CS281 Project Proposal

Scalable Signal Region Identification with applications to the ATLAS $W^\pm W^\pm W^\pm$ Analysis

Nicolo' Foppiani, Jonah Philion, Baojia(Tony) Tong

This project will research a scalable algorthm for the identification of regions of significant signal content in a high background parameter space. Our motivation for studying this problem comes from our field of high energy particle physics. In general, a discovery in particle physics is equivalent to statistically significant aberrations in count data from a given model over the experimental parameter space. First, we will build a robust algorithm which identifies these "signal regions" efficiently. Second, we will apply and evaluate the algorithm on simulation data from the ATLAS $W^\pm W^\pm W^\pm$ analysis.

Problem Statement   The search project is to try to measure the triboson production for the first time at the LHC using the ATLAS detector, one of the two main experiments at the LHC. Specifically, the signal process under study is the following: $pp \to W^\pm W^\pm W^\mp$. This measurement probes the so-called "quartic coupling": a quantity which is predicted by the current theory of the fundamental interactions, called Standard Model (SM) to be small. This process hasn't been observed yet, however, if the current theory was not correct, it would be expected an important difference between the measured value of this quantity and the prediction. This could be an insight toward some new physics process, not yet discovered, which could completely change our understanding of fundamental nature. With the higher center of mass energy of the collisions and the newer and larger datasets collected during the last few years, the probability for this process to happen should be large enough in order to produce several such events.

However, there is a great variety of well known processes which mimic the signal and which therefore are called background. Different other electroweak process ($W^\pm W^\mp$, ZZ, $W^\pm Z$) has about $O(10^3)$ times higher rate and can also generate signatures that are similar to the signal. Other SM process like $t\bar{t}$ has about $O(10^5)$ times higher rate, and can generate decay products that fake the signal signatures in the detector. The challenge is thus the small signal yield compared with the enormous and complicated background. The way to distinguish between the signal and the background is to characterize the events with some features which can be used to classify the events in signal and background. In the 2015 analysis, a selection based on cuts on some of these features has been used, and sensible results have been achieved.

Therefore this is an ideal challenge for testing machine learning methods. The goal is to build a classifier able to efficiently distinguish the signal events from the background, which contains a much larger number of events, up to $O(10^5)$ times the signal yield. The main features, or input variables, will be the raw kinematic measurements of decayed particles. The decay products from signal and background will have different yet complicated correlations and hence could be explored for identification.

Evaluation   Two important parameters in this case are the *signal efficiency* and the *background fake rate*. They are combined together through the ROC curve, which will be one of the metric for evaluating

the result. A naive parameter can be the area under the ROC curve. A more physics-driven metric will be the signal significance, which is what is required for announcing a discovery in the scientific world. In this case also the statistical uncertainty on the signal significance will be taken into account. The previous analysis' selection will be the honest baselines, which can be reproduced and then compared with.

Approach Both signals and backgrounds will be simulated, normalized to the expected yield of current data size. All inputs will be based on simulation and converted to numpy arrays. Standard cleaning and normalization will be applied after checking. Logistic and Softmax regression will be tested, but more interesting options will be decision trees and neutral networks. An important effort will be devoted to extracting new features, and to develop method to find the most powerful features given the datasets.

Milestone 1.0 The proof of concept test could be done using signal and only one background (WZ). It is expected that machine learning methods will do much better than previous selections. This requires baseline framework building, mostly in Python, and utilizing machine learning packages, such as Tensorflow and Pytorch. This is estimated to be finished before November.

Milestone 2.0 From there, more backgrounds could be added. There will be around ten different kinds of backgrounds, and it is challenging to have one identifier that will simultaneously reject all backgrounds while keeping the signal acceptance. Lots of optimization and model selections will be necessary. This is estimated to be finished by December.

Milestone 2.1 Parallel to Milestone 2.0, the optimizer's output needs to be interpretable. In high energy physics, and in science in general, the "black-box"type of machine learnt output is disfavored, not only because it is hard to interpret, but also because it is hard to exam and validate the method. Hence, a challenging part of this project will be map the neural network/decision tree outputs to variables that can not only be computed analytically, but also be easily validated in the future with data. We can try to understand the selections by reducing the number of inputs or employ a data planing method. This is also estimated to be finished by December.

Milestone 3.0 The ultimate goal would be have a software package that could take input as signal and backgrounds with raw input variables, and return a list of higher order **re-combined yet interpretable** variables that will distinguish the signal and background. This could then be widely applied not only in high energy physics research, but in general on all kinds of real world problems. Due to the complexity, this goal may be hard to achieve by the end of this year but will worth continued exploration based on the result of this project.

Collaboration Plan Tony (G6) will be in charge of sample production and will work for reproducing the previous analysis cuts. He will also build NNs, offer support and guidance on the project. Jonah (senior) will focus on understanding the most important features and building frameworks which are able to identify them for the different background. Nicolo' (G1) will be mainly working on the implementation of different algorithms and to assessing the results with the proper metrics.

Double-dipping This project is related to the research interest of all the three students. This result, if successful, will be documented as an ATLAS internal note, and will be used by the official analysis.