PROCEEDINGS OF SCIENCE

ONLINE ICRC 2021
THE ASTROPARTICLE PHYSICS CONFERENCE
Berlin | Germany
37th International Cosmic Ray Conference
12–23 July 2021

Machine learning applications on event reconstruction and identification for ISS-CREAM

Monong Yu,[a,*] **Tyler B. Anderson,**[a] **Yu Chen,**[a] **Stéphane Coutu,**[a] **Tyler LaBree,**[b] **Jason T. Link,**[c,d,†] **John W. Mitchell,**[c] **S.A. Isaac Mognet,**[a] **Scott Nutter,**[b] **Kenichi Sakai**[c,d] **and Jacob Smith**[c,d]

[a] *Dept. of Physics, Pennsylvania State University, University Park, PA 16802, USA*

[b] *Dept. Of Physics, Geology, & Engineering Technology, Northern Kentucky University, Highland Heights, KY, 41076, USA*

[c] *Astroparticle Physics Laboratory, NASA Goddard Space Flight Center, Greenbelt, MD 20771, USA*

[d] *Center for Research and Exploration in Space Science and Technology (CRESST), UMBC, Baltimore, MD 21250, USA*

[†] *No longer affiliated with NASA or CRESST as of May 2021.*

E-mail: mzy49@psu.edu

A supervised machine learning algorithm is applied to the visual representations of the energy deposits in two orthogonal views of the calorimeter of ISS-CREAM. Convolutional Neural Networks (CNNs) backed by Tensorflow are used to calibrate the sampled energy of the calorimeter and reconstruct the total primary energy of cosmic rays (CR), as well as for CR identification. The CNN regression models are trained on detailed Monte Carlo simulated events reproducing the behavior of the ISS-CREAM instrument suite, and the results indicate that a calorimeter energy reconstruction resolution of as good as 25% is achieved. The energy sampled in the calorimeter is determined with a resolution as good as 8%. The CNN classification model can reach a CR identification accuracy of up to 93%. The CR primary energy reconstruction results from machine learning methods are consistent with a simple scaling of the sampled energy. The increased accuracy of this CNN energy reconstruction comes from the additional information of the longitudinal and lateral energy deposit profiles. This machine learning approach is widely applicable to a range of particle physics and astrophysics problems.

37th International Cosmic Ray Conference (ICRC2021),
12-23 July 2021
Berlin, Germany - Online

*Presenter

PoS(ICRC2021)061

1. Introduction

Machine learning techniques, especially the convolutional neural network (CNN), have been successfully applied to image-related scenarios, such as handwriting recognition [1], house price prediction [2], or medical risk suggestion [3]. In the field of high energy physics (HEP), the detectors can be used as imaging devices, such as particle trackers or imaging calorimeters, and machine learning techniques are well-suited for analyzing the results from HEP experiments. From the classification of the hadronic decays [4] to the discovery of Higgs boson [5], machine learning techniques have a wide application in HEP studies.

In this work, we investigate the applications of CNN on analyzing ISS-CREAM data. We design the networks to 1) reconstruct the total CR primary energy, 2) check and calibrate the sampled energy of the calorimeter, and 3) identify CR events from among noise events.

2. The Input Data

2.1 The ISS-CREAM Calorimeter

The ISS-CREAM Calorimeter (CAL) detector consists of two carbon targets and a sampling electromagnetic calorimeter with 20 layers of high purity tungsten and scintillating-fiber ribbons. It measures the energy of incident CR particles in the range of $10^{12} \sim 10^{15}$ eV. The energy of an incident particle that interact in the carbon targets is proportional to the secondary particle shower sampled in the calorimeter. The scintillation light generated from the shower particles is transmitted to hybrid photodiodes (HPD), and is digitized with an analog-to-digital converter (ADC). The conversion factor of raw ADC signal values to the measured energy (in MeV) deposit in each ribbon is estimated by Monte Carlo simulations [6]. The output from the CAL for analysis can be represented by two images which provide a three-dimensional track reconstruction: energy deposition in the X-Z plane and in the Y-Z plane, with Z the vertical direction. Layers in the X-Z view are perpendicular to those in the Y-Z view. For more information about ISS-CREAM, see [7]. Figure 1 shows typical CR-induced showers developing in the CAL, adapted from the ISS-CREAM Event Display [8]. Figure 1(a) and Figure 1(b) display CAL showers for events from Monte Carlo simulations, and Figure 1(c) displays CAL showering events from ISS-CREAM on-orbit data.

