Deep Learning for Astrophysics, Understanding the Impact of Attention on Variability Induced by Parameter Initialization

Mikaël Jacquemont, Thomas Vuillaume, A Benoit, Gilles Maurin, Patrick Lambert

Deep Learning for Astrophysics, Understanding the Impact of Attention on Variability Induced by Parameter Initialization*

Mikaël Jacquemont[1,2] [0000−0002−4012−6930], Thomas Vuillaume[1] [0000−0002−5686−2078], Alexandre Benoit[2] [0000−0002−0627−4948], Gilles Maurin[1] [0000−0002−6970−0588], and Patrick Lambert[2] [0000−0003−0478−9443]

[1] CNRS, LAPP, Univ. Grenoble Alpes, Université Savoie Mont Blanc, Annecy, France
{jacquemont,vuillaume,maurin}@lapp.in2p3.fr
[2] LISTIC, Univ. Savoie Mont Blanc, Annecy, France
{alexandre.benoit,patrick.lambert}@univ-smb.fr

Abstract. In the astrophysics domain, the detection and description of gamma rays is a research direction for our understanding of the universe. Gamma-ray reconstruction from Cherenkov telescope data is multi-task by nature. The image recorded in the Cherenkov camera pixels relates to the type, energy, incoming direction and distance of a particle from a telescope observation. We propose γ-PhysNet, a physically inspired multi-task deep neural network for gamma/proton particle classification, and gamma energy and direction reconstruction. As ground truth does not exist for real data, γ-PhysNet is trained and evaluated on large-scale Monte Carlo simulations. Robustness is then crucial for the transfer of the performance to real data. Relying on a visual explanation method, we evaluate the influence of attention on the variability due to weight initialization, and how it helps improve the robustness of the model. All the experiments are conducted in the context of single telescope analysis for the Cherenkov Telescope Array simulated data analysis.

Keywords: multitasking · artificial neural networks · gamma rays · attention · visual explanation.

* We gratefully acknowledge financial support from the agencies and organizations listed here: www.cta-observatory.org/consortium_acknowledgment. This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 653477, and from the Fondation Université Savoie Mont Blanc. This work has been done thanks to the facilities offered by the Univ. Savoie Mont Blanc - CNRS/IN2P3 MUST computing center and HPC resources from GENCI-IDRIS (Grant 2020-AD011011577) and computing and data processing ressources from the CNRS/IN2P3 Computing Center (Lyon - France). We gratefully acknowledge the support of the NVIDIA Corporation with the donation of one NVIDIA P6000 GPU for this research.

1 Introduction

Gamma-ray astronomy aims to study astronomical phenomena (supernova remnants, dark matter annihilation...) based on the gamma radiation generated by these phenomena. The analysis of this radiation is performed through the observation of telescope images of the particle shower resulting from the penetration of high-energy particles in the atmosphere (Cherenkov effect [9], see Fig. 1).
The purpose of the image analysis is twofold:

1. identifying the gamma rays in the cosmic ray background mainly composed of protons (with a signal-to-noise ratio typically lower than 1/1000), which is a classification problem,
2. estimating the energy and direction of the identified gamma rays, which is a regression problem.

The Cherenkov Telescope Array (CTA)[1] is the next generation of Imaging Atmospheric Cherenkov Telescopes (IACTs). Composed of ~ 100 telescopes of different sizes, it will improve sensitivity and accuracy in gamma-ray analysis. However, the huge amount of data (210 PB of raw data per year when in full operation) requires a shift towards new methods of analysis, in particular deep neural network approaches. The work presented in this paper is carried out on the large-scale simulation data shared within the CTA international collaboration. As ground truth is not available in the field of astrophysics, the analysis toolchains are prepared with very high-quality simulation relying on a well-known physical model of the phenomenon and on the detector simulation [2]. Besides, the first real data are just available.

In this paper, we first propose a new deep multi-task architecture, named γ-PhysNet, taking into account physics considerations and designed for single telescope gamma event analysis. We evaluate this model on the simulation data of the Large Size Telescopes 1 (LST1 [1]), the first prototype installed at the Northern CTA site in La Palma. In a second step, with the help of a visual explanation method for neural networks, we analyze the impact of attention (mechanism that reinforces relevant features) on the variability introduced by model weights initialization, and show that augmenting the network with attention improves the robustness of the model.

The rest of the paper is organized as follows. In Section 2, we give a short state of the art related to this work. Section 3 is a presentation of γ-PhysNet. Understanding the impact of attention mechanism and weight initialization are discussed in Section 4. Finally, Section 5 gives conclusions and some perspectives for future works.

2 Related Work

Over the past decade, deep learning has emerged as the leading approach in many computer vision tasks. IACT data has not escaped this trend [16,18,20,25].

