Accelerating dark matter search in emulsion SHiP detector by deep learning

S K Shirobokov[1,2], A E Ustyuzhanin[2,4] and A I Golutvin[1,3]

[1] Imperial College, Prince Consort Road, London SW7 2BZ, UK
[2] National Research University Higher School of Economics, 20 Myasnitskaya Ulitsa, Moscow, Russia
[3] CERN, 1211 Geneva 23, Switzerland
[4] National University of Science and Technology MISIS, 119049, Leninsky pr., 4, Moscow, Russia

E-mail: sergey.shirobokov@cern.ch

Abstract. We introduce a novel approach for the reconstruction of particle properties for the SHiP detector. The SHiP experiment significantly focuses on finding effects of dark matter particle interaction. A characteristic trace of such an interaction is an electromagnetic shower. Our algorithm aims to reconstruct the energy and origin of such showers using online Target Tracker subdetectors that do not suffer from pile-up. Thus, the online observation of the excess of events with proper energy can be a signal for a dark matter. Two different approaches were applied: classical, using Gaussian Mixtures and machine learning based on a convolutional neural network. We've refined the output of the previous step by clusterization techniques to improve transverse coordinate estimation. The obtained results are 25% for energy resolution, 0.8 cm for position resolution in the longitudinal direction and 1 mm in the transverse direction, without any usage of the emulsion.

1. Introduction

The SHiP experiment [1] is dedicated to the search for Beyond the Standard Model physics. In particular, one can search for Dark Matter (DM) scattering on electrons in the SHiP scattering detector. The detector consists of lead, an emulsion and target trackers (TT). In order to search for DM signatures, one must detect electromagnetic showers (EM) in either the emulsion or target trackers. The emulsion can only be analysed after half a year of exposure, thus making it sensitive to background pile up. Techniques for EM location and separation in the emulsion were studied in [2].

On the contrary, TT is an online detector and thus does not suffer from events pile up. Nonetheless, it is challenging to identify energy and position of the initial particle in a particular scattering event in TT since the sampling frequency of TT is much smaller than of the emulsion. We investigate the possibility of reconstructing energy and position of the initial particle, using only TT.

2. Problem statement

2.1. Data sample

The electromagnetic shower is a cone-like structure, consisting of particles (hits), detected in emulsion. Each hit in the emulsion has six initial features: X, Y, Z coordinates, XZ and YZ

planes projection angles - θ_x and θ_y, which fully determine the direction of the hit and χ^2 , which determines the goodness of fit of the hit. When an electromagnetic shower passes through the target tracker, the transverse proportion of the shower is detected. The resulted response of the TT is a 2D map of hits.

The resulting 2D map has a resolution of 1500x1800 pixels. The relative shape of the pixels in the picture and their intensity can indicate the energy and the position(X, Y, Z) of the vertex of the initial particle. Precise knowledge of the position and the energy is very important, since it facilitates the discrimination of a lot of background hits, as stated in [2]. Moreover the collaboration is currently considering the possibility to identify events associated with Dark Matter, using target trackers, without emulsion. The above task is even more important for such a study.

We will work with approximately 450000 events generated by Monte Carlo simulation, with training set size being 90% of the above data and rest 10% being test set size. The events are determined by the energy of the initial particle and the vertex location. The vertex location will be determined by the distance d to the first TT plane. This distance will vary from 0 to -7.5 cm, and all events are uniformly distributed in it. The energy of the particle will be denoted by E, it varies from 1 to 100 GeV and all events are also uniformly distributed in that range. The response of the detector is described by the "picture of hits".

2.2. Performance metrics

The energy of the shower is proportional to the number of hits, thus the reconstructed energy is defined as

$$E_{reco} = aN_h + b, \tag{1}$$

where N_h is the number of reconstructed hits in the detector. The quality of the algorithm is assessed by energy resolution, which is defined as

$$\sigma_E = \sigma\left(\frac{E_{true} - E_{reco}}{E_{true}}\right), \tag{2}$$

where E_{reco} is the predicted energy, E_{true} is the true energy of the shower and $\sigma(\cdot)$ denote the standard deviation function. This metric, of course, assumes that the predicted values are unbiased. The same metrics are useful to measure the performance of the vertex prediction, namely,

$$\sigma_x = \sigma(X_{true} - X_{reco}). \tag{3}$$

The ideal algorithm will be unbiased and has zero resolution.

3. Related work

Classical techniques to identify eletromagnetic shower energy and position are presented in [3, 4, 5]. There are also some recently discussed techniques [6, 7] based on manually created features, describing electromagnetic showers. Unfortunately, none of these techniques can be applied to our scenario. The reason for this, is that in this study tracking detectors are used. These detectors by construction are not suitable for reconstruction of energy of the shower. To the best of our knowledge, there have been no attempts to estimate the parameters described above using tracking detectors.

