Astronomical Data Analysis Software and Systems XXX
ASP Conference Series, Vol. 532
J. E. Ruiz, F. Pierfederici, and P. Teuben, eds.

Transfer Learning in Large Spectroscopic Surveys

Ondřej Podsztavek,[1] Petr Škoda,[2,1] and Pavel Tvrdík[1]

[1]*Faculty of Information Technology, Czech Technical University in Prague, Czechia;* `podszond@fit.cvut.cz`

[2]*Astronomical Institute of the Czech Academy of Sciences, Ondřejov, Czechia*

Abstract. Transfer learning is a machine learning method that can reuse knowledge across spectroscopic archives with different distributions of observations. We applied transfer learning based on a convolutional neural network to spectra from Large Sky Area Multi-Object Fiber Spectroscopic Telescope and Sloan Digital Sky Survey archives. Taking advantage of known quasars in LAMOST DR5 version 3, we wanted to discover yet unseen quasars in SDSS DR14. Our transfer learning approach reaches 99.6% precision and 98.9% recall. We found examples of quasars previously classified as stars.

1. Introduction

Current multimillion spectroscopic archives of the Sloan Digital Sky Survey (SDSS) and the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST) provide enough data for analysis with the most advanced machine learning methods. However, it is not straightforward to reuse a previously gained knowledge from one archive to another due to statistical properties. Each archive has a different strategy for target selection based on particular scientific goals. Diverse strategies for target selection result in different distributions of observations. This breaks the assumption of most machine learning algorithms that data are independent and identically distributed (Goodfellow et al. 2016, chapter 5).

The distribution mismatch problem has arisen when we wanted to discover quasars (QSOs) in SDSS data using LAMOST data (Sect. 2). It would be senseless to use only SDSS data to train a machine learning model because the model would not learn anything new. However, the utilization of LAMOST data should bring new knowledge to the model due to the distribution difference between LAMOST and SDSS data. The model based on both LAMOST and SDSS data can potentially lead to a discovery of yet unknown QSOs.

Transfer learning is a possible approach to reuse knowledge for discovery (Sect. 3). In our case, we applied transfer learning to reuse knowledge from the LAMOST DR5 version 3 archive for discovery in the SDSS DR14 archive. We experimented with transfer learning based on a convolutional neural network (ConvNet) that revealed new QSOs in SDSS data (Sect. 4).

2. Data

At the time of our experiments, the most current catalogs of QSOs were the LAMOST DR4&5Q catalog (Yao et al. 2019) and the SDSS DR14Q catalog (Pâris et al. 2018). Unlike SDSS DR14Q, LAMOST DR4&5Q is not cumulative, so we also included QSOs from LAMOST DR1Q (Ai et al. 2016) and DR2&3Q (Dong et al. 2018) catalogs. To complete our data, we downloaded all spectra from LAMOST DR5 version 3 (Luo et al. 2019) and SDSS DR14 (Abolfathi et al. 2018).

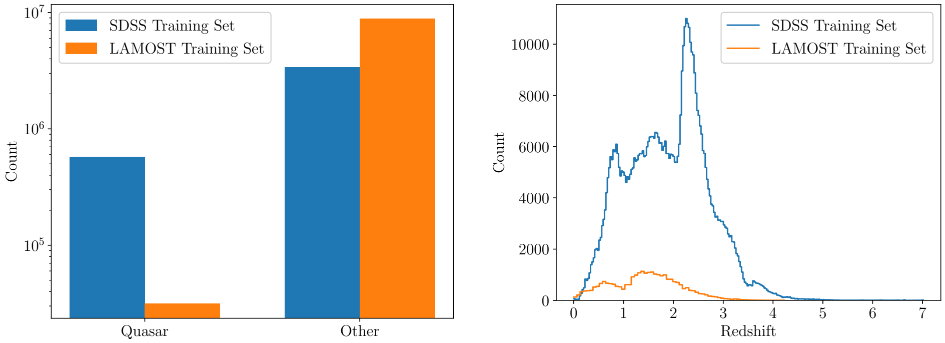

Figure 1. Different distributions of training sets. *Left:* Proportion of classes *QSO* and *Other*. *Right:* Redshift distributions of QSOs.