[1] https://www.cta-observatory.org/

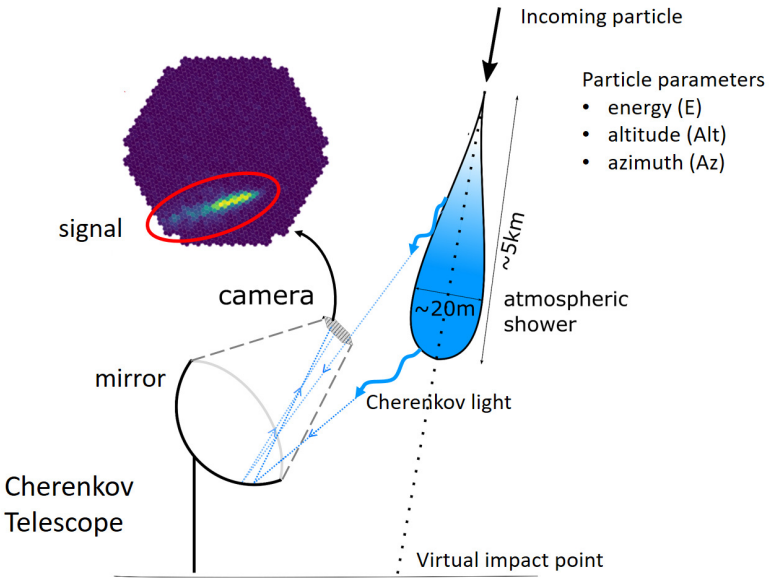

Fig. 1. Imaging Atmospheric Cherenkov Telescope. When a gamma particle enters the atmosphere, it generates an electromagnetic shower of secondary particles emitting Cherenkov light. This blue light is captured by the telescope on the ground and recorded as the signal by the camera. The event reconstruction consists in analyzing this signal to retrieve the physical parameters (type, energy, altitude, azimuth) of the primary particle.

These papers present promising results, especially for gamma/proton classification. However, they have handled the different reconstruction problems as single tasks, without considering their strong interdependence. Further, multiple single task models can increase the computational cost, thus making their use complicated in our massive data context.

Multi-Task Models. When dealing with a multi-task problem, it has been shown [29] that transferring knowledge across related tasks improves the generalization of the deep models with fewer data. Thus, in Multi-Task Learning (MTL), the different tasks to address are trained simultaneously, using a partially shared modeling. In hard parameter sharing architectures, the most frequently used, a whole part of the network (generally the encoder or its first layers) is shared between all tasks [22]. In soft parameter sharing architectures [4], each task is learned with its own network. Some additional layers are shared and constrained in order to encourage their weights to be similar. In MTL, balancing the tasks is the critical point. For most of the MTL related papers [15, 21], this is done manually, requiring an extensive optimization process. However, adaptive methods have been proposed, with different balancing strategies such as modeling homoscedastic uncertainty for each task [12], using the task loss gradients [5],

using learning progress signals as key performance indicators [6] or regarding the problem as a multi-objective optimization [24].

Complementary to MTL, the development of deep learning models integrates attention mechanisms.

Attention Mechanisms. Attention is a mechanism that helps deep learning model focus on relevant features based on a defined context through trainable weights. A distinction can be made between restricted self-attention, focusing on local spatial neighborhoods and global self-attention as a non-local operation [31]. Local and global self-attention can be considered as spatial attention mechanisms, as they capture short- and long-range dependencies in data, by weighting each pixel. On the contrary, Hu *et al.* [10] introduce a lightweight channel-wise attention denoted Squeeze-and-Excitation. The squeeze operation produces descriptors for each input channel, and is followed by an adaptive recalibration, the excitation, and a scale operation that weights the input channels. Dual Attention [28] has been proposed to improve model interpretability and robustness of U-Net models for a semantic segmentation task. It combines Squeeze-and-Excitation and a simple spatial attention path. The latter first compresses the number of input channels to one, and then applies a sigmoid function to the resulting pixel values, and adds one to them to produce an attention map. This way, the spatial attention map can only increase pixel values. This map is then used to rescale the output of the Squeeze-and-Excitation. Such attention strategy then allows for the weighting of both spatial and channel-wise information in a low-cost operator.

Model Explanation. While traditional expert systems are highly interpretable and explainable, deep neural networks are often perceived as *black boxes*. Recently, the deep learning community has put an increasing effort on opening the black box. Some methods explore the role of individual neurons or linear combination of units through ablation [17,32] or optimization [19]. Others estimate the importance of input features for a particular output activation. They produce saliency maps [3,26,27] also called heatmaps in [23] for network visualization. In the rest of this paper we rely on Grad-CAM [23], a highly class-discriminative localization method that produces a heatmap of the region that is the most relevant for the model's decision. It is applicable to any network structure. The heatmaps are computed based on the last convolutional layer. In the context of multi-task learning, Grad-CAM is then especially adapted to hard parameter sharing networks that share a convolutional encoder between all tasks. In addition, when available, attention maps, produced by attention mechanisms (see previous section), can provide insights on the model regions of interest that can explain the predictions.

3 Proposed Architecture and Performance

To achieve full event reconstruction from IACT data, we propose a MTL architecture, denoted γ-PhysNet. We then compare the vanilla model to a version

augmented with attention modules. Our aim is to understand model robustness against weight initialization, and the regularization effect provided by attention mechanisms. Model performances are assessed relying on different random seeds.

