PROCEEDINGS OF SCIENCE

# Overview of Machine Learning and Big Data tools at HEP experiments

**A. Castaneda***
*Universidad de Sonora*
*E-mail:* castaned@cern.ch

Following the preparation for the High Luminosity era of the Large Hadron Collider (LHC) and the imminent increase on the frequency of collisions by one order of magnitude it is evident the need for the development and implementation of new tools to optimize several tasks such as particle identification, reconstruction, data storage and processing. Many of these implementations will be based on machine learning algorithms that have the potential to process signals in a smarter way than current technologies allowing to fully exploit the detector capabilities of the LHC experiments and increase the probability to find new physics phenomena.

*7th Annual Conference on Large Hadron Collider Physics - LHCP2019*
*20-25 May, 2019*
*Puebla, Mexico*

*Speaker.





PoS(LHCP2019)234

## 1. Introduction

There is a growing interest in the implementation of machine learning and big data tools to be used by the modern particle physics experiments, in particular those associated with the Large Hadron Collider (LHC) at the CERN laboratory. The LHC has successfully collected data from 2010 to 2018, having the discovery of the Higgs boson among its greater achievements. Recently the LHC has started a maintenance and upgrade period towards the starting of its Phase-2, which is also known High Luminosity era in which the frequency of collisions will increase one order of magnitude compared to the current operation. This will allow scientist to fully exploit the detector capabilities and increase the probabilities for the discovery of new physics phenomena. New analysis techniques and data processing algorithms need to be developed to cope with this increase on luminosity. In the past Multivariate techniques have been used mostly to discriminate between signal and background processes, however with the development of deep learning it seems possible to optimize several tasks as for instance those related to particle identification, reconstruction and triggering. Also the use of rather new architectures such as the Generative Adversarial networks (GANs) seems promising to achieve a substantial reduction on the computational resources for the production of Monte Carlo simulated samples.

## 2. Particle Identification and Reconstruction

Classification of jets is one of the most important tasks in high energy physics experiments. Identifying the nature of a jet through its internal structure has broad applications to searches for new physics and Standard Model (SM) measurements at the LHC. According to the SM a jet could be initiated by a gluon or a quark, a key difference between these production mechanisms is that gluon jets tend to have more constituents and a broader radiation pattern than quark jets, traditional approaches use a set of key observable to achieve such separation, however a rather new method that uses state-of-the-art image classification techniques has proven to have similar or better performance. Experimentally jets are identified by the energy deposited by its constituents in the calorimeters. This energy deposition could be interpreted as a 2D image as shown in the Figure 1, where a comparison between a gluon initiated jet and a quark is presented. A kind of deep neural network architecture known as Convolutional Neural Network (CNN) is used to train a model that recognize whether a jet is originated by a quark or gluon. The CNN uses an array of filters that capture patters in the image and reduce dimensionality, afterwards the information is sent to a fully connected neural network that uses that information to classify the image. This architecture is shown in Figure 2 (left). The CNN was trained and the results evaluated using Monte Carlo simulation samples. The performance in terms of Quark jet efficiency and gluon jet rejection is shown in Figure 2 (right), more details about the neural network and samples used can be found in [1].

Another relevant tasks concerns the classification of jets according if they were originated from heavy or light quarks. Rare and interesting processes such as some of the Higgs decay modes, decays of the top quark and searches for new physics phenomena such as Supersymmetry (SUSY) contain b-quarks in the final state, therefore it is extremely important to tag these kind and events in order to separate from contribution of events with light quarks such as QCD processes. The CMS

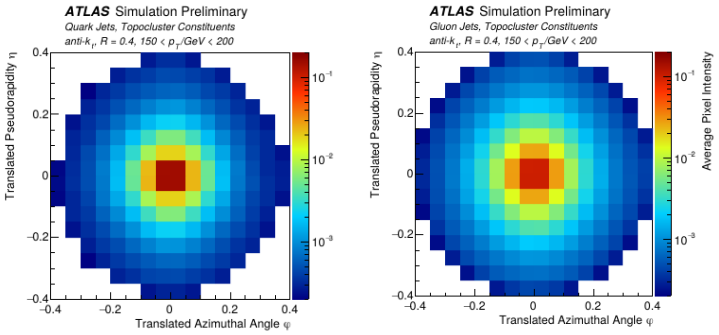


**Figure 1:** Quark jet image (left) and gluon jet image (right) produced with Monte Carlo simulation. Quark-jets are more collimated than gluon ones

