Contribution ID: **112**   Type: **Talk**

Doing More With Less; Label-Efficient Learning for Euclid and Rubin

Wednesday, November 29, 2023 3:31 PM (15 minutes)

Deep learning is data-hungry; we typically need thousands to millions of labelled examples to train effective supervised models. Gathering these labels in citizen science projects like Galaxy Zoo can take years, delaying the science return of new surveys. In this talk, I'll describe how we're combining simple techniques to build better galaxy morphology models with fewer labels.

First [1], we're using large-scale pretraining with supervised and self-supervised learning to reduce the number of labelled galaxy images needed to train effective models. For example, using self-supervised learning to pretrain on unlabelled Radio Galaxy Zoo images halves our error rate at distinguishing FRI and FRII radio galaxies in a separate dataset.

Second [2], we're continually retraining our models to prioritise the most helpful galaxies for volunteers to label. Our probabilistic models filter out galaxies they can confidently classify, leaving volunteers able to focus on challenging and interesting galaxies. We used this to measure the morphology of every bright extended galaxy in HSC-Wide in weeks rather than years.

Third [3], we're using natural language processing to capture radio astronomy classes (like "FRI" or "NAT") through plain English words (like "hourglass") that volunteers use to discuss galaxies. These words reveal which visual features are shared between astronomical classes, and, when presented as classification options, let volunteers classify complex astronomical classes in an intuitive way.

We are now preparing to apply these three techniques - pretraining, active learning, and natural language labels - to provide day-one galaxy morphology measurements for Euclid DR1.

Primary author: WALMSLEY, Mike (University of Toronto)

Presenter: WALMSLEY, Mike (University of Toronto)

Session Classification: Contributed talks

Track Classification: New York