PROCEEDINGS OF SCIENCE

Efficient Label Gathering for Machine Training: Results from Muon Hunter 2

M. Laraia[a]*, **for the VERITAS Collaboration** [†‡] **& D. Wright**[a], **H. Dickinson**[a], **A. Simenstad**[a], **K. Flanagan**[b], **S. Serjeant**[c]

[a] *University of Minnesota*
[b] *University College Dublin*
[c] *The Open University, UK*
E-mail: larai002@umn.edu, lfortson@umn.edu

In 2017, the Muon Hunter project on the Zooniverse.org citizen science platform successfully gathered more than two million classification labels for nearly 140,000 camera images from VERITAS. The aim was to select and parameterize muon events for use in training convolutional neural networks. The success of this project proved that crowdsourcing labels for IACT image analysis is a viable avenue for further development of advanced machine-learning algorithms. These algorithms could potentially lend themselves to improving class separation between gamma-ray and hadronic event types. Nonetheless, it took two months to gather these labels from volunteers, which could be a bottleneck for future applications of this method. Here we present Muon Hunters 2.0: the follow-on project that demonstrates the development of unsupervised clustering techniques to gather muon labels more efficiently from volunteer classifiers.

International Cosmic Ray Conference
July 24 – August 1, 2019
Madison, Wisconsin, USA

*Speaker.
†for collaboration list see PoS(ICRC2019)1177
‡http://veritas.sao.arizona.edu

PoS(ICRC2019)678

1. Introduction

The upcoming Cherenkov Telescope Array (CTA) will produce over 50 TB per night [1]. CTA and other projects generating large amounts of data can benefit from the development of algorithms to identify the phenomena under investigation in real time. Deep Neural Networks (DNN) can be used to this end [2]. Once a deep learning model has been trained, the algorithm is capable of processing large volumes of data in a short period of time and on relatively inexpensive hardware. Developing such a model, however, requires large amounts of annotated data.

Citizen science is a method that can be used to generate the initial set of training data for the deep learning model by distributing the task of annotating images to a large crowd of volunteers (see e.g. [3,4]). Many projects use this method of label gathering through the Zooniverse platform, the world's largest platform for citizen science research [5]. Muon Hunters is one project hosted on the Zooniverse where the images are derived from Imaging Atmospheric Cherenkov Telescope (IACT) data provided by the VERITAS collaboration. While muons are typically considered background for IACTs, telescope images produced by muons exhibit a characteristic arc or ring, and are therefore useful for calibrating the optical throughput of the telescopes [6, 7].

The first iteration of the Muon Hunter project followed a traditional model for Zooniverse projects: a volunteer was shown one image at a time and asked whether it contains a muon [8, 9]. In the second iteration, Muon Hunters 2.0 (MH2), we explore a new method of collecting classifications. The volunteer is shown a group of images simultaneously and asked whether most of the images contain a muon. Once establishing the majority class, the volunteer is asked to identify images that are not members of the majority class. Our hypothesis is that using this method will increase the efficiency of gathering classifications. This should also be related to how well an initial clustering algorithm separates into different clusters images with and without muons.

Xie et al. [10] developed a method for clustering data by using an unsupervised DNN to learn a deep embedding of the input data, and clustering within that space. Wright et al. [11] developed a method to further improve the purity of the clustering by using a set of labeled images to update the embedded space. An iterative feedback loop can be established whereby querying volunteers for image labels using the grid interface improves volunteer efficiency, and the labels provided by volunteers improve the clustering purity. Since cluster purity implies the majority class is more dominant this improves efficiency for volunteers in the future and the cycle continues.

Wright et al. in [11] used existing labels collected in the Supernova Hunters project to simulate how using the clustering method would affect the efficiency of gathering labels. In this current paper we demonstrate that the method generalizes well to another dataset. An interface for collecting volunteer classifications on a group of images was implemented using the clustering method. This allows measuring the actual gained efficiency, as well as how the classification performance of volunteers is affected by the new interface.

2. Methods

2.1 Clustering

Deep Embedded Clustering (DEC) [10] is an unsupervised DNN architecture that assigns input data to clusters based on a learned feature representation [10]. It consists of three fully connected layers

followed by a clustering layer. The fully connected layers are initialized using stacked denoising autoencoders.The clustering layer is trained by minimizing the Kullback-Leibler (KL) divergence to a target distribution. DEC was configured with layer dimensions 500, 500, 2,000 and 10. The clustering layer was configured for either 10 or 50 clusters. Stochastic gradient descent was used as the optimizer with a learning rate of 0.1 and momentum 0.9. DEC was trained for 80 epochs, with batch size 256.