The most modern techniques, connecting calorimetry and machine learning were discussed in [8, 9]. Still, both of these approaches utilise proper calorimeters and are dedicated to generating the detector response with GANs. For example, the authors of Ref. [9] also discuss the problem of sparsity and how to approach it.

4. Gaussian fit approach

Since we are basically solving a regression problem, we can use some simple methods, like linear regression or SVM and simple features, describing the signal. For example, one can fit a Gaussian function in the middle of the TT response picture and use its variance and covariance as descriptors for the response. The idea behind this is that the two variables we want to predict, energy and Z coordinate, are correlated and the shape of the response depends on both of them.

If we now independently fit three 2D Gaussians to each of the target tracker planes, we will get 3x3 features and the number of hits in each of the tracker - a total of 12 features.

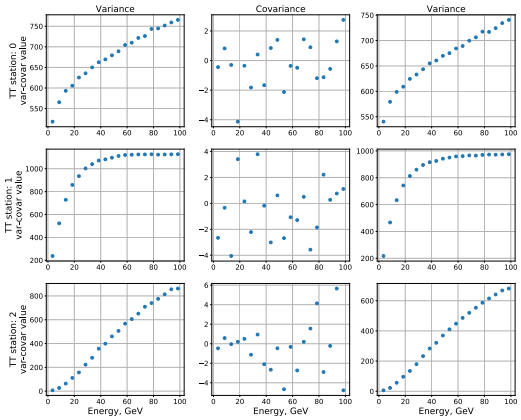


Figure 1: Features from fitting Gaussian to the detector response. Rows represent Target Tracker stations, columns represent variance in X, covariance and variance in Y respectively.

The dependence of the features on energy are presented in Fig 1. We can then fit a simple regression on these features, such as a Linear Regression or a XBoost regression. The results are presented in the Fig. 2. The corresponding RMSE are 7 GeV for energy and 0.92 cm for distance.

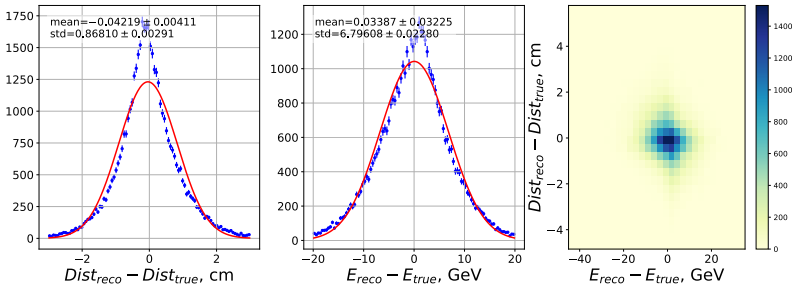


Figure 2: Distribution of errors for Gaussian regressor.

5. Convolutional Neural Network approach

Since our response is basically an image, we can represent it by stacking detectors as "color" maps. Then, we can devise an architecture of the network and a loss function, that will perform well. With the basic Convolutional Neural Network(CNN) architecture we get 7 GeV RMSE for the energy and ~ 0.87 cm for the distance.

With minimising the Huber loss as the objective, we were able to achieve better performance. The Huber loss is defined as

$$Loss_{Huber} = \frac{1}{n}\sum_i \begin{cases} 0.5(\hat{y}_i - y_i)^2, & \text{if } |\hat{y}_i - y_i| < 1 \\ |\hat{y}_i - y_i| - 0.5, & \text{otherwise} \end{cases}$$

Due to faster convergence and better robustness, the CoordConv [10] layers have been used in the final architecture. The benefit of such a convolution is that it can learn local aware features, but unlike locally-connected convolution, it adds only $O(K)$ additional parameters, where K is the size of the kernel. The net was trained with $Loss_{Huber}$, where targets were initially scaled to the same range. The training process with gradual L_2 weight regularisation removal was applied. The initial regularisation constant is set to 0.01 and decreased by half every ten epochs. The training loss and validation metrics for the training process are presented in Fig. 3

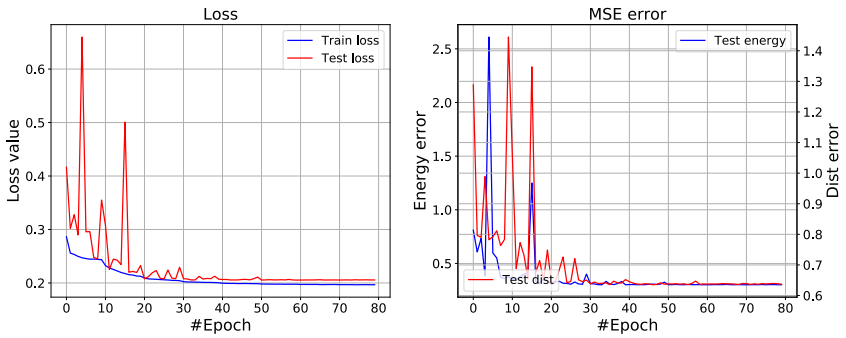


Figure 3: Training loss and validation metrics dynamics for CNN with CoordConv.

