2023 IEEE 23rd International Symposium on Computational Intelligence and Informatics (CINTI) | 979-8-3503-4294-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CINTI59972.2023.10382059

Galaxy detection and classification in sky images with neural network

Alex Szabó*, Ádám Pintér†

* John von Neumann Faculty of Informatics, Óbuda University, Budapest, Hungary
† John von Neumann Faculty of Informatics, Óbuda University, Budapest, Hungary, pinter.adam@nik.uni-obuda.hu

***Abstract*—Galaxy detection and classification play an important role in astronomical research. The ongoing sky surveys produce enormous data, which makes a need for automation. Knowing the main class is important, but the uncommon-ringed galaxy shapes are especially interesting. Two methods were developed to detect and classify these in sky images. The first approach was a modified Faster R-CNN and the new one is DBSCAN – CNN-based network. SDSS imaging data were used for training and test purposes. Classification tests were done and the results for the new method showed 99.7% accuracy on main classes and 94.9% F1 score on ringed galaxies.**

***Index Terms*—galaxy, DBSCAN, Faster R-CNN, neural network**

I. Introduction

Astronomers use, among others, galaxy catalogs for their research. These store information about galaxies, like their coordinates or type. Sky surveys collect data about the Universe with telescopes. Since there are more than 2 trillion galaxies and we obtain an enormous amount of data each year, we need an automatic detection and classification solution to speed up this process. Along with the identification of the main type, we also want to know about special features, like rings.

II. Related Work

In the past decade several projects aimed at this. In 2014, a galaxy classification challenge took place, where classifications were done by computers. At this competition all the top methods were CNNs. The winning team, Sander et al. [1] developed multiple CNN-s and used their averaged output as a result. They predicted the user vote probabilities of images with 88% accuracy on main types and 91% precision on ringed shapes. A few years later H. Domínguez Sánchez et al. [2] trained a CNN that achieved an impressive 98.2% accuracy on the main types. Xiao-Pan Zhu et al. [3] in 2019 used a residual type of network to classify 5 galaxy morphologies. Their network and the other famous networks (e.g. Inception), yielded an F1 score of around 95%. To both detect and classify, Roberto E. González et al. [4] using YOLO neural network achieved a 72% and 84% precision on spiral and elliptical respectively with a 0.7 IOU. Colin J. Burke et al. [5] trained a Mask R-CNN with simulated sky images to detect and classify objects as stars or galaxies. Their network yielded an 84% average precision with 0.5 IOU.

III. Methodology

A. Dataset

First, using the debiased Galaxy Zoo Catalogue [6], [7], galaxies were sorted by their number of votes. From this $20-20$ thousand main class and $5-5$ thousand ringed/non-ringed galaxies were selected. To enhance the generalization capabilities of the network, images from Data Release 9 were downloaded in addition to those from Data Release 7.

Differences such as color or brightness are shown in Fig 1. To train the Faster R-CNN [8] network, sky images were generated using these galaxies by putting them on a *"galaxy-less"* sky image after they have been randomly rotated and rescaled. This way the coordinates for the Faster R-CNN training are already available. The inserted galaxy image edges were noticeable on the generated sky image and multiple background noises, such as stars, appeared around the galaxy. To solve this, first, a logarithmic-based intensity correction was applied to the galaxy images. In this, the first maximum intensity of the image is calculated, which will be the base of the logarithm, then at every pixel, each color's value will be changed to its logarithmic value. In the next step then the images are split into tiles and their average intensity is calculated. From the center of the image, a search starts to find the edges in eight directions. When a change exceeds a certain value, it is considered to be the edge.

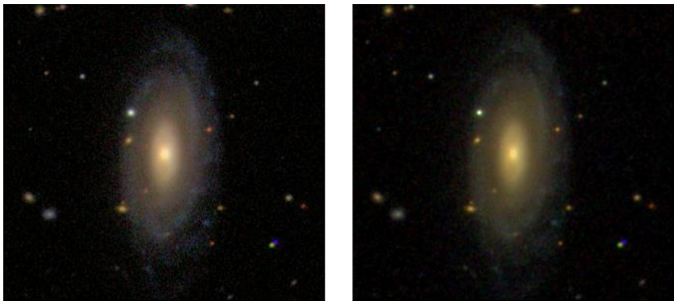

Fig. 1. Example of Data Release 7 (left) and Data Release 9 (right) images from SDSS.