After the clustering step, each cluster was sampled and the ensuing subset of data uploaded to the Zooniverse as the first batch for volunteer labeling. Specifically, the trained DEC network was used to assign each image to one of M clusters; each cluster was then randomly split into N-image sets. Duplicate images were sampled from the cluster to fill the last set to the same size. Each image set was used to create a $n \times n$ grid image. Each grid image was then uploaded to the MH2 project as a subject. Figure 1 shows the new MH2 interface which was developed using the Zooniverse Project Builder [12].

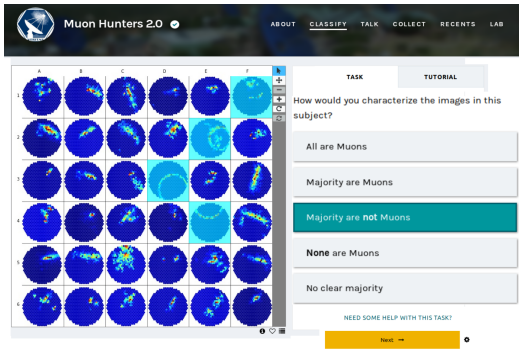

Figure 1: A screenshot of the MH2 interface. Volunteers are asked if the image mostly contains muons or not. Then they are asked to label all images that belong to the minority class.

The volunteers are then asked to classify each grid image in two steps. The first step asks them to characterize the makeup of the grid image as mostly muon, mostly non-muon, all muons, all non-muons, or no clear majority class. In the second task the volunteer is asked to highlight those images in the grid that do not belong to the majority class. For example, if the volunteer answers the first task that the grid image is dominated by non-muons, then they will be asked in the second task to highlight all the muons. If the volunteer answers the first task that the grid is entirely dominated by a single class, then the second task is skipped. For the case where there is no clear majority in the grid, the volunteer is asked in the second task to highlight muons in the image. By adding this question the volunteer is never asked to highlight more than half the images in a grid. This guarantees that even in the worst case of clustering performance, where every cluster contains exactly equal proportions of muons and non-muons, we can expect volunteers to be twice as efficient.

The MH2 Zooniverse project is set to retire a grid-image after it has been classified by 10 independent volunteers. After the first batch of grid-images has been retired the process moves to the multitask and reclustering steps described in [11]. The grid-image classifications are aggregated to produce a single label per telescope image in each grid-image. These labels are then used to retrain the clustering model to improve the clustering performance of the model. The next batch of data is then clustered and chunked into grid-images using the new clustering model. These grid-images are then uploaded to the Zooniverse for classification.

2.2 Aggregation

Grid-image classifications need to be aggregated to a single label per image. Each grid classification can be decomposed into N image classifications, where N is the number of images in the grid. We explored two methods of aggregating the decomposed classifications. The first is majority

vote, where a label is assigned to whatever class most volunteers assigned to the image. The second uses the Space Warps Analysis Pipeline (SWAP) [13], a Bayesian algorithm that considers both the prior probability of an image as well as the volunteer's performance. The volunteer performance is determined by maintaining a confusion matrix for a volunteer using a small set of images where the true label is known. SWAP increases the weight of classifications from high-performing volunteers while limiting the weight from low-performing volunteers. Labels aggregated using SWAP are typically more accurate than labels aggregated using majority vote, especially when volunteers are inconsistent or noisy (see e.g. [14, 15]).

3. Data

MH2 was implemented in two stages the first of which was a beta test of the new classification interface. The second stage involved procuring a new dataset entirely independent from the first iteration of the Muon Hunters project. 2.9M of these images were sourced from real VERITAS data, derived from data runs between January 2017 and January 2018. The VERITAS Gamma-ray Analysis Suite (VEGAS) can be configured to identify images containing muons where a radius cut must be tuned in order to produce a reliable sample of muon detections [16]. Clean high confidence true labels were generated from the set of images that VEGAS identified as muons with radius cuts where $r < 0.4$ were labeled as non-muons, and images where $r > 0.6$ were labeled as muons. This created 20,281 labeled non-muons, and 3,987 labeled muons. We should note that these images are not a representative sample of the data, as we are restricting only to images initially identified as muons by VEGAS. Nonetheless, we use these labels to gauge volunteer performance.

An additional 95,000 images were sourced from Corsika simulations [17]. We estimate that the real data contains approximately 10,000-20,000 muons, and the simulated data contains approximately 1,000-2,000 muons. The clustered real events were used to create 79,266 grid-images, and the clustered simulated events were used to create 2,640 grid-images. Of these images, 25% were reserved as the test set, and 75% were used to train the DEC clustering model.

