Using Convolutional Neural Networks for Muon detection in WCD tank

B.S. González *, R. Conceição [†] **, B. Tomé** [†] **, M. Pimenta** [†] **, L.J. Herrera *, A. Guillén ***

* Computer Architecture and Technology Department, University of Granada, Granada, Spain.
[†] LIP/IST, Lisbon, Portugal.

E-mail: `borjasg@correo.ugr.es, aguillen@ugr.es`

Abstract. The aim of this paper is to study the possibility of improving the gamma/hadron discrimination in extensive air showers. For this purpose, the identification of hadronic extensive air showers is carried out by means of the detection of muons in water Cherenkov detectors (WCDs). Machine learning algorithms have proven to be useful in a wide variety of fields, and due to their outstanding performance in problems involving complex data, Convolutional Neural Networks (CNNs) have been used in the analysis of the signals measured by the WCDs. Taking simulated events, different approaches were proposed attending to the balance of the classes in the training stage. The results obtained are promising and show that machine learning algorithms provide a powerful tool for muon detection and gamma/hadron discrimination to be considered in future gamma-rays detectors like The Southern Wide-field Gamma-ray Observatory (SWGO) to be built in South America.

1. Introduction

High-energy gamma-rays detection, whose energies span from 10 GeV up to 100 TeV, is crucial in the research field of high-energy astrophysics. Their neutral nature allow them to travel long distances along the Universe without being affected by galactic magnetic fields. Thus, they can be used to track some of the most extreme, non thermal, phenomena in the Universe which produce gamma-rays, such as: fast rotating neutron stars or supermassive black holes [1].

The direct detection of primary gamma-rays at low energies is possible using satellite-based detectors, like *Fermi*-LAT in the Very-High-Energy region. However, for gamma-rays with energy above a few hundred GeV the flux becomes too small, and cannot be detected by this kind of detectors as their effective area is limited by their size due to the high cost of space technology. At higher energies, above some tens of GeV, the primary cosmic photon –by its interactions with the Earth's atmosphere– is able to produce a

cascade of secondary particles, which are known as Extensive Air Showers (EAS). With enough energy to reach the Earth's surface. In this framework, ground-based array detectors are effective, but, one has to deal with a huge background signal produced by cosmic rays, which are constantly bombarding the Earth.

Gamma-ray ground-based array detectors must be placed at high altitude to minimized the atmosphere's attenuation and the most successful are based on a dense array of Water Cherenkov Detectors (WCDs). These detectors consist on tanks filled with water and with photomultipliers placed inside them. Hence, they are able to measure the Cherenkov light produced when they are reached by relativistic particles [2].

There are different techniques related to ground-based detectors with the aim of selecting gamma-rays out of the background produced by cosmic rays. On one side, regarding to the ground patterns produced, contrary to pure electromagnetic showers, hadronic showers produce high transverse momentum particles which lead to the transverse broadening of the shower and the creation of clusters. In this field, Machine Learning techniques have been able to surpass the classical statistical approaches [3, 4] and may unveil unknown features. On the other side, as muons are only found in hadronic cascades, the identification of this particle in WCDs provide a powerful discriminator between gamma and hadron induced showers at high energies. In this analysis, fast direct Cherenkov light pulse of a single muon is seen mainly in only a part of the readout matrix whilst the spread signal of several photons/electrons entering at random positions in the WCD is seen across the whole readout matrix. Given those spatial characteristics, Machine Learning models like Convolutional Neural Networks (CNNs) can exploit this information and become a powerful tool in the analysis of such complex data (signals).

Convolutional Neural Networks (CNNs) constitute the state-of-the-art in many Machine Learning applications, such as signal analysis and image classification. In this work it is proposed a model using one dimensional CNNs to analyse the signal traces measured by WCDs and get the probability of having a muon in order to identify muons, and then be able to reduce the hadronic background. CNNs are feed-forward Artificial Neural Networks (ANNs) whose typical structure consist on dividing its hidden-layers in two blocks: a first set of convolutional layers that extract features from the data, and a second set of fully-connected layers which classify by means of the information learned by the previous convolutional layers. Thus, the main advantage of using CNNs is their ability to combine the process of feature extraction (crucial when working with complex data like signals) and classification in a single process [5].