2.2 Data Pre-Processing

The convolutional neural network [9] is a deep learning algorithm in computer vision. It can extract features of images by dividing the image into sub-regions with a small sliding filter (e.g., 3×3 pixels) traversing the whole image. As shown in Figure 1, particle showers measured by detectors can be represented as pixelized images which CNN can be used to analyze. In this sense, the CAL gives a 50×40-pixel image, which is a joining of 50×20 pixels from the X-Z view, and 50×20 pixels from the Y-Z view.

To make the model stably converge, reshaping the input data is necessary. For the input images, the intensity of each pixel (detector channel) is re-scaled and normalized. The particle primary energy and sampled energy in the CAL are transformed logarithmically. The tracking direction, i.e., the zenith angle, is naturally normalized.

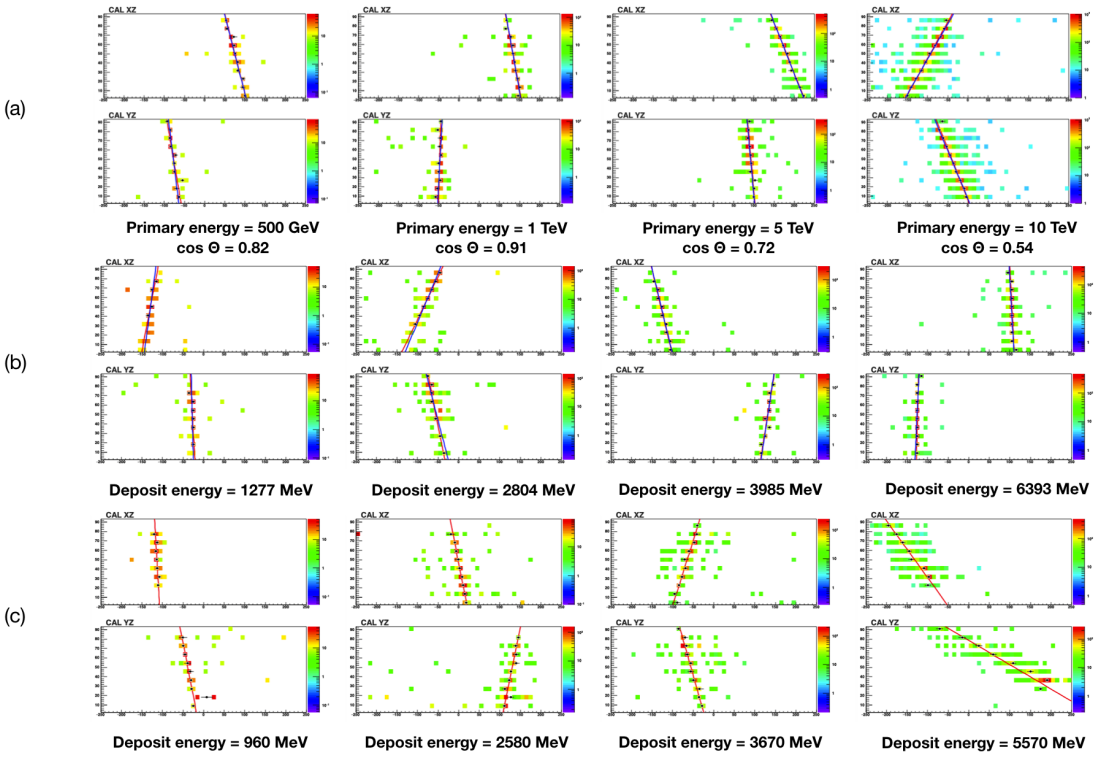

Figure 1: Row (a): examples of CAL showers for Monte Carlo events used in primary CR energy reconstruction. Row (b): examples of CAL showers for Monte Carlo events used in CAL energy scale determination. Row (c): examples of CR events used in CR identification from ISS-CREAM on-orbit data; these cleanly showering events are categorized as CR induced. Non-shower-like events are categorized as induced by noise in the CAL.

