Application of Machine Learning Techniques on Experimental High Energy Physics Real Data

Contreras Cossio Michelle, Universidad de Sonora. Duarte Gonzalí Luis Fernando, Universidad de Sonora.
Tapia Takaki Daniel, University of Kansas. KC. Kong, University of Kansas.

Introduction

The data analysis through the use of Machine Learning (ML), which is seen as a subset of the Artificial Intelligence (AI), is an emerging area in different sciences, more specifically in particle physics. Algorithms used in ML produce mathematical models based on sample data, called training data (TD).

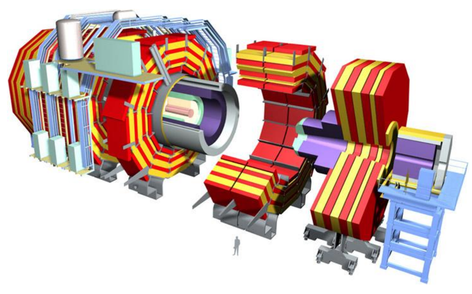


Figure: The CMS experiment at CERN. CMS Detector Layout, website: *http://www.hephy.at/user/friedl/diss/html/node8.html*

When applying ML techniques in real data of CMS, it allows us to identify particles or a certain type of events and the reconstruction of these which make more efficient the process of searching new particles and the study within high energy physics (HEP).

The goal of this project is to identify the background and signal of different runs in the Large Hadron Collider (LHC) collected by the Compact Muon Solenoid (CMS) experiment by using a ML algorithm with simulated data for the signal and real data for the background.

Methodology

A common structure of ML are Neural Networks (NN). The most important thing we need to build a NN is data. There are two types of data we will need: an already classified dataset (training data) and a dataset ready to be classified (test data). In this case there are two classes Signal and Background.

The test data was collected from CMS Open Data, an open source of real collision data from CERN. This was achieved with the use of the CERN Virtual Machine and ROOT (software used in experimental HEP).

The Signal Class data was obtained from a simulation made with MADGraph + Pythia + Delphes, a software that allows us to recreate Z events that look the most similar to reality.

The Background Class data was collected from the real collision data.

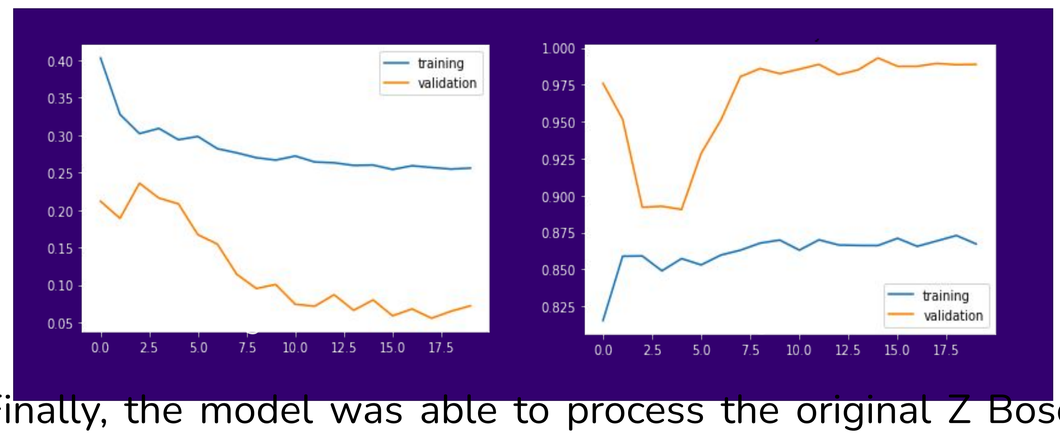
Once we collected all the data we needed, we built a Deep Neural Network (DNN) in Python using two dense layers with 128 neurons each, an Adam Optimizer and a Softmax Activation function running it with 20 epochs.

The network was trained with the training data and then used in test data in order to classify the real collision data and clean the Z Boson invariant mass signal removing the background.

