Galaxy morphology classification using unsupervised machine learning techniques

Author: Regina Sarmiento *(Instituto de Astrofísica de Canarias, La Laguna, Spain)*
Contact: reginas@iac.es

Abstract

Upcoming deep surveys (e.g. LSST, JWST, EUCLID) will provide high quality imaging at unprecedentedly high red shifts, allowing the study of galaxy morphology at different cosmic times. The processing of such data will be necessarily automatized due to its enormous volume. Deep Learning has proven to be a powerful tool in these situations. Previous publications ([1] [2] [3] [4]) have shown the effectiveness of supervised learning algorithms for galaxy morphology classification. In the case of future surveys there is an additional challenge: the data will be completely unlabelled. This limits the use of a standard supervised approach since a subsample of data of which its classification is known will not be available for training and labelling data is expensive. Therefore, we explore an unsupervised approach: Simple framework for Contrastive Learning of visual Representations [5]. We test this algorithm on SDSS images of galaxies with $z<0.15$. We present preliminary results.

Context

The study of galaxy morphology at different cosmic times allows a further understanding of Galaxy evolution. Upcoming deep surveys (e.g. LSST, JWST, EUCLID) will complement previous Galaxy observations with high quality imaging at unprecedentedly high red shifts.

In order to process the large volumes of data that these surveys will provide, an automatized approach is required. Deep Learning has proven to be a powerful tool in these situations. Previous publications ([1] [2] [3] [4]) have shown the effectiveness of supervised learning algorithms for galaxy morphology classification. In the case of future surveys there is an additional challenge: the data will be completely unlabelled. This limits the use of a standard supervised approach since a subsample of data of which its classification is known will not be available for training and labelling data is expensive.

However, classifying unlabelled data is being addressed by a recent field of research: unsupervised learning. The aim is to learn features that are intrinsic to the data and classify the data based on these representations. Recent progress in other fields has shown that these algorithms can obtain a similar accuracy to standard supervised approaches ([5] [6]).

In this presentation, we explore the use of an unsupervised learning framework: Simple Contrastive Learning of visual Representations (SimCLR) [5]. We test this algorithm on SDSS images of galaxies with $z<0.15$ for which a set of labels is known. Although the training process is completely unsupervised, we use these classifications to test the algorithm's performance.

Description of work

A CLR algorithm is trained to detect similarities between two images that were generated from the same base image, but were subject to different augmentations before being contrasted (e.g. rotation, colour distortion, crop and resize, cut-out). This way, the program is self-taught to recognize patterns present in the data and to produce meaningful representations of the images.
In our case, we adapt the SimCLR [5] to images of galaxies. The dataset consists on SDSS 64x64 cut-outs of galaxies in the g, r and i band. Additionally, we perform a 2x2 binning of the images and apply a normalization that spans the values of the images in the range [0; 1]. During training, augmentations that are specific to astronomical images are applied, as illustrated in Fig. 1 for one of our galaxies.

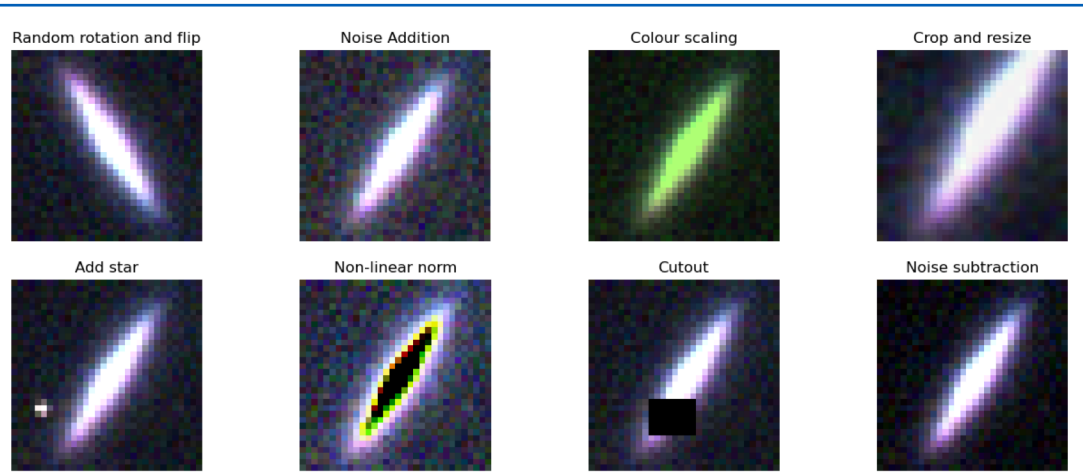

Fig. 1: Transformations applied to images during training. Each augmentation has a 50% chance of being applied to an image, except for "random rotation and flip" that is always applied.

