Citizen Scientist Amplification

Darryl E. Wright
University of Minnesota
116 Church St SE
Minneapolis, MN 55455

Lucy Fortson
University of Minnesota
116 Church St SE
Minneapolis, MN 55455

Chris Lintott
University of Oxford
Keble Road
Oxford, OX1 3RH

Mike Walmsley
University of Oxford
Keble Road
Oxford, OX1 3RH

Abstract

We introduce the idea of Citizen Scientist Amplification applying the method to data gathered from the top 10 contributing citizen scientists on the Supernova Hunters project. We take a novel approach to avail of the complementary strengths of deep learning and citizen science achieving results that are competitive with experts.

1 Introduction

Citizen scientists are enabling research only possible with the scale offered by crowdsourcing (Lintott et al. 2008; Swanson et al. ; Sullivan et al. 2009; Trouille, Lintott, and Fortson 2019). In these crowdsourcing projects, we gather multiple volunteer annotations per image that are aggregated into a single classification. The aggregation aims to compensate for individual mistakes or biases. Once aggregated, the results are used to answer scientific questions or train machine learning to automate the process (Dieleman, Willett, and Dambre 2015; Arteta, Lempitsky, and Zisserman 2016; Norouzzadeh et al. 2017; Zevin et al. 2017).

But what if, instead of aggregating volunteer classifications, we trained a model to replicate each individual volunteers classifications? If volunteers classifications can be exactly predicted then we could achieve the same performance as the citizen scientist crowd. It is unlikely that volunteer votes could be exactly predicted, but that the models as they attempt to learn a series of systematic "rules" could learn to ignore random noise ((Reid et al. 2016); (Rolnick et al. 2017)). The aggregation of these models might then result in even higher performance than citizen scientists alone. This leads to the idea of "Citizen Scientist Amplification" where, a model trained to replicate a citizen scientist, can be applied to annotate data they have not seen and faster. Many projects experience a surge in volunteer involvement shortly after launch, but experience diminishing engagement after a few days ((Spiers et al. 2019)). Classifications of long-term contributors to a project can be amplified by combining their votes with votes from models trained to replicate their peers who may no longer be engaged with a project. This approach differs from the "expertise amplification" of (Keshavan, Yeatman, and Rokem 2019) as we are not just seeking to amplify the number of labelled training examples for machine learning, but rather we seek to amplify performance (classification accuracy of the crowdsourcing system), efficiency (reducing time to achieve accurate classifications) and the long-term legacy of individual citizen scientists contributions to a project (developing a model of the individual citizen scientist that can be used to classify future data and augment the citizen science crowd).

In experiments below, we will demonstrate these benefits with data gathered by the Supernova Hunters citizen science project[1]. We find that amplifying the citizen science crowd with models for the top ten contributors in terms of the number of classifications submitted to the project (not the top ten performing volunteers) leads to performance gains that are competitive with models trained on expert labels. This has important implications for the adoption of machine learning for citizen science platforms.

Presented in the Work in Progress and Demo track, HCOMP 2019.

Table 1: Data set structure.

Set	Real	Bogus	Total
Training	2307	4609	6916
Test	766	1537	2303
Total	3073	6146	9219

2 Methods

Data Set Our data set consists of 20×20 pixel greyscale images extracted from Pan-STARRS1 (Kaiser et al. 2010) difference images (see Section 2.1 of (Wright et al. 2017) for details). We take 9219 images which have been labelled by experts to identify real detections of astrophysical transients from image artefacts. Table 1 shows the structure of this data set. Images are divided into two fixed partitions, the training partition contains roughly 75% of the data with the remaining 25% for testing. The test set partition is only used for measuring performance metircs, where predicted test set labels are compared against expert labels. These data were also uploaded to the Supernova Hunters project. The top ten

[1] https://www.zooniverse.org/projects

contributors each submitted more than three thousand classifications with a median of 3590 across the entire data set and a median of 867 classifications for the test set.

Volunteer vote aggregation Volunteers provide votes for the class membership of each image (real or artefact in this case). We use vote fractions to aggregate votes into a single label per image. The vote fraction is calculated as the fraction of votes that were assigned to each class. The vote fraction can be interpreted as the probability the crowd thinks each image has of belonging to each class. Vote fractions can be converted to hard (real or artefact) class labels by choosing the class with the highest vote fraction.