In this paper we restrict our analysis to a specific model with optional attention modules. This model is a performing baseline that originates from preliminary extensive comparative studies, and the scope of the paper is to study its robustness towards weight initialization variability.

3.1 γ-PhysNet Architecture

As illustrated in Fig. 2, γ-PhysNet is a hard parameter sharing architecture composed of a backbone encoder and a multi-task block inspired by the physics of the reconstruction. The network is fed with two-channel IACT data discussed in Section 3.2. It performs gamma rays from background noise separation, and primary particle energy and arrival direction regression as the altitude and azimuth. The regression of the virtual impact point on the ground of the particle is an auxiliary task (see 1 for the meaning of these characteristics). Even though it is not required for higher-level analysis, physics shows that this parameter provides meaningful information to solve the other tasks. The baseline backbone of γ-PhysNet is the convolutional part of a ResNet-56 [7, 8], CIFAR-10 version, with full pre-activation implemented with IndexedConv [11]. The latter allows the direct processing of the hexagonal pixel images of the LST1 data used in this paper without transforming them to square pixel ones. We also propose a refined version of the model, named γ-PhysNet DA (DA for Dual Attention), that makes use of attention. Dual Attention modules are inserted into the backbone after every stage, i.e., processing scale, to benefit from attention at each feature scale. As mentioned in Section 2, Dual Attention is composed of a spatial attention path and a channel-wise attention path. The latter consists of a Squeeze-and-Excitation module that has a ratio parameter to control its bottleneck. An extensive study of this hyperparameter has shown that, in the context of the experiments carried out for this paper, the default ratio of 16 allows obtaining the best results. The physically inspired multi-task block finalizes the model. It is composed of a global feature network and a local feature one, both based on fully connected layers. The global feature part, dedicated to energy regression, starts with a global average pooling. This strategy follows the physics of the phenomenon, as for a given arrival direction and impact point, the amplitude of the image is roughly proportional to the primary gamma ray energy [30]. The local feature part is devoted to gamma/proton classification, and regression of the arrival direction and impact point. It is fed with the flattened feature maps produced by the backbone. The aim is to exploit local (the shape of the signal in the image) and spatial (the position and orientation of the signal) information that is more deeply related to the particle type, its arrival direction and virtual impact point.

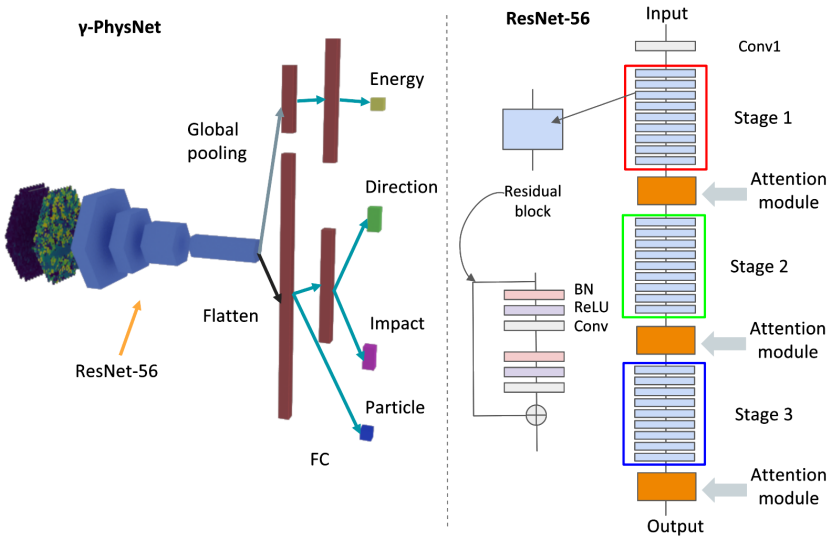

Fig. 2. γ-PhysNet. *Left:* the model architecture composed of a convolutional backbone (ResNet-56) and a physically inspired Multi-task head block based on fully connected layers (FC). The latter is divided into two paths: a global path for the energy regression and a local path for the particle type classification, the direction and the virtual impact point regression. *Right:* refined model backbone with Dual Attention modules inserted after each stage.

3.2 Experiments

Dataset. As ground truth is impossible to obtain from real data, γ-PhysNet and γ-PhysNet DA are evaluated on the *LST4 mono-trigger production* (from 2019/04/15), the reference large-scale Monte Carlo production generated by the LST international consortium for the LST1 commissioning. This dataset has been calibrated and integrated with DL1DataHandler [13]. Each sample corresponds to a single event (particle) and is composed of two-channel images: one with pixel intensities in number of photoelectrons and the other one containing per-pixel temporal information in nanoseconds. Data amplitude is not normalized since it is related to the energy of the detected particles. A data selection step is applied, following the standards in the domain and the project collaboration. It consists of a series of relatively loose selection cuts on image amplitude, shower size and truncated showers applied to the data in order to discard bad quality events that would not be reconstructed by standard methods either. The training set is finally composed of $874k$ gamma events and $506k$ proton events, the validating set $201k$ and $136k$, and the test set $209k$ and $38k$.

