Science Bulletin 66 (2021) 2145–2147

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

News & Views

Artificial intelligence for celestial object census: the latest technology meets the oldest science

Baoqiang Lao [a,*], Tao An [a,*], Ailing Wang [a,b], Zhijun Xu [a], Shaoguang Guo [a], Weijia Lv [a], Xiaocong Wu [a], Yingkang Zhang [a]

[a] Shanghai Astronomical Observatory, Key Laboratory of Radio Astronomy, Chinese Academy of Sciences, Shanghai 200030, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China

Astronomy is the oldest natural science based on observation, and a census of objects in the sky map to create a catalog is the basis for further research. This effort is achieved through astronomical object detection, also known as "source finding", which aims to identify individual objects in an astronomical image and then retrieve the properties of those objects to form a catalog. The completeness, reliability, and accuracy of the resulting catalog has a profound impact on astrophysical research.

We are currently in an era of explosive information growth, where big data is revolutionizing human life, as well as changing the paradigm of scientific research. For example, the Large Synoptic Survey Telescope (LSST)[1] under construction will generate up to 20 terabytes of raw data per night, a scale comparable to that of the Sloan Digital Sky Survey (SDSS)[2] in a decade! The Euclid space mission[3] is expected to create approximately tens of petabytes of total data. The Square Kilometre Array (SKA)[4] is taking the scale of astronomical big data to a new level, generating raw data in a rate of several Tera bits per second in its first phase (10% of the total scale) and 700 petabytes of scientific data per year [1,2]. The challenge for astronomers around the world is how to access and utilize this massive amount of information.

Sky surveys with modern telescopes have led to a dramatic increase in image size and quality, presenting enormous challenges as well as opportunities for new discoveries. As an example, the Australian SKA Pathfinder (ASKAP) all-sky survey is expected to detect 70 million radio galaxies [3], and the classification and morphology of these radio sources provide key information for understanding the formation and evolution of the Universe. However, it is impossible to identify such a vast amount of objects by visual inspection, and classifying the extracted sources is even challenging. To tackle these challenges, algorithms for automatic source finding and classification need to be developed [4].

Early source finding algorithms were integrated in data processing software packages. To process large astronomical data, a number of stand-alone source finding software packages have been developed and offered higher reliability and accuracy than the old ones. Artificial intelligence (AI) technology has been widely used in industries, such as geological monitoring, robotics, autonomous driving, face recognition, medical image analysis, etc. Compared to standard (non-AI) source finders in radio astronomy, AI-based automated methods, especially deep learning (DL) methods, aided by the acceleration of graphics processing unit (GPU) devices, offer pronounced advantages in terms of operating speed. Moreover, machine learning can also analyze data without our instructions, i.e., it can identify unexpected patterns, e.g., identifying more types of galaxies. This new discovery capability will certainly improve our understanding of the Universe.

Convolutional neural networks (CNNs) are the best performing image recognition classifiers in academia and industry. Recently, many DL-based astronomical source detectors or classifiers have been developed ([5,6] and references therein), and they are divided into single-stage and two-stage algorithms: the former is faster but less accurate and suitable for fast detection; the later is more accurate but slower. A suitable CNN architecture needs to be chosen according to the task requirements, with a trade-off between recognition precision and computational cost.

Among the region-based CNN (R-CNN) family, Faster R-CNN has the advantage of faster training speed and easier data annotation. CLARAN [5] v0.1 is a detector built on top of Faster R-CNN, and it can locate and associate the radio source components with ~90% precision, making it one of the highest-precision CNN-based classifiers. However, the performance of CLARAN v0.1 is limited by the backbone network it uses, Visual Geometry Group Network (VGGNet). When the number of network layers of VGGNet increases, the model complexity increases and the recognition performance decreases accordingly; in addition, increasing parameters significantly increases the computational complexity, training time, and GPU memory usage. Another shortcoming of CLARAN v0.1 is that it can only classify radio sources by peak and component, and lacks relevance for extended sources. In fact, most of the commonly used source finders can only identify compact point-like sources, rather than directly identifying and classifying

* Corresponding authors.
E-mail addresses: lbq@shao.ac.cn (B. Lao), antao@shao.ac.cn (T. An).

[1] LSST to explore the deepest and widest optical sky: https://www.lsst.org/.
[2] SDSS is one of the most successful surveys in the history of astronomy: https://www.sdss.org/.
[3] Euclid is an ESA medium class astronomy and astrophysics space mission: https://www.euclid-ec.org/.
[4] SKA is the flagship telescope being built in the field of radio astronomy: https://www.skatelescope.org/.

https://doi.org/10.1016/j.scib.2021.07.015

extended sources, which is done by visual inspection in post-processing.