Preliminary results

The CNN outputs a 64-parameter representation of an image. In order to visualize this high dimensional space, the representations are mapped on to a 2D space with uMAP. The colour coding in Fig. 2 corresponds to externally provided labels (Early/Late type galaxies, in colours blue/red respectively).
The model was trained on 377419 SDSS cut-outs of galaxies at z < 0.15 during 1000 epochs, using a batch size of 1024.

We test the model's performance at classifying early and late galaxies. For this purpose, we train a supervised fully connected network on the representations and compare these results with a supervised CNN and PCA. We find that the SimCLR outputs are good representations of the data since they reach similar results to the supervised CNN (~85% accuracy, recall and precision) after the supervised training on 10000 galaxies.

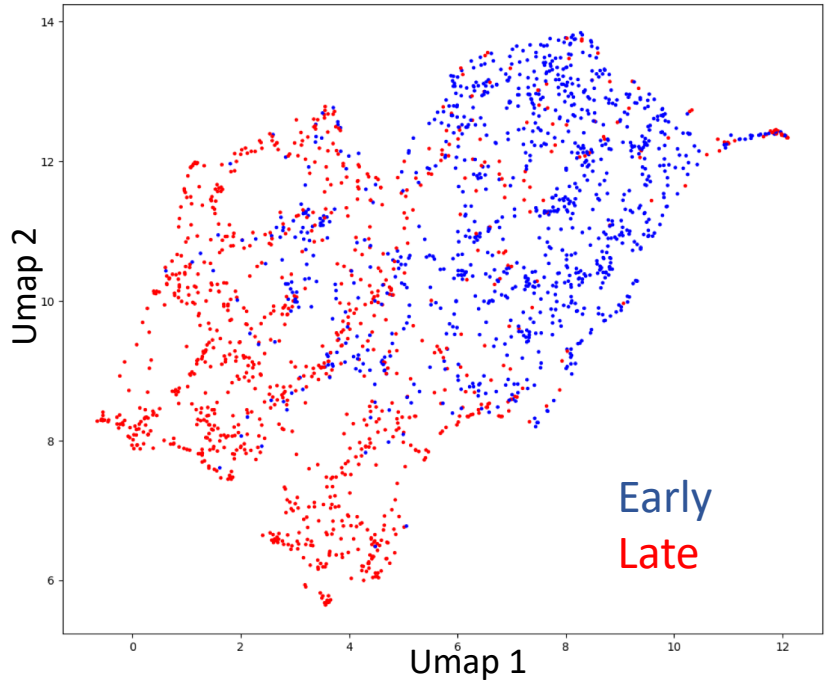

Fig. 2: Visualization of the CNN outputs of 1980 galaxy images.

Conclusions

We adapted the SimCLR framework to astronomical data. The representations of galaxy images produced by the network in this initial stage are promising since we can recover a classification comparable to a supervised CNN.

Work in progress

We aim to improve the representations obtained and expect to benefit from them by minimizing the labelled data required to accurately classify galaxies according to their morphology.

References:

[1] M. Huertas-Company, J.A.L Aguerri, M. Bernardi, S. Mei, J. Sánchez Almeida. (2010) Revisiting the Hubble sequence in the SDSS DR7 spectroscopic sample: a publicly available bayesian automated classification.

[2] M. Huertas-Company et al. (2015) A catalog of visual-like morphologies in the 5 CANDELS fields using deep-learning.

[3] M. Huertas-Company et al. (2016) Mass assembly and morphological transformations since z∼3 from CANDELS.

[4] H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D. Tuccillo, J. L. Fischer. (2017) Improving galaxy morphologies for SDSS with Deep Learning.

[5] T. Chen, S. Kornblith, M. Norouzi, G. Hinton. (2020) A Simple Framework for Contrastive Learning of Visual Representations.

[6] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick. (2020) Momentum Contrast for Unsupervised Visual Representation Learning.