Astronomical Data Analysis Software and Systems XXVII
ASP Conference Series, Vol. 522
Pascal Ballester, Jorge Ibsen, Mauricio Solar, and Keith Shortridge, eds.

AstroCV - A computer Vision Library for Astronomy

Roberto E. González[1,2] and Roberto P. Muñoz[1]

[1]*Centro I+D MetricArts, Santiago, Chile;* `regonzar@astro.puc.cl`

[2]*Centro de Astro-Ingeniería, Pontificia Universidad Católica de Chile, Santiago, Chile*

Abstract. We present AstroCV, a computer vision library for processing and analyzing big astronomical datasets. The goal of AstroCV is to provide a community repository of high performance Python and C++ algorithms used in the areas of image processing and computer vision.

The current AstroCV library includes methods for the tasks of object recognition, segmentation and classification, with emphasis in the automatic detection and classification of galaxies. The underlying models were trained using convolutional neural networks and deep learning techniques, which provide better results than methods based on manual feature engineering and SVMs. The detection and classification methods were trained end-to-end using public datasets such as the Sloan Digital Sky Survey (SDSS) and Galaxy Zoo, and private datasets such as the Next Generation Virgo (NGVS) and Fornax (NGFS) surveys.

The detection and classification methods were trained using the deep learning framework DARKNET and the real-time object detection system YOLO. These methods are implemented in C and CUDA languages and makes intensive use of graphical processing units (GPU). Using a single high-end Nvidia GPU card, it can process a SDSS image in 50 milliseconds and a DECam image in less than 3 seconds.

We provide the open source code, pre-trained networks and python tutorials for using the AstroCV library.

1. Introduction

Astronomical datasets are constantly increasing in size and complexity. The new generation of integral field units (IFUs) are generating about 60GB of data per night while imaging instruments are generating 300GB per night. The Large Synoptic Survey Telescope (LSST) is under construction in Chile and it will start full operations in 2022. It will be equipped with a 3 Gigapixel camera and will generate 15TB of data per night. Dealing with these increasing and more complex amounts of data is one of the main challenges in Astroinformatics.

Computer Vision is an interdisciplinary field that focuses in how machines can emulate the way in which human's brains and eyes work together to visually process the world around them. For many years, the detection of objects was computed using manual feature engineering and descriptors such as SIFT and HOG. Thanks to the advent of large annotated datasets and gains in computing power, deep learning methods has become the favorite for doing detection and classification of objects.

In addition, these efforts are complemented with new machine learning frameworks and new specialized hardware, in particular NVIDIA has developed new specialized hardware with that purpose, and new libraries/platforms based on CUDA, sucha as CUDNN or TensorRT.

We present AstroCV, a computer vision library for processing and analyzing big astronomical datasets. The goal of AstroCV is to provide a community repository for fast Python implementations of common tools and routines used in the areas of image processing and computer vision. In particular, it is focused in the tasks of object recognition, segmentation and classification, applied to astronomical objects.

In this paper we will focus in the automatic detection and classification of galaxies. The detection and classification methods were trained end-to-end using public datasets from the Sloan Digital Sky Survey (SDSS), (Alam et al. 2015), and Galaxy Zoo (Lintott et al. 2008, 2011) explained in section 2.1 We use YOLO method, (Redmon et al. 2015), for object detection which is explained in section 2.2. , and results are shown in section 3.

The open source code, training datasets, documentation and python notebooks of AstroCV are freely available in a Github repository[1].

2. Data and Training

2.1. Training dataset

Galaxy Zoo [2] (Lintott et al. 2008, 2011) is the most successful citizen project in Astronomy. It consists of a web platform for doing visual inspection of astronomical images and morphological classification of galaxies. Hundreds of thousands of volunteers classified images of nearly $900,000$ galaxies drawn from the SDSS survey. The Galaxy Zoo classification consists of six categories: elliptical, clockwise spiral, anticlockwise spiral, edge-on, star/do not know, or merger.

We extracted the galaxy classification for a sub-sample of $22,873$ galaxies and downloaded their respective gri-band images from the SDSS fields. For each field, we generate color RGB images using Lupton et al. (2004) method, and we select the galaxies with a petrosian radius larger than 20 pixels. We split this dataset into a training and validation datasets (See Table 1).