The final metrics for this approach resulted in 7 GeV RMSE for the energy and ~ 0.8 cm for the distance. The results are shown in the Fig 4.

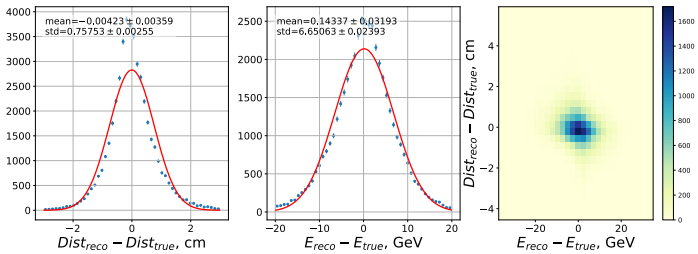


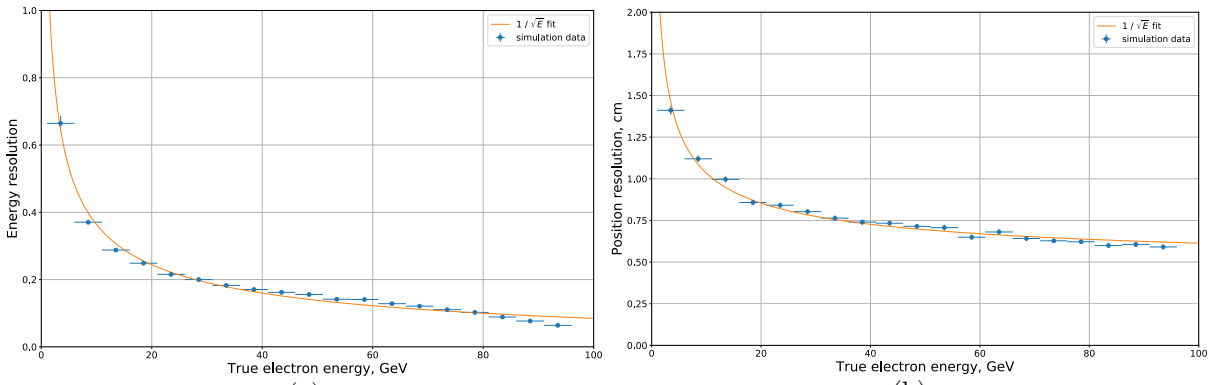
Figure 4: Distribution of errors for CNN regressor with weights regularisation coefficient decay.

In the Fig. 5a and Fig. 5a the final physics metrics are shown. As determined by physical laws, the energy resolution should follow $\sim 1/\sqrt{E}$ dependency. The same rule applies to the position resolution due to common ideas. As one can see, the fit of both metrics lies within the uncertainty error of the bins. This verifies that the predictions of the regressor follows the above physical laws.

The comparison of the algorithms described above is presented in table 1.

6. Transverse shower origin position reconstruction

As discussed in the problem statement, one more task was to reconstruct the (X,Y) coordinates of the vertex. This task is much simpler, since the response of the detectors explicitly contains information about the (X,Y) coordinates.



(a) (b)

Figure 5: (a) Position resolution, (b) energy resolution as a function of energy of the shower. The orange line shows $1/\sqrt{E}$ fit.

Algorithm	RMSE: E, GeV; d, cm	Resolution: E, %; d, cm
Gaussian fit	7.23; 0.92	26; 0.92
CNN, MSE loss	7.00; 0.87	25; 0.87
CNN, Huber loss	6.84; 0.79	24; 0.79
CoordConvNN, Huber loss	6.84; 0.79	24; 0.79

Table 1: Algorithms comparison

We have applied the (H)DBSCAN algorithm [11, 12] to each map of the response picture. The algorithm is presented in alg. 1.

