Classification of Optical Transients at the MeerLICHT Telescope using Deep Learning

Zafiirah Hosenie[*]
Jodrell Bank Centre for Astrophysics
Department of Physics and Astronomy
The University of Manchester, Manchester M13 9PL, UK.
zafiirah.hosenie@gmail.com

Paul J. Groot
Department of Astrophysics
Radboud University, IMAPP
Nijmegen, The Netherlands P.O. 9010, 6500 GL.
p.groot@astro.ru.nl

Robert Lyon
Department of Computer Science
Edge Hill University
Ormskirk Lancashire L39 4QP, UK
lyonro@edgehill.ac.uk

Benjamin Stappers
Jodrell Bank Centre for Astrophysics
Department of Physics and Astronomy
The University of Manchester, Manchester M13 9PL, UK
Ben.Stappers@manchester.ac.uk

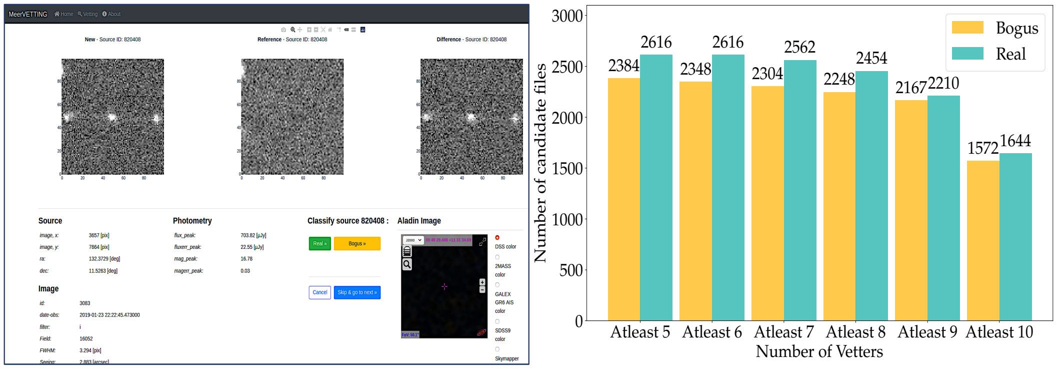
The MeerLICHT Team
Collaboration between South Africa-Netherland-United Kingdom
http://www.meerlicht.uct.ac.za/

Abstract

Astronomers require efficient automated detection and classification pipelines when conducting large-scale surveys of the optical sky. Such pipelines are fundamentally important as they permit rapid follow-up and analysis of those detections most likely to be of scientific value. We present a deep learning framework based on a convolutional neural network model known as `MeerCRAB`. It is designed to filter out the so called "bogus" detections from true astrophysical sources in the transient detection pipeline of the MeerLICHT telescope. Optical candidates are described using a variety of 2D images and numerical features extracted from those images. The relationship between the input images and the target classes is unclear, since the ground truth is poorly defined and often the subject of debate. This makes it difficult to determine which source of information should be used to train a classification algorithm. To proceed we deployed variants of `MeerCRAB` that employed different network architectures trained upon different combinations of input images and different training set choices based on volunteer's classification labels. We found the deepest network worked best with an accuracy of 99.2% and Matthews correlation coefficient (MCC) value of 0.984. The best model is integrated in the MeerLICHT transient vetting pipeline, hence providing a contextual classification of detected transients that allows researchers to select the most promising candidates for their research goals.

[*] https://zafiirah13.github.io/zafiirah-hosenie/

Third Workshop on Machine Learning and the Physical Sciences (NeurIPS 2020), Vancouver, Canada.



Source

image, x:	3657 [pix]
image, y:	7864 [pix]
ra:	132.3729 [deg]
dec:	11.5263 [deg]

Photometry

flux_peak:	703.82 [μJy]
fluxerr_peak:	22.55 [μJy]
mag_peak:	16.78
magerr_peak:	0.03

Image

id:	3083
date-obs:	2019-01-23 22:22:45.473000
filter:	i
Field:	16052
FWHM:	3.294 [pix]
Seeing:	2.883 [arcsec]

(a) `MeerVETTING` web-interface.