3. The Neural Network

3.1 The Network Architecture

We build networks for three purposes, and the details for each aspect are described below.

- *Primary CR energy reconstruction*

  A regression model is built for this task. The network architecture backed by Tensorflow is shown in Figure 2(a). The Input part takes the training images along with the tracking angle, then characterizes and reshapes the images by convolutional layers with a 3×3 size filter, and max-pooling layers which reduce by half the parameters of each dimension (pool size is 2×2). The densely connected part is the main body of the network. Dense layers redistribute parameter weights of the image features. A regularization layer is used to restrict overfitting with a dropout rate of 0.2. The final dense layer predicts the output.

- *CAL sampled energy determination*

  The model used in this task is similar to the previous one, except that the tracking information is not used, see Figure 2(b).

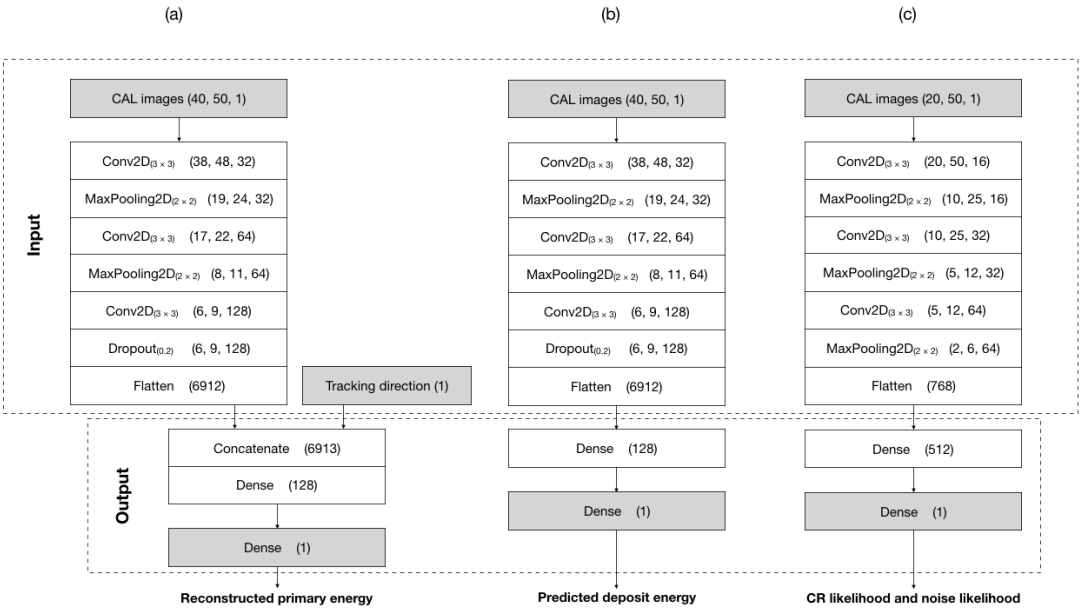

Figure 2: Schematics of the network architectures used. Column (a): network for energy reconstruction. Column (b): network for CAL energy scale determination. Column (c): network for CR identification.

- *CR identification*

  For this purpose we use a classification model. The training labels, "CR like" and "noise like", are identified by human eyes based on the image details revealed using the Event Display. The X-Z view and Y-Z view images are trained separately. The densely connected layers predict the the label of input images. The structure of this network is shown in Figure 2(c).

3.2 Network Training

We train the networks for the three aspects separately.

- *Primary CR energy reconstruction*

  This energy reconstruction task is set up as a regression problem, using the mean absolute error between predicted value and target value as a loss function. The rectified linear activation function (ReLU) is applied to the input layers, and the linear activation function is applied to the output layer. The optimizer is ADAM, with a learning rate of 0.001, exponential decay rate for the first moment estimates of 0.9 and second moment estimates of 0.999. A dropout rate of 0.2 is applied after the input layers.