We generated two different sets of RGB images: Train1 was created using the default parameters of Lupton et al. (2004) method, while Train2 was created using higher contrast and brightness[3]. The purpose of this second training set is to test color scale effects and to resemble a sample with similar background color from typical HST images converted to RGB.

In the training process, we look for the optimal number of iterations where the network converges. We look for maximum recall ratio (fraction of detections recovered), and IOU (Intersect over union, overlap fraction between the detected and real bounding box). Results are shown in figure 1, where we get for Train 1, maximum recall ratio and IOU of 85.7% and 65.8% respectively. In the case of Train 2, maximum recall ratio and

[1] https://github.com/astroCV.

[2] https://www.galaxyzoo.org/.

[3] Contrast and Brightness enhanced by 2 using ImageEnhance python library.

Table 1. Annotations in training and validation datasets.

Dataset	Elliptical	Spiral	Edge-on	DK	Merge	Total	Number images
Training	10366	4535	4598	223	381	20103	6458
Validation	1261	714	723	27	45	2770	921

IOU rise to 90.23% and 70.3% respectively. Bear in mind that a larger training set may still increase these numbers, and that Train 2 returns better results since higher contrast and brightness images built from FITS images gives more information to the training.

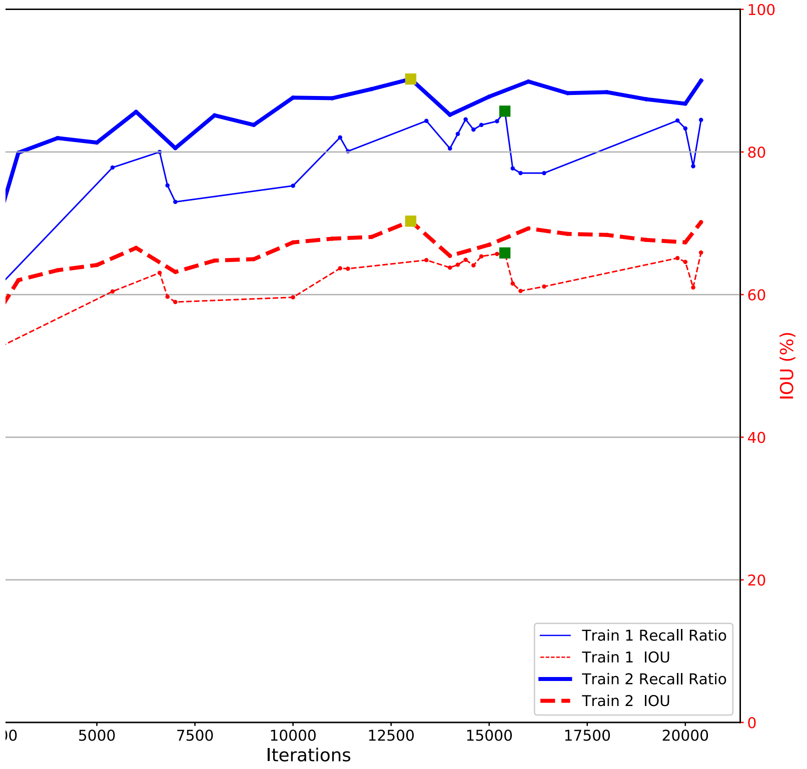

Figure 1. Convergence of the training set using nearly 20000 classified galaxies for two trainings. Green and yellow squares indicate optimal recall ratio and IOU for both trainings.

2.2. YOLO

You only look once (YOLO) method (Redmon et al. 2015; Redmon & Farhadi 2016), is a Single Shot Detector (SSD), it means it computes in a single network the region proposal and classifier. This method run the image on a Convolutional Neural Network (CNN) model and get the detection on a single pass. The network is composed with 13 convolution layers shown in figure 2, and it is programmed on Darknet, an open source neural network framework in C and CUDA. It is very fast and take full advantage of graphical processing units (GPU). This method formerly developed for person/object detection, is configured for the training and detection of galaxies.