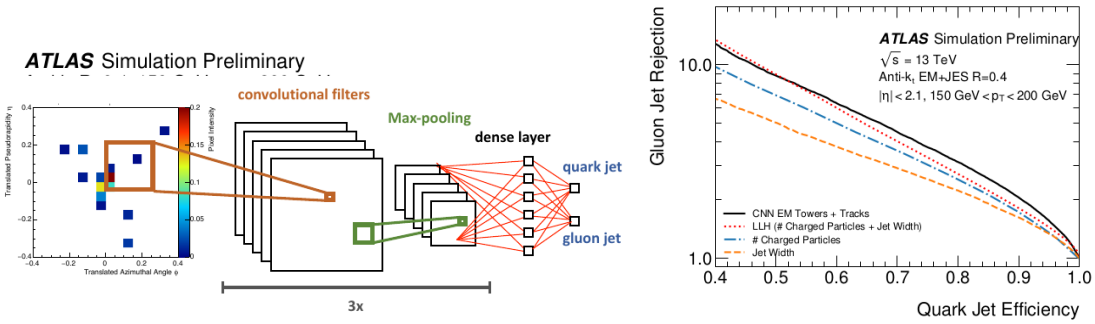


**Figure 2:** Illustration of a deep convolutional neural network architecture (left) and the performance using simulated samples (right).

experiment has implemented deep neural networks to perform jet flavor separation, this has been implemented in a dedicated tool (DeepCSV) used by the CMS collaboration [2]. DeepCSV is base on a deep neural network with four hidden layers each one with 100 nodes.

Reconstruction of fake tracks is a consequence of the limitations of the detector configuration or reconstruction algorithms. It is desired to reduce the fraction of these events as much as possible. The LHCb experiment has implemented methods based on a multilayer neural network to perform fake track rejection. It uses information about the pattern of fake tracks to train a model. This algorithm retains more that 99% of well reconstructed tracks while reducing the number of fake tracks by 60% as reported in [3].

## 3. Generative methods for Monte Carlo simulation

The production of Monte Carlo simulated samples is one of the crucial tasks in High Energy Physics. Currently this production represents more than 50% of the computing load of the LHC system(GRID). Detailed simulation is needed to optimize the event selection for Standard Model and new physics searches, evaluate the performance of the detector systems. In the case of rare processes large samples are needed in order to reduce the statistical uncertainties. The most com-

putationally expensive step in the simulation pipeline of a typical experiment at the LHC is the detailed modeling of the full complexity of physics processes such as the evolution of particle showers inside calorimeters. Generative Adversarial Networks (GANs) [4] are a type of deep neural network composed of two nets, the generator and the discriminator. GANs are trained with a known dataset, the generator usually starts with a seeded randomized input that is evaluated by the discriminator. Using a backpropagation algorithm the generator is able to learn from the training dataset until it reaches a point the discriminator is not able do distinguish between the true information and the one coming from the generator. This principle is used for the simulation of showers in high energy physics. The true showers are generated with a full simulation using GEANT4 package and compared with the output of a CaloGAN, as shown in Figure 3.

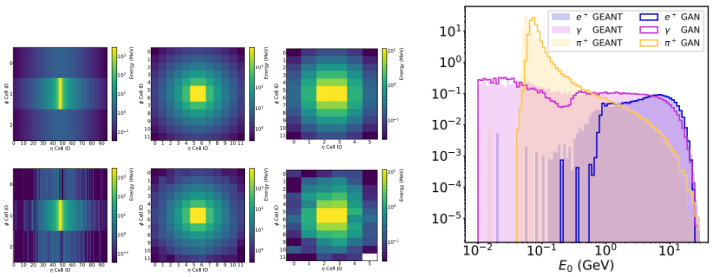


**Figure 3:** Quark jet image and gluon jet image (left) produced with Monte Carlo simulation. Quark-jets are more collimated than gluon ones. Comparison of showers between Geant4 and CaloGan approach (right).

## 4. Big data tools

The LHC collide protons with a frequency of 40MHz, This rate of data is strongly reduced with a set of triggers applied at online and offline levels, however storing the data output, processing and analyzing remains a challenge for the collaboration. Big data tools developed by private companies are currently explored and implemented in the system of CERN experiments. Examples of these tools are Hadoop and Spark which are tools that can be used for very specific projects, some of the applications include: A next generation archiver for accelerator logs, a new event index for the ATLAS experiment, Analytics in the cloud using the SWAN platform for CERN. These big data services are complemented by the use of GPU cluster that speed up different process including the training of neural networks in machine learning.

## References

[1] *Quark versus Gluon Jet Tagging Using Jet Images with the ATLAS Detector*, CERN, ATLAS Collaboration, "ATL-PHYS-PUB-2017-017",Jul 2017, ATL-PHYS-PUB-2017-017, https://cds.cern.ch/record/2275641

[2] *Identification of b quark jets at the CMS Experiment in the LHC Run 2*, CERN, CMS Collaboration, CMS-PAS-BTV-15-001, 2016, CMS-PAS-BTV-15-001

[3] *Fast neural-net based fake track rejection in the LHCb reconstruction*, CERN, Geneva, LHCb-PUB-2017-011, Mar, 2017,

[4] *CaloGAN: Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks*, PhysRevD.97.014021,Paganini, Michela and de Oliveira, Luke and Nachman, Benjamin, Phys. Rev. D, volume 97, issue = 1, pages = 014021, numpages = 12, year = 2018, month = Jan, doi = 10.1103/PhysRevD.97.014021