Particle Physics Domain Result Replication Project

Bryan Ambriz, Rui Lu, Charul Sharma

Halıcıoğlu Data Science Institute, University of California San Diego

DSC 180B: Capstone Project 2

Dr. Aaron Fraenkel, Dr. Javier Duarte

March 6, 2022

Abstract

In the field of particle physics, there are a myriad of methodologies for model-making decisions which we need to balance. Due to the properties of the particle-tracking (like the incredibly short lifetime of particles), the process of distinguishing jets and particles requires some helps from artificial intelligence or deep learning. Our project utilized elements of Graph Neural Networks, including convolutional layers such as EdgeConv and GENConv. In Quarter 1 of our project, we looked at how the Deep Sets Neural Network (DSNN) & Fully-connected Neural Network (FCNN) perform using a Reciever-Operating Characteristic (ROC) curve and AUC% (Area Under the Curve) of the aforementioned ROC. In this quarter (Q2), we had hoped to extend and out-perform the previous models utilizing a graph neural network, which we believed was able to capture the inherent structure of the data generating process (DGP), aka. the particle collisions. A ROC curve helps visualize the false positive and true positive rate of our model, while the AUC provides a point estimate that tells us how our model is doing overall. We also used Designed Decorrelated Taggers (DDT) to decorrelate the predictions from the jet mass using a mass-varying threshold for cuts on the output predictions of our model. This procedure helped further cement the finding that jet mass is a particularly good estimator for discovering H->bb jets, and therefore Higgs Boson fields/particles.

Introduction

Before having more details of our project, some terms need to be explained. A jet is a spray of particles that goes in the same direction. And there are tons of jets when we collide the protons in the Large Hadron Collider (LHC) like q/g jet, b jet, W/Z -> qq jet and H->bb jet. Due to the properties of jets, it reproduces the Higgs boson more likely. We need to distinguish the H-> bb jet from all the others. As for the importance of Higgs boson particles, it is the first and only spineless elementary particle observed and the building blocks of the universe while it is the mass-giver of other particles. Diving deep into the unknown properties of the Higgs boson would definitely help us to understand the existence of the universe and something relevant to dark matter.

Also, false positive rate (FPR) and true positive rate (TPR) are measurements for accuracy. Taking medical diagnosis as an example, we suppose there is an anomaly detection test that checks the existence of a certain type of disease. If the patient is sick, there would be two different outcomes. One is the patient is sick and the test is positive (the test considers the patient is sick) and another outcome is that the patient is sick but the test is negative (the test considers the patient is not sick). Thus, there are four situations and in this case, false positive is that the patient is not sick but the test considers it as sick, false negative is that the patient is sick but the test considers it as not sick, true positive is the patient is sick and the test considers it as sick and true false is the patient is not sick but the test considers it as not sick. TFR means $TP/TP + FN$ which represents the probability that an actual positive will be test positive. FPR means $FP/FP + TN$ which represents a positive result will be given when the true value is negative.

Also, signal in our project is actually the H->bb jets (Higgs events) which is rare and

its products are created by other protons decaying and the QCD events are backgrounds or noise which are hard to avoid due to the data generation issues, data collection issues and some other experiment issues. To improve this condition, we need to filter these out when we process our models.

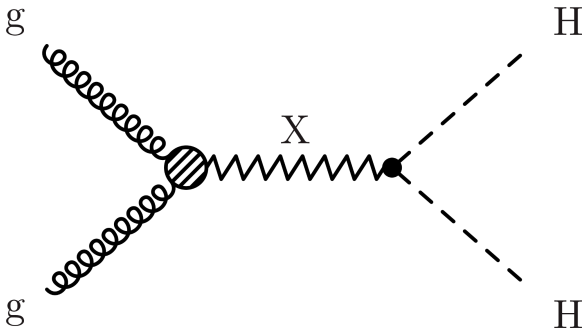


Figure Signal (Higgs) Events

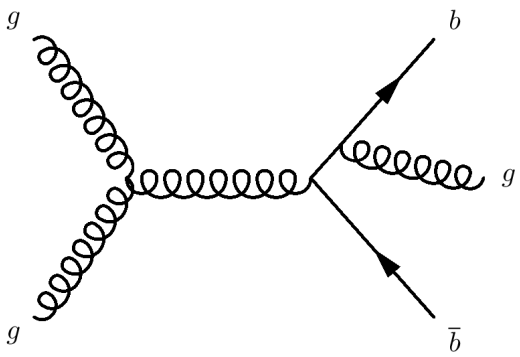


Figure Backgrounds(QCD) events

Motivations

In the previous quarter, we tried graph neural network model with deep sets and adversarial neural network. And in this project, we want to try something new and different like some fresh convolutional layers like GENConv and EdgeConv to see if whether our model could perform better or not. Meanwhile,