4. Results

The F1 score is used to measure classification performance and is the geometric mean of the purity and completeness of the predicted labels. It is preferred over accuracy when there is a class imbalance such as in both datasets for this experiment. To measure the clustering F1 score, each cluster member inherits the class label of the majority of images in that cluster.

The relative efficiencies of the classification interfaces can be inferred from the number of clicks a volunteer must make to classify an image. In the standard single-image interface the number of clicks is twice the number of subjects being classified: one click to assign a label and one click to submit the classification. In the grid interface a volunteer makes one click to select which class dominates the grid-image, one click per image in the grid that does not belong to the majority class, and one click to submit. The efficiency is calculated from the effort expended by volunteers, measured in terms of the number of clicks, and is defined as: $\varepsilon = n_{\text{single}}/n_{\text{grid}}$.

The full DEC-Multitask-Reclustering pipeline was run for the first stage of the MH2 project. The entire pipeline was trained five times on the data, and the results of each step were averaged.

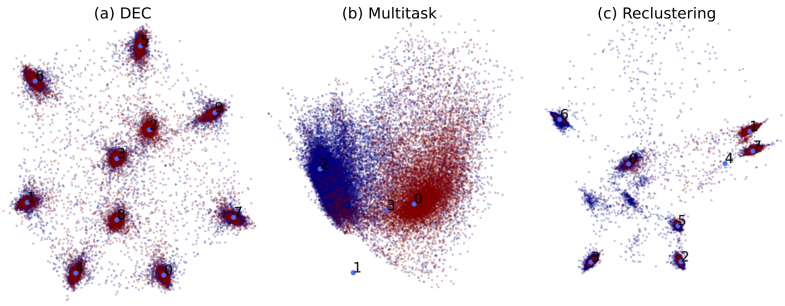

Figure 2: The PCA projections of the various clustering model steps described in the text. The blue dots represent non-muon images, and the red dots represent muon images.

First the DEC model is trained on all the data not in the test set. In the Multitask step the model is trained until the performance measured on the training development set stops decreasing. In the Reclustering step the model is trained until the model converges, which is defined as when less than 0.1% of images in the training and training-development sets are assigned a new cluster between training iterations. These models were trained with $M = 10$ clusters. Grid-images were generated from the DEC clustering with size $N = 100$. The F1 score of volunteers measured on the test set is F1 $= 0.22$ when the labels are aggregated by majority vote, and F1 $= 0.83$ when their labels are aggregated with SWAP. Because of the large difference in performance the labels aggregated with SWAP were used to train the multitask and reclustering models.

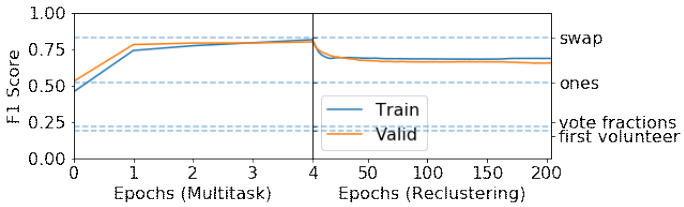

Figure 3: Multitask and Reclustering training performance.

The clustered space can be projected into two dimensions using Principal Component Analysis (PCA). The PCA projections for the DEC, Multitask, and Reclustering models are shown in Figure 2. The learning curves for the Multitask and Reclustering models are shown in Figure 3. The clustering F1 score for the training and validation sets are shown in blue and yellow respectively. Benchmark scores are shown as the F1 score measured on various sets of labels. The SWAP benchmark is the set of labels aggregated from volunteer classifications using SWAP. The *all-ones* benchmark is labeling all images as muons. The vote fractions are labels aggregated from volunteer classifications using majority vote, and the first volunteer benchmark is the set of labels from the first volunteer that labeled each image. Figure 3 shows that the performance of the Multitask model approaches the SWAP benchmark performance.

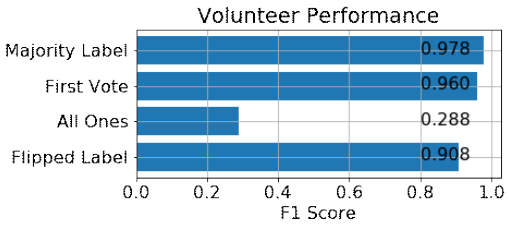

Figure 4: Performance of the volunteer classifiers for the second stage of the project, benchmarked by *First Vote* (label assigned by the first volunteer that sees the image), *All Ones* (muon label assigned to every image), and *Flipped Label* (flip every volunteer majority label).