(b) *Thresholding method*: removing noisy labels.

Figure 1: (a) `MeerVETTING` web-interface to label candidates as either real or bogus, based on three images (NRD), (b) Thresholding method applied to label candidates based on the 10 volunteers' labels. The x-axis represents thresholding criteria applied (*atleast5 (T5)* to *atleast 10 (T10)*). For *atleast 9 (T9)*, this indicates that 9 out of 10 volunteers have agreed on the labelling.

1 Introduction

Current large-scale survey telescopes such as the Skymapper [8], the MeerLICHT telescope [2] and the Zwicky Transient Factory [1] are already generating a plethora of transient events. For these surveys to be feasible, it is imperative that we automate the transient search process including the separation of likely real transient events, from those "bogus" detections originating from sources that are not of interest. Bogus detections may be caused by instrumental errors or by data processing errors. In this paper, we present an application of deep learning to a new data set obtained at the MeerLICHT telescope for the classification of "real versus bogus" (`MeerCRAB`). We deal pragmatically with the problems raised during real-world application of deep learning in a domain where labels are of poor quality and incomplete, reflecting gaps in our knowledge. With new telescopes coming on-line we need a viable solution that can overcome these problems (or at least mitigate them) that meet our operating constraints – auditability, runtime efficiency to name a few.

We constructed three models based on convolutional neural networks (CNN, [10]) and we employed two techniques to label our data (i) *thresholding* that removes noisy labelling and (ii) *Latent class model*, L_{lcm} [4] that incorporates the labelling uncertainty in our model. In addition, the networks can incorporate various inputs: new, reference, difference and significance images. These networks are telescope agnostic in nature and are currently implemented in the MeerLICHT transient-vetting pipeline to classify candidates in real time. Hence, `MeerCRAB` reduces the need for manual verification by humans, which is an expensive and most likely, impossible process to conduct given the volumes of data to be dealt with.

2 Data

MeerLICHT is an optical wide-field telescope that is operated robotically [2]. Each image captured by this telescope begins as a matrix $n \times m$ in size. These images are downsampled by a data processing pipeline producing "reduced" images that form the inputs to a classification model. There are four distinct forms of image inputs to `MeerCRAB` – (i) the new (**N**) image which is the latest science image fully reduced, (ii) the reference (**R**) image which is the first image of the field, (iii) the difference (**D**) which is the residual after the new and reference image are subtracted from each other, (iv) the significance (**S**) image which is constructed based on the difference image using `ZOGY` [12], but with an additional noise model taken into account to calculate significances.

In a supervised context, the success of deep neural networks depends highly on the availability and accessibility to high-quality labelled training data. In addition, the data set needs to be representative, else machine learning (ML) algorithms tend to be biased towards the majority class [6; 7]. We therefore construct a large representative training dataset (5000 candidates) for the Real-Bogus challenge by manually vetting a selection of transients, using a web-interface, known as MeerVETTING (Figure. 1(a)). Each candidate is vetted by 10 volunteers, who are shown three images (**N**, **R**, **D**) during vetting. Each volunteer's ability to classify a particular candidate may vary according to the class, images and criteria. Unfortunately, this leads to large training datasets that will almost always contain examples with inaccurate labels. We therefore test the performance of the MeerCRAB models by (i) removing noisy labels using a *thresholding* method, and (ii) including the entire dataset with noisy labels based on L_{lcm}.