Data

The data collected at CERN (European Organization for Nuclear Research) was gathered by using simulated proton collision events. These events were used to gather information about track, secondary vertex, and jet feature data which could be used to inform the machine learning algorithms which can classify Higgs Bosons as they decay to bottom quark-antiquark pairs.

Labels

Labels in the dataset are used to differentiate between H-bb jets and all other jet types resulting from the strong interaction

between quarks and gluons (e.g. quantum chromodynamics, aka QCD).

Track Features

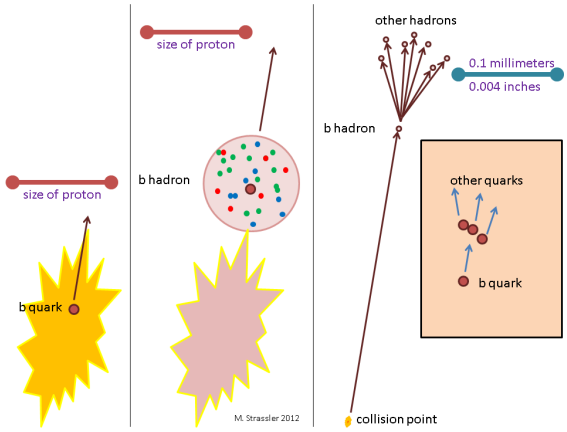
Track features in the dataset are properties of tracks (e.g. transverse momentum / maximum relative momentum, number of tracks, maximum signed 3D impact parameter value, etc.). Tracks themselves are simply the paths which the newly created hadrons (resulting from the particle collisions) took as they exited the collision.

Secondary Track (SV) Features

SV Features in the dataset are properties of the decay occurring and jet originating at the secondary vertex (e.g. transverse momentum / maximum relative momentum, number of secondary vertices … etc.). The secondary vertex is the location where a bottom hadron decays into other hadrons; a significant number of hadrons from a jet initiated by the production of a bottom quark come not from the collision point, but from the secondary vertex.

Jet Features

Jet features in the dataset are properties of the jet itself, a collection of particles which emanate and collimate from the initial collision point (e.g. transverse momentum / maximum relative momentum, sdmass, mass. … etc.). The jet itself is the target of classification for our purposes of detecting h->bb jets.



Methods

Graph Neural Network

Graph neural network is a type of machine learning which takes or extracts key information from a graph that contains nodes and edges like jets of particles. Based on these information, graph neural networks make predictions. Taking the social relation

as an example, the nodes would be individuals and the edges could be the relationships between each individual.

GENConv

GENConv is a model for implementation of GCNs (Graph Convolutional Networks). Regular GCNs model suffers from the vanishing gradient, overfitting and over-smoothing when the layers go deeper and deeperGCN here differentiate generalized aggregation functions like mean function and max function while the creators of it proposed MsgNorm as a new normalization layer and another pre-activation for GCNs. In other words, for exact operation and mechanism behind, the model combines information from the regular node and its neighbor by aggregating then updating the node feature by giving the aggregated value. As for regular GCNs, it just gives information about connected neighbors for updating the node feature. Moreover, the specialty for the method is that the creators made a novel generalized aggregation function to cover all common sorts of aggregation functions. And according to their essay, the model boosted performance significantly. (Li, Xiong, Thabet, & Ghanem, 2020)

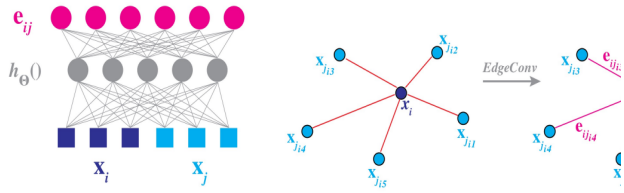
As for the results of GENConv so far, we have finished this and we plan to figure it out at the end of week 5.

