DEEP LEARNING AND UNSUPERVISED FEATURE EXTRACTION FOR GALAXY SED

Keywords. Deep learning, unsupervised feature extraction, denoising autoencoders, galaxy classification.

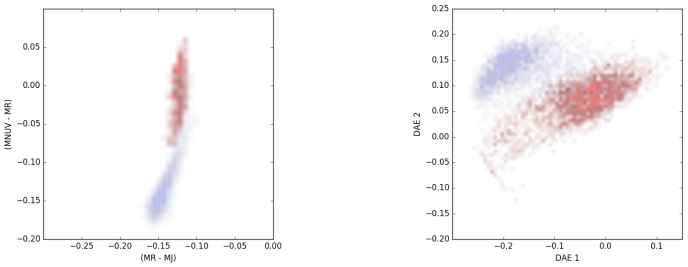
Context. With the increasing number of multi-wavelength galaxy surveys, astrophysicists have access to a wealth of data that provides crucial information to investigate the formation of galaxies as well as their evolution. More precisely, most galaxy properties are generally analyzed from physical measurements describing the distribution of the energy across different spectral bands. From this spectral energy distribution (SED), two major galaxy populations can be differentiated : one family referred to as star-forming or blue galaxies ; and another tagged as quiescent or red galaxies (see Fig. 1). This bi-modality has been a key discovery of the last decade.

The most widely spread methods in astrophysics usually represent the galaxies in a diagram built from only two or three colors of the SED. Therefore, **most of the information of the spectra is not exploited**. In addition, the color boundaries separating the two classes are empirically established and depend on the bands chosen to build the diagram. From a statistical point of view, **these representations do not ensure to efficiently extract relevant information** from the data and will lead to results largely biased by the definition of the two classes.

Goal and method. In the meantime, recent advances in machine learning, such as Deep Learning methods, introduced highly sophisticated data analysis tools that are promising candidates to build unsupervised data-driven alternatives to SED color diagrams. These statistical methods (see for e.g. [1, 2] and references therein) have already proved their efficiency to solve supervised data classification tasks in applications as diverse as computer vision (e.g. [3, 4]), speech recognition, natural language processing, etc.

More specifically, denoising autoencoders, introduced in [5], have been investigated to retrieve information for galaxy SEDs in an unsupervised manner. Preliminary results have demonstrated the interest of these methods as well as their capability to retrieve the galaxy bi-modality and extract astrophysically relevant information from the data that standard SED color diagrams do not exhibit. Fig. 1 illustrates the bi-modality defined in classical color/color diagrams on the left, and the representation obtained with the denoising autoencoders on the right.

Classical color/color planes and denoising autoencoder diagrams are both derived from galaxy SED obtained through best-fit parameterized SED models. However, these models can be quite imperfect and only carry a portion of the input information. **The goal of this project is to investigate alternative deep**



(a) (b)

FIGURE 1. Example of (a) the classical color/color plane and (b) the DAE diagram. Star-forming galaxies are shown in blue, and quiescent are represented in red.

learning architectures for SEDs information extraction on the magnitudes directly measured by the instrument. Due to the increased complexity of the data, this requires adapting the learning method. Ultimately, the learnt SED representations will be used in classification tasks in order to properly separate star-forming galaxies from passive populations.

Candidate. The candidate should be a Master 2 student (or equivalent) and should have a good knowledge in signal/image processing. Knowledge in machine learning is a plus.
During this internship, the candidate will acquire knowledge in various fields of machine learning : i) theoretical models of deep learning and their application for unsupervised representation learning, ii) implementation of deep learning architectures using widespread libraries like TensorFlow, iii) application of classification techniques to astrophysics real data.

Contact information.

- *Contact :* joana.frontera-pons@cea.fr or jbobin@cea.fr
- *Lab :* CEA/IRFU in Saclay
- *Duration :* at least 4 months
- *PhD :* yes
- **Applications are expected before the 28th of february 2017.**

Bibliography. [1] Y. Bengio, A. Courville, and P. Vincent, **"Representation learning : A review and new perspectives "**, IEEE Tr. on Pattern Analysis and Machine Intelligence, 2013.
[2] Y. Bengio, **"Learning deep architectures for AI"**
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, **"Imagenet classification with deep convolutional neural networks"**, in Advances in neural information processing systems, 2012.
[4] D. Ciresan, U. Meier, and J. Schmidhuber, **"Multi-column deep neural networks for image classification "**, in CVPR, 2012.
[5] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, **"Extracting and composing robust features with denoising autoencoders "** in Proceedings

of the 25th international conference on Machine learning, 2008.