ConvNets assume that fluxes of spectra are sampled from the same wavelength grid. Therefore, we resampled both LAMOST and SDSS spectra to a logarithmically spaced grid of 2048 bins between 3839 and 8915 Å using a flux conserving resampling (Carnall 2017). The wavelength range keeps as many LAMOST spectra as possible. Then, we min-max normalized flux axis of spectra to the interval $[-1, 1]$ in order to enable classification based primarily on the shape of spectra and added labels *QSO* or *Other* to spectra according to their presence in the corresponding catalogs of QSOs.

Finally, we split the data into LAMOST and SDSS training, validation, and test sets. With inspiration from sizes of splits of ImageNet Large Scale Visual Recognition Competition (Russakovsky et al. 2014), sizes of our validation and test sets are 50 and 100 thousand spectra. The remaining spectra are in training sets (see Table 1).

Figure 1 demonstrates the distribution mismatch between LAMOST and SDSS training sets with respect to the number of QSOs and their redshift distributions.

Table 1. Counts of LAMOST and SDSS spectra and QSOs in splits of datasets

Data Release	Training Set	Validation Set	Test Set
LAMOST DR5 v3	8876365	50000	100000
No. of QSOs	31236	164	355
SDSS DR14	3990515	50000	100000
No. of QSOs	577712	7434	14447

3. Transfer Learning

Transfer learning is a machine learning method that reuses previously gained knowledge to learn a new problem. In the context of ConvNets (Goodfellow et al. 2016, chapter 9), which are a specialized kind of neural networks for processing data with grid-like topology (e.g. images), transfer learning is carried out by initializing with pretrained weights followed by fine-tuning with new data. The idea is supported by Yosinski et al. (2014), who showed transferred weights to be better (in terms of generalization) than random initialization.

For our classification to *QSO* and *Other* classes, we started from VGG Net-A ConvNet (Simonyan & Zisserman 2014) chosen from all their ConvNets because it provided the best F_1

score on the SDSS DR14 validation set. However, VGG Net-A was originally designed for images and not one-dimensional spectra, so we replaced its convolutions with one-dimensional counterparts. This one-dimensional ConvNet will be denoted in the following text as **1D-ConvNet**.

Our transfer learning procedure has the following steps:

Step 1 Xavier (i.e. random) initialization (Glorot & Bengio 2010) of 1D-ConvNet.

Step 2 Training of 1D-ConvNet with the LAMOST DR5 v3 training set using binary cross-entropy loss, Adam optimizer (Kingma & Ba 2014), batch size of 256 spectra, and early stopping when the loss on LAMOST DR5 v3 validation set is not improved during an epoch.

Step 3 Reusing of the trained 1D-ConvNet with transferred weights for training with the SDSS DR14 training set.

Step 4 Fine-tuning of the pretrained 1D-ConvNet with the SDSS DR14 training set with the same training characteristics as in Step 2, but the early stopping is based on the SDSS DR14 validation set.

4. Results

Our application of 1D-ConvNet based on transferred weights to the SDSS DR14 test set resulted in 14277 correctly predicted QSOs, 170 missed QSOs, and 84330 spectra classified as *Other* class. Moreover, there were 1223 false positives (spectra predicted as QSOs but not included in SDSS DR14Q).

We visually inspected the false positives, compared their spectroscopic classification by the SDSS pipeline (Bolton et al. 2012) with the prediction by 1D ConvNet. This analysis confirmed that 1164 of the 1223 spectra are QSOs. If we incorporate this result of our analysis, 1D-ConvNet achieves 99.6% precision and 98.9% recall.

Furthermore, 1D-ConvNet found 49729 false positives in all SDSS DR14 spectra (train, validation, and test sets combined). The SDSS pipeline classifies 47946 of them as QSOs, but they are not listed in SDSS DR14Q. We also visually inspected false positive spectra classified as *Star* by the SDSS pipeline and found that 1D-ConvNet correctly predicted at least 10 of them as QSOs (see examples in Fig. 2).