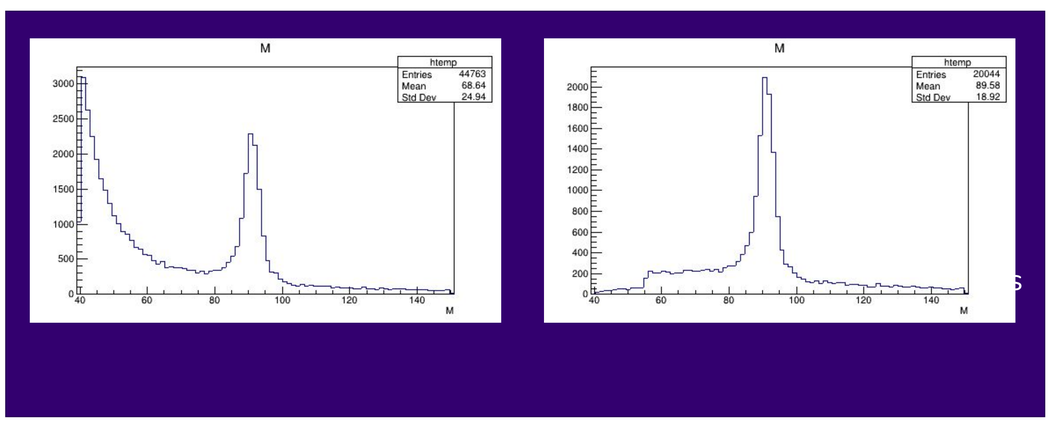
Results

The efficiency of the NN was pretty high. A way to measure this is calculating the loss function and accuracy.

In a good - or efficient- NN the loss function has to be tending to zero as epochs go. Shown in Diagram A, the loss function has this behavior. On the other hand, accuracy should get the closest to one, as it happens in Diagram B.



Finally, the model was able to process the original Z Boson signal (Diagram C) and classify between signal and background to get a cleaner look at the signal (Diagram D).



Conclusions

This research opens the possibilities to continue applying ML techniques on high energy and particle physics focused in different ways. In our case it has been focused on cleaning real data from CMS Open Data to find a signal in an easier way. The project could be a starting point on the generalization of the method for many more particles and different events or reconstruction of these. In the future it is sought to apply ML techniques for the search of new particles without having knowledge of them, that would create an agnostic model that doesn't depend in the information previously obtained. Also we are trying to configure algorithms to add physical variables and do it nearest to the reality.

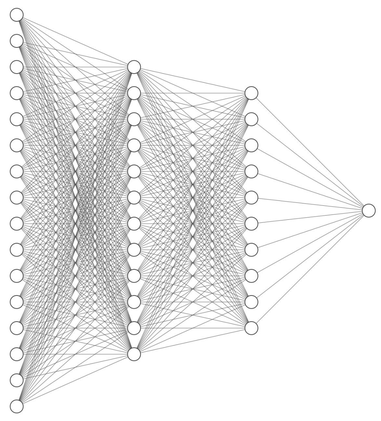


Figure: Example of a Fully Connected Neural Network Diagram.

Aplicación de Técnicas de Machine Learning en Datos Reales de la Física Experimental de Altas Energías

Duarte Gonzalí Luis Fernando, Universidad de Sonora. Contreras Cossio Michelle, Universidad de Sonora. Ramírez Álvarez César Omar, Universidad de Sonora. Valencia Palomo Lizardo, Universidad de París-Sur. Tapia Takaki Daniel, University of Kansas. KC. Kong, University of Kansas.

