Astronomical Data Analysis Software and Systems XXXI
ASP Conference Series, Vol. 535
B. V. Hugo, R. Van Rooyen, and O. M. Smirnov, eds.

Prototype of Interactive Visualisation Tool for Bayesian Active Deep Learning

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. In the era of big data in astronomy, we need to develop methods to analyse the data. One such method is Bayesian active deep learning (synergy of Bayesian convolutional neural networks and active learning). To improve the method's performance, we have developed a prototype of an interactive visualisation tool for a selection of an informative (contains data with high predictive uncertainty, is diverse, but not redundant) data subsample for labelling by a human expert. The tool takes as input a sample of data with the highest predictive uncertainty. These data are projected to 2-D with a dimensionality reduction technique. We visualise the projected data in an interactive scatter plot and allow a human expert to label a selected subsample of data. With this tool, she or he can select a correct subsample with all the previously mentioned characteristics. This should lower the total amount of data labelled because the Bayesian model's performance will improve faster than when the data are selected automatically.

1. Introduction

Astronomy faces an avalanche of data. There are so many data that most of them will never be visually inspected by astronomers. Therefore, we need to develop (e.g. machine learning) methods to analyse them. These methods should help us understand the data or make discoveries. For example, the Sloan Digital Sky Survey archive (SDSS; Blanton et al. 2017) contains millions of astronomical spectra processed only by an automated pipeline. Furthermore, the Gaia mission (Gaia Collaboration et al. 2016) will survey more than 1 billion stars, and Large Synoptic Survey Telescope (LSST; Ivezić et al. 2019) will observe about 40 billion galaxies and stars.

When such big data are available, deep learning models are the state of the art to analyse them. Deep learning (LeCun et al. 2015) is a subfield of machine learning that builds deep models by composing them from several processing layers. Convolutional neural networks (CNNs) are deep learning models that process data with a grid structure (e.g. spectra or imaging data). However, CNNs will work adequately only if two conditions are met: (1.) there is a sufficiently large human-labelled training set and (2.) the training set is representative of the target test data (i.e. the data we want to analyse).

Astronomical data mostly do not satisfy both conditions. First, human-labelled datasets are scarce in astronomy. Data in astronomy are mostly unlabelled, although they commonly have machine labels. The performance of CNNs depends on human-labelled datasets. Otherwise, they would only replicate the algorithm that produced the machine labels with all its errors. Second, the scarce human-labelled astronomical datasets are diverse due to different instruments, scientific goals, or their observations come from different parts and depths of the sky.

Bayesian active deep learning (synergy of active learning and Bayesian CNNs) can help us mitigate the problems. Here, we present a prototype of an interactive visualisation tool to improve the method's performance further.

2. Bayesian Active Deep Learning

Bayesian active deep learning combines active learning (Tong 2001) with Bayesian CNNs (Gal 2016). Active learning solves the problem of lack of a large, human-labelled, representative training set. Bayesian CNN aids active learning with theoretically founded predictive uncertainty. This method has already been successfully applied to galaxy morphological classification by Walmsley et al. (2020).

