Multiband image classification of astronomical objects

Ana Martinazzo, Nina S. T. Hirata
Institute of Mathematics and Statistics
University of São Paulo, São Paulo, Brazil
{amartina, nina}@ime.usp.br

***Abstract*—Astronomy has entered the era of large digital sky surveys, transitioning from a relatively data-scarce field of study to a very data-rich one. The images coming from these new surveys are hyperspectral (having up to a few dozen bands) and noisy (due to limitations on telescope resolution and atmospheric conditions), present faint and saturated signals, and can amount to tens of terabytes. This unique set of characteristics make them very attractive for trying out deep learning methods. In this paper, we present a multiband image classifier for stars, galaxies and quasars, and propose steps towards a semi-supervised scheme that could enable the discovery of new objects.**

I. Introduction

Unlike other natural sciences such as physics and biology, in which various kinds of experiments can be designed to validate theories, Astronomy relies almost entirely on images. Each photon captured by the lens or the mirrors of a telescope encodes information about its time of arrival, its spatial position and its energy content. Those three pieces of information should provide us with all there is that we can learn about the universe.

The development of observational astronomy at first naturally followed advances in photography. The first well-known attempt at taking a picture of an object at the sky happened in the 1830s. Later on, astronomers transitioned to using telescopes, which become larger and capable of capturing more photons at better spatial and temporal resolutions as technology evolves. This results in increasingly large amounts of images to process and analyze. Even nowadays, it is common practice for experts to pick a few dozens or hundreds of objects and then analyze their images by eye with the aid of specific image processing tools. However, with the newest generation of telescopes scanning our sky at rates of the order of terabytes per night [1], this kind of manual work is becoming unfeasible.

Much like observational astronomy, computer vision is also transitioning from rather small-scale, task-specific to large-scale, generalist techniques. Deep neural networks automatically learn high-level information without the need for designing specific feature extractors. All it needs to learn from data is lots of it. And current sky surveys are already providing us with a flood of high-quality data across a rich range of positions, time intervals and bandwidths.

Some researchers believe that the next step towards more data-driven science would be to recognize patterns in tabular data generated by a few well-stablished pieces of software. But couldn't we do better if we just let machine-learning models see the data in a more raw format – that is, in images?

In this short paper, we present a multiband image classifier trained on labeled samples from the Southern Photometric Local Universe Survey (S-PLUS) and propose steps toward a semi-supervised classification scheme that could, among other things, enable the discovery of new objects. The remainder of this paper is structured as follows: in Section II, a brief overview of related works is given; in Section III, our dataset, its pre-processing pipeline and our models are described; in Section IV, preliminary results achieved with these data and these models are presented; and, finally, in Section V, possibilities for further developments are enumerated.

II. Related Work

The most usual approach to analyzing astronomical data nowadays is running statistical or more traditional machine-learned models on tabular data generated from raw telescope images through a close-sourced, fixed pipeline. These data are referred to as catalogs of objects. Each row of a catalog includes an object's unique identification, its sky coordinates, and some of its physical properties, which are computed from the images using physical or probabilistic models. Those properties can be used to learn object classes, as in [2]. Two of the main limitations of this approach are: (i) its dependence on the quality of properties inferred through the pipeline, which often have large uncertainties due to saturated or faint signals, and (ii) its lack of information on the morphology of the objects. In [3], authors try to include detailed morphological information by fitting each object to a mathematical model, but it is a very slow method, unfeasible for the large amounts of data available.

Another approach is to analyze three-color composite images. Based on the kinds of properties that are to be emphasized and analyzed, bands from multispectral images are combined into a fake but representative RGB image. There are a few well-stablished algorithms for this, such as the Lupton algorithm [4]. A popular dataset of astronomical RGB images is the one from the Galaxy Zoo [5], a crowdsourced project in which volunteers classify galaxies based on their morphologies. Many works were built upon either the composite images or the Galaxy Zoo dataset [6], [7]. Using three-color images has obvious advantages, such as the possibility of fine tuning powerful models pre-trained on large datasets. However, one

may ponder why not use all the bands available, specially after so much effort was put in for designing proper instruments and collecting all these data.

A third approach is to analyze spectroscopic data. They are the most reliable ground truths and can be used to readily identify known physical processes that happen at specific wavelengths, which in turn can be used to separate objects into fine-grained subclasses. The main drawbacks of spectra are that they are considerably more expensive to collect, and thus are way less abundant than images; and that, like catalogs, they lack morphological information.

In this work, a method that uses all 12 bands available in our images is presented. We compare it with approaches based on catalogs, and verify that it is significantly more robust. To the best of our knowledge, directly using multispectral images is a novel approach to large-scale classification of galaxies, stars and quasars.

