Draft version March 5, 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. (2020) 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. (2020) (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. DATA

3.1. *Catalog Data*

GZ2 catalogs are available online[2] in multiple formats, with 231 columns and nearly 300k rows as described in Willett et al. (2013).

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

3.2. *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.

4. ZOOBOT CODE

Code: All the Python/Tensorflow code is on Github[3] (Walmsley 2019), claiming to be an exact copy of that used for W+20.

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

5. COMPUTATION

W+20 reports that training was carried out on a p2.xlarge EC2 instance with K80 GPU, taking about

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

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

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

8 hours. AWS pricing for GPU-based instances is complex and Google is more opaque, but a budget under $50 for this sort of run looks plausible.

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).

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. 2020, \mnras, 491, 1554, doi: 10.1093/mnras/stz2816

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. 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