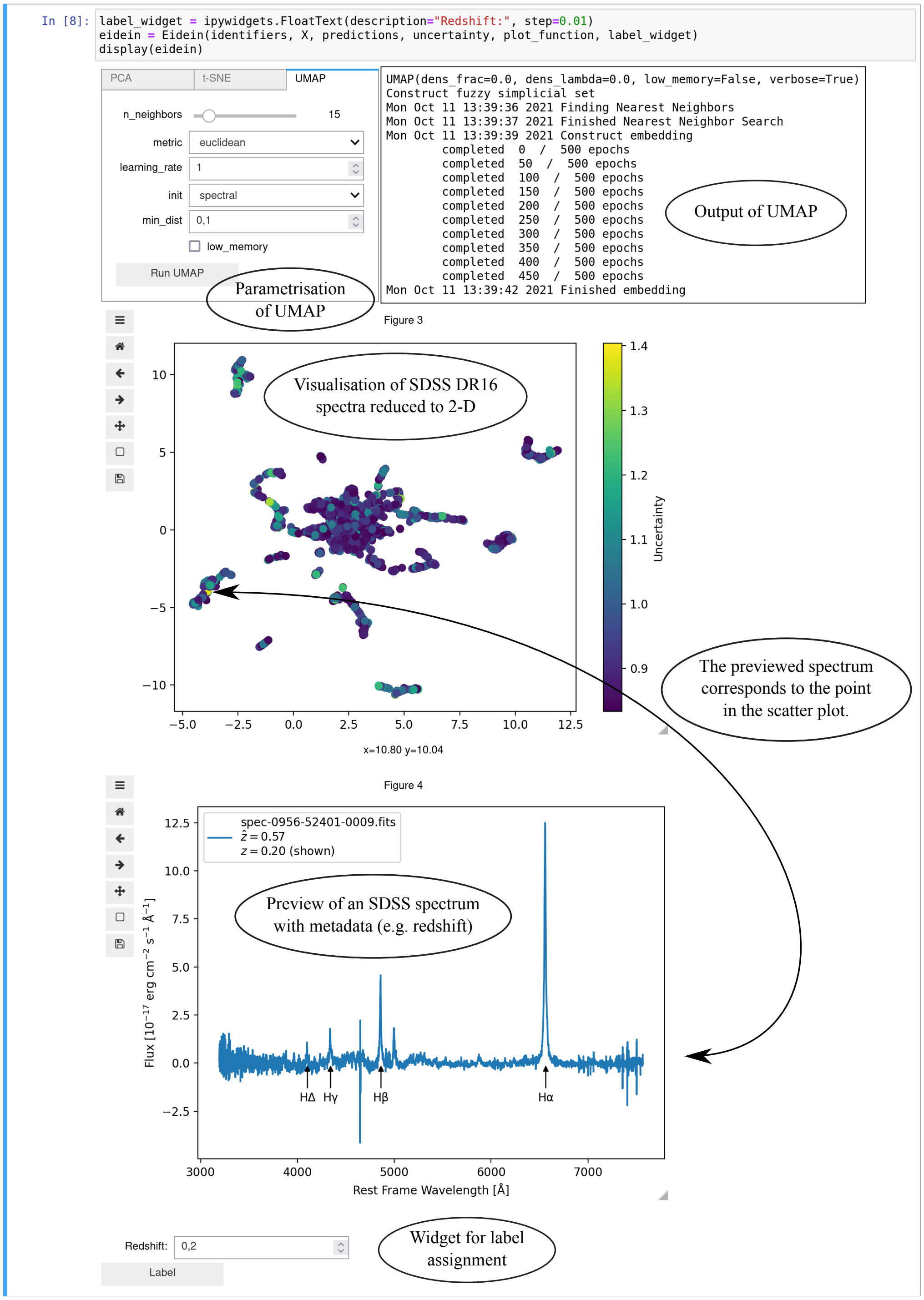
Active learning is based on the idea that a machine learning model will perform better and with fewer data if it can choose its training data. The model queries a subsample of unlabelled data to be labelled by a human expert (astronomer). The subsample usually consists of data with the most uncertain predictions. The initial training set of the model is extended with the labelled subsample. Then, the model is retrained, and the whole process repeats until we are satisfied with the model's performance.

Selecting an informative subsample for labelling is crucial to improve Bayesian active deep learning performance. We want a subsample that will improve the model's performance as much as possible. That means we need a subsample that (1.) contains data with high predictive uncertainty, (2.) is diverse, (3.) but not redundant. A correctly selected subsample for labelling is essential for active learning because it can speed up the method. After all, labelling by a human expert is often time-consuming and expensive.

Bayesian CNNs will ensure that the selected subsample contains data with high predictive uncertainty. These CNNs are deep learning models that bring theoretically founded predictive uncertainty into deep learning. Unlike classical CNNs, they also capture the uncertainty in the model parameters. Therefore, we know when the model is certain and when it is guessing at random. Gal et al. (2017) have shown that Bayesian CNNs used with active learning are better than classical CNNs.

