Astronomical Data Analysis Software and Systems XXIX
ASP Conference Series, Vol. 527
R. Pizzo, E.R. Deul, J-D. Mol, J. de Plaa and H. Verkouter, eds.

VO-supported Active Deep Learning as a New Methodology for the Discovery of Objects of Interest in Big Surveys

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

[1]*Faculty of Information Technology, Czech Technical University in Prague, Czech Republic*

[2]*Astronomical Institute of the Czech Academy of Sciences, Ondřejov, Czech Republic;* `skoda@sunstel.asu.cas.cz`

Abstract. Deep neural networks have been proved a very successful method of supervised learning in several research fields. To perform well, they require a massive amount of labelled data, which is challenging to get from most astronomical surveys. To overcome this limitation, we have developed a novel active deep learning method.

It is based on an iterative training of a deep network followed by relabelling of a small sample according to a qualified decision of an oracle (usually a human expert). To maximise the scientific return, the oracle brings to the decision the domain knowledge not limited only to the data learned by the network. By combining some external resources to extract the key information by an expert in a field, much more relevant labels are assigned.

Setup of an active deep learning platform thus requires incorporation of a Virtual Observatory (VO) client infrastructure as an integral part of a machine learning experiment, which is quite different from current practices. As proof of concept, we demonstrate the efficiency of our method for discovery of new emission-line stars in a multimillion spectra archive of the LAMOST DR2 survey.

1. Introduction

Discovery of objects of interest (often rare cases) in large archives of astronomical spectra would be a standard machine learning task if a large and representative labelled data sample of a given archive were available. With such a training set, it would be straightforward to train a supervised learning model and classify the whole archive with high accuracy. However, if there is no proper labelled training dataset (which is a common case in astronomy due to exponential growth of data and a limited number of human annotators), standard machine learning methods often provide poor results with a high rate of both false and missed candidates, so other machine learning approaches need to be developed to get reasonable discovery results.

2. Active deep learning method

Deep learning is a type of machine learning that allows computers to learn a good data representation by building complicated representations out of more simple ones (Goodfellow et al. 2016). Nowadays, convolutional neural networks (CNNs) (LeCun

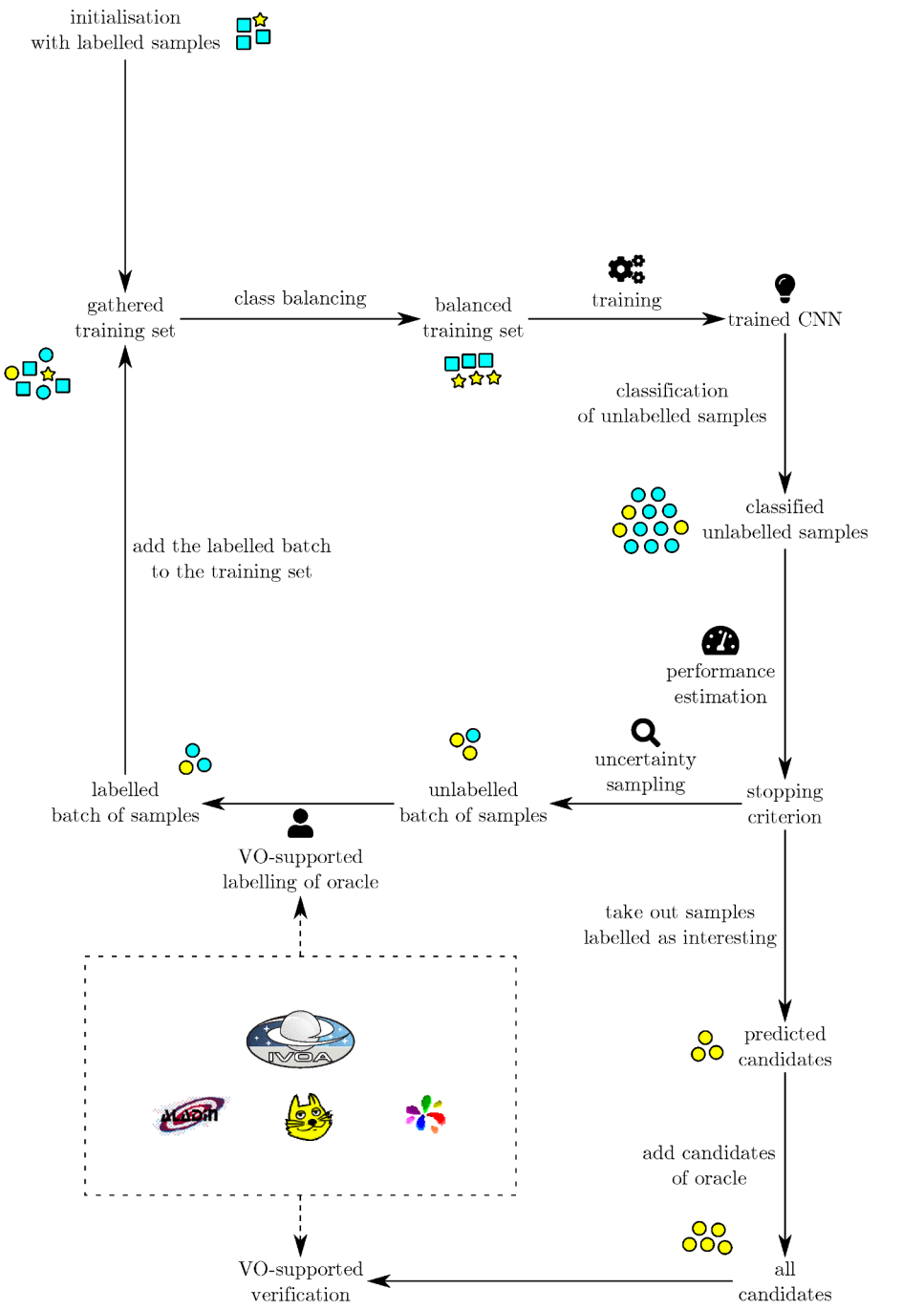

Figure 1. Flowchart of our active deep learning method.

et al. 1989) are the state-of-the-art deep learning method performing well in many astronomical tasks, but this comes with some caveats.