Algorithm 1: (X,Y) coordinates of vertex location

```
TT_stations = sort(TT_stations, key=number of hits);
for station in TT_stations do
    n_clusters = FindClusters(station);
    if n_clusters > 0 then
        for cluster in n_clusters do
            cluster_centers.add(FindCenter(cluster));
        break;
if len(cluster_centers) > 0 then
    SelectBestCluster(cluster_centers);
else
    reject event;
```

where

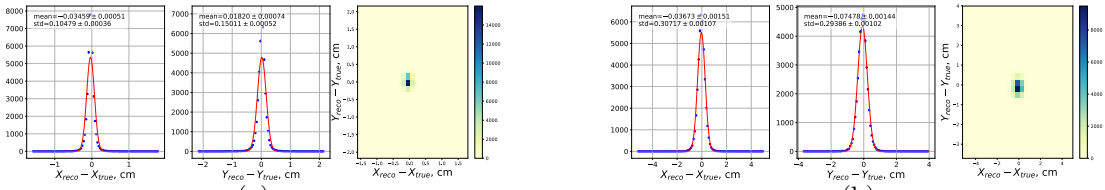
$$FindCenter(cluster)_{x,y} = \frac{\sum_{i \in cluster}(x,y) * I_i}{\sum_{i \in cluster} I_i},$$

with I_i being the intensity of the hit. SelectBestCluster algorithm works as follows: It performs a linear fit between all possible combinations of the cluster centers in all the TT stations. It selects those cluster centers in each TT station, which minimises MSE error of the fit. Nevertheless we perform a combinatorial number of operations to perform this fit, this is feasable in pracite, since the number of clusters is of order 2 to 3 in each TT station.

As a FindClusters function we have tested DBSCAN, HDBSCAN and GaussianMixtures and

selected DBSCAN as the best option.

The obtained position resolution is ~ 0.1 cm, whereas the baseline solution (locating center of the cluster just as a weighted mean of all the hits the TT station) provides only ~ 0.3 cm resolution. Results are shown in Fig. 6a, 6b.



(a) (b)

Figure 6: (a) XY resolution with clustering. (b) XY resolution without clustering.

7. Conclusion

We provide particle energy and position reconstruction algorithms for the SHiP experiment given only online detector output. Our approach shows comparable resolution in energy and position to classical emulsion algorithms [13] for high energy events. A Shallow convolutional neural network with special (CoordConv) convolutional layers, trained with gradual regularisation, was used. A resolution of 24% in energy and 0.8 cm in longitudinal direction was obtained. The (H)DBSCAN was used for the prediction of the transverse coordinates. A resolution of ~ 0.1 cm in the transverse direction was obtained.

Acknowledgement

The research leading to these results has received funding from Russian Science Foundation under grant agreement n 17-72-2012.

[1] W. Bonivento *et al.*, "Proposal to Search for Heavy Neutral Leptons at the SPS," 2013.

[2] A. Ustyuzhanin, S. Shirobokov, V. Belavin, and A. Filatov, "Machine-Learning techniques for electromagnetic showers identification in OPERA datasets," *ACAT 2017 conference proceedings*, 2017.

[3] G. Akopjanov et al., "Particle identification in a hodoscope Cherenkov spectrometer," *Nucl. Instr and Meth., pp. 441*, 1977.

[4] J. Gomez et al., "Particle identification in modular electromagnetic calorimeters," *Nucl. Instr and Meth., pp. 284*, 1987.

[5] T. Awes et al., "Charged-particle distributions in 16O induced nuclear reactions at 60 and 200 A GeV," *Nucl. Instr and Meth., pp. 130*, 1987.

[6] A. Babeanu, "Electromagnetic Shower Recognition with a Forward Calorimeter for the ALICE Experiment," *Master Thesis, UU(SAP) 13-7*, 2013.

[7] M. Mazouz, L. Ghedira, and E. Voutier, "Determination of shower central position in laterally segmented lead-fluoride electromagnetic calorimeters," *JINST*, vol. 11, no. 07, p. P07001, 2016.

[8] L. de Oliveira, M. Paganini, and B. Nachman, "Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis," *Comput. Softw. Big Sci.*, vol. 1, no. 1, p. 4, 2017.

[9] M. Paganini, L. de Oliveira, and B. Nachman, "CaloGAN : Simulating 3D high energy particle showers in multilayer electromagnetic calorimeters with generative adversarial networks," *Phys. Rev.*, vol. D97, no. 1, p. 014021, 2018.

[10] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," *CoRR*, vol. abs/1807.03247, 2018.

[11] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," in *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, SIGMOD '99, (New York, NY, USA), pp. 49–60, ACM, 1999.

[12] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[13] D. Di Ferdinando, "Nuclear emulsions in the OPERA experiment," *Radiat. Meas.*, vol. 44, pp. 840–845, 2009.