



Contribution ID: 14

Type: **Oral Presentation**

Galaxy Zoo: Practical Methods for Large-Scale Learning

Tuesday, 31 May 2022 12:35 (20 minutes)

Deep learning is fundamental to creating Galaxy Zoo's latest catalogs. In this talk, we explore the methods we've developed to best exploit large-scale human labels and how other researchers can benefit from them.

We open by presenting Galaxy Zoo LegS - new deep-learning-powered detailed morphology measurements for 8 million galaxies imaged by the DESI Legacy Surveys. Our models are trained on human labels collected over 8 years, during which time different volunteers answered different questions and followed different instructions. We describe how we overcome the resulting label distribution shift to learn from more human responses than any previous astronomical model.

We next show how answering every Galaxy Zoo question simultaneously forces the resulting models to learn meaningful semantic representations of galaxies. These representations can then be directly used for similarity search and to outperform a recent approach at personalized anomaly-finding. Further, and crucially for other researchers, because the models are trained on a diversity of tasks (answering every GZ question), the trained models make excellent base models to finetune to new tasks. We demonstrate this by finetuning to find ringed galaxies. Models pretrained on all GZ questions are better able to find rings than models pretrained on a single GZ question or on ImageNet. We go on to exploit this to create the largest ringed galaxy catalog to date by an order of magnitude. Our trained models are available for the community to finetune for their own tasks at www.github.com/mwalmsley/zoobot (in both TensorFlow and PyTorch).

Finally, we describe our very latest work combining self-supervised approaches with broad supervised pre-training on Galaxy Zoo to classify galaxies better than with either alone. We believe such approaches are ideally suited to Euclid and Rubin because they allow us to leverage both the millions of human labels collected over the last decade and the raw scale of unlabelled images these new surveys will produce.

Presenter: WALMSLEY, Mike (University of Manchester)

Session Classification: Anomaly Detection