2.1 Thresholding

We assign a probability $\mathcal{P}\left(Real\right)$ and $\mathcal{P}\left(Bogus\right)$ to each vetted candidate as follows:

$$\mathcal{P}\left(\text{Real}\right) = \frac{n\left(R\right)}{n\left(T\right)}; \mathcal{P}\left(\text{Bogus}\right) = \frac{n\left(B\right)}{n\left(T\right)}, \tag{1}$$

where $n\left(R\right)$ is the total number of vetters who classified a candidate as real, $n\left(B\right)$ is the total number of vetters who classified a candidate as bogus, $n\left(T\right)$ is the total number of vetters classifying a particular candidate and in this case $n\left(T\right) = 10$. The volunteer's classification results are illustrated in a bar plot in Figure 1(b). On the x-axis, for example "*atleast* 9" (**T9**) implies that all candidates with $\mathcal{P}\left(Real\right) \geq 0.9$, are labelled as real or if $\mathcal{P}\left(Bogus\right) \geq 0.9$, they are labelled as bogus. Candidates with $\mathcal{P} < 0.9$ are considered as noisy labels and removed from the data.

2.2 Labelling data with Latent Class Model, $\mathbf{L}_{lcm}$

Latent class model is a statistical technique used to classify candidates into mutually exclusive, or latent classes. It is mostly based on their pattern of answers on a set of categorical data. When observed data in the form of a series of categorical responses, for example, individual-level voting data as in the case of real-bogus classification, it is often an interesting analysis to identify and characterize clusters of similar cases. In this paper, some confused sources[2] were removed from the data when using the *thresholding* method. However, it is useful for determining how the system will perform in a real-world scenario. Therefore, confused examples will also be used during the evaluation phase, and this is achieved using L_{lcm} to assign them their most likely labels.

3 MEERCRAB Models

We employed a CNN as it has been proven by various studies to have excellent classification performance [1; 3; 5]. In this work, we construct three CNN models: MeerCRAB1, MeerCRAB2, and MeerCRAB3 as illustrated in Figure 2. During training, the binary cross-entropy loss function, Adam optimizer [9] with a low learning rate (lr = 0.0002) and a batch-size of 64 were used. We then split our data into 50% training, 25% validation and 25% testing. As input to the MeerCRAB models, we cropped the images from centre to a size of (30×30) pixels that were vetted by volunteers.

The best strategy for better generalisation of a ML model is to train with a large amount of data. However, we simply do not have access to such large volumes of labelled data. Therefore, we apply data augmentation techniques to create new labelled training samples at each training step, thus the images are augmented by flipping randomly in a horizontal and/or vertical direction. Moreover, to avoid any over-fitting during training, we employ an early stopping technique to stop the training process if no further decrease in validation loss is observed for several epochs. The models are trained for epochs varying from 50 to 150.

4 Experiments and results

In this paper, we investigated various scenarios for training and evaluating the pipeline. We analysed the performance based on MeerCRAB1, MeerCRAB2 and MeerCRAB3. We also varied the number of input images and we investigated the effect of noisy data labels.

Effects of noisy data labelling on performance: We used various *thresholding* criteria '*atleast 8, 9, 10*' (**T8**, **T9** & **T10**) to remove noisy labels. We also investigate the effect on the performance of

[2] 5 vetters labelled them as bogus and the other 5 as real.