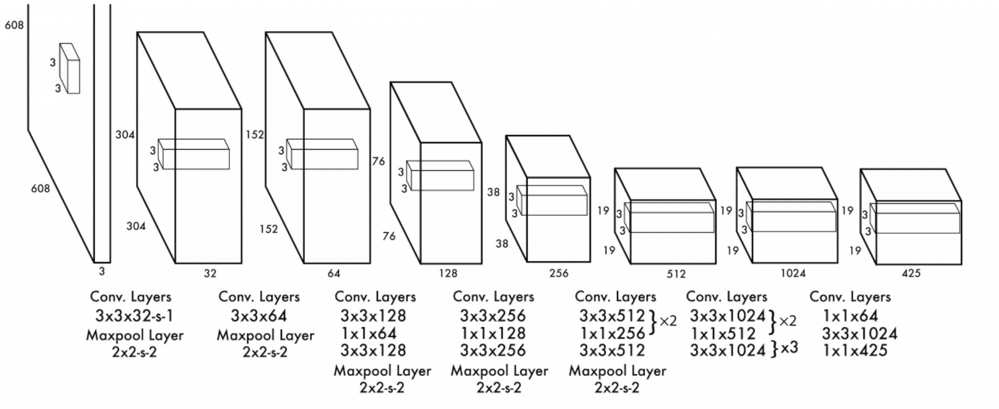

Figure 2. YOLO network schema.

3. Results and Discussion

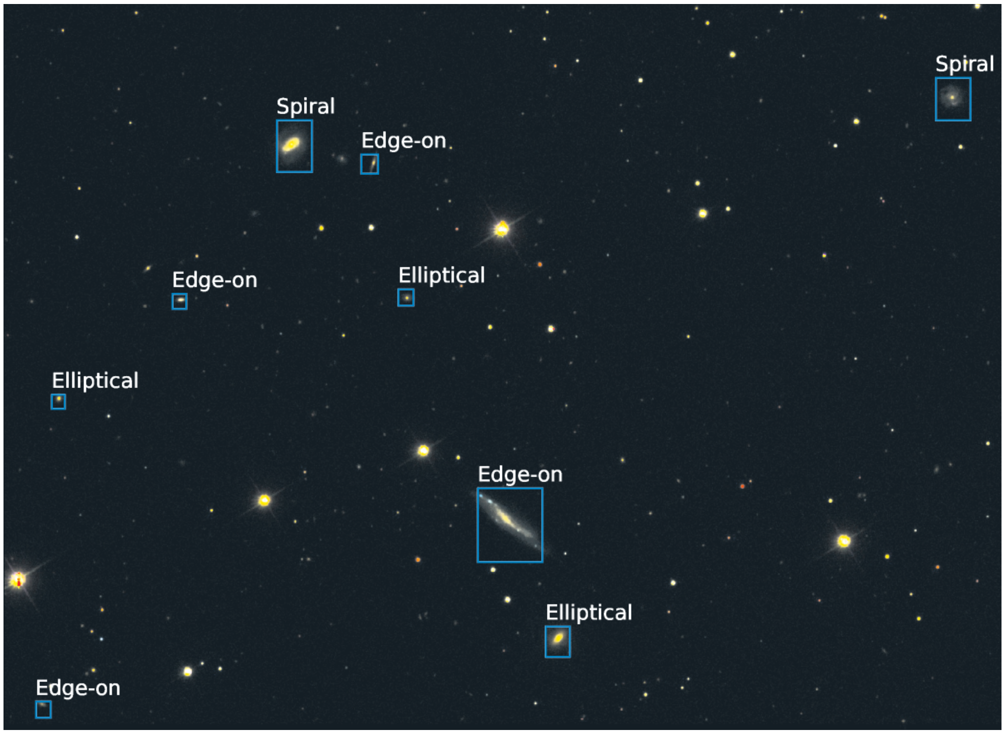

Figure 3. Detection and Classification of galaxies in a 2048x1489 image from SDSS; i, r, and g filters converted to rgb color space.

Table 2 show the confusion matrix, precision and recall for each class using the validation set. Notice that numbers does not match exactly with the validation set from table 1, this is because we are counting matches between true and predicted detections which include non-maximum suppression(NMS) issues, this correspond to the post-processing algorithm responsible for merging all detections that belong to the same object. From table 2, we show there is very little confusion between the three major classes elliptical, spiral and edge-on. In the case of DK and Merge classes the number of annotation is very small compared to the other classes resulting in poor performance for those classes. The total classification accuracy is 80%, which is high taking into account the nature of Galaxy Zoo classification based on votes, where some objects have divided votes but are assigned to a single classification (i.e. Dwarf spheroidal or S0 galaxies).