Training. In order to analyze the robustness of both models to parameter initialization, we repeat the experiments with six different random seeds. We define the following optimization criteria: cross-entropy loss for the classification task

and the $L1$ loss for each regression task. Models are trained for 25 epochs with Adam [14] as the optimizer, and a weight decay of 10^{-4} for regularization purpose. The learning rate is set to 10^{-3}, decayed by a factor of 10 every 10 epochs. The different tasks are balanced with the uncertainty estimation method presented in [12]. In this setup, both models reach their best performance plateau on the validation set. In gamma-ray astronomy, proton events are considered as background noise. To prevent them from penalizing the learning of energy and direction regression for gamma events, we rely on a masked loss method, setting to zero the loss of the regression parameters (energy, arrival direction and impact point) when particles are protons.

Evaluation Metrics. The performance on energy and direction reconstruction tasks is measured through resolution curves. The angular resolution is defined as the 68% containment radius of the point-spread (distribution) function and the energy resolution as the 68% containment radius of the relative absolute deviation. Lower resolutions are better.

Weight initialization plays an important role in neural network performance. We then repeat the experiment six times for both models, and we illustrate the variability of these different runs by drawing the average resolution curve per energy bin, the surfaces representing the standard deviation. The latter, referred to as dispersion in this paper, serves as measure of the robustness of the models.

For the gamma/proton classification task, the overall performance of the network is given by the area under the ROC curve (AUC), the precision and the recall.

Results. As shown in Table 1, both models with and without attention have comparable results on the classification task, within the standard deviation range. However, for the energy and direction regression, as illustrated in Fig. 3, the model with Dual Attention (γ-PhysNet DA) obtains slightly better average results above 100 GeV. Furthermore, we observe that the network with attention has significantly less spread results. In particular, γ-PhysNet DA has a constantly lower dispersion on the direction reconstruction task. On the energy reconstruction one, at energies above 200 GeV, the model without attention has dispersion up to three times higher. This lower dispersion of the results of γ-PhysNet DA denotes a better robustness to parameter initialization. This will lead to a more reliable estimation of the particle parameters when we analyze real data.

4 Understanding the Impact of Dual Attention

As we have seen in Section 3.2, the addition of Dual Attention modules to γ-PhysNet improves the robustness of the model, especially for the energy and arrival direction reconstruction tasks, by reducing the dispersion due to parameter initialization during the training step. However, it has no clear impact on the classification task. To understand how and why attention acts on the model predictions, we carefully observe the Grad-CAM heatmaps produced by γ-PhysNet

Table 1. Classification performance of both models with and without attention. The AUC represents the overall performance, while the precision shows the ability of the model to discard protons (the background noise), and the recall highlights the ability of the model to retrieve gammas.

Model	AUC	Precision	Recall
γ-PhysNet	0.882±0.001	0.929±0.001	0.935±0.007
γ-PhysNet DA	0.882±0.001	0.929±0.001	0.935±0.005

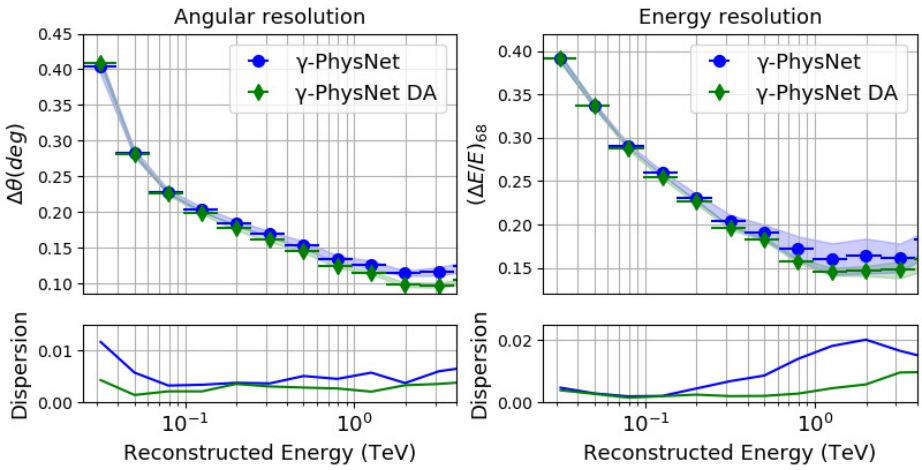

Fig. 3. Angular (*left*) and energy (*right*) resolution as a function of the energy in the LST1 energy range (lower is better). Both resolution curves represent the error of the model for the regression of respectively the direction and the energy of the detected gamma particle. The dispersion, representing the variability induced by weight initialization, is a measure of the robustness of the models.

and γ-PhysNet DA for 25 well and badly reconstructed events of the test set. We also observe the spatial attention maps of the three Dual Attention modules. Additionally, for each input data and model seed, we combine these three spatial attention maps with the Hadamard product to obtain a global representation of the spatial attention in the model. Finally, to highlight the variability brought by the different initialization seeds for a particular event, we compute the mean and standard deviation of Grad-CAM heatmaps and spatial attention maps at the pixel level (to lighten the paper, only these means and standard deviations will be presented in the following figures).

