ML Challenge 4

Thomas Britton, David Lawrence

February 2020

1 Introduction

In particle physics calorimetry is used in conjunction with tracking detectors (see Challenge #1). Calorimeters seek to measure the energy of particles, typically by employing a high density material that absorbs the energy of the particle and produces a small amount of measureable light in the process. When a particles hit the material they tend to create cascades of particles and light, referred to as showers. The detectors are often segmented so that the showers spread over multiple blocks allowing one to reconstruct the position of the shower as well as its energy. Given that individual cells report being hit or not there exists a challenge to group all hits of a calorimeter according to the showers that produced them. These groups of hits are referred to as "clusters". To complicate matters further showers can be widely grouped into two classes; hadronic (showers produced from particles made from quarks) and electromagnetic (showers produced by particles like photons or electrons). The class of shower can be determined by looking at the profile of a given shower. Simulation of atmospheric showers (Figure 1) shows how different the shower profile can be between protons (hadronic) and photons (electromagnetic).

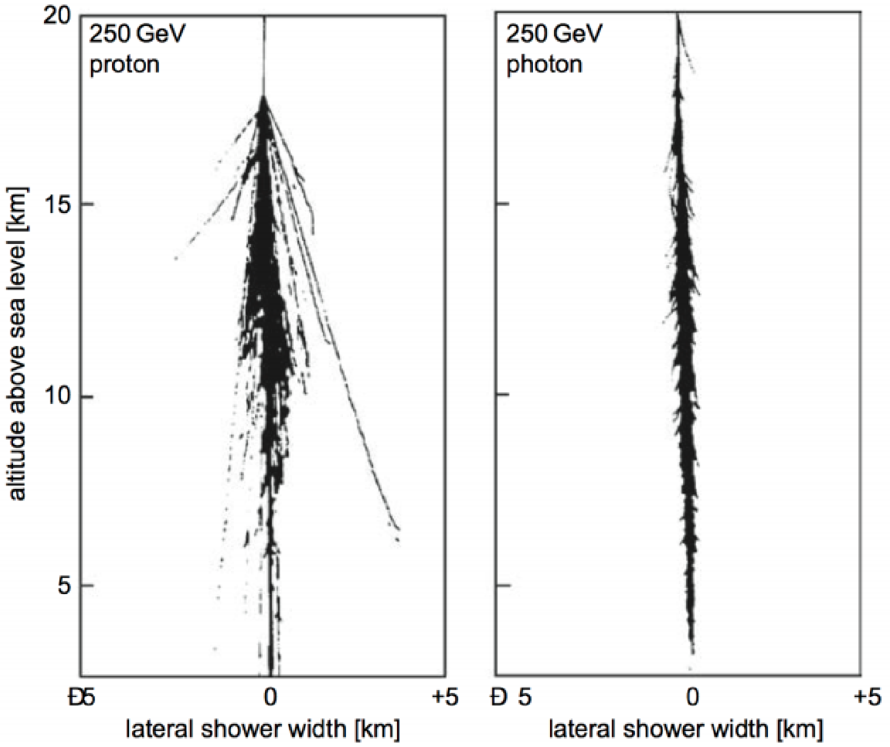

Figure 1: An illustration of the difference between hadronic and electromagnetic showers via a simulation of showers in the atmosphere. Here, the initial particle is coming in at the top of the picture and going towards the bottom.

2 The Challenge

2.1 The Problem

Contestants will be given the hit information from GlueX's Forward Calorimeter (FCAL) with the goal of counting the number of hadronic and electromagnetic showers present in each event.

2.2 Materials

Approximately 400k events contained in two different files (one labeled 002 and one labeled 003) each event contains at least 1 shower (as found through more traditional means). These events are given to contestants as a csv file. Each row in this csv file is a different event. Each row also contains the hit information from the detector as comma-separated 4-valued blocks (x,y,E,t) for each hit. Note that there may be differing number of blocks between rows. David Lawrence has provided a Jupyter notebook (ML4_DataCheck.ipynb) to help in visualizing some the data. To guide contestants better a tiny set of 18 events will also be provided. These 18 events will contain labels placed in a separate file; the labels will be in the same format as a valid submission and would be worth a total of 270 points (see below for scoring). Note the set 400k events will NOT have labels.

2.3 Judging

Contestants will submit a single file containing one row per event. Each row will consist of 3 comma separated values indicating, from left to right, the number of electromagnetic showers, the number of hadronic showers, the total number of showers. Points will be given for each column of each event in the following manner:

- 5 point for getting 0 showers/categories off;
- 4 points for getting 1 showers/categories off
- 2 points for getting 2 showers/categories off
- 1 point for getting 3 showers/categories off
- 0 points for getting 4 or more categories off

For example take an event with 2 electromagentic showers, 3 hadronic showers, and 5 total showers. A valid submission could be 2,1,5 (perhaps there is a confidence cut for classifying the shower). This submission would be worth 5+2+5=12 points. Please note that you will not be required to properly label electromagnetic and hadronic showers, only separate them into two groups and give the counts. On judging the counts of each class of shower will be swapped and the greater value taken. Taking the above example the possible point values would be 12 and 4+4+5=13 (obtained by swapping the electromagnetic and hadronic counts); thus the submission of 2,1,5 for an event containing 2 electromagentic showers, 3 hadronic showers, and 5 total showers would net 13 points. The winner is the submission that obtains the greatest total number of points across all events of the test set.

2.4 What is due

Like in the past a test set will be released on April 29th with final submissions due May 1st 2020. Submission will be all scripts/codes used in the training and testing (these should be in working order) as well as a text file containing the produced values in the order: electromagnetic shower count, hadronic shower count, total number of showers. If a submitter does not wish to try to split the count into the electromagnetic and hadronic showers -1 should be entered in those fields, though this is not recommended as a random guess may be right at least once.

Please direct any questions to[1]:
Thomas Britton: tbritton@jlab.org
David Lawrence: davidl@jlab.org

[1] Or even better, bring them to the ML lunch meeting on Wednesdays at noon in CEBAF Center F324-325!