University of São Paulo

Self-supervised Learning for Astronomical Image Classification

Ana Martinazzo, Mateus Espadoto, Nina Hirata
amartina | mespadoto | nina@ime.usp.br

January 13, 2021

Modern Astronomy is a data-intensive discipline

Astronomical data are abundant, rich and unique, but mostly unlabeled

Automated data analysis techniques, such as ML and DL, are needed to extract useful information from these data
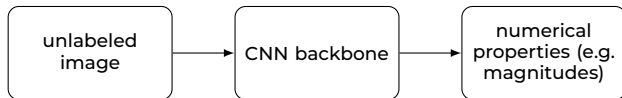
Astronomical Data

Science-ready data products may come in various formats:

- raw images (may have dozens of channels)
- RGB composite images (generated from raw images)
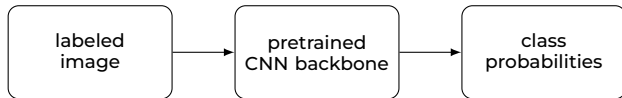- tabular catalogues of properties (generated from raw images)
- spectra

The core idea of our work is combining images and astronomical properties into a single representation.

Our self-supervised approach

Pretext (regression) task

```
┌──────────┐      ┌──────────────┐      ┌──────────────────┐
│ unlabeled│ ───► │ CNN backbone │ ───► │ numerical        │
│ image    │      │              │      │ properties (e.g. │
│          │      │              │      │ magnitudes)      │
└──────────┘      └──────────────┘      └──────────────────┘
```

Downstream (classification) task

```
┌──────────┐      ┌──────────────┐      ┌──────────────────┐
│ labeled  │ ───► │ pretrained   │ ───► │ class            │
│ image    │      │ CNN backbone │      │ probabilities    │
└──────────┘      └──────────────┘      └──────────────────┘
```

Classification datasets

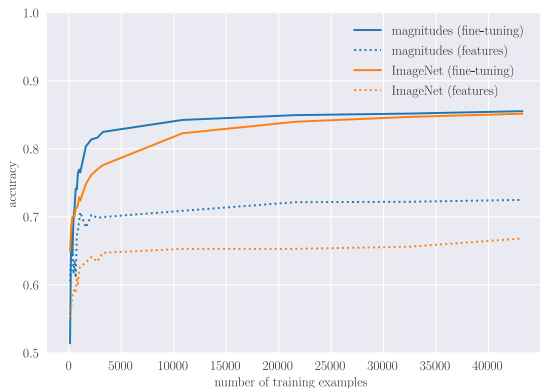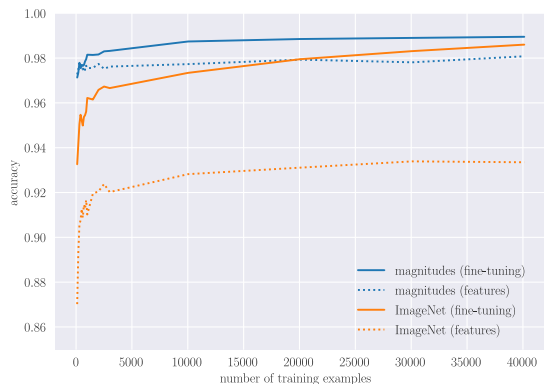|	
Star/Galaxy (SG)	2 classes, 50090 images
Star/Galaxy/Quasar (SGQ)	3 classes, 54000 images
Merging Galaxies (MG)	2 classes, 15766 images
Galaxy Morphology (EF-2)	2 classes, 3604 images
Galaxy Morphology (EF-4)	4 classes, 4389 images
Galaxy Morphology, (EF-15)	15 classes, 4327 images

Figure: Accuracy curves as a function of the number of training samples for classifiers trained on the SG (left) and SGQ (right) datasets.

Figure: t-SNE projections of learned representations for the SG, SGQ and MG datasets.

Conclusions

Our method is advantageous when:

1. there are few labeled examples
2. ImageNet weights for the CNN architecture of interest are not available
3. extracting features for unsupervised tasks, such as clustering

Future work

Possible directions for future work are:

1. extending our analyses for raw images
2. using different properties as targets for self-supervised learning
3. evaluating learned representations in unsupervised tasks

Thank you