EdgeConv

EdgeConv is the implementation of deep learning models which could tremendously increase the performance of these models because first, it acts on graphs dynamically computed in each layer of the model, incorporates local neighbor information and can be used for learning the global properties. Moreover, the EdgeConv helps the model to handle irregularity easily but the major difference between this and other methods is that it captures the local geometric structure while still maintaining

its own permutation invariance. Besides that, EdgeConv could generate edge features that describe the relationship between a node and the neighbors and is capable of grouping nodes both in Euclidean space and in semantic space which would be extremely useful for particle physics.

(Wang et al., 2019)



Left: Computing an edge feature, ei j (top), from a point pair, xi and xj (bottom). In this example, hΘ() is instantiated using a fully connected layer, and the learnable parameters are its associated weights. Right: The EdgeConv operation. The output of EdgeConv is calculated by aggregating the edge features associated with all the edges emanating from each connected vertex.

Designed Decorrelated Tagger

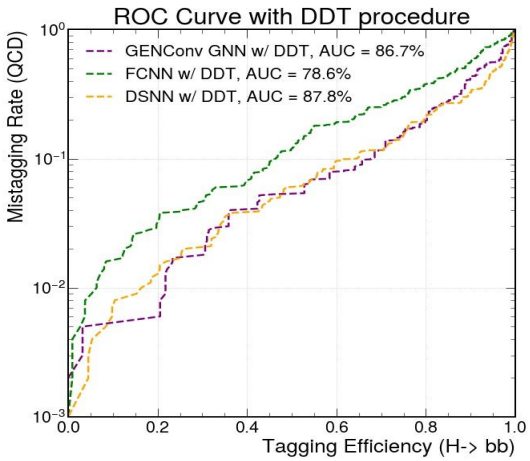
DDT could provide a simple approach to substructure decorrelation that the DDT transform yields a jet substructure discriminant which is decorrelated from the jet mass. (The ATLAS Collaboration 2018) Meanwhile, it could also define a mass-dependent threshold to reduce the impact. Decorrelating the jet mass is useful as the model can learn to distinguish signal from noise using the difference in mass.

Measurement

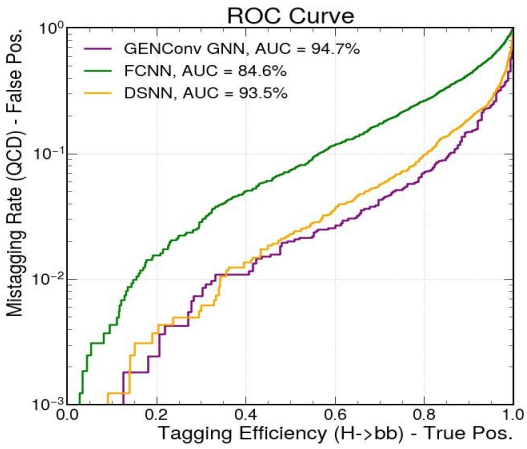
As for measurement, we still choose to use the plots of AUC and the plots of ROC to represent the final performance of models. ROC(Receiver operating characteristic) here helps to measure the performance of our GNNs model at all classification thresholds using 2 parameters, True Positive Rate(TPR) and False Positive Rate(FPR).

RESULTS

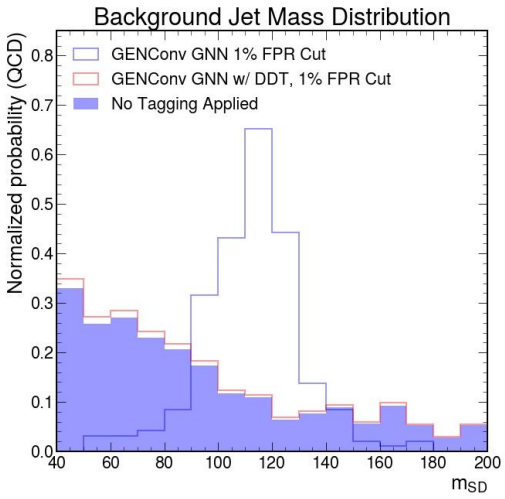
Visualizing the Accuracy of Predictions after DDT Decorrelation



