MNRAS **000**, 1–4 (2020) Preprint 7 December 2020

Morphological Classification of Galaxies based on Residual Neuronal Networks and Transfer Learning

A. F. Gómez-Muñoz,[1]★
[1]*Instituto de Física, Facultad de Ciencias Exáctas y Naturales, Universidad de Antioquia (UdeA), Medellín, Colombia*

7 December 2020

ABSTRACT
The morphological classification of galaxies plays a fundamental role in the construction of astrophysical theories. Nevertheless, the classification task has its difficulties: there are large numbers of galaxies and it is very time consuming. Fortunately, two facts make the classification labor perfect to be tackled by a supervised machine learning algorithm: each galaxy image can be map to a given class and there are already ≈ 1e6 of them classified with a 90% expert accuracy. The construction and evaluation of such algorithm is the purpose of the present paper. For that we first analyze the symmetries present in the images to apply data augmentation. Then, we pass to explain the network architecture used to solve the problem and some motivations behind it. After that, we justify the choice of the Adam optimizer and the hyper-parameters for the training. From it we expose the results obtained for the mean square error (mse) and the accuracy of the model. Finally, we end concluding: that our model is consistent with other ones previously made, that within the approach taken preceding investigations are actually improved, that free tools can be used to tackle complex scientific problems, and with a set of ideas that could be used to improve the results and insights obtained.

Key words: galaxies: general – galaxies: spiral – galaxies: elliptical and lenticular – ResNet – galaxies: automated classification – galaxies: morphology classification – Galaxy Zoo Project– Transfer Learning.

1 INTRODUCTION

The morphological classification of galaxies plays a fundamental role in astrophysics, as it helps to test theories and find new conclusions to explain: the physical processes that govern galaxies, the formation of stars and the evolution of the universe. The central idea in this type of classification is simple: galaxies are divided into categories based on their morphology, generally equivalent to the visual appearance of the galaxy in the sky. Thus, the importance of these classification systems comes from the fact that the visual appearance broadly correlates with important physical parameters.

These types of classifications began with Hubble in 1936 Hubble. (1936) and most of the time have been a matter of expert astronomers, since the intensity and spectra of galaxies had to be carefully analyzed. However, despite the efforts of astronomers like Lahav et. al. (1995); Schawinski et al. (2007); Fukugita et. al. (2007) to classify more than 45,000 galaxies, these approaches have proven to be impractical. The main reason is that the classification process takes a long time and it must be done for millions of galaxy images; which we now have available thanks to projects like the SDSS (Sloan Digital Sky Survey).

These limitation have off course lead to new approaches. One of them, named Galaxy Zoo, have had great success in classifying around a million of galaxies from the SDSS images, with an expert accuracy of 90% Lintott et al. (2008); Willett et al. (2013). This insurmountable task was done with the help ≈ 100000 volunteer citizen scientists that past a simple test in order to make a guided classification by inspecting a set of images. Therefore, the results of this effort somewhat contain a prior distribution of the galaxy type with the image as a random variable.

These last ideas, together with the current outstanding development in machine learning object recognition (see Goodfellow et al. (2016) introduction), motivate the creation of a supervised classification algorithm for the present classification problem. The construction of such algorithm is the intention of the present article.

To begin, it must be recognized that this problem has been already tackled by hundreds of people. Fact that is well evidenced by the competition Sedielem. et. al. (2013) and the recent work Khalifa et al. (2017) that talks about the current more accurate approaches. Despite of this, we intend to tackle the problem with a transfer learning approach, which we have only found in Ackermann et al. (2018) and Batra (2019), and that permits the reduction of computational power needed for training. The core idea is to take a pretrained neuronal network that solves a similar

★ E-mail: andres.gomez27@udea.edu.co (UdeA)

problem and then apply additional training with the galaxy images, each associated with 37 category probabilities.