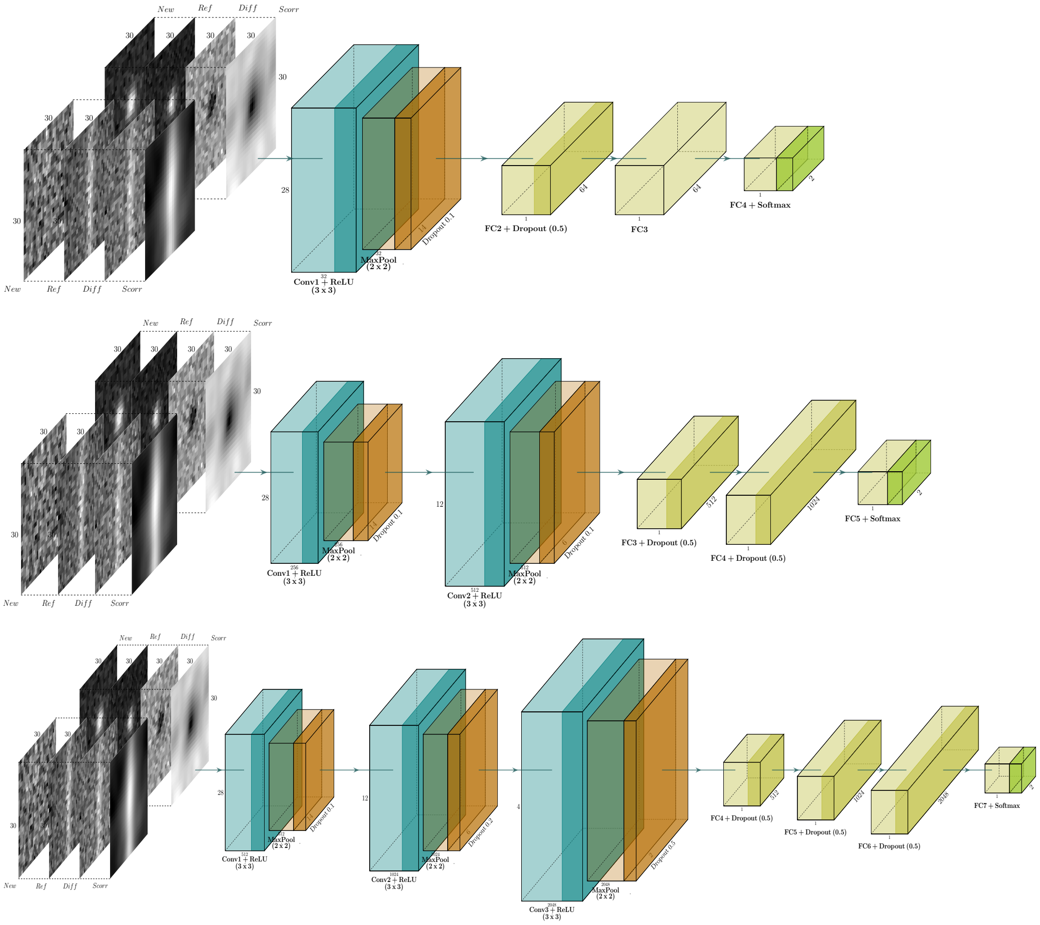
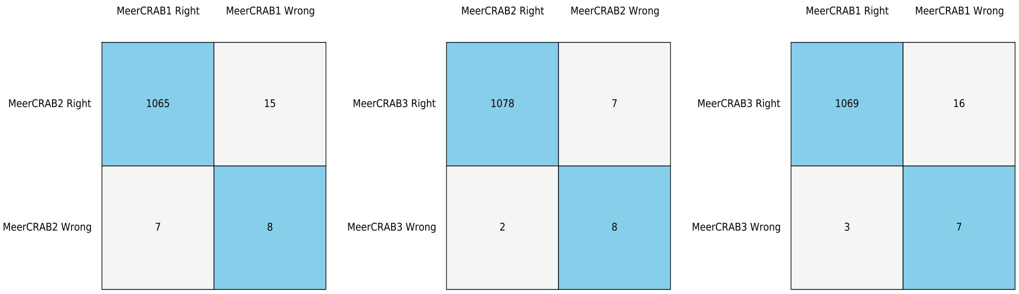


Figure 2: The three network architectures considered in this work: `MeerCRAB1` (top), `MeerCRAB2` (middle) and `MeerCRAB3` (bottom). Real and bogus sources are represented by four images stacked together (*NRDS*) to form the input of the networks, followed by convolutional layers, max-pooling, dropout and dense layers. At the end, the network outputs a probability whether a candidate is either real or bogus in a particular test set.