2. Data description and model design methodology

2.1. Simulations

The investigation carried out in this paper is done at WCDs level, and the purpose is to differentiate stations with muons from those with photons or electrons. Thus, the analysis can be done using simulations of hadronic EAS, which provide events with the particles stated above. The hadronic cascades simulations have been generated with CORSIKA [6], and the water Cherenkov detectors proposed with Geant4 toolkit [7]. The data set is composed by 3693 hadronic showers events caused by a proton as primary particle, with energies of a few TeVs.

As stated before, ground-based array detectors are based on a dense array of Water Cherenkov Detectors in the Earth's surface. Thus, there are several possible configurations for these detectors. The design chosen for this work consist on a circular array area of 80000 m^2, placed at 5000 meters above sea level and composed of cubic WCDs with 9 (3×3) Silicon Photomultipliers at the bottom and white diffusive walls. The design maximizes the collection of direct light, important for the shower geometry reconstruction, and a good calorimetric measurement.

2.2. Methodology and model Design

2.2.1. Data preparation
As the data analysis is carried out at station level, there is the need of separating the showers beforehand to avoid classifying a station from an already seen shower. Thus, data set is partitioned into 3 independent subsets: train, validation (10%) and test (1500 showers). For the train subset, due to the complexity in the storing and processing of the 9 signals sampled at high frequency, 300 showers are used containing roughly 760000 station events. Regarding to the train data set two approaches have been used: *All muons* and *Single muons*. Both approaches use the same input values, but *Single muons* data set only contains stations with muons and no electromagnetic components. With this strategy we intend to train the CNN with the cleanest muon signal as possible. Additionally, all the events used are normalized using z-score normalization to ensure the maximum fairness.

2.2.2. Data balance
From Physics it is known that the number of stations with muons is low. What is more, at low energies many hadronic showers do not have any muons. Thus, for this disjoint binary data set it is necessary to study the class balance before training the CNN. After taking a closer look at the simulations, the proportion of stations with muons is roughly a 1% in both (single and all muon) data sets. For this reason, using strategies to balance the data set is a must if we want to identify any station with muons. Different strategies have been used: Random undersampling (UD), Random oversampling (OS) and Synthetic Minority Over-sampling Technique (SMOTE). UD balances the data set eliminating samples of stations without muons, whilst OS and SMOTE create new samples of stations with muons.

2.2.3. Model design
The model's output is defined as the probability of having a muon in tank. Afterwards, it is necessary to design the pipeline necessary to obtain such result. For each station reached by a particle there are nine signals (one per SiPM) to store. Our approach is proposed in analogy to the use of CNNs for image classification, where three channels per pixel are used (one channel per primary color). In our case, there are nine values for each temporal instant (nanosecond), one per SiPM in the station. Thus, each channel contains the signal trace recorded by the SiPM, and each signal trace value is attached to its temporal value by its position in the array storing the signal. The length of the signal trace to work with has been established to the first 30 nanoseconds due to the study of the average signal trace.

The design of the CNN architecture has been based on the experience tackling other problems related to extensive air showers using CNNs like [8].Three convolutional layers were used to cover the inner complexity of the data. As seen from the average traces,

there is a maximum in the signal within the first 3 nanoseconds and it does not maintain long in time. Thus, small filters of size 3 are enough to locate them spatially. The number of filters for convolutional layers are 20, 15 and 10 respectively. By experimentation the two first fully-connected layers with 20 and 10 neurons were proposed, and a final fully-connected layer with a neuron. The hidden layers use ReLU activation function, but, as it is a regression problem the last layer use a linear function to get the probability of having a muon from 0 to 1 in the station evaluated.

3. Experiments

The CNN was developed using Keras framework in Python3.7. The models were trained using Adam Optimizer [9] with the following parameters: learning rate = 0.001, betas = (0.09, 0.999) during 100 epochs with a batch size of 100. Mean Square Error is used as loss function. Pooling layers have been discarded as they reduce the average specificity, though higher average sensitivity is achieved when using them.

Table 1 shows the results obtained by the model using different approaches to balance the data set. As the data set is imbalanced the fitness of the models have been evaluated using F1-score. The performance shown at the table corresponds to an average where the threshold to determine the muon/no muon classification varies in the range $[0.05, 0.95]$ in steps of 0.05. Regarding the results shown at the table 1, it can be seen that it is not advisable to fully balance the data set as there are roughly a 1% of samples with muons. Additionally, training with the single muons approach can be advisable in many configurations. To select the best result, the one with highest F1-score is chosen, this is the model UD_{20} which was obtained by training the CNN using All Muons with Undersampling and a class ratio $N_\gamma/N_\mu = 20$. The average accuracy achieved by this model is 96.64%, and setting a threshold of probability to 0.75, it corresponds to a 99.13% accuracy in stations without muons and a 23.02% in stations with muons.