In many cases we face the *class imbalance problem* (Prati et al. 2009). Labelled instances of rare objects of interest will usually be in the minority. To overcome this problem, the Synthetic Minority Over-sampling Technique (SMOTE) proposed by Chawla et al. (2002) is used, that allows us to enlarge the number of labelled spectra of interest to the same size as the more abundant uninteresting ones.

Another problem of deep learning is the need for a very large and representatively labelled training set. Unfortunately, such a training set is not available in most cases of the discovery of the rare objects of interest. We show that the introduction of active learning techniques helps. Active learning (Settles 2009) is a machine learning technique based on the idea that an algorithm will perform better if it is allowed to choose data for its training. A machine learning algorithm combined with active learning queries unlabelled data samples to be labelled by an *oracle* (usually a human expert). In our method, samples are queried in batches from a large pool of data using *uncertainty sampling*, which selects data with the least certain labelling (based on information entropy).

In the majority of scientific applications, the oracle should decide about the correct label after checking all relevant information. As this must be done in every iteration, the rapid access to global databases in the VO infrastructure becomes an integral part of the whole active deep learning workflow. The whole algorithm of our active deep learning method is shown in Fig. 1.

In summary, the active deep learning takes the labelled data as the initial training set and balances it. Having a balanced training set, we train a CNN and use it to classify all the unlabelled spectra. Then, we use the uncertainty sampling query strategy to get batches of samples for labelling by an oracle. The labelled samples are moved to the training set These steps are repeated until the performance of our CNN is satisfactory. When we are satisfied, the unlabelled samples that were predicted as interesting, become new candidates of objects of interest. Finally, we add the samples labelled by the oracle as interesting from the training set to the candidate set.

3. Search of emission-line spectra in LAMOST DR2

To illustrate the application of our active deep learning method, we have performed experiments with the discovery of objects with signatures of Hα emission in the LAMOST DR2 survey using labelled data from the Ondřejov 2 m Perek telescope.

The publicly available LAMOST DR2 survey contains over four million spectra with a spectral resolution power around 1 800 covering the range 3 690–9 100 Å. Although most of them have a spectral classification assigned by an automatic pipeline, there is no special class of emission-line stars defined there. The homogeneous sample of spectra of emission-line stars, namely Be and B[e], is collected in the archive of the Coudè spectrograph of the 2 m Perek telescope at the Ondřejov Observatory of the Astronomical Institute of the Czech Academy of Sciences. Therefore, we used 12 936 spectra exposed in spectral range 6 250–6 700 Å with spectral resolving power about 13 000 as the initial training set. Those spectra were first transformed to look as exposed by LAMOST (reducing the resolution by Gaussian blurring) and classified according to the visual shape of the Hα into three classes: single-peak, double-peak, and absorption (see the Ondřejov dataset on `https://doi.org/10.5281/zenodo.2640971`).

The labelled spectra entered our active deep learning workflow. In our case, the oracle (human annotator) classified spectra into target classes (with single-peak or double-peak profile) and the non-target (absorption spectra plus noisy or somehow corrupted data) in a batch of 100 spectra with the highest information entropy.

Our method identified over four thousand of candidate spectra with signatures of emission-line profiles in more than four million of LAMOST DR2 spectra with an estimated error less than 6%. Thanks to the VO technology we were able to cross-match most of them with known emission-line objects in SIMBAD, but still there are more that one thousand of yet unknown objects where the emission is identified by LAMOST for the first time.

The final catalogues of all our spectra of emission-line candidates obtained by active deep learning are available as an on-line supplement on `https://doi.org/10.5281/zenodo.3241521` for further investigation.

4. Conclusions

We have introduced a new promising method for discovery of objects of interest in large archives based on active deep learning. This technique supported by interactive visual classification of a small sample of suggested target classes turned out to be very efficient, leading to a discovery of many new unknown emission-line stars.

The main advantage of the method is the possibility to identify target classes with characteristic spectral features in cases where the classical deep learning fails due to the insufficient number of labelled examples.

Unlike the current machine learning workflows, the active learning is based on the iterative visualization of predicted candidates followed by labelling by a domain expert (in the role of an oracle). The oracle must thus have complex information about the given candidate to decide correctly. Here is the place where the complex queries in global databases of VO are necessary and the VO clients become an integral part of active learning setups.

Acknowledgments. This research has been 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.

The work is based on spectra from the Ondřejov 2 m Perek telescope and the public LAMOST DR2 survey. We are namely grateful to Dr. Chenzhou Cui for support of our research by Chinese VO and Dr. Miroslav Šlechta for reducing all spectra in the Ondřejov archive of the 2 m Perek telescope.

References

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, Journal of Artificial Intelligence Research, 16, 321

Goodfellow, I., Bengio, Y., & Courville, A. 2016, Deep Learning (MIT Press). `http://www.deeplearningbook.org`

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. 1989, Neural Computation, 1, 541

Prati, R. C., Batista, G. E., & Monard, M. C. 2009, in IICAI, 359

Settles, B. 2009, Active Learning Literature Survet, Computer Sciences Technical Report 1648, University of Wisconsin–Madison