Although our results seem promising to confirm the benefits of transfer learning for astronomy, we have yet to evaluate 1D-ConvNet trained from random initialization with SDSS DR14 training set and compare results.

Acknowledgments. This work was supported by project OP VVV, Research Center for Informatics, CZ.02.1.01/0.0/0.0/16 019/0000765 of the Czech Ministry of Education, Youth and Sports; and by the Grant Agency of the Czech Technical University in Prague, grant No. SGS20/212/OHK3/3T/18. Astronomical Institute of the Czech Academy of Sciences is supported by the project RVO:67985815.

References

Abolfathi, B., et al. 2018, ApJS, 235, 42. 1707.09322
Ai, Y. L., et al. 2016, AJ, 151, 24. 1511.01647
Bolton, A. S., et al. 2012, AJ, 144, 144. 1207.7326
Carnall, A. C. 2017, arXiv e-prints, arXiv:1705.05165
Dong, X. Y., et al. 2018, AJ, 155, 189. 1803.03063
Glorot, X., & Bengio, Y. 2010 (Sardinia, Italy: JMLR Workshop and Conference Proceedings), vol. 9 of Proceedings of Machine Learning Research, 249
Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (MIT Press)
Kingma, D. P., & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980

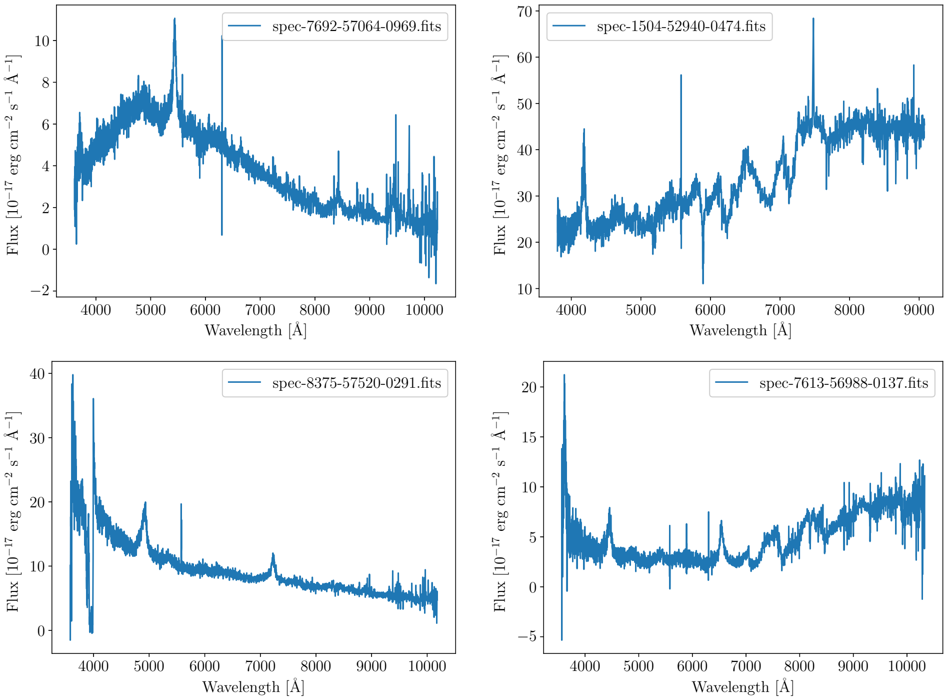

Figure 2. Four examples of discovered QSOs that were classified as stars by the SDSS pipeline.

Luo, A.-L., et al. 2019, VizieR Online Data Catalog, V/164
Pâris, I., et al. 2018, A&A, 613, A51. `1712.05029`
Russakovsky, O., et al. 2014, arXiv e-prints, arXiv:1409.0575
Simonyan, K., & Zisserman, A. 2014, arXiv e-prints, arXiv:1409.1556
Yao, S., et al. 2019, ApJS, 240, 6. `1811.01570`
Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. 2014, arXiv e-prints, arXiv:1411.1792