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 of the Czech Technical University in Prague, Czech Republic

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

Abstract

Deep neural networks have been proved as 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 present a novel active deep learning method.

It is based on an iterative training of a deep network followed by re-labelling of a small sample of predicted target classes according to a qualified decision of an expert in the role of an oracle. 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, a much more relevant label is assigned.

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

1 Deep Learning

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

Balanced classes In many cases of machine learning experiments 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) may be used, that allows one to enlarge the number of labelled samples of interest to the same size as the more abundant uninteresting ones.

Massive labelled training set Another problem of deep learning is the need for a very large and representative labelled training set. Unfortunately, such set is not available in most cases of the discovery of the rare objects.

2 Active Learning

To overcome the need for a large and representative training set an active learning was invented. Active learning (Settles, 2009) is a machine learning technique based on the idea the 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). The samples are selected in batches of given size according to a certain informativeness measure. Commonly used is the *uncertainty sampling*, which selects data with the least certain labelling (based on information entropy)

$$H = -\sum_i p_i \ln p_i, \tag{1}$$

where p_i is the probability of class i in all samples of the large data pool.
This measure is evaluated in every iteration on the whole pool of data entities (it is called *pool-based strategy*).

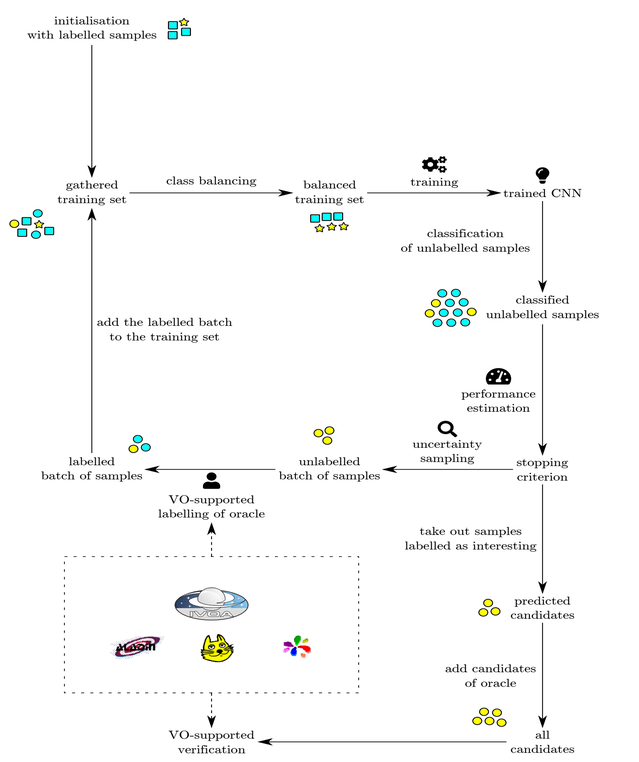
3 Know-how of the oracle

The role of the oracle (usually a human annotator) is to assign the correct label to the queried sample. It is expected he can make the decision immediately just by looking at the visual properties of the sample.
In real cases, namely in the majority of scientific applications, the decision may be difficult just based on a information seen in a presented sample. This opens question, whether the oracle should decide about the correct label after checking all relevant information available about the investigated sample. For example, in astronomical spectra classification needs the annotator to check what is about given object written in other catalogues, look at its place in all-sky surveys or see other spectra in extended spectral range than this presented.
As this must be done in every iteration, **the rapid access to global databases becomes an integral part of the whole active deep learning workflow**.
This is different in current machine learning practice. Here the training set is already labelled when entering the process and all the training and prediction runs without the need of human intervention automatically. The data expert only needs to check the performance and adjust the hyper-parameters to achieve a good convergence and prevent the over-fitting.
In order to exploit the power of active learning in real astronomical research current Big-Data machine learning infrastructures based on cloud computing, distributed nodes, GPUs, Spark-based parallelisation and other modern IT technologies need to incorporate the *iterative labelling GUI and complex VO clients*. **This opens the space for a new type of Big Data science platforms**.

4 Active Deep Learning Cycle

Our method combines all components presented above in a novel active learning procedure. The whole algorithm of our active deep learning method is shown below. Note that the access to the global VO infrastructure is needed twice:

- To support decision of an oracle about correct class.
- To verify the final candidates and identify interesting objects.