`MeerCRAB` models with the introduction of noisy labelling based on L_{lcm} method. Our first analysis involves comparing `MeerCRAB3` model with NRD as input where the results are summarised in Table 1. We observe that as the threshold increased from **T8** to **T10**, the accuracy of the model increases from 0.988 to 0.998. However, when using L_{lcm} method, we note a significant drop in accuracy as deep networks tend to memorize training label noise, resulting in poorer model performance. Therefore, it is necessary to have good labelling for CNN to work appropriately, thus removing noisy labelling from the model shows a better model performances.

Input Images: We use different inputs independently to the networks: *NRDS*, *NRD*, *NRS*, *NR*, *D*, and *S*, to see whether a competitive performance can be achieved with less or more input data. Focusing on **T9** and `MeerCRAB3`, we note that the NRD input yields the best performance model with an accuracy of 99.2%. With only *NR* as input, we note that `MeerCRAB3` yields a competitive performance and indicates that a reduced set of images is sufficient for approaching the problems. However, with only *D* or *S* as input, the classification performance is worse, thus indicating that information only from the difference or significance imaging is not enough for CNN to solve the tasks.

Network architectures From Table 1, with **T9** and *NRD* we note that `MeerCRAB1` predicts an accuracy of 97.9%. With deeper networks (`MeerCRAB2` and `MeerCRAB3`), we obtain a higher performance with an accuracy of 98.6% & 99.2% respectively. However, how would one decide which network performs best? In this case, we employ the McNemar statistical test [11] which is based on contingency



	MeerCRAB1 Right	MeerCRAB1 Wrong
MeerCRAB2 Right	1065	15
MeerCRAB2 Wrong	7	8

	MeerCRAB2 Right	MeerCRAB2 Wrong
MeerCRAB3 Right	1078	7
MeerCRAB3 Wrong	2	8

	MeerCRAB1 Right	MeerCRAB1 Wrong
MeerCRAB3 Right	1069	16
MeerCRAB3 Wrong	3	7

Figure 3: The contigency tables based on the test set for the three models: `MeerCRAB1` vs `MeerCRAB2` (left), `MeerCRAB2` vs `MeerCRAB3` (middle), `MeerCRAB1` vs `MeerCRAB3` (right). We observe that `MeerCRAB3` is a better model compared to `MeerCRAB1`, having a lower misclassification rate.

tables as shown in Figure 3. We compute the p-value from a binomial distribution as the misclassified sample size is relatively small (< 25). If the p-value is less than 0.05, we reject the null hypothesis that both models perform equally well on the test set, else we accept the null hypothesis. The p-value for `MeerCRAB1` vs `MeerCRAB2` is 0.134 and for `MeerCRAB2` vs `MeerCRAB3` is 0.180. For both cases, their p-values are greater than 0.05. This implies that the models have equal performance. However, when comparing `MeerCRAB1` vs `MeerCRAB3`, the p-value is 0.004, which is less than 0.05. This indicates one model is favored. `MeerCRAB3` has less misclassified instances compared to `MeerCRAB1` as shown in Figure 3(c), therefore, we conclude `MeerCRAB3` is the best model architecture.

5 Summary

Being able to filter out bogus events with an automatic technique will enable the labeling of real events saving human experts from going through a painstaking process. The proposal of a deep learning framework (`MeerCRAB`) integrated in the MeerLICHT facility is a step forward in the automation and improvement of the transient vetting process. In practice, by using `MeerCRAB` we can significantly reduce the number of missed transients per night and this may have a great impact on detecting and classifying the unknown unknowns of our universe. Our code is made publicly available at https://github.com/Zafiirah13/meercrab.