Modelling Citizen Scientists To model each individual citizen scientist's votes, we train a Convolutional Neural Network (CNN). Throughout experiments the neural network architecture was held constant. The training set was divided into five stratified folds. Each model was trained for five trials using four folds for training and the fifth fold as a validation set. Trials were run to average performance across any single training-validation split that might be chosen. The loss on the validation set is monitored for model check-pointing and early-stopping. Individual models were trained on the images from the training partition that each volunteer had labelled. Crucially, the targets for training are the labels provided for each image by each individual. The model therefore, aims to predict how the volunteer would classify a particular image, which need not be the correct label. The model will predict a volunteers behaviour as a function of an image and all other images in their training set.

Citizen Scientist Amplification We simulate amplifying citizen scientists by augmenting the classifications made by the top ten contributors with classifications made by the individual models. Specifically, for every image in the test set, if all ten volunteers have not classified that image, we add classifications from individual models corresponding to those volunteers who have not labeled that image. The additional machine votes are considered when calculating the vote fractions for the test set.

3 Results

We track two performance metrics, the F1-score and the Missed Detection Rate (MDR) at a 1% False Positive Rate (refered to as the MDR for brevity always assuming the 1% FPR). The former tracks the performance of the hard label assignments, while the latter is more aligned with the scientific goals of the project.

Table 2 shows the classification performance results from our experiments. Considering only the classifications submitted by the top ten contributing citizen scientists we measure a F1-score of 0.851 using vote fractions (*vote fractions (10)* in Table 2) on the test set. The MDR is measured as 1.000. Since each image has only received on average four classifications from the top ten contributors, the vote fractions result in a few discrete values. As a result the 1% FPR condition cannot be met with any threshold. Next, we report results on some machine learning benchmarks. These

Table 2: Experimental results.

method	MDR	F_1-score
vote fractions (10)	1.000	0.851
single CNN - expert	0.200(0.016)	0.929(0.003)
single CNN - vote frac.	0.431(0.052)	0.824(0.014)
10 CNNs - individuals	0.287(0.019)	0.906(0.006)
10 CNNs - vote frac.	0.295(0.031)	0.906(0.014)
Amplified cit-sci	**0.200(0.026)**	**0.917(0.006)**
Bagging + cit-sci	0.213(0.004)	0.918(0.003)

benchmarks are designed to present alternative approaches that could be taken for machine learning. These are namely, an expert model trained on expert labels and an aggregated votes model trained on vote fraction labels produced by the top ten contributors. We find that training the model on expert labels only (denoted *single CNN - expert* in Table 2) achieves a mean MDR of 0.200 across five trails. The aggregated votes model (*single CNN - vote frac.*) performs marginally worse (MDR=0.431) than the vote fractions (10) F1-score. This is expected since the model is trained to replicate the crowd and we anticipate it will make the same mistakes.

We found that training 10 models (*10 CNNs - individuals*), one for each of the top ten contributors results in a mean MDR of 0.287. This is an improvement of 14.4% on the CNN trained on vote fractions and is 8.7% worse than the CNN trained on expert labels. One possible explanation for the performance improvement is bagging (Breiman 1996). To test this hypothesis, we train 10 models on bootstrap samples drawn from the training set using vote fractions as training labels. This approach achieves an MDR of 0.295, 0.8% higher than the aggregated individual models, suggesting that most (if not all) the performance gains are realised through the effect of bagging. It therefore appears that aggregating volunteers classifications provides no benefit.

Finally, we report the results of amplifying the citizen scientist crowd with individual models as *Amplified cit-sci* in Table 2) measuring a MDR of 0.200. This performs as well as the model trained on expert labels. For these results we amplified those volunteers who classified each image with classifications from the models of their peers who had not classified that image. We also tested amplifying the crowd with ten bootstrap models (*Bagging + cit-sci*) which resulted in an MDR of 0.213, performing 1.3% worse than the proposed amplification approach.

4 Conclusions

We have explored the possible advantages of training models to replicate individual citizen scientists and shown how these models can amplify the citizen science crowd. We expect it to be more challenging to model individual citizen scientist behaviour for larger, more complex images or more intricate tasks (such as drawing tasks). Nonetheless, since we still expect humans to differ, our method could prove beneficial across many tasks and data types.

References

Arteta, C.; Lempitsky, V.; and Zisserman, A. 2016. Counting in the wild. In *European conference on computer vision*, 483–498. Springer.

Dieleman, S.; Willett, K. W.; and Dambre, J. 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly notices of the royal astronomical society* 450(2):1441–1459.

Kaiser, N.; Burgett, W.; Chambers, K.; Denneau, L.; Heasley, J.; Jedicke, R.; Magnier, E.; Morgan, J.; Onaka, P.; and Tonry, J. 2010. The pan-starrs wide-field optical/nir imaging survey. In *Ground-based and Airborne Telescopes III*, volume 7733, 77330E. International Society for Optics and Photonics.