The coordinates for the crop are calculated from these edge coordinates, where values further from the center are chosen. In the end, the center of the image, consisting mostly of only the galaxy is cropped out. These cleaned galaxy images, as

shown in Fig 2, were used to train and test the new method as well.

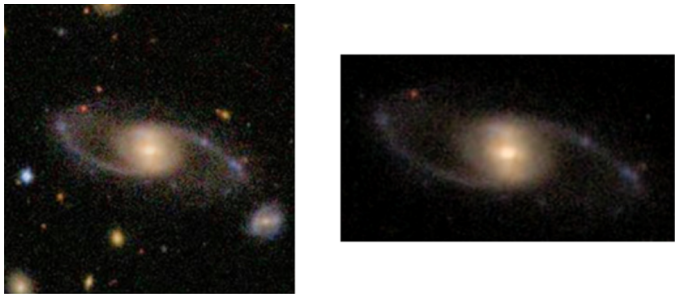

Fig. 2. The original (left) and cropped (right) images of a galaxy.

B. Faster R-CNN

At first, a Faster R-CNN type network has been developed, which follows mostly the original paper. This network consists of two main modules: Region Proposal Network and Fast R-CNN. The former is responsible for selecting regions where objects could be found. This is done by using anchors, which are placed evenly all over the image.

An anchor has multiple bounding boxes with different sizes and ratios, which is shown in Fig 3. By calculating the overlaps between these and the ground-truth boxes, foregrounds and backgrounds can be selected for training. At inference, the RPN calculates an objectness score for each anchor's bounding box. To reduce overlaps between these a non-maximum suppression is used. Then for training the overlaps between the remaining proposed regions and the ground-truth boxes are calculated, from which are the foregrounds and backgrounds selected for the Fast R-CNN. This network uses the feature maps of these RoIs (region of interest). The main difference from the original paper is that for the Fast R-CNN inputs, the feature map selection was switched from RoI pooling to RoIAlign. This new method uses bilinear interpolation that yields less information loss. This can be seen in Fig 4. The

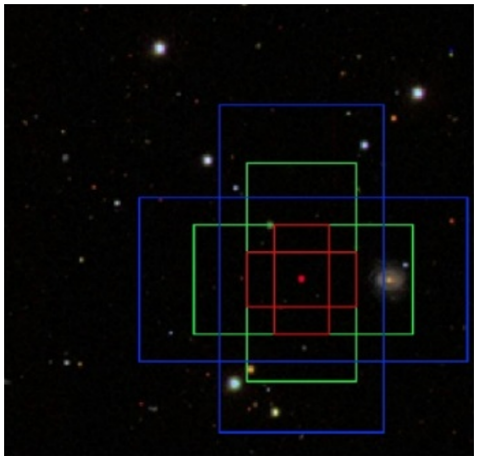

Fig. 3. The ground-truth boxes for an anchor.

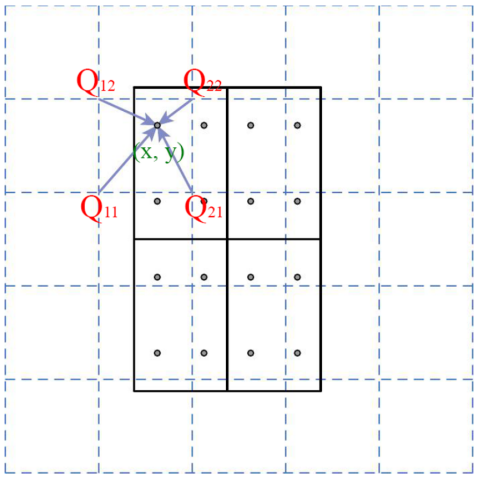

Fig. 4. The process of bilinear interpolation.

region is split into that many cells, which is the shape of the feature map, here i.e. 2×2. For every cell, 4 points are calculated with bilinear interpolation. Then the maximum of these four is inserted into the final feature map.

3000 sky images were generated for the train and test. Each image contained $12 - 12$ main classes and half of the spiral galaxies were ring-shaped. The computational intensive methods, i.e. RoiAlign and the calculation of the overlaps between ground-truth boxes and anchor boxes, were implemented on GPU with CUDA to make the process faster. While using a Ryzen 5 3600x and an Nvidia Geforce RTX 3060 Ti, the process speed-up was around 400x.