The test set for these data contains 4,948 images; the traditional interface would require 9,896 clicks to classify all images. Random assignment and a supervised model are used as benchmarks for the clustering models in this stage of the project. For random assignment, each image is assigned at random to one of M clusters, and the metrics are calculated for this clustering in the same way. For the supervised model benchmark, a supervised model is trained to classify the images as containing a muon using the same architecture as the DEC model. This model is trained on all images in the training set using the aggregated SWAP label from the collected volunteer labels. The F1 score can be calculated directly from this model's output, and the number of clicks can be determined in the same way as described above by using the model's two output classes as clusters.

		F1	Clicks	Efficiency
Stage 1	Random Assignment	0.053	1,861	5.32
	Supervised	0.771	717	12.05
	DEC	0.407	1,127	8.00
	Multitask	0.803	634	13.33
	Reclustering	0.633	723	11.90
Stage 2	Random Assignment	0.035	3,990	21.74
	DEC	0.010	3,990	21.74

Table 1: F1 performance and gained efficiency of the clustering methods. *Stage 1* and *Stage 2* refer to the first and second stages of the MH2 project. DEC, Multitask, and Reclustering steps were all trained for stage 1, but only DEC was trained for stage 2. Random assignment clustering was used to benchmark the results of the models. An additional supervised model was trained to benchmark the stage 1 results.

For the second stage of MH2, which included a public launch of the project on March 14, 2019, the DEC model was trained with $M = 50$, and the grid-images were generated from this model with size $N = 36$. At the time of writing, 174,134 of the 2.9M images were classified by at least 5 volunteers. A majority of volunteers labeled 168,283 of these as non-muons, and 5,851 as muons. Of the labeled images in the test set, 1,756 are muons and 49,304 are non-muons.

The F1 score of the volunteer labels measured against the cleaned VEGAS labels is F1 $= 0.978$ when aggregated by majority vote. There is no significant gain in performance when the labels are instead aggregated with SWAP. These results are summarized in Figure 4 where we also show benchmark F1 scores of (1) the label assigned by the first volunteer that viewed each image, (2) labeling all images as muons, and (3) flipping the labels aggregated by majority vote. We note that the volunteers disagreed with the cleaned labels for only 7 images where ring-structure is present but were mislabeled by VEGAS. The volunteer labels aggregated by majority vote are thus used as the true labels for measuring the clustering performance for the second stage of MH2. The metrics for the various clustering algorithms for both stages of the project are summarized in Table 1.

5. Discussion and Conclusions

The results described above offer a number of conclusions for the first stage of the project. The most obvious of which is that the volunteers did a very poor job classifying the grid-images. Because that stage of MH2 was released as a Beta test on Zooniverse, there are a number of possible explanations for this. Often when a project is in Beta volunteers focus more on providing feedback on auxiliary items of the project, like the tutorial, field guide, and ease of using the interface, rather than on providing quality classifications. For example, the feedback from the Beta prompted implementing the custom interface shown in Figure 1 rather than using more generic drawing tools already available on the platform. Additionally, Beta feedback led to a reduction in the grid size from 10×10 to 6×6. Though there were enough volunteers who provided quality classifications,

even these were much improved by aggregating the classifications with SWAP. In the second stage of MH2, implementing SWAP made little difference as the labels were already of high quality.

In Table 1, the DEC, Multitask, and Reclustering models all show better performance than the random assignment benchmark. The Multitask model improved on the performance of the DEC model, while the Reclustering model lost much of the gained performance. It is evident that the Reclustering model can be improved. One idea currently being explored is to add dropout [18] to the model so the model does not overfit to the training data.

The Multitask model does very well compared to the benchmarks. The clustering F1 score of the Multitask model approaches the F1 score of the volunteer labels aggregated with SWAP suggesting that this model can identify muons almost as well as the volunteers. Furthermore, the Multitask model outperforms the Supervised model by a significant margin. This shows that first training the unsupervised model followed by retraining with a batch of imperfect labels from volunteers, offers better performance than training the same architecture directly on the cleaner aggregated labels produced by SWAP. Additional efficiency gains could be realized by clustering a new batch of images with the Multitask model for classification on the Zooniverse.

Finally, Table 1 shows that all forms of clustering showed significant gains in efficiency. Even the case of random assignment is *5.32* times more efficient compared to the traditional method of gathering classifications. The results offer slightly different conclusions for the second iteration of the project. The DEC clustering performance was actually worse than the random assignment clustering, due likely to the extreme skew of the dataset. Only 3 in 100 images contained a muon making it difficult for the unsupervised algorithm to learn anything significant related to the presence of a muon. A possible solution would be to artificially skew the data using the cleaned VEGAS labels to remove some of the images that are definitely not muons. Nevertheless, the extreme skew of the data means that the clustering interface offers *21.74* times the efficiency of the traditional classification interface with volunteers making only 3,990 clicks to classify 174,134 images.