III. Method

A. Dataset

Three classes of astronomical objects are considered: stars, galaxies and quasars. Stars are the majority of point sources we are able to see in night sky with naked eye. Galaxies are systems of stars, gas and dust which are bound together by gravitational forces. Because they are made of many stars, they can look like extense objects. Quasars, or quasi-stellar objects, are more mysterious: they are the brightest objects known in the universe (with the exception of occasional stellar explosions), but they seem faint because they are so much further away than observable stars and galaxies. Quasars and stars can look very much alike, and devising trusty methods for distinguishing them is an active topic of research. It is of great interest to be able to identify quasars because they encode information about the distant universe.

Each object in our dataset is represented by three different sources of data: 12-band images, photometric catalogs, and spectroscopic catalogs. Our dataset corresponds to a region of the sky known as Stripe 82, which is very well-studied and has been imaged by various telescopes, making it ideal for developing new techniques that may later be extended to more regions of the sky. The multiband images and the photometric catalogs were obtained from the first Data Release of the Southern Photometric Local Universe Survey (S-PLUS) [8] through a collaboration with researchers from the Institute of Astronomy and Geophysics of the University of São Paulo (IAG-USP), whereas the spectroscopic catalogs were downloaded from the Sloan Digital Sky Survey (SDSS) [9], whose data is mostly public. Object classes derived from spectra are used as ground truths for training and testing our models.

The multiband images from the S-PLUS are 11000x11000px images collected by five broad-band filters, which are widely adopted by various sky surveys, and seven narrow-band filters, which were designed by the S-PLUS team to capture processes that happen at specific wavelengths, trying to mimick what could be captured by a spectrometer. These images, in their most raw format, would be a pixelwise count of photons captured in each band. The images made available by the S-PLUS team have already been preprocessed in the telescope pipeline, which includes tasks such as stacking images of the same area of the sky, calibrating the signals between bands, and subtracting the background signal. This yields images with 32-bit float values, as opposed to the usual 8-bit unsigned integer images.

To use these images as inputs to deep learning models, their values were scaled to the [0,1] range using the minima and maxima per band. Moreover, information from the catalogs (pixel coordinates and full-width half-maximum[1] of the objects) was used to generate square crops of the objects. Knowing that 99% of the objects in our catalog have full-width half-maximum smaller or equal to 20px, we found that 32x32px crops were adequate to frame the majority of the objects, keeping enough context (e.g. visible dust around the center) while avoiding cluttering. For the few objects that ended up larger than 32px, a squared area of the size of the object was cropped and resized to 32x32px.

The photometric catalogs are also generated in the preprocessing pipeline of the telescope. This pipeline relies on an astronomical tool that detects objects using traditional approaches, such as background subtraction and adaptive thresholding, which works sufficiently well for this case where there is a dark background with bright point objects. Then, it generates a catalog of detected objects with information such as sky coordinates, magnitudes (a measure of brightness) in each band, signal-to-noise ratios, and full-width half-maximum. The 12 magnitudes (one per band) were picked as our feature set.

The spectroscopic catalogs were downloaded from the SDSS server and matched against the photometric catalogs by searching for objects whose sky coordinates were equal within a tolerance of one arcsec. This yields a single catalog including photometric properties and spectroscopic object classes. After the catalog matching step, our dataset was significantly reduced: from the nearly three million objects detected in S-PLUS images, only about 115.000 have a spectroscopic counterpart. This was expected, given that collecting spectra is an expensive and time-consuming procedure, and cannot be performed in large scale. In this work, only the labeled samples were used.

In previous machine-learning-based works, a subset of the objects is selected, for instance, discarding objects that are too bright or too faint, that have low signal-to-noise ratio, or that are overlapping. This may give a false impression that the models were accurate, when in fact they had not been trained nor tested with harder examples. In this work, two datasets are compared: a filtered set of objects whose magnitude is in the range [16, 19], which excludes saturated and faint objects, and the complete set. For the filtered set, the prevalence of classes is 49% galaxies, 49% stars and 2% quasars, whereas

[1]The full-width half-maximum refers to the width of a Gaussian at half of its maximum value; it is used to approximate the diameter of point sources that do not have sharp edges.

for the complete set, it is 50% galaxies, 39% stars and 11% quasars.

B. Models

Neural networks were used to train classifiers based on catalogs and images. For catalogs, a simple fully-connected network with two hidden layers having 512 nodes each was chosen. We experimented with the number of hidden layers and the number of nodes per layer, and empirically found the values which yielded complex and expressive enough models just before the onset of overfitting.