C. DBSCAN - CNN

Although early results from the Faster R-CNN were already promising, another method was developed for comparing its results. In this new method, the detection is done using a density-based clustering algorithm, the DBSCAN [9]. This algorithm has 3 hyperparameters: intensity cutoff, epsilon, and minPts. The first one defines at which pixel intensity the points are split into zeroes and ones. Next, the epsilon defines the radius in which at least as many points have to be ones as the minPts, to be considered as part of the cluster. Large images contain a huge amount of pixels and so to make the detection faster, the images were split into tiles. The process is shown in Fig 5.

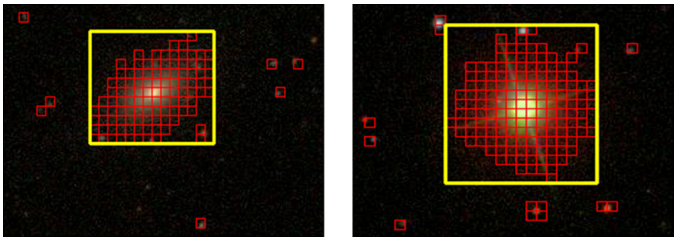

Fig. 5. DBSCAN detection of a galaxy (left) and a star (right) indicated as yellow. The red blocks are the tiles of the image and the search points of the algorithm.

DBSCAN has a few advantages compared to the Faster R-CNN. First and most importantly, it is more robust than the RPN, which means that it can detect any arbitrary-sized galaxy contrary to the RPN, which is bound to an interval caused by the anchors. Secondly, it can be used directly without any previous training. Lastly, an advantage can be, that there are no overlaps between the bounding boxes. However this could be a minor issue, if there is another galaxy in the same area of the image plane as another galaxy, then the algorithm detects them as one. Finding the optimal hyperparameters can be a challenge as well. If they are set too permissively, although it detects more galaxies, but also the false positive detections increase with it. Even if the optimal hyperparameters are set, there could be many false positives, because it is based only on the intensity. However, this issue can be solved easily by training the CNN [10]–[13] networks to classify stars or other background noises as well. For this, 20 thousand background images were cropped out from the *"galaxyless"* sky images.

IV. Evaluation

The new method result is shown in Table 1. As seen in Table I, every metric shows that the new method is better than any other solution. Compared to the Faster R-CNN, there has been an increase in the main class accuracy and more importantly a significant increase in the ringed shape F1 score. The latter is also noticeable in Fig 6 and Fig 7 precision–recall curves of the models. It is important to mention that Dieleman and Sánchez used different data and Dieleman predicted the vote probabilities itself, instead of the galaxy type of the image. The proposed methods have been also tested on sky images that were downloaded from the SDSS server, such that the galaxy is always in the center with a large surrounding sky area. The model was able to accurately detect these galaxies and classify them nearly as well as shown before with only a few percent difference. Galaxies detected by the two networks on an SDSS image are shown in Fig 8 and Fig 9. Although the DBSCAN did not detect two small galaxies (which is caused by the chosen size limit), it has fewer false positives and there are no overlaps between the bounding boxes.

V. Conclusion

The aim of this paper was to present the detection and classification of galaxies in sky images. In addition to determining the main category (spiral and elliptical), the goal was also to recognize the galaxy ring. Based on the review of the literature, it is most appropriate to implement this with neural networks.

TABLE I
Comparison of models

Model	Main class accuracy	Ring accuracy	Ring precision	Ring recall	Ring F1 score
CNN	99.7%	94.9%	95.7%	94.0%	94.9%
Faster R-CNN	92.8%	89.3%	58.1%	74.3%	65.2%
Dieleman	87.8%	-	90.1%	91.6%	91.3%
Sánchez	98.2%	-	-	-	-

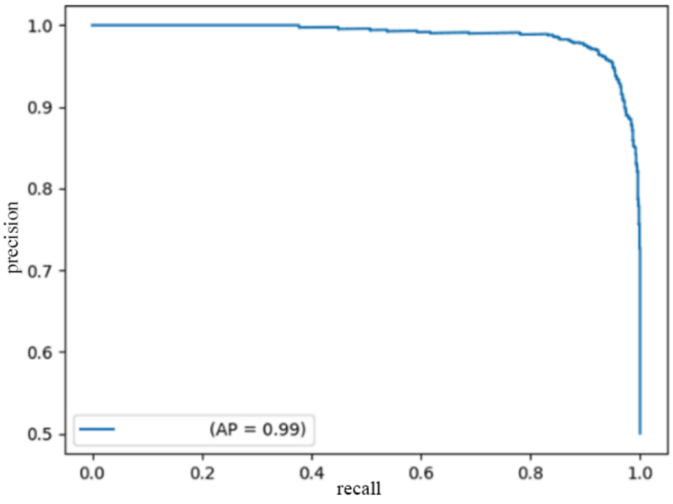

Fig. 6. Precision – Recall curve of the new CNN for ring classifications.