To overcome these shortcomings, we construct a new source finder, named HETU[5], based on an improved CNN model and a new classification method. HETU uses a combined network structure with Residual Network (ResNet) and Feature Pyramid Networks (FPN) as the backbone network. It exploits the advantage of ResNet in balancing recognition precision and computational cost and the advantage of FPN in multi-feature object detection. As a result, HETU not only increases the network depth, but also provides multi-scale feature maps without causing a significant decrease in running speed. In this study, we used two different layers (50 and 101 layers) of ResNet, the generated models are called HETU-50 and HETU-101, respectively. The backbone network ResNet50-FPN is also used in CLARAN v0.2 [4]. The workflow of HETU is depicted in Fig. S1 (online) and the HETU network is discussed in details (see the Supplementary materials Sections 2.3 and 2.4).

We run three experiments to verify the performance of HETU: (1) the training experiment; (2) the testing experiment; (3) the predicting experiment. All experiments were conducted on the China SKA Regional Centre Prototype [7] (see the Supplementary materials Section 3).

In the training experiment, we used the same datasets and the same source classification scheme as CLARAN [5] in order to compare the results. The metric of the mean Average Precision (*mAP*) [8] increases from 78.4% for CLARAN to 86.7% for HETU-50 and to 87.6% for HETU-101 (Table S3 online), indicating a significant improvement in the recognition performance of HETU compared to CLARAN. *mAP*s obtained from HETU are also much higher than those derived from the ResNet models alone (Tables S3–S5 online), validating the higher performance of the combined ResNet-FPN network. The deeper HETU-101 network increases the precision by 0.9% over HETU-50, therefore, HETU-101 is used for both testing and predicting experiments. We also found that HETU's performance is not strongly dependent on the dataset used and it is therefore widely adaptable. HETU supports parallel execution using multiple GPU devices, and in our training experiment the training speed is 2.5 times faster than without parallelism.

In the testing experiment (see the Supplementary materials Section 3.2), we used a different source classification scheme from CLARAN. HETU automatically locates radio sources in the images and at the same time assigns them to one of the four classes according to their morphology: compact point-like sources (CS), Fanaroff-Riley type I (FRI) sources characterized with a central core and prominent two-sided jets which are weaker further from the core, Fanaroff-Riley type II (FRII) sources characterized by two prominent terminal components with symmetric shapes, and core-jet (CJ) sources showing a bright core component at one end of an elongated weaker jet feature. This classification scheme encompasses most of the radio sources with practical astrophysical meaning. We re-labelled all images of the training dataset by visual recognition according to the new classification scheme. To avoid overfitting due to the imbalance of the different classes, we used the data augmentation technique to enlarge the FRI, FRII and CJ samples. It took 4.9 h to train the workflow of HETU-101 over 40,000 steps for the re-labelled augmented dataset on 8 GPU devices. The processing time is about 5.4 ms per image, two orders of magnitude faster than the visual recognition. *mAP* is 94.2% for the re-labelled augmented dataset, 4.3% higher than the unaugmented dataset (Table S7 online). The average precisions (*AP*s) for some source classes are as high as 0.994 (CS) and 0.981 (FRII). After augmentation, increasing the network depth did not greatly improve the recognition performance. Moreover, the total loss curves show that the training model for the augmented dataset is stable for all classes (Fig. S9 online).

Based on the successful establishment of the training set and CNN model from the training experiment, we applied HETU to the practical astronomical data processing (the predicting experiment, see details in the Supplementary materials Section 3.3). We used HETU for source detection and classification on the images from the all-sky survey GLEAM [9] observed with the SKA-low precursor telescope MWA [10], and compared the results with those obtained with the traditional source finding software AEGEAN [11]. HETU's detection (and classification) speed is 100 ms per image, 21 times faster than AEGEAN. If only the identification task is performed without classification (Gaussian fitting), HETU's runtime is even ~2.5 times faster. We cross-matched the sources detected by HETU and AEGEAN with a search radius of 30 arcsec. The cross-matching fraction varies when different detection thresholds are adopted (Table S8 online). For example, when the detection threshold is 6σ, the cross-matched CS objects account for 94.5% of the HETU-detected CS sources and 94.3% for the AEGEAN-detected CS sources. If the detection threshold is set to 5σ, the cross-match rates change to 96.9% and 89.2% for the HETU and AEGEAN CS catalogs. A lower detection threshold results in more weaker sources detected, but at the cost of introducing more fake sources. A large fraction of the un-matched sources are found at the image edges (e.g., Fig. S14 online), and they are discarded by HETU since HETU considers them morphologically incomplete. At lower thresholds, AEGEAN detects fake sources associated with sidelobes of very bright sources, which are not identified by HETU. The predicting experiment shows that HETU not only has high recognition precision, but also has excellent ability in identifying weak sources (Fig. S16 online).