3. Prototype of the Interactive Visualisation Tool

We have developed a prototype of an interactive visualisation tool for the selection of the informative subsample. The tool is implemented in the Python programming language as a widget for the Jupyter Notebook. It takes as input a sample of data with the highest predictive uncertainty. These data are projected to 2-D with a dimensionality reduction technique. Currently, the user can choose between (1.) principal component analysis (PCA), (2.) t-distributed stochastic neighbor embedding (t-SNE; van der Maaten & Hinton 2008), and (3.) uniform manifold approximation and projection (UMAP; McInnes et al. 2018). We visualise the projected data in an interactive



```
In [8]: label_widget = ipywidgets.FloatText(description="Redshift:", step=0.01)
        eidein = Eidein(identifiers, X, predictions, uncertainty, plot_function, label_widget)
        display(eidein)
```

n_neighbors 15

metric euclidean

learning_rate 1

init spectral

min_dist 0,1

□ low_memory

```
UMAP(dens_frac=0.0, dens_lambda=0.0, low_memory=False, verbose=True)
Construct fuzzy simplicial set
Mon Oct 11 13:39:36 2021 Finding Nearest Neighbors
Mon Oct 11 13:39:37 2021 Finished Nearest Neighbor Search
Mon Oct 11 13:39:39 2021 Construct embedding
        completed  0  /  500 epochs
        completed  50  /  500 epochs
        completed  100  /  500 epochs
        completed  150  /  500 epochs
        completed  200  /  500 epochs
        completed  250  /  500 epochs
        completed  300  /  500 epochs
        completed  350  /  500 epochs
        completed  400  /  500 epochs
        completed  450  /  500 epochs
Mon Oct 11 13:39:42 2021 Finished embedding
```

Redshift: 0,2

Parametrisation of UMAP

Output of UMAP

Visualisation of SDSS DR16 spectra reduced to 2-D

The previewed spectrum corresponds to the point in the scatter plot.

Preview of an SDSS spectrum with metadata (e.g. redshift)

Widget for label assignment

Figure 1. Screenshot of the protype of the interactive visualisation tool that is implemented in Python as a widget for the Jupyter Notebook.

scatter plot and allow the user to select and label a subsample of diverse but not redundant data.

There are several advantages in selecting the subsample from the visualisation of data with reduced dimensionality. Visualisation will reveal clusters of similar data while separating dissimilar clusters. Therefore, the human expert can select a diverse subsample that consists of data from different clusters. Additionally, more data can be labelled at once because similar spectra (i.e. with the same label) cluster together. Finally, the human expert will ignore noisy or bad data that would not serve the learning of the Bayesian CNN. With this tool, a human expert can select the correct subsample with all the previously mentioned characteristics: high predictive uncertainty, diversity, non-redundancy. This should lower the total amount of data labelled because the Bayesian CNN's performance will improve faster than when the data are selected automatically.

Figure 1 illustrates an application of the prototype to spectroscopic redshift prediction in the SDSS data release 16 (Ahumada et al. 2020). The tool is given 2048 data items with the highest predictive uncertainty of the Bayesian CNN. Visualisation of the data with reduced dimensionality using UMAP displays many clusters of spectra. The bottom plot shows a spectrum selected by clicking into the visualisation in the top plot. Now, a human expert can correct the spectrum's predicted redshift to the true redshift and add the spectrum to the subsample.

4. Conclusions

We have presented a prototype of an interactive visualisation tool for Bayesian active deep learning. The tool empower a human expert to select and label a subsample that is diverse, non-redundant (because of properties of dimensionality reduction techniques) and contains data with uncertain predictions (because of Bayesian CNNs). Such an informative subsample will be better for the Bayesian CNN's learning than data selected for labelling automatically. Therefore, we will be able to analyse large unlabelled data in astronomy because fewer labelled data will be needed.

Acknowledgments. This research is supported by the 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 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

Ahumada, R., et al. 2020, ApJS, 249, 3. 1912.02905
Blanton, M. R., et al. 2017, AJ, 154, 28. 1703.00052
Gaia Collaboration, et al. 2016, A&A, 595, A1. 1609.04153
Gal, Y. 2016, Ph.D. thesis, University of Cambridge
Gal, Y., Islam, R., & Ghahramani, Z. 2017, arXiv e-prints. 1703.02910
Ivezić, Ž., et al. 2019, ApJ, 873, 111. 0805.2366
LeCun, Y., Bengio, Y., & G., H. 2015, Nature, 521, 436
McInnes, L., Healy, J., & Melville, J. 2018, arXiv e-prints. 1802.03426
Tong, S. 2001, Ph.D. thesis, Standford University
van der Maaten, L., & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579
Walmsley, M., et al. 2020, MNRAS, 491, 1554. 1905.07424