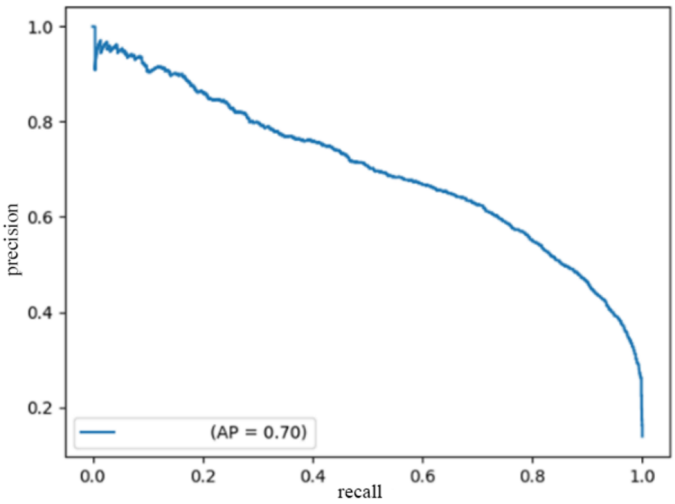

Fig. 7. Precision – Recall curve of the Faster R-CNN ring classifications.

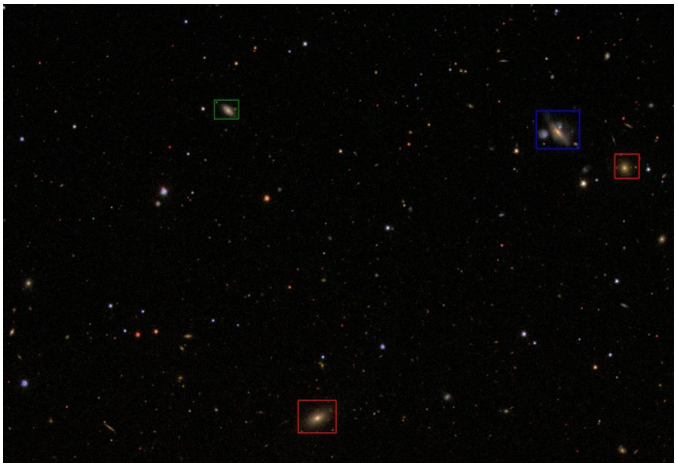

Fig. 8. Detected galaxies by the DBSCAN – CNN network on a real SDSS image. Classes are marked as blue for spiral, red for elliptical and green for ringed spiral.

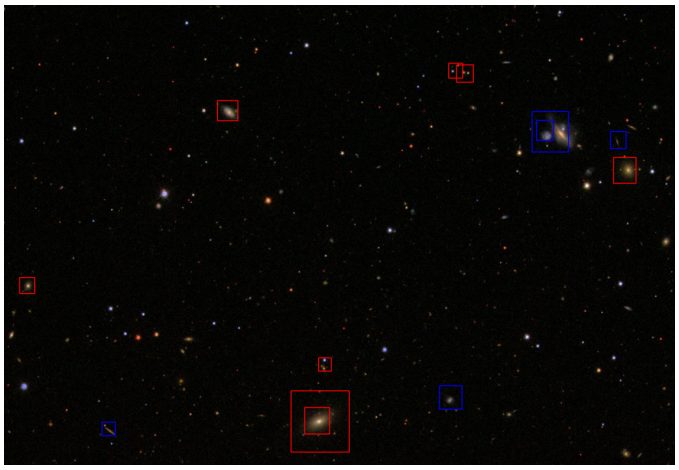

Fig. 9. Detected galaxies by the Faster R-CNN network on a real SDSS image. Classes are marked as blue for spiral, red for elliptical and green for ringed spiral.