Observation of the Grad-CAM Heatmaps and the Spatial Attention Maps. For all the 25 events analyzed, we observe a common trend in the averaged heatmaps produced by the Grad-CAM. Fig. 4, Fig. 5 and Fig. 6 illustrate 3 typical examples respectively, a well-reconstructed but partially truncated gamma shower, a

well-reconstructed and centered gamma shower and a badly reconstructed one. It is worth noticing that the different maps are represented using different color scales. For the classification task (denoted "class" in the figures), with or without attention, the most relevant pixels highlighted by Grad-CAM are located in the signal area. On another hand, for the regression tasks (denoted "Energy", "Altitude" and "Azimuth" in the figures) we observe different behaviors. Without attention, the most relevant ones are situated on the border of the camera, while the signal pixels are of less importance. In our understanding, this phenomenon serves as an evaluation of the signal outside of the camera, and thus not acquired, that is important for regression tasks. With attention, all the relevant pixels are located in the shower area, thus better taking into account the signal pixels and relevant information. Besides, for all tasks, the model with attention focuses on a larger part of the image. It is worth noticing that, in addition to the signal pixels themselves, pixels situated in the signal neighborhood contain useful information about the shower shape. Then, the deviation measures (denoted "std" in the figures) of Grad-CAM heatmaps highlight the same general trend related to the robustness against the initialization. Although the variability of pixel relevance is quite important in both cases, for γ-PhysNet with attention the relevant pixels fluctuate in the signal area, while without attention they vary on a larger extent between the shower area and the image boundaries.

Then, the observation of the spatial attention maps of γ-PhysNet DA shows that the action of Dual Attention is different depending on the feature scale. The output of stage 1 has the same resolution as the input data. At this scale, the attention module mainly focuses on the shower pixels, and strongly rescales their value. The value of the signal pixels is multiplied by up to 1.8 in average, while, by construction, the rescaling factor computed by the spatial attention path of Dual Attention modules ranges in [1, 2], as explained in Section 2. However, the attention maps are quite noisy at this scale, as also highlighted by the deviation measures. After stage 2, the attention modules also strongly favor the pixels in the signal area, and the attention maps are less noisy. Then, the last attention modules, after stage 3, have a lighter impact on the feature maps values. Finally, the observation of the combined spatial attention maps highlights that attention strongly helps the model focus on the signal pixels, which is consistent with the observation of the Grad-CAM heatmaps.

Impact of the Attention on the Classification Task. Table 1 shows no significant effect of the attention of the classification task. This can be explained by the attention and Grad-Cam heatmaps that always focus on the event shower area. More into the details, attention allows for a larger extent of the heatmaps around the shower and also avoids interest on the noisy boundaries of the images. However since both models report similar predictions performances, one can expect that they rely on the same most contributing, shower centered, features.

Impact on the Regression Tasks. For the energy and direction reconstruction tasks, Dual Attention forces the model to focus on the shower area instead of the image border, by strongly rescaling the signal pixels and their close neigh-

bors. This is significant enough to improve the results of both tasks presented in Section 3.2, and to reduce the dispersion introduced by the parameter initialization.

Besides, it is worth noticing that for the event 16 represented in Fig. 4, the most important pixels are still on the border of the image. Indeed, the signal is truncated, and the network has learned that it is an important information to take into account for the estimation of the energy and the direction of the gamma ray. However, focusing on the signal part of the image does not automatically imply a good reconstruction, as exemplified by the badly reconstructed event 24 shown in Fig. 6.

5 Conclusion

In this paper we have presented γ-PhysNet, a deep multi-task architecture for gamma-ray full-event reconstruction for IACT single telescope images. We have shown that augmenting the model with attention allows for a reduction of the performance variability induced by parameter initialization. Relying on a visual explanation method, we have then realized the first steps to understand how attention modifies the behavior of the model. In a future work, in order to deepen this analysis, we will define statistical criteria to quantify this effect. Correlating the pixels highlighted by Grad-CAM and the true signal over the whole test set is an option. We also consider analyzing statistically the distance of the relevant pixels to the shower centroid. Finally, it would also be interesting to study if the robustness brought by attention mechanisms also holds with slightly different datasets (different levels of noise, altered images, etc.).