In summary, a method was developed to gather classifications of images from a citizen science platform using a grid-based interface, allowing volunteers to classify many images simultaneously. This method was found to require 81% fewer clicks from volunteers in the worst case of clustering by random assignment. The DEC-Multitask pipeline was shown to significantly improve the performance of clustering compared to the benchmarks, rivaling the performance of volunteers in classifying images. The addition of the reclustering step requires additional work to realize gains in performance. This method can realize significant gains in efficiency for classifying images, even when the data is heavily skewed.

Acknowledgments: This research is supported by grants from the U.S. Department of Energy Office of Science, the U.S. National Science Foundation and the Smithsonian Institution, and by NSERC in Canada. This research used resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science, and resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility operated under Contract No. DE-AC02-05CH11231. We acknowledge the excellent work of the technical support staff at the Fred Lawrence Whipple Observatory and at the collaborating institutions in the construction and operation of the instrument.

References

[1] "Cherenkov Telescope Array." https://www.cta-observatory.org/project/technology/data/.

[2] Y. LeCun, Y. Bengio and G. Hinton, *Deep learning*, *Nature* **521** (2015) 436.

[3] L. Trouille, C. J. Lintott and L. F. Fortson, *Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems*, *Proceedings of the National Academy of Sciences* **116** (2019) 1902 [https://www.pnas.org/content/116/6/1902.full.pdf].

[4] L. Fortson, D. Wright, C. Lintott and L. Trouille, *Optimizing the Human-Machine Partnership with Zooniverse*, *arXiv e-prints* (2018) arXiv:1809.09738 [1809.09738].

[5] "Zooniverse Project." https://www.zooniverse.org/.

[6] D. Hanna, *Calibration Techniques for VERITAS*, *International Cosmic Ray Conference* **3** (2008) 1417 [0709.4479].

[7] J. Tyler and for the VERITAS Collaboration, *Muon Identification with VERITAS using the Hough Transform*, *arXiv e-prints* (2013) arXiv:1307.8361 [1307.8361].

[8] Q. Feng, J. Jarvis and VERITAS Collaboration, *A citizen-science approach to muon events in imaging atmospheric Cherenkov telescope data: the Muon Hunter*, *International Cosmic Ray Conference* **301** (2017) 826 [1708.06393].

[9] R. Bird, M. K. Daniel, H. Dickinson, Q. Feng, L. Fortson, A. Furniss et al., *Muon Hunter: a Zooniverse project*, *arXiv e-prints* (2018) arXiv:1802.08907 [1802.08907].

[10] J. Xie, R. Girshick and A. Farhadi, *Unsupervised deep embedding for clustering analysis*, in *International conference on machine learning*, pp. 478–487, 2016.

[11] D. Wright, M. Laraia, L. Fortson, C. Lintott and M. Walmsley, "Help me to help you: Machine augmented citizen science." In-prep, 2019.

[12] "Zooniverse Project Builder." https://www.zooniverse.org/lab.

[13] P. J. Marshall, A. Verma, A. More, C. P. Davis, S. More, A. Kapadia et al., *Space Warps — I. Crowdsourcing the discovery of gravitational lenses*, *Monthly Notices of the Royal Astronomical Society* **455** (2015) 1171 [http://oup.prod.sis.lan/mnras/article-pdf/455/2/1171/18509531/stv2009.pdf].

[14] D. E. Wright, C. J. Lintott, S. J. Smartt, K. W. Smith, L. Fortson, L. Trouille et al., *A transient search using combined human and machine classifications*, **472** (2017) 1315 [1707.05223].

[15] M. R. Beck, C. Scarlata, L. F. Fortson, C. J. Lintott, B. D. Simmons, M. A. Galloway et al., *Integrating human and machine intelligence in galaxy morphology classification tasks*, **476** (2018) 5516 [1802.08713].

[16] P. Cogan, *VEGAS, the VERITAS Gamma-ray Analysis Suite*, *International Cosmic Ray Conference* **3** (2008) 1385 [0709.4233].

[17] R. Engel, D. Heck, T. Huege, T. Pierog, M. Reininghaus, F. Riehn et al., *Towards a Next Generation of CORSIKA: A Framework for the Simulation of Particle Cascades in Astroparticle Physics*, *arXiv e-prints* (2018) arXiv:1808.08226 [1808.08226].

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, *Dropout: a simple way to prevent neural networks from overfitting*, *The Journal of Machine Learning Research* **15** (2014) 1929.