Among these, Faster-RCNN seemed to be the best solution, so a network of this kind was built for the first time. Based on the first test results, Faster R-CNN achieves an accuracy of 93% for main categories and 89.3% for ring recognition. Then, another method was developed, in which the detection was done with DBSCAN, and the classifications were done with one CNN. Comparing the two methods, the newer one based on DBSCAN and CNN proved to be the better one, which achieved 99.7% main class, 94.9% ring accuracy, and 94.91% F1 score for rings.

The biggest change from other models focusing on galaxy detection is the use of DBSCAN. This allows the detection of arbitrary-sized objects, compared to the restricted networks. The reasons behind better classification are not clear. It could have been caused by a different dataset or network. However, feeding the raw data of the detected area to the CNNs, possibly makes the classification better as well. To improve classifications, images from other telescopes would be useful for generalization. That way, galaxies from a new telescope could be classified more accurately. This method, with minor modifications, could be able to detect and classify desired galaxy types when applied to large datasets. To shorten the required time for this, some form of parallelization would be necessary.

VI. Acknowledgments

The authors would like to thank the Hungarian National Talent Program (NTP-HHTDK-22) for its valuable support.

References

[1] Sander Dieleman, Kyle W. Willett, Joni Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction", *Monthly Notices of the Royal Astronomical Society*, Vol. 450, No. 2, pp. 1441–1459, 2015.

[2] H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D. Tuccillo and J. L. Fischer, "Improving galaxy morphologies for SDSS with Deep Learning", *Monthly Notices of the Royal Astronomical Society*, Vol. 476, No. 3, pp. 3661-3676, 2018.

[3] Xiao-Pan Zhu, Jia-Ming Dai, Chun-Jiang Bian, Yu Chen, Shi Chen, Chen Hu, "Galaxy morphology classification with deep convolutional neural networks", *Astrophysics and Space Science*, Vol. 364, No. 4, pp. 1-12, 2019.

[4] Roberto E. González, Roberto P. Munoza, Cristian A. Hernández, "Galaxy detection and identification using deep learning and data augmentation", *Astronomy and Computing*, Vol. 25, pp. 103-109, 2018.

[5] Colin J. Burke, Patrick D. Aleo, Yu-Ching Chen, Xin Liu, John R. Peterson, Glenn H. Sembroski, Joshua Yao-Yu Lin, "Deblending and Classifying Astronomical Sources with Mask R-CNN Deep Learning", *Monthly Notices of the Royal Astronomical Society*, Vol. 490, No. 3, pp. 3952-3965, 2019.

[6] GalaxyZoo [Online], Available at: https://data.galaxyzoo.org/ [Accessed: 28 July 2023].

[7] Ross E. Hart, Steven P. Bamford, Kyle W. Willett, Karen L. Masters, Carolin Cardamone, Chris J. Lintott, Robert J. Mackay, Robert C. Nichol, Christopher K. Rosslowe, Brooke D. Simmons, Rebecca J. Smethurst, "Galaxy Zoo: Comparing the demographics of spiral arm number and a new method for correcting redshift bias", *Monthly Notices of the Royal Astronomical Society*, Vol. 461, No. 4, pp. 3363-3682, 2016.

[8] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, 2015.

[9] Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226-231, 2019.

[10] Fatemeh Rashidi Fathabadi, Janos L. Grantner, and Ikhlas Abdel-Qader, "Box-Trainer Assessment System with Real-Time Multi-Class Detection and Tracking of Laparoscopic Instruments, using CNN", *Acta Polytechnica Hungarica*, Vol. 19, No. 2, 2022.

[11] Iris Iddaly Méndez-Gurrola, Abdiel Ramírez-Reyes, "A Review and Perspective on the Main Machine Learning Methods Applied to Physical Sciences", *Acta Polytechnica Hungarica*, Vol. 19, No. 10, 2022.

[12] Wang, Fei-Yue, Ding, Wenwen, Wang, Xiao, Garibaldi, Jon, Teng, Siyu, Imre, Rudas, Olaverri-Monreal, Cristina, "The DAO to DeSci: AI for Free, Fair, and Responsibility Sensitive Sciences", *IEEE INTELLIGENT SYSTEMS* Vol. 37, No. 2, 2022.

[13] Biro, Attila, Tunde Janosi-Rancz, Katalin, Szilagyi, Laszlo, Ignacio Cuesta-Vargas, Antonio, Martin-Martin, Jaime, Miklos Szilagyi, Sandor, "Visual Object Detection with DETR to Support Video-Diagnosis Using Conference Tools", *APPLIED SCIENCES-BASEL*, Vol. 12, No. 12, 2022.