References

1. Ambrosi, G., Awane, Y., Baba, H., et al., for the CTA Consortium: The Cherenkov Telescope Array Large Size Telescope. Proceedings of the 33rd International Cosmic Ray Conference pp. 8–11 (2013). https://doi.org/10.1117/12.2054605
2. Bernlöhr, K., Barnacka, A., Becherini, Y., Bigas, O.B., Carmona, E., Colin, P., Decerprit, G., Di Pierro, F., Dubois, F., Farnier, C., et al.: Monte carlo design studies for the cherenkov telescope array. Astroparticle Physics **43**, 171–188 (2013)
3. Cao, C., Liu, X., Yang, Y., et al.: Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2956–2964 (2015)
4. Cao, J., Li, Y., Zhang, Z.: Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4290–4299 (2018)
5. Chen, Z., Badrinarayanan, V., Lee, C.Y., Rabinovich, A.: GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 794–803. PMLR (10–15 Jul 2018)

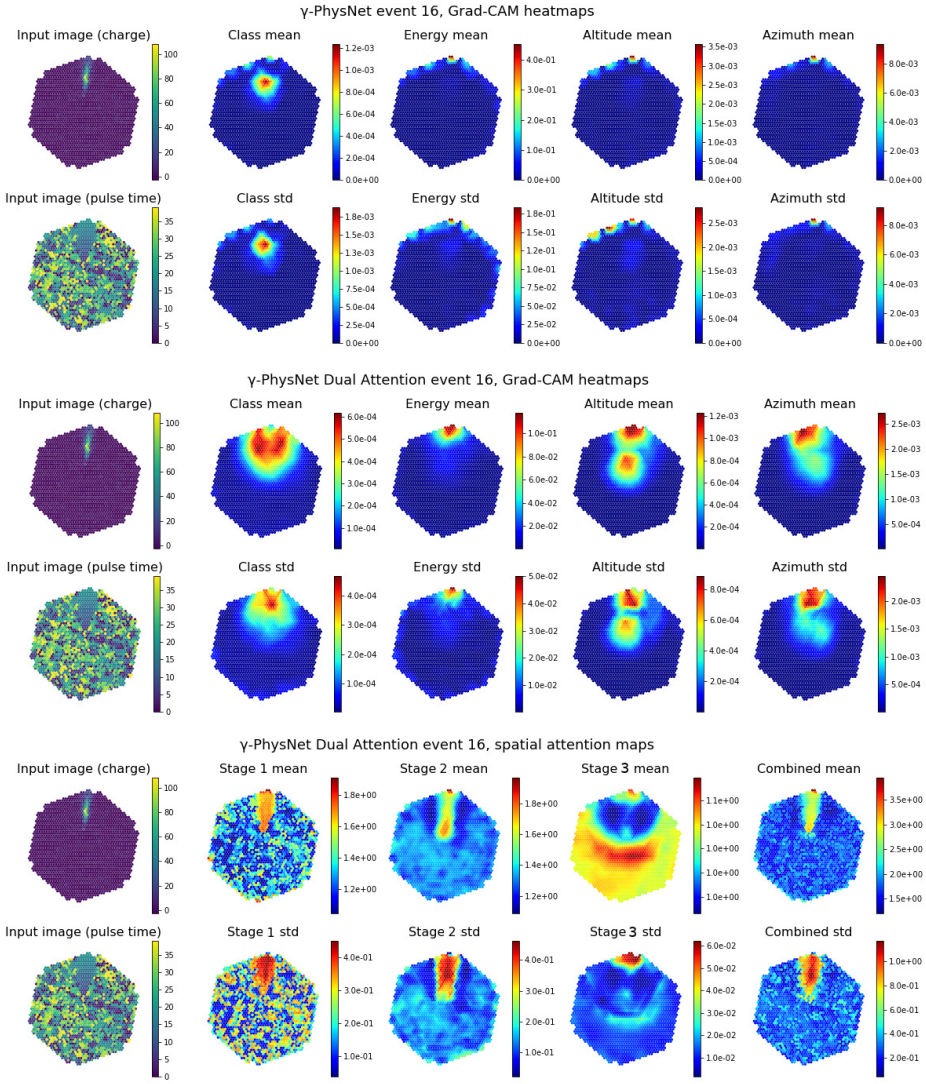

Fig. 4. Well-reconstructed gamma, event 16. The two upper rows represent the Grad-CAM heatmaps (mean and standard deviation) obtained for the vanilla γ-PhysNet, the next two rows correspond to the Grad-CAM heatmaps of γ-PhysNet DA (with attention), and the last two rows show the corresponding spatial attention maps.