Introducción

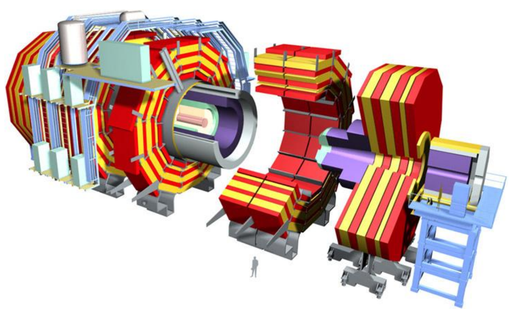


Figure: The CMS experiment at CERN. CMS Detector Layout, website: *http://www.hephy.at/user/friedl/diss/html/node8.html*

El análisis de datos a través del uso de Machine Learning (ML), el cual es visto como un subconjunto de la Inteligencia Artificial (IA), es un área emergente en diferentes ciencias y sus ramas, más específicamente en la física de altas energías y de partículas.

Los algoritmos usados en ML producen modelos matemáticos basados en datos de ejemplo, llamados datos de entrenamiento. Cuando aplicamos las técnicas de ML en datos reales del CMS nos permite identificar partículas o cierto tipo de eventos y la reconstrucción de los mismo lo que hace más eficiente el proceso de búsqueda de nuevas partículas y el estudio de la física de altas energías.

El objetivo de este proyecto es identificar el ruido de la señal de diferentes corridas en el Gran Colisionador de Hadrones (LHC) recolectados por el experimento CMS (Compact Muon Solenoid) usando un algoritmo de ML con datos simulados para la señal y datos reales para el ruido.

Metodología

Una estructura común de ML son las Redes Neuronales (NN). La cosa más importante que necesitamos construir u obtener son los datos. Existen dos tipos de datos que vamos a necesitar: un conjunto de datos de ejemplo (training data) y un conjunto por clasificar (test data). En este caso tenemos dos clases, la Señal y el Ruido.

Los datos por clasificar fueron recolectados desde CMS Open Data, datos open source de colisiones reales del CERN. Esto fue realizado con el uso de la CERN Virtual Machine y ROOT (software usado en HEP experimental).

Los datos de la clase “Señal” fueron obtenidos de una simulación hecha con MADGraph + Pythia + Delphes, un software que nos permitió recrear los eventos de un bosón Z lo más parecido a la realidad.

Los datos de la clase “Ruido” fueron recolectados de los datos reales de las colisiones.

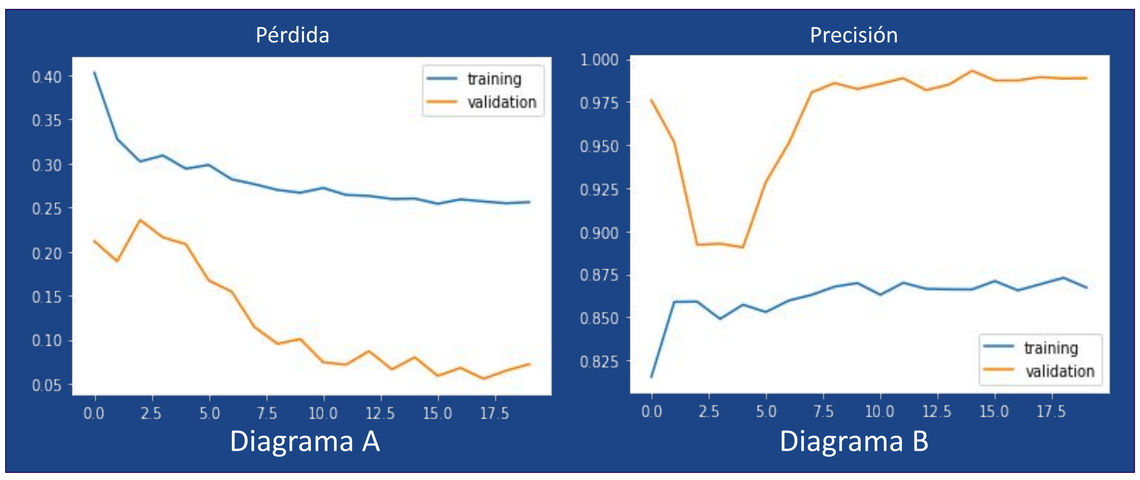
Una vez recolectados todos los datos necesarios, nos dedicamos a construir una Red Neuronal Profunda (DNN) en Python usando dos capas densas de 128 neuronas cada una, un optimizador Adam y función de activación Softmax corriendo por 20 épocas.

