Estimating Galactic Distances From Images Using Self-supervised Representation Learning

Md Abul Hayat[1], Peter Harrington[2], George Stein[2], Zarija Lukić[2], Mustafa Mustafa[2]

[1] University of Arkansas, Fayetteville, AR 72701 [2] Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Abstract

We use a contrastive self-supervised learning framework to estimate distances to galaxies from their photometric images. We incorporate data augmentations from computer vision as well as an application-specific augmentation accounting for galactic dust. We find that the resulting visual representations of galaxy images are semantically useful and allow for fast similarity searches. It can be successfully fine-tuned for the task of redshift estimation.

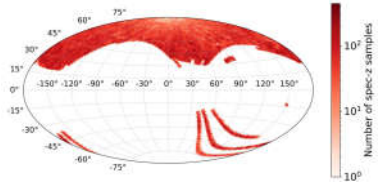
We show that (1) pretraining on a large corpus of unlabeled data followed by fine-tuning on some labels can attain the accuracy of a fully-supervised model which requires 2-4x more labeled data, and (2) that by fine-tuning our self-supervised representations using all available data labels in the Main Galaxy Sample of the Sloan Digital Sky Survey (SDSS), we outperform the state-of-the-art supervised learning method.

Background & Dataset

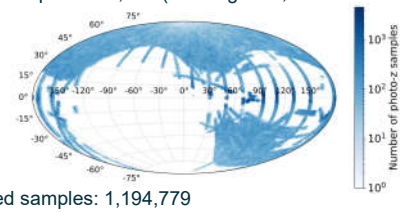
Redshift is a measure of distance to galaxies and is an important cosmological probe.

- Spectroscopic redshift (spec-z) estimation is accurate but expensive
- Photometric redshift (photo-z) is less accurate but less expensive

- Sloan Digital Sky Survey data has been used in this study (DR12)
- Each image is of size 64x64 pixels with 5 channels (*ugriz*)
- Redshift of labeled images are within range of 0 to 0.4

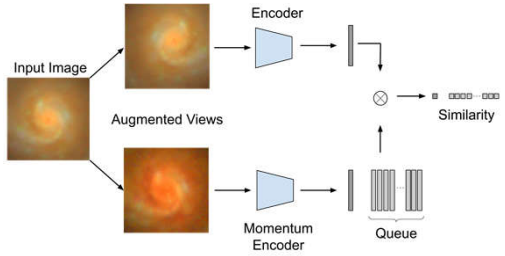


Labeled samples: 501,997 (Training: 399,984 Validation: 102,993)



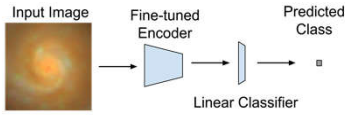
Unlabeled samples: 1,194,779

Self-supervised Learning



Step 1: Pretraining (MoCo)

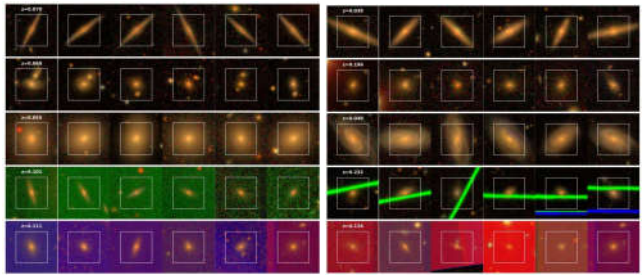
- Two augmented views of the same image are fed to encoder and momentum encoder networks with same architecture (ResNet50)
- InfoNCF loss tries to maximize cosine similarity between the two embeddings of the same image compared to the embeddings stored in the queue



Step 2: Self-supervised fine-tuning

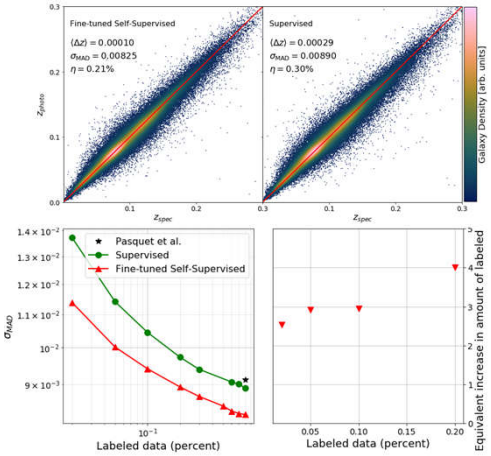
- The encoder network is fine-tuned using labeled data with a 10x smaller learning rate for the convolutional portion of the network followed by a fully-connected layer

Similarity Search



- By computing the dot products of a query vector with those of the unlabeled training samples, and sorting by the largest values, galaxies are returned sorted by their similarity to the query
- This methodology could be used for galaxy similarity identification or for anomaly detection

Results



If z_p and z_s are photometric and spectroscopic redshift respectively

- $\Delta z = \frac{z_p - z_s}{1+z_s}$ and $\mathrm{MAD}(\Delta z) = median(|\Delta z - median(\Delta z)|)$
- MAD deviation, $\sigma_{MAD} = 1.4826 \times MAD(\Delta z)$

- We trained the labeled data in fully-supervised way using ResNet50
- Self-supervised network, fine-tuned on 100% of the labeled training data, outperforms both our fully-supervised baseline and the previous state-of-the-art result (Pasquet et. al., 2019)
- Self-supervised network achieves the equivalent accuracy of a fully-supervised network while using 2-4x less labeled data

Conclusion

- Even in the case of limited data labels our model performs well, This is extremely valuable given the relative cost of acquiring spectroscopic measurements for a large sample of galaxies

- Our results show great promise for self-supervised learning methods to assist in deriving more precise photometric redshift estimates, helping address fundamental physics and cosmology questions on the nature and properties of dark energy, dark matter and gravity