For that, we first explore the possible symmetries present in any galaxy image in order to increase the training data; this is known as data augmentation. Then, we explain why do we use Resnet50, based in analytical and experimental results. After that, we justify the election of the Adam optimizer and the hyper-parameters for the training.

Finally, we expose the results obtained for the mean squared error (mse) and the accuracy of the model. From that, we conclude that our model is consistent with the results obtained by Ackermann et al. (2018); Batra (2019). Actually improving the mse for Ackermann et al. (2018) and Sedielem. et. al. (2013), and being near the result of Batra (2019). We also conclude that a model which classify with a $\approx$ 72% expert accuracy was successfully constructed, being this result still very likely to be improved in consideration of Khalifa et al. (2017). In addition, we remark the careful use of theoretical insights and the use of free tools to solve a complex science problem. We then end talking about the things that could be done to improve the actual work; between them some that could speak more precisely about the unique characteristics of the problem in question.

2 GALAXY DATA SET AND AUGMENTATION

The training data consist of 677358 labeled images (see Sedielem. et. al. (2013)). All of them classified according to the decision tree observed in figure 1. The training and the validation proportions are set to 80 and 20 percent as is suggested in the literature. Each images has 37 corresponding labels, where each signifies the probability of belonging to a certain class.

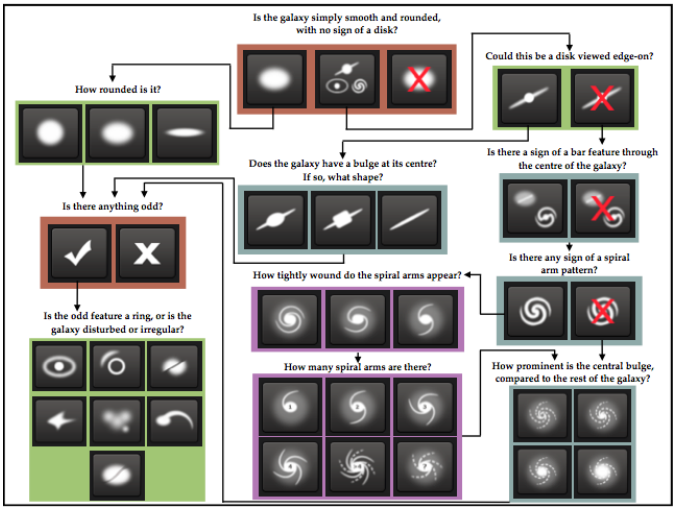

Figure 1: The image illustrate the steps that must be follow in order to classify a given galaxy.

Also, as it is observed from the image, the classification is based purely in the visual appearance. Fact that permits us to fully concentrate in the morphology of the galaxy. This is a new possibility since even in the classification made by experts before the observed color of the galaxy was used for a preselection.

Now, from the 37 classes that are deduced from the image, it is clear that a rotation by any angle, a reflection about the horizontal or vertical axis, and a small translation still leads to a type within the ones considered. We therefore considered the random application of this transformations to the images in order to increase the variety and quantity of the training data.

3 ARCHITECTURE AND TRAINING CONFIGURATION

First, due to our computational power limitations to achieve better results that some of networks already out there, we decided to use a transfer learning approach. That is, we took a pretrained neuronal network, add an additional layer, and then refined all the network weights through training. The consistency of the approach is clear from the universal approximation theorem, which states that given enough neuronal units a feed forward network with a single hidden layer is sufficient to represent any function. So, in the worst case scenario we would only need only to increase the number of units in some layer before the last one.

Second, based in the short overview in Vicent. Fung. (2017) about Residual Neuronal Networks (ResNet), we choose to tackle our problem using them principally because they encourage feature reuse in well separated parts of the network. Fact that go along with transfer learning, since given that we already adjust some features with the pretrained election, we would only need to adjust the residuals (see figure 2 to clarify this ideas). In addition, we demanded two additional conditions. The network must classify great variety of objects to avoid biases and it must have been trained with millions of images, this makes more probable that important features of galaxy images are already took into account and agrees with the data intensive nature of the problem.