Table 2. Confusion matrix for galaxy classification using Train 2. Predicted values(columns) vs actual values(rows).

n=2756	Elliptical	Spiral	Edge-on	DK	Merge	Recall
Elliptical	1172	33	57	1	2	0.92
Spiral	176	469	69	0	3	0.65
Edge-on	96	60	554	0	3	0.78
DK	6	9	3	3	0	0.14
Merge	26	2	3	0	9	0.23
Precision	0.79	0.82	0.81	0.75	0.53	Accuracy=0.8

The AstroCV library provides computer vision and image processing tools for processing big astronomical datasets. These tools are focused in the tasks of object recognition, segmentation and classification of astronomical objects using deep learning and neural network frameworks. Most of these methods take advantage of GPUs capabilities, improving performance for real-time applications such as data reduction pipelines. As an example, running galaxy detection using a Titan-X Nvidia GPU card on a typical full HD image takes nearly 50ms, and for a DECAM 500Mpx images it takes less than three seconds.

The current classification model was created using a sample of SDSS galaxies from the Galaxy Zoo Project. In Figure 3, we show results of the method on an image from the SDSS. There are several other interesting datasets to be trained such as Galaxy Zoo 2, (Willett et al. 2013) and the Next Generation Fornax Survey Muñoz et al. (2015, NGFS). The Galaxy Zoo 2 extends the classification for nearly $300,000$ of the largest Galaxy Zoo 1 galaxies, including detailed features such as discs, bars, and arms. The NGFS survey extends the classification to low surface brightness galaxies down to $\mu_i =$ 28 mag arcsec^2.

The next generation of astronomical observatories, such as the Large Synoptic Survey Telescope (LSST) and the Extreme Large Telescope (ELT), will observe hundreds of thousands of galaxies and will generate Terabytes of data every night. We are planning to improve the scalability of AstroCV to do real-time processing of tomorrow's big astronomical datasets. Another interesting application we are exploring is the identification and classification of transient objects.

4. Acknowledgements

The authors acknowledges support from the NVIDIA developer program. The Centro I+D MetricArts acknowledges support from the NVIDIA Inception program and the Microsoft BizSpark Program. We are grateful to Patricio Cofre, Wilson Pais and Adrian Fernandez for helpful discussions.

Funding for SDSS-III has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, and the U.S. Department of Energy Office of Science. The SDSS-III web site is http://www.sdss3.org/.

References

Alam, S., Albareti, F. D., Allende Prieto, C., Anders, F., Anderson, S. F., Anderton, T., Andrews, B. H., Armengaud, E., Aubourg, É., Bailey, S., & et al. 2015, ApJS, 219, 12. 1501.00963

Lintott, C., Schawinski, K., Bamford, S., Slosar, A., Land, K., Thomas, D., Edmondson, E., Masters, K., Nichol, R. C., Raddick, M. J., Szalay, A., Andreescu, D., Murray, P., & Vandenberg, J. 2011, MNRAS, 410, 166. 1007.3265

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., & Vandenberg, J. 2008, MNRAS, 389, 1179. 0804.4483

Lupton, R., Blanton, M. R., Fekete, G., Hogg, D. W., O'Mullane, W., Szalay, A., & Wherry, N. 2004, PASP, 116, 133. astro-ph/0312483

Muñoz, R. P., Eigenthaler, P., Puzia, T. H., Taylor, M. A., Ordenes-Briceño, Y., Alamo-Martínez, K., Ribbeck, K. X., Ángel, S., Capaccioli, M., Côté, P., Ferrarese, L., Galaz, G., Hempel, M., Hilker, M., Jordán, A., Lançon, A., Mieske, S., Paolillo, M., Richtler, T., Sánchez-Janssen, R., & Zhang, H. 2015, ApJ, 813, L15. 1510.02475

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. 2015, ArXiv e-prints. 1506.02640

Redmon, J., & Farhadi, A. 2016, ArXiv e-prints. 1612.08242

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R. V., Edmondson, E. M., Fortson, L. F., Kaviraj, S., Keel, W. C., Melvin, T., Nichol, R. C., Raddick, M. J., Schawinski, K., Simpson, R. J., Skibba, R. A., Smith, A. M., & Thomas, D. 2013, MNRAS, 435, 2835. 1308.3496