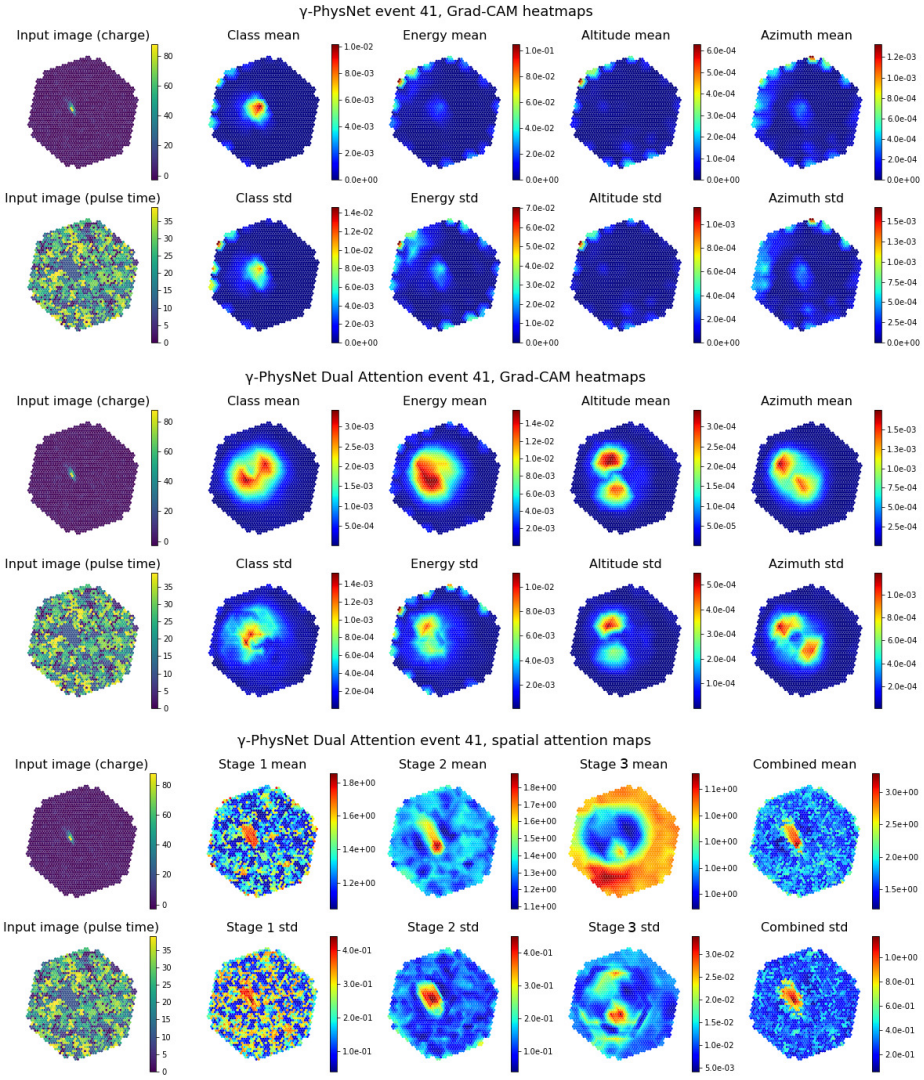

Fig. 5. Well-reconstructed gamma, event 41. The two upper rows represent the Grad-CAM heatmaps (mean and standard deviation) obtained for the vanilla γ-PhysNet, the next two rows correspond to the Grad-CAM heatmaps of γ-PhysNet DA (with attention), and the last two rows show the corresponding spatial attention maps.