Broader Impact

`MeerCRAB` is a first step for MeerLICHT to discover an unprecedented number of transients, expanding the new era of big data in optical astronomy. With the streaming data coming from MeerLICHT, the vast majority of astrophysical phenomena are challenging to classify efficiently and effectively. Therefore, `MeerCRAB` will enable the rapid identification of promising astrophysical sources in timely-manner. In addition, `MeerCRAB` can be adapted to be a system that disentangles interesting objects from a noisy background. We have already implemented similar models in radio astronomy that distinguish Single Pulses from Radio Frequency Interference for the MeerKAT telescope (`FRBID`: Fast Radio Burst Intelligent Distinguisher). `MeerCRAB` is a flexible software, thus we were able to easily modify it to integrate different images as its inputs and as result, achieved high levels of performance when using it for radio astronomy images.

Given the performance of `MeerCRAB` on both optical and radio image sources in astronomy, the method may have utility for those working in related areas, for instance, in pipeline monitoring and oil leakage detection technologies. Pipelines are widely utilised to transport hydrocarbon fluids over long distances. However, leaks in pipeline networks can result in serious human casualties, financial loss, climate and ecological disasters. With the availability of enough training data taken from drones at various locations and at several time-spans, `MeerCRAB` can be implemented to tackle this challenging task. This will provide an exact location of leakage occurrences and will help to monitor the problem quickly and efficiently.

Acknowledgment

ZH acknowledges support from the UK Newton Fund as part of the Development in Africa with Radio Astronomy (DARA) Big Data project delivered via the Science & Technology Facilities Council (STFC). BWS acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 694745). This work used data obtained with the MeerLICHT telescope, located at the SAAO Sutherland station, South Africa. The MeerLICHT telescope is run by the MeerLICHT consortium, on behalf of Radboud University, the University of Cape Town, the Netherlands Foundation for Scientific Research (NWO), the National Research Facility of South Africa through the South African Astronomical Observatory, the University of Oxford, the University of Manchester and the University of Amsterdam.

References

Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002

Bloemen, S., Groot, P., Woudt, P., et al. 2016, 9906, 990664

Cabrera-Vives, G., Reyes, I., Förster, F., Estévez, P. A., & Maureira, J.-C. 2017, ApJ, 836, 97

Formann, A. K. 1984, Die latent-class-analyse: Einführung in Theorie und Anwendung (Beltz)

Gieseke, F., Bloemen, S., van den Bogaard, C., et al. 2017, MNRAS, 472, 3101

Hosenie, Z., Lyon, R., Stappers, B., Mootoovaloo, A., & McBride, V. 2020, MNRAS, 493, 6050

Hosenie, Z., Lyon, R. J., Stappers, B. W., & Mootoovaloo, A. 2019, MNRAS, 488, 4858

Keller, S. C., Schmidt, B. P., Bessell, M. S., et al. 2007, PASA, 24, 1

Kingma, D. P. & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980

Lecun, Y., Haffner, P., Bottou, L., & Bengio, Y. 1999

McNemar, Q. 1947, Psychometrika, 12, 12

Zackay, B., Ofek, E. O., & Gal-Yam, A. 2016, ApJ, 830, 27

Appendix

Table 1: The results for various labelling methods are presented in terms of precision, recall, accuracy and mathew correlation coefficient (MCC) values using *NRD* as input to the three models.

Methods of labelling	**Precision**	**Recall**	**Accuracy**	**MCC**
		`MeerCRAB1`		
L_{lcm}	0.96	0.96	0.960	0.920
T8	0.98	0.98	0.980	0.958
T9	0.98	0.98	0.979	0.958
T10	0.99	0.99	0.991	0.983
		`MeerCRAB2`		
L_{lcm}	0.97	0.97	0.967	0.936
T8	0.99	0.98	0.977	0.953
T9	0.99	0.99	0.986	0.973
T10	0.99	0.99	0.994	0.988
		`MeerCRAB3`		
L_{lcm}	0.97	0.97	0.968	0.936
T8	0.99	0.99	0.988	0.976
T9	**0.99**	**0.99**	**0.992**	**0.984**
T10	1.00	1.00	0.998	0.995