La red fue entrenada con los datos de ejemplo y después fue usada en datos de test para clasificar los datos reales de colisión y limpiar la señal de masa invariante del bosón Z removiendo el ruido de fondo.

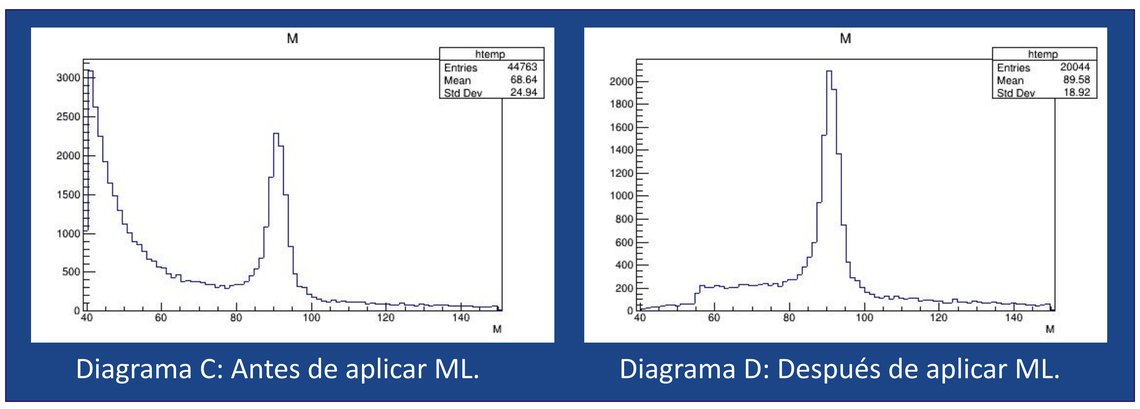
Resultados

La eficiencia de la red neuronal fue bastante alta. Una manera de medirla es calculando la función de pérdida y precisión.

En una buena - o eficiente- red neuronal la pérdida tiene que tender a cero con el paso de las épocas. En el Diagrama A, la pérdida tiene el siguiente comportamiento. Por otra parte, la precisión debe acercarse lo más posible a uno, así como ocurre en el Diagrama B.



Finalmente, el modelo fue capaz de procesar la señal original (Diagrama C) y clasificar entre señal y ruido de fondo para tener una señal más limpia (Diagrama D).



Conclusiones

Esta investigación abre las posibilidades para continuar aplicando técnicas de ML en la física de partículas y de altas energías enfocado de diferentes maneras. En nuestro caso ha sido enfocado a la limpieza de datos reales del CMS Open Data para encontrar una señal de una forma más sencilla.

El proyecto podría ser un punto de inicio en la generalización de este método para muchas más partículas y diferentes eventos o en la reconstrucción de estos. En el futuro se busca aplicar técnicas de ML para la búsqueda de nuevas partículas sin tener conocimiento de ellas, eso crearía un modelo agnóstico que no depende en la información previamente obtenida. Además, estamos tratando de configurar algoritmos para agregar variables físicas y hacerlo lo más cercano a la realidad.

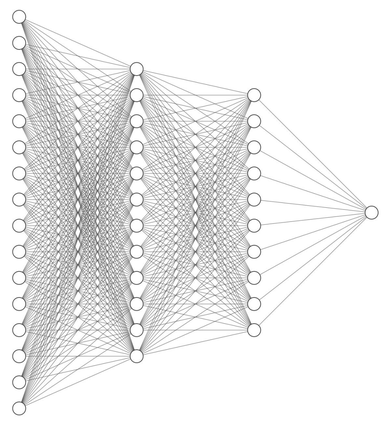


Figura: Ejemplo de una Red Neuronal totalmente conectada.