Icons by Font Awesome are licensed under CC BY 4.0

5 Search of emission-line spectra in LAMOST DR2

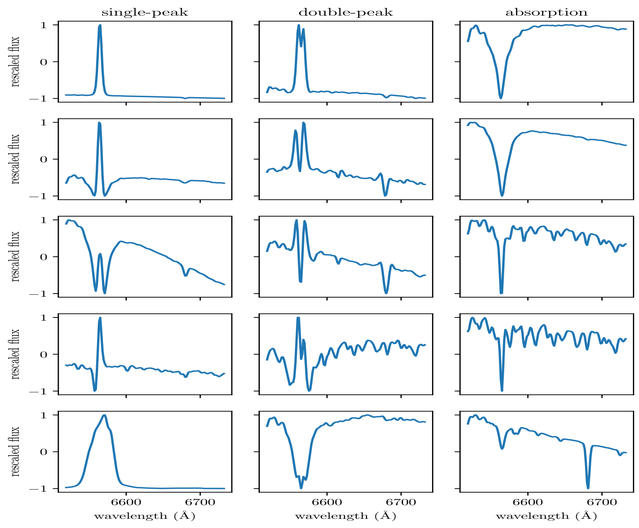
Be stars, cataclysmic variables, young stellar objects or quasars, where a gaseous envelope in the shape of a sphere or a disk is expected, present some spectral lines in emission. These emission lines show single-peak, double-peak or complicated combined emission and absorption profiles. Our goal is to find such objects in a big spectra survey using labelled data from other dedicated archive.

Candidates were searched in LAMOST DR2 containing over 4 million of spectra covering the range 3 690–9 100 Å with spectral resolution power around 1 800.

Training data: 12936 spectra in the archive of coudè spectrograph of the 2m Perek Telescope at Ondřejov observatory exposed in spectral range 6250–6700 Å with spectral resolving power about 13000.

Domain transfer: Gaussian blurring and resampling to the same resolution and wavelength points as in LAMOST spectra.

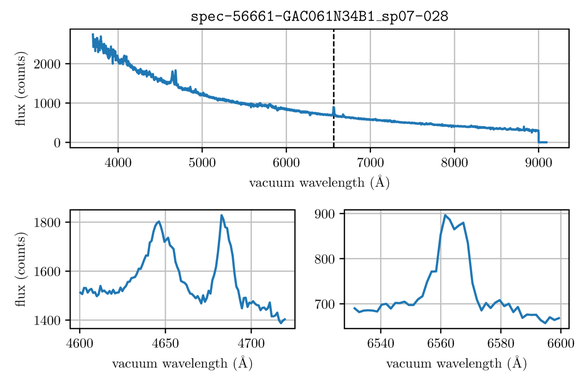
Classification: All spectra from Perek telescope were classified in three classes: single-peak emission, double-peak emission and the absorption, which is not, however, our target class):



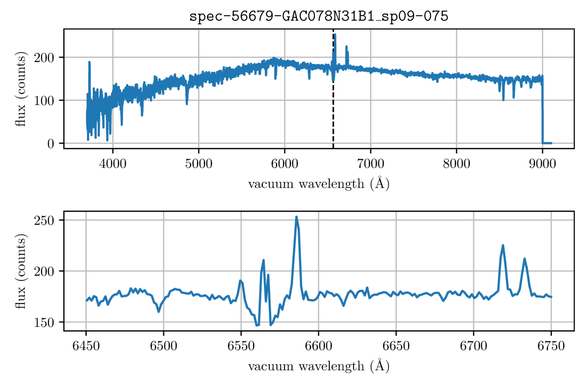
Example of classes of emission line stars in Ondřejov archive

6 Interesting objects discovered

The active deep learning method allowed us to discover hundreds of new emission line objects as Be stars, young stellar objects, T Tau stars, cataclysmic variables and a lot of objects with unknown classes. We have cross-matched part of them with SIMBAD but over one thousand objects discovered are not listed there. Tables of all 4321 emission line objects discovered by our method (1013 yet unknown) are at *https://doi.org/10.5281/zenodo.3241521*. Two interesting examples are given below.



Candidate Wolf-Rayet WN star LAMOST J040901.83+323955.6.



A unknown star with complex line profile

7 Conclusions

We have introduced a new promising method for discovery of objects of interest in large archives based on active deep learning, which allowed us to discover many yet unknown emission-line stars.
Its main advantage 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 visualisation 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.

Acknowledgements

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 archive of Ondřejov 2 m Perek telescope.

References

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

R. C. Prati, G. E. Batista, and M. C. Monard. Data mining with imbalanced class distributions: concepts and methods. In *IICAI*, pages 359–376, 2009.

B. Settles. Active learning literature survet. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

Poster presented at ADASS 2019, 6 – 10th October 2019, Groningen, the Netherlands

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research, Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

RESEARCH CENTER FOR INFORMATICS
rci.cvut.cz