Draft version March 15, 2022
Typeset using LaTeX **twocolumn** style in AASTeX631

Term Project
Galaxy Zoo: Probabilistic Morphology through Bayesian CNNs and Active Learning

Colin Leach

ABSTRACT

Astronomical survey data has expanded impressively since the era when professional astronomers could keep up with it by themselves. As an early enhancement, Galaxy Zoo used large numbers of amateur volunteers for classification of SDSS results, more recently extended to HST, CANDELS and DECaLS images. To scale further for the Rubin/Euclid era, that approach needs to be supplemented with ML techniques to use the volunteers more efficiently. Walmsley et al. (2020a) attempts to develop such a hybrid human/ML system. The current term project attempts to reproduce and (perhaps) extend this work.

1. INTRODUCTION

The Galaxy Zoo started as an attempt to scale manual classification of SDSS images by recruiting citizen scientists (Lintott et al. 2008). This succeeded beyond expectations, but is struggling to keep up with new data sources: DES, Rubin, Euclid, etc. Volunteer input is increasingly regarded as a finite and valuable resource, which needs to be used more efficiently (Dickinson et al. 2020).

Sorting galaxies by color has been done for decades (blue spirals, red ellipticals), though this has been criticized as inaccurate (Smethurst et al. 2022). Other approaches include radial brightness curves, looking for central bulges and bars. Attempts to use neural networks to classify morphology go back at least to a Kaggle challenge[1] in 2014, won by Dieleman et al. (2015). The concept of transfer learning, using older surveys to train models for a newer one, was explored by Domínguez Sánchez et al. (2019) and later by W+20, discussed in more detail in Walmsley et al. (2021). These all focus on visual images (or their redshifted equivalents), but Fielding et al. (2021) discusses an exchange of techniques with radio astronomy. A broader review of ML in astronomy is given in Fluke & Jacobs (2020).

GZ2 (Willett et al. 2013; Hart et al. 2016) is based on SDSS DR7. Later catalogs include Galaxy Zoo: Hubble (Willett et al. 2017), CANDELS (Simmons et al. 2017) and DECaLS (Walmsley et al. 2022).

2. AIMS

In Walmsley et al. (2020a) (hereafter W+20), an attempt is described to develop a human-machine hybrid strategy for galaxy morphology:

- Use the large Galaxy Zoo 2 (GZ2) catalog to train a CNN that can classify SDSS images.
- Use this model as a starting point to classify new data sources and formats, using only modest amounts of labeling from human volunteers to fine-tune the model.

3. CODE

3.1. *Zoobot Code*

Code: All the Python/Tensorflow code is on Github[2] (Walmsley 2019), claiming to be an exact copy of that used for W+20.

Perhaps more interesting is the zoobot repo[3], a fork which is still under active development. This extends the project to DECaLS (Dark Energy Camera Legacy Survey) data, as described in Walmsley et al. (2021) (hereafter W+21). It also has much better documentation[4] than the earlier code.

3.2. *Code for Term Project*

Python code and documentation associated with ASTR 502 is available on Github[5]. This aims to cover both GZ2, as in W+20, and DECaLS, as in W+21.

4. DATA

4.1. *Catalog Data*

GZ2 catalogs are available online[6] in multiple formats, with 231 columns and nearly 300k rows as described in Willett et al. (2013).

[1] https://www.kaggle.com/c/galaxy-zoo-the-galaxy-challenge

[2] https://github.com/mwalmsley/galaxy-zoo-bayesian-cnn

[3] https://github.com/mwalmsley/zoobot

[4] https://zoobot.readthedocs.io/

[5] https://github.com/colinleach/proj502

[6] https://data.galaxyzoo.org/

The table used in this work was based on Hart et al. (2016), downloaded as a gzipped csv file. This is Table 5 in Willett et al. (2013) and the column format is described in an accompanying file.

4.2. *Image Data:*

The GZ team do not make their images library publicly available. However, each 512 × 512 image is available from the SDSS cutout service, using the ra/dec coordinates in the catalog table.

Before analysis, the images need to be downsampled to 256 × 256 monochrome pixels and stored as uint8.

5. COMPUTATION

W+20 reports that GZ2 training was carried out on a p2.xlarge EC2 instance with K80 GPU, taking about 8 hours. For DECaLS, the GPU was upgraded to a V100.

Experiments with the GPUs available at the start of this project rapidly proved that 2GB of GPU memory is wholly inadequate for training a CNN. Upgrading to a 6GB GTX 1660 (far from state of the art, but only $450 and compatible with the existing motherboard and PSU) allowed some progress. However, this still proved limiting for batch size as discussed below.

