Shadura, *Analysis Grand Challenge Documentation*, https://agc.readthedocs.io/en/latest/ (2023)

[5] J. Shlomi, P. Battaglia, J.R. Vlimant, *Graph Neural Networks in Particle Physics* (2020), arXiv:2007.13681v2 [hep-ex]

[6] S. Thais, P. Calafiura, G. Chachamis, G. DeZoort, J. Duarte, S. Ganguly, M. Kagan, D. Murnane, M.S. Neubauer, K. Terao, *Graph Neural Networks in Particle Physics: Implementations, Innovations, and Challenges* (2022), arXiv:2203.12852v2 [hep-ex]