Keshavan, A.; Yeatman, J. D.; and Rokem, A. 2019. Combining citizen science and deep learning to amplify expertise in neuroimaging. *Frontiers in neuroinformatics* 13:29.

Lintott, C. J.; Schawinski, K.; Slosar, A.; Land, K.; Bamford, S.; Thomas, D.; Raddick, M. J.; Nichol, R. C.; Szalay, A.; Andreescu, D.; Murray, P.; and Vandenberg, J. 2008. Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society* 389:1179–1189.

Norouzzadeh, M.; Nguyen, A.; Kosmala, M.; Swanson, A.; Packer, C.; and Clune, J. 2017. Automatically identifying wild animals in camera trap images with deep learning. *arXiv preprint arXiv:1703.05830*.

Reid, R. S.; Nkedianye, D.; Said, M. Y.; Kaelo, D.; Neselle, M.; Makui, O.; Onetu, L.; Kiruswa, S.; Kamuaro, N. O.; Kristjanson, P.; Ogutu, J.; BurnSilver, S. B.; Goldman, M. J.; Boone, R. B.; Galvin, K. A.; Dickson, N. M.; and Clark, W. C. 2016. Evolution of models to support community and policy action with science: Balancing pastoral livelihoods and wildlife conservation in savannas of east africa. *Proceedings of the National Academy of Sciences* 113(17):4579–4584.

Rolnick, D.; Veit, A.; Belongie, S.; and Shavit, N. 2017. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Spiers, H.; Swanson, A.; Fortson, L.; Simmons, B. D.; Trouille, L.; Blickhan, S.; and Lintott, C. 2019. Everyone counts? design considerations in online citizen science. *Journal of Science Communication* 18(1).

Sullivan, B. L.; Wood, C. L.; Iliff, M. J.; Bonney, R. E.; Fink, D.; and Kelling, S. 2009. ebird: A citizen-based bird observation network in the biological sciences. *Biological Conservation* 142(10):2282 – 2292.

Swanson, A.; Kosmala, M.; Lintott, C.; and Packer, C. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology* 30(3):520–531.

Trouille, L.; Lintott, C. J.; and Fortson, L. F. 2019. Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems. *Proceedings of the National Academy of Sciences* 116(6):1902–1909.

Wright, D.; Lintott, C.; Smartt, S.; Smith, K.; Fortson, L.; Trouille, L.; Allen, C.; Beck, M.; Bouslog, M.; Boyer, A.; Chambers, K.; Flewelling, H.; Granger, W.; Magnier, E.; McMaster, A.; Miller, G.; O'Donnell, J.; Simmons, B.; Spiers, H.; Tonry, J.; Veldthuis, M.; Wainscoat, R.; Waters, C.; Willman, M.; Wolfenbarger, Z.; and Young, D. 2017. A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society* 472(2):1315–1323.

Zevin, M.; Coughlin, S.; Bahaadini, S.; Besler, E.; Rohani, N.; Allen, S.; Cabero, M.; Crowston, K.; Katsaggelos, A. K.; Larson, S. L.; et al. 2017. Gravity Spy: integrating advanced ligo detector characterization, machine learning, and citizen science. *Classical and Quantum Gravity* 34(6):064003.

5 Acknowledgments

DEW and LF gratefully acknowledge partial support through the US National Science Foundation grants IIS 1619177 and PHY 1806798. CJL acknowledges support from STFC under grant ST/N003179/1. MW acknowledges funding from the Science and Technology Funding Council (STFC) Grant Code ST/R505006/1. The Pan-STARRS1 Surveys have been made possible through contributions of the Institute for Astronomy, the University of Hawaii, the Pan-STARRS Project Office, the Max-Planck Society and its participating institutes, the Max Planck Institute for Astronomy, Heidelberg and the Max Planck Institute for Extraterrestrial Physics, Garching, The Johns Hopkins University, Durham University, the University of Edinburgh, Queen's University Belfast, the Harvard-Smithsonian Center for Astrophysics, the Las Cumbres Observatory Global Telescope Network Incorporated, the National Central University of Taiwan, the Space Telescope Science Institute, the National Aeronautics and Space Administration under Grant No. NNX08AR22G issued through the Planetary Science Division of the NASA Science Mission Directorate, the National Science Foundation under Grant No. AST-1238877, the University of Maryland,you and Eotvos Lorand University (ELTE) and the Los Alamos National Laboratory. We also wish to acknowledge the dedicated effort of our citizen scientists who have made this work possible.