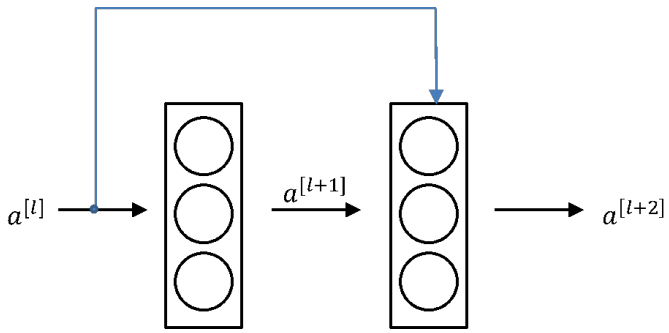

Figure 2: The image illustrates a residual neuronal network in which the features $a^{[l]}$ is directly summed to the output of the layer $l+2$ (before the activation function is applied). So, if the features of the initial input are already well characterized by $a^{[l]}$ (as the universal approximation could make possible) the $l+2$ layer will try to adjust its weights, $W^{[l+2]}$, for the identification of further subtle features or to simply apply the identity to the input; the output that only comes from the layer units $l+2$ is known as the residual. In addition, information from other layers (different than l) could also be added to the output of layer $l+2$, this fact with a similar argument leads to the reuse property mentioned above.

Based in this we chose the network called ResNet50, trained on more than a million images from the ImageNet database. The network is 50 layers deep and can classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. In addition, we added 37 nodes as a final layer, in order for the algorithm to estimate the belonging probability.

Now, to train the algorithm we decided to use Adam since it has solve with great success similar problems Yelan. et. al. (2018). In fact, it has won his rightful place as default optimizer for neuronal networks. Similarly, the batch size was taken as 64, since the literature states that 32 is usually a good minimum default. In addition, we use the mean square error (mse) as cost function for the expected probabilities of the 37 labels. Moreover, it is important to add that we used early stopping: if the mse of the validation data did not improve in 4 iterations we stopped the process. So that running the algorithm several times and choosing the best one we refined the initial free parameters of the model.

With all this set, the program was ran whenever possible in Colab. The total running time end up being of about 16 hours. For this to be done, Colab was configured with Google Drive, so that the training data as well as the checkpoint models were saved there.

4 MAIN FINDINGS

The program ran in the free GPUs available in Colab and it always halted because of early stopping. Figure 3 illustrate the best results of the mse for the training and validation data. From the image is clear that the training data is still not over-fitted, because in such case we would have a considerable greater error for the validation data. So, it can be said that it is still valuable to relax the condition of early stopping. In fact, this is consistent with the fact that the mse could still be decreased as the work of Batra (2019) shows, obtaining a lower value by about 33 %. For the present case the mse error obtained was ≈ 0.0091, which is about 10 times better that the one obtained by the winners of the competence Sedielem. et. al. (2013) and 3 times better that the one obtained by Ackermann et al. (2018).

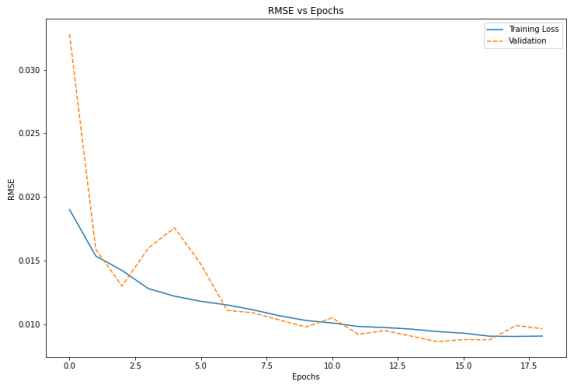

Figure 3: The image illustrate the results of the mse as a function of iterations for the training and the validation data.