  The primary energies of events are selected in the range of 500 GeV $\sim$ 100 TeV, and a pre-selection of deposited energy > 600 MeV is applied to guarantee a sufficient CAL response. The elemental species have a mixed composition of H, He, C, O, Fe with equal fractions. All events are generated by Monte Carlo. The distributions of true energy of events used in this task are shown in Figure 3. The model is trained on 40,179 showers and uses an additional 7091 samples as validation to tune the parameters. The validation process shows a learning

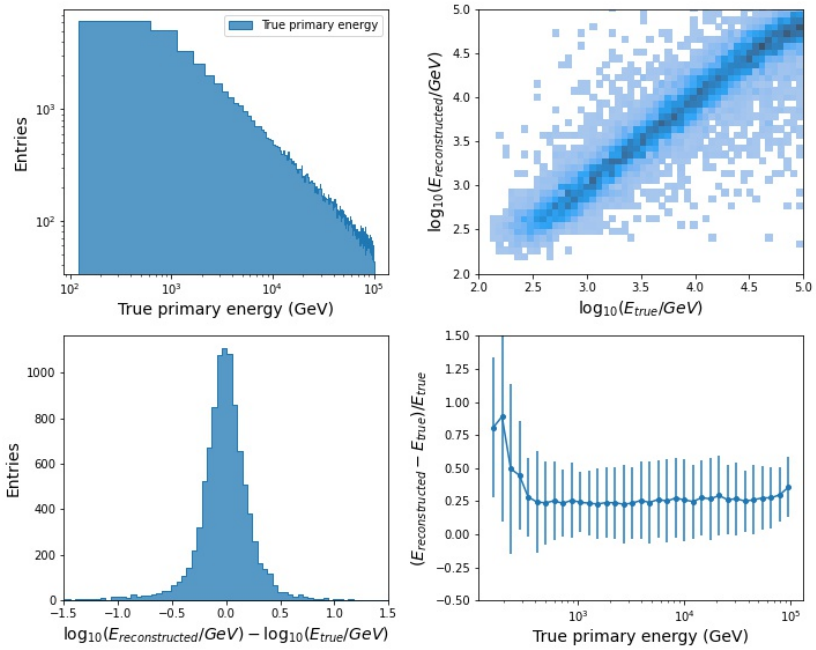

Figure 3: *Upper left:* primary true energy distribution of events used in this task. *Upper right:* distribution of reconstructed energy as a function of true energy. *Lower left:* residual distribution of the logarithm of the reconstructed energy. *Lower right:* relative energy resolution as a function of true energy.

curve that reaches a plateau after 50 epochs. Each training epoch takes ~ 37s on an Intel Core i7-6400 @3.40GHz × 8 computer with a Ubuntu 16.04 operating system. A test set of a further 10,000 samples, which does not participate in the training process, is used to provide an unbiased evaluation of the model's performance. The model's performance is illustrated in Figure 3.

- *CAL sampled energy determination*

  This task is also set up as a regression problem. The loss function is the mean absolute error plus a modified term to reduce bias and suppress overfitting. Optimizer and environment settings are the same as for the previous task. The training epoch is 100, and each epoch takes ~ 45s.

  The energy is trained as CAL deposited energy, from 100 MeV ~ 100 GeV. We also require that the training events pass certain conditions, for example, that the CAL should have 6 consecutive layers each with higher than 10 MeV energy deposit, and that the tracking position should be in geometry of other ISS-CREAM detectors. Such pre-selections improve the overall quality of CAL images by removing events that are noisy or out of geometry for this instrument. The elemental species have a mix of components from B to Fe, and the weighting of each element follows the relative abundances measured [10]. We use 22,037 events for training, 5510 events for validation, and an additional 6630 events for testing. All those events are generated by Monte Carlo. We apply stricter pre-selections for the test set