TODO Colab and AWS

6. GOALS

My time is less valuable than for faculty or grad students, so goals are open-ended depending on energy, enthusiasm and (hopefully) competence. Roughly:

1. Get the published code running on my local machine, using whatever cut-down training set proves viable.
2. Deploy the code on either AWS or Google.
3. Extend the model to other data such as Hubble, CANDELS, DECaLS, for which there is already some GZ classification.
4. Think about newer CNN algorithms. The W+20 paper was submitted in 2019, but software decisions were made well before then and the authors admit it is not the latest technology.
5. Rewrite using other frameworks, for my education. Most obviously PyTorch, but (unlike most astronomers!) I would also be interested to try Julia with Flux. As a stretch goal, I may try getting it working in F#/ML.NET, but don't hold your breath waiting for that.

I think we can assume that not all of this will be done before the end of the semester (an understatement).

7. WORKFLOW

In outline, these are the steps required:

1. Get the Galaxy Zoo catalog data for each survey of interest.
2. Get the image files (JPG or PNG), one per galaxy in the classification.
3. Make a combined catalog, including a path to the image on disk plus the data fields relevant to the model.
4. Split the galaxies into train, evaluate and test sets. For each, prepare a binary tensor (tfrecord) file containing image and classification data.
5. Train the model on the train and evaluate sets.
6. Predict results with the test set and compare with the GZ classification.

The following subsections address each of these in more detail.

7.1. *GZ data*

GZ2 files were downloaded from the Galaxy Zoo website. There are a total of 243,500 rows in the table. For better consistency, only those marked 'original' in the sample field were used in subsequent analyses, a set of 211,922.

Extensive DECaLS data is available from Zenodo (Walmsley et al. 2020b). For this study the file 'gz_decals_volunteers_5.parquet' was used, a total of 253,286 rows.

For maximum flexibility (and because old habits die hard), all this data was stored in a PostgreSQL database, running locally.

7.2. *Images*

The RA and Dec fields in the GZ2 dataset were used to fetch 424 × 424 JPG cutouts from the SDSS SkyServer. Because Zoobot is currently configured to use PNG images, the Python code converted each file with PIL. The PNG files totaled around 33 GB, much more than the corresponding JPG files.

DECaLS DR5 images were downloaded from Zenodo (Walmsley et al. 2020b) as 4 ZIP files, unpacked to 272,725 424 × 424 PNG files totaling 83 GB.

File paths and some metadata was stored in PostgeSQL.

REFERENCES

Dickinson, H., Fortson, L., Scarlata, C., Beck, M., & Walmsley, M. 2020, in Panchromatic Modelling with Next Generation Facilities, ed. M. Boquien, E. Lusso, C. Gruppioni, & P. Tissera, Vol. 341, 99–103, doi: 10.1017/S1743921319001418

Dieleman, S., Willett, K. W., & Dambre, J. 2015, MNRAS, 450, 1441, doi: 10.1093/mnras/stv632

Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., et al. 2019, MNRAS, 484, 93, doi: 10.1093/mnras/sty3497

Fielding, E., Nyirenda, C. N., & Vaccari, M. 2021, arXiv e-prints, arXiv:2111.04353. https://arxiv.org/abs/2111.04353

Fluke, C. J., & Jacobs, C. 2020, WIREs Data Mining and Knowledge Discovery, 10, e1349, doi: 10.1002/widm.1349

Hart, R. E., Bamford, S. P., Willett, K. W., et al. 2016, MNRAS, 461, 3663, doi: 10.1093/mnras/stw1588

Lintott, C. J., Schawinski, K., Slosar, A., et al. 2008, MNRAS, 389, 1179, doi: 10.1111/j.1365-2966.2008.13689.x

Simmons, B. D., Lintott, C., Willett, K. W., et al. 2017, MNRAS, 464, 4420, doi: 10.1093/mnras/stw2587

Smethurst, R. J., Masters, K. L., Simmons, B. D., et al. 2022, \mnras, 510, 4126, doi: 10.1093/mnras/stab3607

Walmsley, M. 2019, Zenodo, doi: 10.5281/zenodo.2677874

Walmsley, M., Smith, L., Lintott, C., et al. 2020a, \mnras, 491, 1554, doi: 10.1093/mnras/stz2816

Walmsley, M., Lintott, C., Tobias, G., et al. 2020b, Galaxy Zoo DECaLS: Detailed Visual Morphology Measurements from Volunteers and Deep Learning for 314,000 Galaxies, 0.0.2, Zenodo, doi: 10.5281/zenodo.4573248

Walmsley, M., Scaife, A. M. M., Lintott, C., et al. 2021, arXiv e-prints, arXiv:2110.12735. https://arxiv.org/abs/2110.12735

Walmsley, M., Lintott, C., Géron, T., et al. 2021, Monthly Notices of the Royal Astronomical Society, 509, 3966–3988, doi: 10.1093/mnras/stab2093

—. 2022, \mnras, 509, 3966, doi: 10.1093/mnras/stab2093

Willett, K. W., Lintott, C. J., Bamford, S. P., et al. 2013, Monthly Notices of the Royal Astronomical Society, 435, 2835, doi: 10.1093/mnras/stt1458

Willett, K. W., Galloway, M. A., Bamford, S. P., et al. 2017, MNRAS, 464, 4176, doi: 10.1093/mnras/stw2568