HETU is able to classify the detected sources into relevant classes while recognizing them; in contrast, AEGEAN only identifies the components of a source and can not directly determine whether there is a connection between adjacent components, leading to the classification of extended sources to be done in an offline manner by visual inspection. After associating the HETU-detected extended sources with the brightest component of the corresponding AEGEAN-detected sources, we found that the cross-match rate is 100% for FRII, 97.4% for FRI and 97.6% for CJ classes (Table S8 online), respectively, indicating that HETU performs very well in identifying extended sources.

The ongoing and upcoming large radio continuum survey projects using the SKA pathfinder telescopes[6] and SKA itself will produce a tremendous amount of images. Automated and accurate source finding and classification tools are particularly important to support these large sky surveys and to mine the data archive. Future predicting experiments will be performed to further improve HETU's recognition performance and speed, to support larger-scale images and to focus more on extended sources.

Neural networks have a deeper understanding of data than expected, but require large data sets for training (learning), and the vast Universe provides neural networks with a naturally enormous amount of data, and AI will undoubtedly have a profound impact on astronomy. However, it is important to note that AI can only perform certain tasks well if there is a large, correctly labelled dataset to learn from, and the trained model performs a single type of tasks. In other words, AI is not an "all-around champion". But even so, the speed and efficiency of AI is increasingly shaping our understanding of the natural world. The network framework of HETU is used not only for astronomical source identi-

[5] HETU is named after two mysterious patterns handed down from ancient China, which contains profound cosmic astrology.

[6] See latest advances of the SKA pathfinder telescopes at 2021 SKA Science Meeting: https://www.skatelescope.org/skascicon21/.

fication and classification but also in other fields such as medical CT image analysis (e.g., automated tumor detection).

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This work was supported by the National Key R&D Program of China (2018YFA0404603), the International Partnership Program of Chinese Academy of Sciences (114231KYSB20170003) and the Youth Association for Promoting Innovation. We thank Chen Wu, Qi Dang and Xiaofeng Li for their help on experiment implementation. We thank Natasha Hurley-Walker, Sumit Jaiswal, Ivy Wong, and Xiaolong Yang for helpful discussion.

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.scib.2021.07.015.

References

[1] An T. Science opportunities and challenges associated with SKA big data. Sci China Phys Mech Astro 2019;62:989531.
[2] Quinn P, van Haarlem MP, An T, et al. SKA regional centres white paper v1.0 2020.
[3] Norris RP, Hopkins AM, Afonso J, et al. EMU: evolutionary map of the universe. Publ Astron Soc Aust 2011;28:215–48.
[4] Bonaldi A, An T, Brueggen M, et al. Square kilometre array science data challenge 1: analysis and results. Mon Not Roy Astron Soc 2021;500:3821–37.
[5] Wu C, Wong OI, Rudnick L, et al. Radio galaxy zoo: claran – a deep learning classifier for radio morphologies. Mon Not Roy Astron Soc 2019;482:1211–30.
[6] Becker B, Vaccari M, Prescott M, et al. Cnn architecture comparison for radio galaxy classification. Mon Not Roy Astron Soc 2021;503:1828–46.
[7] An T, Wu XP, Hong X. SKA data take centre stage in China. Nat Astron 2019;3:1030.
[8] Lin TY, Dollar P, Girshick R, et al. Feature pyramid networks for object detection. Proc IEEE Conf Compu Visi Patt Recog 2017:936–44.
[9] Hurley-Walker N, Callingham JR, Hancock PJ, et al. GaLactic and extragalactic all-sky murchison widefield array (gleam) survey-I. a low-frequency extragalactic catalogue. Mon Not Roy Astron Soc 2017;464:1146–67.
[10] Tingay SJ, Goeke R, Bowman JD, et al. The murchison widefield array: the square kilometre array precursor at low radio frequencies. Publ Astron Soc Aust 2013;30:e007.
[11] Hancock PJ, Trott CM, Hurley-Walker N. Source finding in the era of the SKA (precursors): aegean 2.0. Publ Astron Soc Aust 2018;35:e011.

Baoqiang Lao received the M.Sc. degree from Guilin University of Electronic Technology, China in 2015. He is currently a software engineer of Shanghai Astronomical Observatory of the Chinese Academy of Sciences. His research area mainly focuses on the application of artificial intelligence and HPC performance computing in Square Kilometre Array (SKA) data processing.

Tao An is a professor at the Shanghai Astronomical Observatory of the Chinese Academy of Sciences. He is Member of SKA Regional Centre Streering Committee and International Astronomical Society (IAU) Commission B4, and co-chair of SKA VLBI science working group. He is leading the China SKA Regional Centre Prototype Construction. His research fields are astrophysics, radio astronomy, and very long baseline interferometry (VLBI).