Fantastic Five Project Proposal

By: Portia Allen, Connor Caldwell, Johnny Murphy, Arnaud Filliat

Subatomic particle detection has become critical to deepening our understanding of what makes up the universe. Scientists at CERN are colliding protons using the Large Hadron Collider (LHC), and observing these small scale, big-bang-esque collisions to learn more about fundamental particles. However, as observation techniques continue to increase in quality, so does the quantity of data that documents a single collision. The purpose of this project is to use machine learning algorithms to accurately categorize high energy particle collisions, and then predict new particles.

This is possible through a reconstruction of the particle events from the LHC. Using the dataset provided by CERN scientists, it is possible to break down confirmed collision events and categorize them by particle type, and then use that algorithm with unknown collision data to predict the resultant particles. Each event contains a lot of data, such as the coordinates of the collision and the particle charges, and then the trajectories of the post-collision particles. Therefore, it is critical to have a higher degree of understanding of what the data represents, in order to build accurate models and correctly interpret the training data, as well as categorize unknown data using the same parameters.

By creating this machine learning algorithm, we can more efficiently and effectively track and categorize high energy particle collisions. We can gain a better understanding of the particles, how they collide, and what particles they break into after the collision. Additionally, several members of our group have a deep background in physics, so this dataset is of particular interest. Our personal connection to the discipline can be more motivating for the completion of this project. The application of data science to this project is extremely straightforward – given some dataset, we have to parse through it, train a machine learning algorithm with some of it, test it using the rest of the data, and then refining our model to be the most accurate it can be. Through this process, we can develop a better understanding and ability to apply data science/machine learning concepts to real-world data, while also gaining a deeper appreciation for and understanding of current particle physics research.

It should be noted that this dataset was used as part of a machine learning algorithm competition in 2018. The goal was to write an algorithm with the highest accuracy, and then highest throughput (speed relative to accuracy). Several people who placed highly in the competition have posted their answers and solution process online, so that other people can understand how to approach and execute an algorithm like this. Therefore, our group can make a solid attempt at it with our more limited data science skills, and then compare to those who have more experience in the field; from this, we can see what we could improve, as well as gain more insight into applied data science.

We will specifically be using this data set: Particle Tracking from kaggle. The question the competition asked was “can machine learning assist high energy physics in discovering and characterizing new particles. As for related work well we just went through a couple submissions by the teams that were involved in the competition

1st place team was the top quarks. Their solution was based upon speed as a note they said they “did not use much machine learning (only some logistic regression for candidate pruning), but rather classical mathematical modeling with statistics and 3d geometry.” Additionally coding it all in C++ with no dependencies. The team goes on to note that they had a data pruning stage (getting rid of duplicates) and only looked at specific data. They then developed a heuristic to prune the data and select promising pairs of hits. The heuristics are based on “how far the line passing through the two hits

passes from the origin, and the angle between the direction between hits and the direction given by the cells data for each of the hits .” With these hits and modeling a “helix track” for the particles to go they were able to achieve 92% score using 90% of the hits.

2nd place team was outrunner. His solution was to use a neural network to select the hits and then train a statistical model to predict and reconstruct tracks. It also looks like he used 27 parameters for his neural network to select candidates for the training model. His approach was more machine learning oriented and was extremely close to the 1st place solution.

The last one I looked at was the 5th place solution by Edwin Steiner. His solution implements “a combinatoric geometric algorithm which first builds a candidate list of track seeds by picking combinations of two to three hits in adjacent detector layers and then alternately truncates this list to the most likely correct track candidates and extends each candidate track by finding the hit(s) closest to an extrapolated idealized trajectory fitted to the hits found so far.” Also he ends up using python for his implementation as well as numpy and pandas libraries.

It will be incredibly interesting to delve into the specific data and learn more about our subatomic world. Hopefully our project can implement some of the strategies from the competition to create a good solution that categorizes the collisions and can predict new particles.

References:

top quarks - https://www.kaggle.com/c/trackml-particle-identification/discussion/63249

outrunner - https://www.kaggle.com/c/trackml-particle-identification/discussion/63256

Edwin Steiner - https://www.kaggle.com/c/trackml-particle-identification/discussion/67943