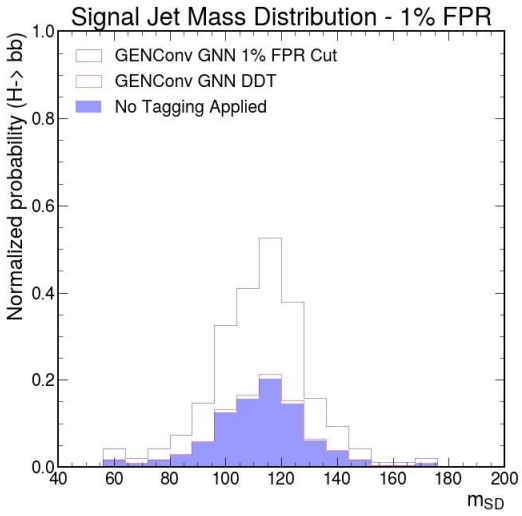
Visualizing Accuracy of Base Predictions



Visualizing Jet Mass Correlation with QCD / Noise



Visualizing Jet Mass Correlation with H-bb> / Signal



Related Works

To improve our model predictions, we decided to use a graph neural network approach. In order to approximate the results of the paper we were expected to replicate, we achieved / surpassed the accuracy of the initial model (Deep Sets NN) we had utilized during quarter 1. In the research process, we conducted a literature review of graph neural networks which specifically utilize convolution, as we believed convolution and convolutional models can discover patterns not only between node features, but also between jets / i.e particle interactions. For example, we specifically looked at papers which go over EdgeConv (Wang et al., 2019) and GENConv (Li. et al., 2020) convolutional layers and their implementation in graph neural networks.

Conclusion

After visualizing the ROC curves and comparing the AUC of our GENConv Graph Neural Network to the previous FCNN/DSNN we found an improvement of about 1.2% to last quarter's highest-performing model, the Deep Sets Neural Network - which was our goal. However, the improvement was much more when compared to the more basic FCNN, as we saw an improvement of about 9.4%, which gives weight to the hypothesis that graph neural networks are an excellent model choice. Applying the DDT procedure (i.e decorrelating the mass from the signal/qcd background) showed us that the results of our model get worse when the influence of mass is removed from the predictions - and the influence is visible across models and the difference *mostly* maintains the order in terms of performance. Ultimately, we were able to replicate figures 4, 5, and 8 in the paper "Interaction networks for the identification of boosted H → bb decays" by Duarte et.al. We discovered that particle jet mass is a

valuable predictor for the Higgs Bosons and that the choice of model is crucial in improving the accuracy. For example, we found the Deep Sets NN model performs about equally to our fairly simple GENConv Graph NN, because the variable nature of Deep Sets is able to capture the inherent structure of the ROOT / Awkward data provided by CERN / LHC. Further, looking at the jet mass distribution and the effect DDT has on our predictions was able to convince us that decorrelating the mass is not a good idea.

Credits

This project was a lot of hard work, and was difficult for us to conceptualize as none of the students on our team had any background in particle physics. However, the process of discovering new methods in conducting data analysis by replicating the results in Dr. Javier Duarte's paper was helpful. We have him to thank for providing resources such as the graph dataset which was used to transform the raw data into a more usable format, as well as early-stage models (Deep Set, Fully-Connected, Adversarial, and Interaction Neural Network) for inspiration. Our TA Farouk Mohktar was also helpful in providing guidance and clarity on our project.

Reference

Duarte, Javier M. “Interaction networks for the identification of boosted H→bb____ decays.” Arxiv, 2019, p. 20. Cornell University, https://arxiv.org/abs/1909.12285. Accessed 6 March 2022.

Li, G., Xiong, C., Thabet, A., & Ghanem, B. (2020, June 13). DeeperGCN: All you need to train deeper gcns. Retrieved February 04, 2022, from https://arxiv.org/abs/2006.07739

Wang, Y., Sun, Y., Liu, Z., Sarma, S., Bronstein, M., & Solomon, J. (2019, June 11). Dynamic graph CNN for learning on point clouds. Retrieved February 04, 2022, from https://arxiv.org/abs/1801.07829