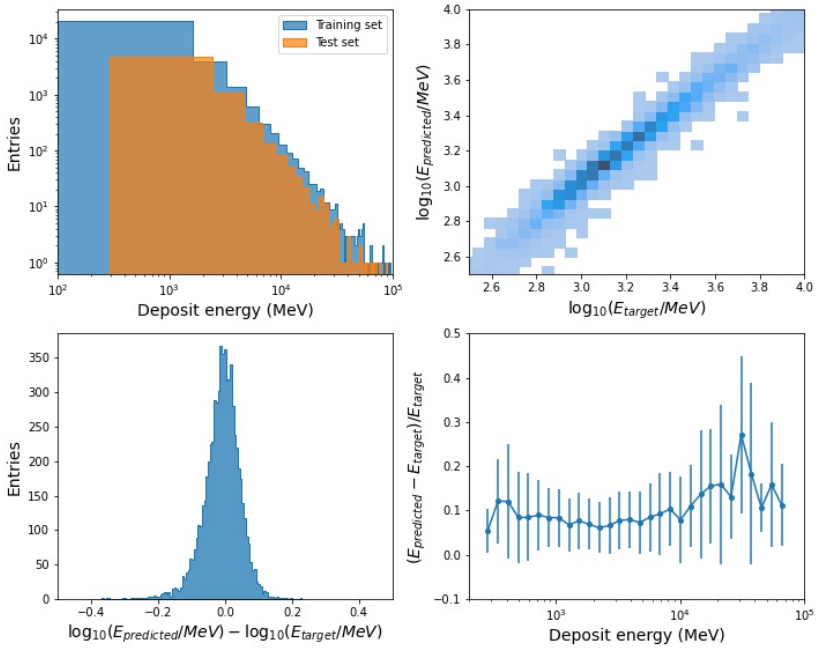

Figure 4: *Upper left:* deposited energy distribution of training events *(blue)* and deposited energy distribution of test events *(orange)*. *Upper right:* distribution of predicted energy as a function of target energy. *Lower left:* residual distribution of the logarithm of the predicted energy. *Lower right:* relative energy resolution as a function of true deposited energy.

to serve our purpose, such as trigger conditions of detectors. The distributions of target deposited energy of training events and test events in this task are shown in Figure 4.

- *CR identification*

  This task is set up as a classification problem. The loss function is sparse-categorical-cross-entropy, which computes the cross-entropy loss between true labels and predicted labels. The ReLU activation function is applied to the input layers, and the sigmoid activation function is applied to the output layer. The output layer returns a likelihood that an input image belongs to a label. Optimizer and environment settings are the same as for the previous two tasks. The training epoch is 30, and each epoch takes $\sim$ 0.5s.

  The events used in this task are selected from on-orbit flight data. We randomly select 1246 CR events and 2524 noise events for training, 624 CR events and 1263 noise events for validation, 353 CR events and 20,000 noise events for testing. The training labels, "CR like" and "noise like", are identified by human eyes based on the image details revealed using the Event Display. Since we train the X-Z view and Y-Z view separately, we define that a CR event is one where both views have likelihoods of 50% or higher being "CR like", otherwise the event is classified as noise. See Figure 5 for the model's performance on test samples.

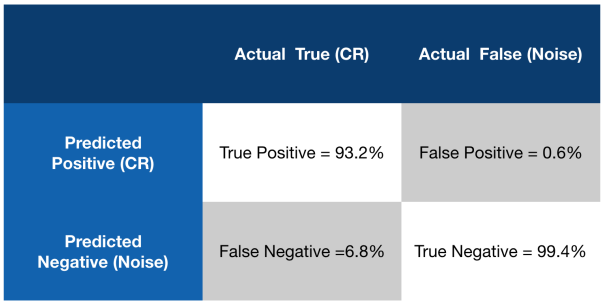

Figure 5: The confusion matrix of test samples for identification of an event as a cosmic ray.

4. Results and Discussion

From Figure 3, we see that the energy reconstruction based on a machine learning method can in principle achieve a resolution of as good as 25%. Compared to the traditional method that requires detailed Monte Carlo simulations and usually gives a resolution of $\sim 50\%$ for on-orbit performance, this method not only saves significant computing resources, but also potentially increases the precision of the data analysis technique. In this study, the results from machine learning methods are consistent with a simple scaling of the sampled energy. The increased accuracy of this CNN energy reconstruction comes from the additional information of the longitudinal and lateral energy deposit profiles. In the 100 TeV region, there is a small bias, as illustrated in the upper right panel of Figure 3. This is because our training set has fewer high-energy samples.