On the other hand, it was obtained that the the model had an accuracy of $\approx$ 80%. Which combined with the results of Willett et al. (2013) permit us to say that a successful program that classify images with an expert level of about 72% was constructed. In this respect the algorithm still have a lot to improve, since as it is shown in Khalifa et al. (2017) a method to classify with an expert level of about 87% has been already constructed.

Now, the fact that the accuracy is not as high as in other works could be related with two factors. One of them is off course the number of iterations done and the parameters adjusted from the validation data. The other has to do with the possible limitations that could arise from the transfer learning approach used. More concretely, we begin from a set of weights of a pretrained neuronal network that we seek to refined. So, could it be possible that that this initialization greatly affects the accuracy of the results? If so, as a solution we propose, in agreement with the above discussion, to study how the accuracy behaves as more layers are added at the end, such that the initial added layer behaves as $a^{[l+2]}$ does in figure 2.

5 CONCLUSIONS

An algorithm for the identification of a galaxy morphology based in its image was successfully constructed. This with an expert accuracy of $\approx$ 72 % and using the concepts of ResNets and transfer learning as fundamental pillars in its construction. In fact the results of mse improved respect similar recent researches and can still be improved to try to surpass other ones.

The use of transfer learning, ResNets, the consideration of data augmentation, and the careful choosing of training parameters show that it is possible to greatly reduce the computational resources needed to solve a complicate scientific problem, as it is the case of galactic classification; this is more evident taking into account that the training of a residual neuronal network from the beginning could take weeks.

The construction of the algorithm exposed is not only justified in a practical approach, instead it appeals to theoretical observations which are done through the article. These considerations permit to devise a bigger picture, and we consider them therefore as the fundamental pillars for the improvement of the present algorithm, as well as for the application of the approach taken to new problems.

With the free tools available nowadays, like Colab and Keras, is possible to tackle complicate data intensive scientific problems. In addition, the fact that this approach surpasses results of other ones illustrates the importance of more computational power and better deep learning theories in order to solve scientific problems.

6 PERSPECTIVES

Study of how the accuracy changes with the addition of layers to the pretrained neuronal network. This is of great interest to understand the limitations of transfer learning.

Relax more the early stopping condition to obtain better results for the training as well as for the initial free parameters of the training. Also, vary the learning rate of the algorithm and use the validation data to take the one with better error.

Study how the algorithm behaves respect to galaxy images that

were classified by experts and other ones that were only classified by participants of the Galaxy Zoo Project.

Improve data augmentation taking into account that the observational resolution plays a fundamental role for a successful classification, as well as considering that the training images have important biases as the Malquimist bias.

Once all of this is done a valuable research path to extend the problem to a semi-supervised methodology would be suitable; this to take advantage of millions of unlabeled data.

ACKNOWLEDGEMENTS

I thank my teacher Jose David Luis, because his continuous feedback through the course as well as his methodology permitted me to improve considerable in the resolution of a computational problem. I also thanks my thesis Advisor Juan Carlos Muñoz, which encouraged me to take this course and has believed in me as a capable scientist.

REFERENCES

Ackermann S., Schawinski K., Zhang C., Weigel A., Turp M., 2018
Batra A., 2019, Winning Kaggle's Galaxy Zoo challenge. Medium
Fukugita et. al. 2007, The Astronomical Journal
Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press
Hubble. E., 1936, Yale University Press
Khalifa N. E. M., Taha M. H. N., Hassanien A. E., Selim I. M., 2017, arXiv e-prints, p. arXiv:1709.02245
Lahav et. al. 1995, Science
Lintott C. J., et al., 2008, MNRAS, 389, 1179
Schawinski et al. 2007, MNRAS
Sedielem. et. al. 2013, "Galaxy zoo - The Galaxy Challenge". Kaggle
Vicent. Fung. 2017, "An Overview of ResNet and its Variants". Towards Data Science
Willett K. W., et al., 2013, MNRAS, 435, 2835
Yelan. et. al. 2018, "Human Protein Atlas Image Classification". Kaggle