For images, ResNeXt [10] was chosen as our feature extractor and adapted to receive our 12-band image tensors as inputs. It was our chosen architecture because it is a flexible, efficient architecture that makes extensive use of grouped convolutions. It has been shown in the AlexNet [11] that each convolution group consistently learns specialized features, and this can be helpful for larger numbers of bands (more experiments are needed to evaluate whether it does help). Two variants of the model were trained: one using only the filtered dataset, and another using the complete dataset. The cost function was adjusted to include class weights inversely proportional to their prevalences. The following parameters were set for the architecture: depth=29, width=16, cardinality=4.

All models were trained from scratch using the Adam optimizer with initial learning rate of 10^{-4}, reduced by a factor of $\sqrt{10^{-1}}$ when plateaus were reached. Data augmentation was not used.

IV. Preliminary Results

Results are reported for models trained and validated on the filtered set and on the complete set. The filtered set contains 52K samples, whereas the complete set contains 115K samples. In both cases, the sets were split into 80% train, 10% validation and 10% test (test sets have not been used yet). Tables I and II show confusion matrices for the validation sets of image and catalog classifiers, respectively. Values inside parentheses are for the filtered dataset, whereas values outside of the parentheses are for the complete dataset. Even though the classification task may seem more challenging when using the complete dataset with saturated and faint examples, it can be seen that both models get significantly better at classifying quasars.

It is of extreme importance to be able to reliably classify faint signals. For that reason, the performance of the models was also evaluated across varying magnitude ranges. Figure 2 shows curves of accuracy as a function of magnitude. Lower values of magnitude indicate brighter objects, whereas higher values indicate fainter objects. Both for images and for catalogs, results are reported only for the complete dataset. In order to generate those curves, the objects in the validation set were put into 22 bins according their magnitudes.

It is noticeable that the image classifier reaches nearly 100% accuracy up until a magnitude of 19 for stars and galaxies, that it remains more stable than the catalog classifier across all magnitudes, and also that class weights helped the model adjust to the least represented class. It can also be seen that the accuracy for stars and galaxies in the image classifier degrades after a magnitude of 21, presumably because of the larger number of quasars in that magnitude range.

TABLE I
Confusion matrix for image classifiers

	galaxy	star	quasar
galaxy	**0.94 (0.98)**	0.03 (0.01)	0.03 (0.01)
star	0.02 (0.01)	**0.95 (0.99)**	0.03 (0.00)
quasar	0.10 (0.16)	0.11 (0.10)	**0.79 (0.74)**

TABLE II
Confusion matrix for catalog classifiers

	galaxy	star	quasar
galaxy	**0.92 (0.91)**	0.06 (0.08)	0.02 (0.01)
star	0.18 (0.13)	**0.81 (0.84)**	0.01 (0.01)
quasar	0.27 (0.36)	0.08 (0.07)	**0.64 (0.57)**

We realized that, in addition to class imbalances, intra-class magnitude imbalances also significantly hurt the performance of the models. Magnitude imbalance is expected: each class of objects has a different nature and a different distribution of brightnesses. However, some of this magnitude imbalance could also be due to biases in the dataset, which need to be more carefully considered. The confusion between quasars and stars or faint galaxies is another challenge. Figure 1 shows an example of a quasar of magnitude 21 that was misclassified as a star with 96% of confidence. Introducing domain knowledge in the model could help it disentangle representations. Further steps to try and mitigate problems with imbalances and confusion are proposed in the following section.

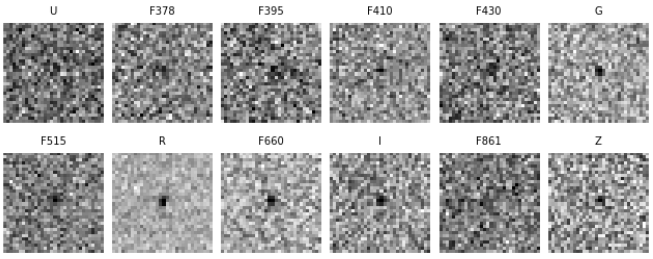

Fig. 1. A quasar that was misclassified as star

V. Next Steps

In this short paper, we presented an alternative approach for large-scale classification of astronomical objects, based on raw telescope images instead of tabular data inferred from the images. We believe that directly using the images to learn from data is more powerful and scalable, and that their potential should be explored together with modern data-driven machine learning techniques. As this is a work in progress, there are