We must recall that, as we want to discard hadronic extensive air showers, the fact of missclassifying a gamma-ray shower is not desired at all. Thus, identify correctly most stations hit by photons or electrons is a must. Meanwhile, in the identification of stations which have muons, as long as we are able to identify some of them it can be enough to discard the observation. In this sense, our best result has been able to achieve a desirable performance identifying correctly the 99.13% of the events with electrons and/or photons, and the 23.02% of the events with muons.

Configuration	Accuracy (%)		F1-Score (%)	
	All muons	Single muons	All muons	Single muons
UD_{20}	96.64	97.33	24.70	22.97
UD_{10}	95.26	96.01	24.32	21.62
UD_1	81.22	80.18	11.15	10.08
OS_1	91.21	95.68	18.40	22.22
$UD_{10}\ OS_1$	89.81	91.66	15.49	16.71
$SMOTE_1$	94.83	96.48	19.89	20.08
$SMOTE_{10}$	97.86	97.47	14.64	14.53

Table 1: CNN performance over 1500 showers of the proton test set when training with all stations with muons available in the simulation vs. using the stations with only single muons.

4. Conclusions

Due to their neutral nature, gamma-ray detection is essential in the investigation of the emitting astrophysical sources of extreme, non-thermal, phenomena in the Universe. In their detection with ground-based detectors, given the huge background produced by cosmic rays, it is necessary to implement strategies to be able to select the gamma-rays out of this background. To this end, this work explores the use of CNNs in the detection of muons in WCDs to improve the EAS gamma/hadron discrimination. Different models were proposed to balance the data set, where the number of stations reached by muons correspond roughly to a 1%. The results obtained using different approaches show a great potential, in which, in spite of the low number of stations with muons, some of these stations were identified, which is enough to discard the hadronic showers. Thus, by taking simulated events, it was shown that this machine learning model provides a powerful tool for muon detection and, consequently, for gamma/hadron discrimination to be considered in future gamma-rays detectors.

Acknowledgements

References

[1] Alessandro De Angelis and Mário Pimenta. *Introduction to particle and astroparticle physics: multimessenger astronomy and its particle physics foundations.* Springer, 2018.

[2] Pedro Assis, Ruben Conceição, Mário Pimenta, Bernardo Tomé, Alberto Blanco, Paulo Fonte, Luis Lopes, Ulisses Barres de Almeida, Ronald Shellard, Benedetto D'Ettorre Piazzoli, et al. Lattes: a novel detector concept for a gamma-ray experiment in the southern hemisphere.

[3] Filipe Assunção, João Correia, Rúben Conceição, Mário João Martins Pimenta, Bernardo Tomé, Nuno Lourenço, and Penousal Machado. Automatic design of artificial neural networks for gamma-ray detection. *IEEE Access*, 7:110531–110540, 2019.

[4] A. Guillén, A. Bueno, J.M. Carceller, J.C. Martínez-Velázquez, G. Rubio, C.J. Todero Peixoto, and P. Sanchez-Lucas. Deep learning techniques applied to the physics of extensive air showers. *Astroparticle Physics*, 111:12 – 22, 2019.

[5] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *arXiv preprint arXiv:1905.03554*, 2019.

[6] D Heck, J Knapp, JN Capdevielle, G Schatz, and T Thouw. A monte carlo code to simulate extensive air showers. *Report FZKA*, 6019, 1998.

[7] Sea Agostinelli, John Allison, K al Amako, John Apostolakis, H Araujo, P Arce, M Asai, D Axen, S Banerjee, G 2 Barrand, et al. Geant4—a simulation toolkit. *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003.

[8] Francisco Carrillo-Perez, Luis Javier Herrera, Juan Miguel Carceller, and Alberto Guillén. Improving classification of ultra-high energy cosmic rays using spacial locality by means of a convolutional dnn. In *International Work-Conference on Artificial Neural Networks*, pages 222–232. Springer, 2019.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.