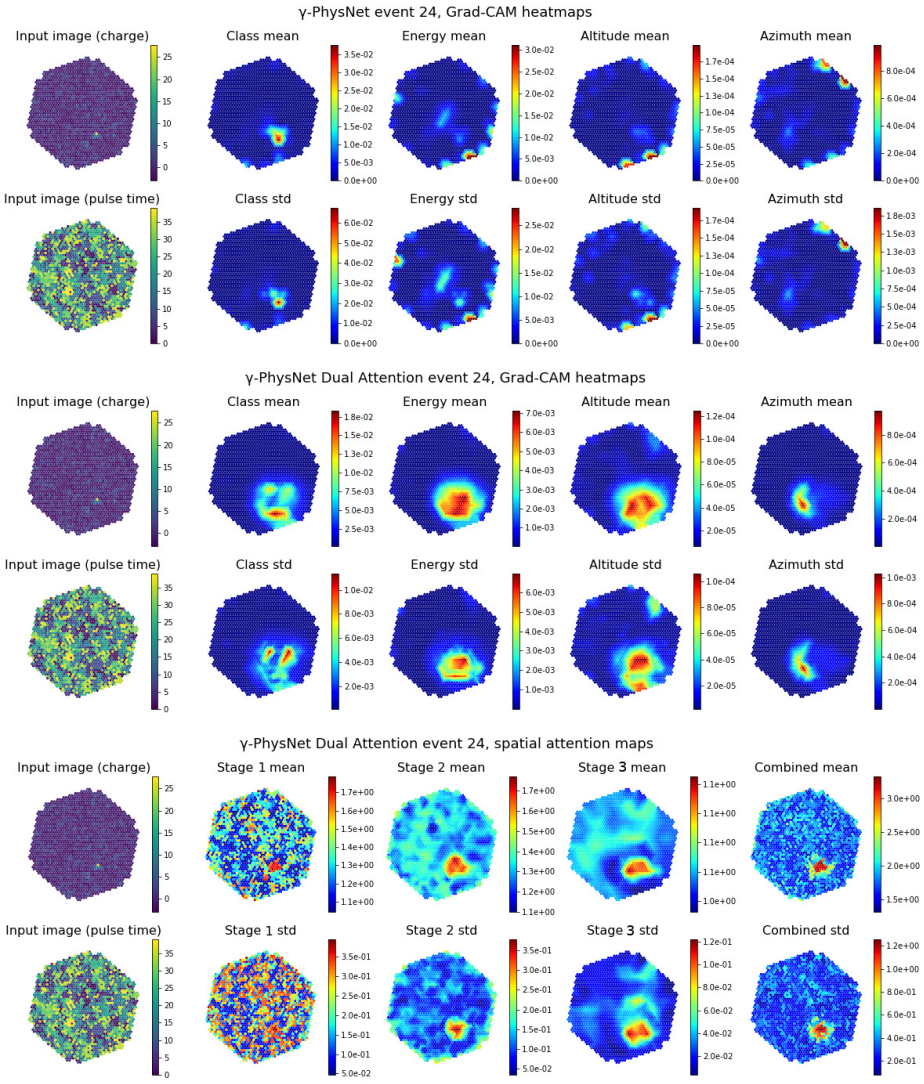

Fig. 6. Badly reconstructed gamma, event 24. The two upper rows represent the Grad-CAM heatmaps (mean and standard deviation) obtained for the vanilla γ-PhysNet, the next two rows correspond to the Grad-CAM heatmaps of γ-PhysNet DA (with attention), and the last two rows show the corresponding spatial attention maps.

6. Guo, M., Haque, A., Huang, D.A., Yeung, S., Fei-Fei, L.: Dynamic task prioritization for multitask learning. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 270–287 (2018)
7. He, K., Zhang, J., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
9. Hillas, A.: Cerenkov light images of eas produced by primary gamma. In: International Cosmic Ray Conference. vol. 3 (1985)
10. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
11. Jacquemont, M., Antiga, L., Vuillaume, T., Silvestri, G., Benoit, A., Lambert, P., Maurin, G.: Indexed operations for non-rectangular lattices applied to convolutional neural networks. In: Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP,. pp. 362–371. INSTICC, SciTePress (2019). https://doi.org/10.5220/0007364303620371
12. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7482–7491 (2018)
13. Kim, B., Brill, A., Miener, T., Nieto, D., Feng, Q.: DL1-DataHandler: DL1 HDF5 writer, reader, and processor for IACT data. https://doi.org/10.5281/zenodo.3336561 (Jul 2019), v0.8.1-legacy
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015)
15. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
16. Mangano, S., Delgado, C., Bernardos, M.I., Lallena, M., Vázquez, J.J.R., Consortium, C., et al.: Extracting gamma-ray information from images with convolutional neural network methods on simulated cherenkov telescope array data. In: IAPR Workshop on Artificial Neural Networks in Pattern Recognition. pp. 243–254. Springer (2018)
17. Morcos, A.S., Barrett, D.G.T., Rabinowitz, N.C., Botvinick, M.: On the importance of single directions for generalization. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)
18. Nieto Castaño, D., Brill, A., Kim, B., Humensky, T.B., Consortium, C.: Exploring deep learning as an event classification method for the Cherenkov Telescope Array. In: 35th International Cosmic Ray Conference. ICRC, vol. 301, p. 809 (Jan 2017)
19. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill **2**(11), e7 (2017)
20. Parsons, R.D., Ohm, S.: Background rejection in atmospheric Cherenkov telescopes using recurrent convolutional neural networks. European Physical Journal C **80**(5), 363 (May 2020). https://doi.org/10.1140/epjc/s10052-020-7953-3
21. Ren, Z., Jae Lee, Y.: Cross-domain self-supervised multi-task feature learning using synthetic imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 762–771 (2018)

22. Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017)
23. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
24. Sener, O., Koltun, V.: Multi-task learning as multi-objective optimization. In: Advances in Neural Information Processing Systems (2018)
25. Shilon, I., Kraus, M., Büchele, M., Egberts, K., Fischer, T., Holch, T.L., Lohse, T., Schwanke, U., Steppa, C., Funk, S.: Application of deep learning methods to analysis of imaging atmospheric cherenkov telescopes data. Astroparticle Physics **105**, 44–53 (2019)
26. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.A.: Striving for simplicity: The all convolutional net. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings (2015)
27. Srinivas, S., Fleuret, F.: Full-gradient representation for neural network visualization. In: Advances in Neural Information Processing Systems. pp. 4126–4135 (2019)
28. Sun, J., Darbeha, F., Zaidi, M., Wang, B.: Saunet: Shape attentive u-net for interpretable medical image segmentation. arXiv preprint arXiv:2001.07645 (2020)
29. Thrun, S.: Is learning the n-th thing any easier than learning the first? In: Advances in neural information processing systems. pp. 640–646 (1996)
30. Völk, H.J., Bernlöhr, K.: Imaging very high energy gamma-ray telescopes. Experimental Astronomy **25**(1-3) (2009)
31. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
32. Zhou, B., Sun, Y., Bau, D., Torralba, A.: Revisiting the importance of individual units in cnns via ablation. arXiv preprint arXiv:1806.02891 (2018)