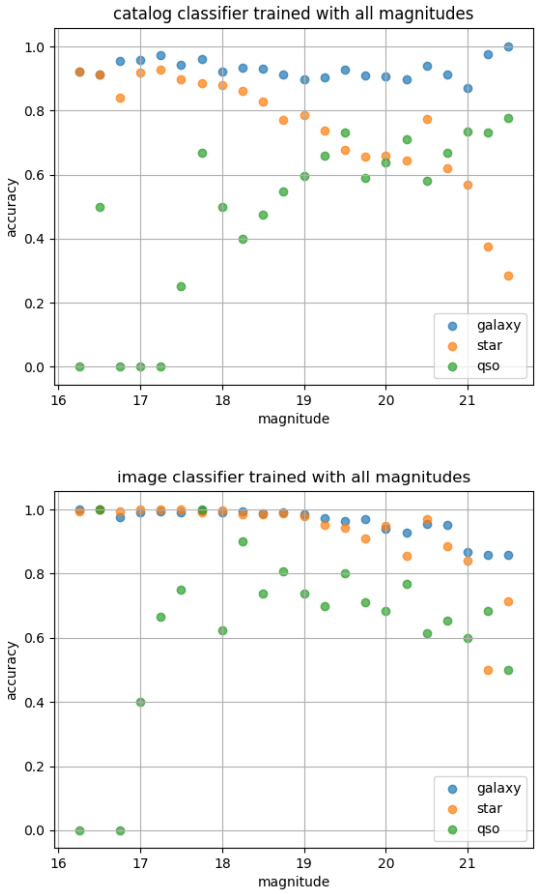

Fig. 2. Accuracy versus magnitude curves. Curve for catalog classifier on the top, and for image classifier on the bottom.

tons of ideas that are still flourishing. We will highlight some of them in this section.

Given that the majority of our samples are unlabeled, it would make sense to try semi-supervised approaches. We suspect that our labeled dataset is biased because objects for which spectroscopic data is available have been handpicked for decades, according to the research interests of the teams involved in the surveys. Thus, including the unlabeled samples in training is a priority. In addition to mitigating class imbalances, we expect that it will also lessen magnitude imbalances. In other words, in the unlabeled data there will be more examples of bright quasars and faint galaxies, which were lacking in the labeled data.

A straightforward step that is in progress is training a deep model that receives all three million objects (labeled and unlabeled) as inputs and learns their magnitudes. This way, the model would be learning to generate catalogs from images, that is, learning all the steps taken in the preprocessing pipeline of the telescope. In a sense, this could mean that the model would implicitly learn relevant physical processes. We'd like to assess whether this model would yield representations that could then be reliably used to separate the objects through clustering. It is not clear yet how we would measure and validate the performance of this approach.

We also intend to experiment more with training techniques and convolutional architectures. A very common technique that we have not employed yet is data augmentation. We can use standard transformations such as flipping and rotating, or a generative model. These augmentations would need to take into account both class and magnitude imbalances. However, we conjecture that higher performance gains could be achieved in our case by tweaking architectures, making them more efficient and sensible for the specificities of our dataset. For instance, we know that our objects are basically point sources (stars and quasars) or slightly elongated sources (galaxies), both of which could be reasonably modelled only by Gaussian kernels. We can use this prior knowledge to constrain the convolution kernels, reducing the number of parameters of the model while mantaining performance. We can also separate image bands in groups and test multi-branch architectures where each group of bands goes through an independent set of convolutions. Cleverly chosen band groups may better distinguish quasars from stars and faint galaxies. We seek to explore these alternatives in further works.

Acknowledgments

Ana Martinazzo receives support from the São Paulo Research Foundation (FAPESP), process 2018/25671-9. This work is also funded by grant 2017/25835-9 from FAPESP and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

We thank Mateus Espadoto and Eric K. Tokuda for assisting us with experiment design. We also thank Claudia M. de Oliveira and Lilianne M. I. Nakazono for guiding us through domain-specific questions, and Fabio R. Herpich and Laura Sampedro for preparing the S-PLUS data.

References

[1] J. A. Tyson and the LSST Collaboration, "Large Synoptic Survey Telescope: Overview," 2003.
[2] M. Duarte, L. M. Sampedro, and A. Molino, "The S-PLUS: the star/galaxy classification based on a Machine Learning approach," 2018.
[3] B. Sesar et al, "Machine-learned Identification of RR Lyrae Stars from Sparse, Multi-band Data: The PS1 Sample," 2017.
[4] R. Lupton et al, "Preparing Red-Green-Blue (RGB) Images from CCD Data," 2003.
[5] C. Lintott et al, "Galaxy Zoo 1: Data Release of Morphological Classifications for nearly 900,000 galaxies," January 2011.
[6] S. Dieleman, K. W. Willett, and J. Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," 2015.
[7] R. E. González, R. P. Muñoz, and C. A. Hernández, "Galaxy detection and identification using deep learning and data augmentation," 2018.
[8] C. M. de Oliveira et al, "The Southern Photometric Local Universe Survey (S-PLUS): improved SEDs, morphologies and redshifts with 12 optical filters," 2019.
[9] D. S. Aguado et al, "The Fifteenth Data Release of the Sloan Digital Sky Surveys: First Release of MaNGA Derived Quantities, Data Visualization Tools and Stellar Library," 2018.
[10] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated Residual Transformations for Deep Neural Networks," 2017.
[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," 2012.