As for the second task, we aim to compare the energy deposition response of the CAL used in on-orbit data analysis to the Monte Carlo simulations. We use an additional 185 showering events from on-orbit data to compare with the test set generated by simulation. Figure 4 shows the performance of the model that uses the best calibration factor as input images. The results of this work show that the original absolute energy calibration is not very consistent with simulations, and the real energy calibration of the CAL is still undetermined and is under study for ISS-CREAM analysis, see [7] for the preliminary results. However, the uncertainty of the energy calibration does not negate the ability of a CNN that predicts the shower profile of deposited energies. We see that the energy resolution could in principle reach a value as good as 8%, based on simulated showering events. For more details about ongoing work to calibrate the on-orbit CAL energy deposition, see [11].

The classification task gives a true positive rate of 93.2% and a true negative rate of 99.4%, as illustrated in Figure 5. This could help us preserve most of the "CR like" events and reject a significant fraction of the noise events that triggered the instrument acquisition electronics. Although the machine-learning based method only relies on CAL images, it is as efficient in rejecting noisy events as traditional methods (trigger combinations and other data cuts), and has the advantage of being independent of any detector calibration. Our continuing ISS-CREAM data analysis uses the machine-learning based method in various event selection exercises.

5. Conclusions

This study shows three machine learning approaches applied to ISS-CREAM data analysis: reconstructing the primary CR energy, checking detector performance - in particular CAL energy scale, and event classification. The results show that these approaches have the same or even better performance compared to traditional methods, and furthermore less computing power and time are needed. This makes the analysis of complex showers straightforward. The power of machine learning tools leads to ever increasing applications in high energy physics and particle astrophysics.

Acknowledgments

This work was supported in the U.S. by NASA grants NNX17AB43G, NNX17AB42G, and their predecessor grants, as well as by directed RTOP funds to NASA GSFC. The authors also thank M. Geske, Penn State, for contributions to the BSD, and K. Wallace at Northern Kentucky University for contributions to Monte Carlo simulations. We also recognize the contributions of past CREAM and ISS-CREAM collaborators.

References

[1] Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. "A fast learning algorithm for deep belief nets." Neural computation 18, no. 7 (2006): 1527-1554.

[2] Géron, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media, 2019.

[3] Kourou, Konstantina, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. "Machine learning applications in cancer prognosis and prediction." Computational and structural biotechnology journal 13 (2015): 8-17.

[4] CMS collaboration. "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques." Journal of Instrumentation 15 (2020).

[5] Baldi, Pierre, Peter Sadowski, and Daniel Whiteson. "Searching for exotic particles in high-energy physics with deep learning." Nature communications 5, no. 1 (2014): 1-9.

[6] Ahn, H. S., P. Allison, M. G. Bagliesi, J. J. Beatty, G. Bigongiari, P. Boyle, J. T. Childers et al. "The cosmic ray energetics and mass (CREAM) instrument." Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 579, no. 3 (2007): 1034-1053.

[7] Poster 696 of Scott Nutter: Analysis Results from the Cosmic Ray Energetics And Mass Instrument for the International Space Station (ISS-CREAM), 37th International Cosmic Ray Conference (ICRC2021).

[8] Poster 1051 by Kenichi Sakai: ISS-CREAM detector performance and tracking algorithms, 37th International Cosmic Ray Conference (ICRC2021).

[9] Le Cun, Yann, Lionel D. Jackel, Brian Boser, John S. Denker, Henry P. Graf, Isabelle Guyon, Don Henderson, Richard E. Howard, and William Hubbard. "Handwritten digit recognition: Applications of neural network chips and automatic learning." IEEE Communications Magazine 27, no. 11 (1989): 41-46.

[10] Grieder, Peter KF, ed. Cosmic rays at Earth. Elsevier, 2001.

[11] Poster 866 by Yu Chen: On-Orbit Energy Calibration of the Calorimeter on the ISS-CREAM Instrument Using the Boronated Scintillator Detector, 37th International Cosmic Ray Conference (ICRC2021).