AAP Contrats doctoraux en Intelligence artificielle

Cofinancé par l'ANR

TITLE: Cosmic web and galaxy evolution with Deep Learning

1. DESCRIPTION OF THE PHD THESIS PROJECT

1.1 OBJECTIVES OF THE PROJECT BASED ON THE CURRENT STATE OF THE ART

This thesis will develop deep learning (DL) methods to estimate the photometric redshifts ("photo-z") of galaxies (i.e. their distances) and their physical properties by exploiting multi-band imaging surveys. As our team (CPPM, LAM, IAP) recently demonstrated, the convolutional neural network (CNN) technique, applied to the SDSS local survey, significantly improves the accuracy of photo-z compared to all previous methods (**Pasquet et al., 2019**). The goal of this thesis is to extend this work at higher redshift with the state-of-the-art HSC-CLAUDS deep imaging survey, which mimics the future LSST survey and where a large spectroscopic redshift training set is available. One major challenge is to deal with unbalanced and incomplete representativity of the training set. New DL methodologies will be developed, based on unsupervised learning and deep generative approaches to better consider poorly represented objects in the training set. The expected gain in photo-z accuracy (dz<0.02) will enable us, for the first time, to reconstruct the cosmic web (CW) in thin redshift slices to high redshift. The science goal will then be twofold: 1- to study the link between galaxy properties and their large scale environment; 2- to measure the connectivity of the filamentary structures (mean number of filaments) around CW peaks at different epochs. This measurement of the local topology near the nodes at different redshifts is a tracer of the growth rate of the cosmic structures (merging and disconnection of filaments) and depends on dark energy which will be used as an independent cosmological constraint. The PhD thesis is part of an existing collaboration between Marseille, Montpellier and Paris, recently reinforced by the approved ANR project DEEPDIP that will provide a dynamic and rich environment for a PhD student.

1.2 METHODOLOGY

The cosmic web (CW) is a complex network of voids, walls, filaments and knots in which galaxies form and evolve. The exchanges of gas and energy (infalls/outflows) between the galaxies and their environment play a crucial role in shaping galaxy properties. Theoretical predictions have emphasized the close link between the large scale cosmic flows and the spin orientation of galaxies, recently observed in the local universe (Tempel et al., 2013). Dependencies between the stellar mass and star formation activity of galaxies and their distances to CW filaments have also been detected (**Kraljic et al., 2018 at z<0.3; Malavasi et al., 2017 at z~0.8**). These preliminary results offer a new way to understand galaxy evolution in a cosmological context with the forthcoming large spectroscopic surveys. Beyond the spectroscopic 3D mapping, the CW can also be reconstructed in thin 2D redshift slices with robust photo-z's, as shown by **Laigle et al. (2018)** with the COSMOS 30 bands survey. This technique will be used to investigate the influence of cosmic web on galaxy evolution to existing imaging surveys (CFHTLS and HSC-CLAUDS) in preparation for the gigantic imaging survey LSST (more than a half hemisphere).

However to reach this goal, a challenge is to get very accurate photo-z measurements (dz<0.02) with a limited set of filters. One limiting factor of current photo-z techniques has been the extraction of photometric quantities (fluxes and colors) used as input, which capture only a fraction of the information present in the images. Taking advantage of the latest DL techniques, of the GPU acceleration and of the large size of spectroscopic training samples in the local universe (SDSS), photo-z derived with DL algorithms (e.g. dealing directly at the pixel level; Hoyle et al., 2016; d'Isanto, 2017; Pasquet et al., 2019) outperform traditional approaches. In particular, we (**Pasquet et al.; 2019**) found that DL photo-zs show no bias with redshift, galaxy inclination, ..., compared to other machine learning methods and provide robust probability distribution functions, crucial for cosmological analyses.

In this thesis, we will exploit the CFHTLS and the HSC+CLAUDS survey (covering 20 deg2 with multi-band UgrizY imaging down to r~27, **Sawicki et al., 2019**). The later is the best existing survey for this purpose, with a unique combination of depth and area, rich ancillary data and a large spectroscopic training set (>100,000 redshifts). This dataset represents a major forerunner to prepare for LSST science. This will be the first time that DL methods are pushed to high redshift and faint magnitudes (i.e. low SNR images). A key point to address will be how to deal with unbalanced and/or incomplete representativity of the training sample.

To this end, the thesis will focus on two complementary ways, each bringing major innovations in DL:

- The first one will address the problem of mismatch between the training and testing databases through unsupervised and semi supervised approaches. Indeed, **Pasquet et al. (2019)** demonstrated that the mismatch problem could be reduced by using an unsupervised pre-learning step on the testing database, followed by a semi-supervised learning step. In this thesis, the student will explore the physical properties of the galaxies that may constrain this training process. We propose to set up an innovative architecture based on a contrastive loss function (Hadsell et al., 2016). An extension will be to generate pairs of galaxies with very distant physical properties. The goal is to create a robust model able to extrapolate the poorly populated areas of the training database. The student will then upgrade this architecture to take into account the testing database during the semi-supervised learning step. To this end, we propose to explore state-of-the-art clustering algorithms, such as DeepCluster (Caron et al., 2018), during the training step. This phase will allow the model to use the testing database to train the network, using the physical properties of the galaxies. In addition, these clusters could be used to separate the training data in different classes and overcome the mismatch problem as shown in (**Pasquet et al., 2014**). The convergence time of such a structure is an important technological barrier and this will be a challenge for the thesis.
- The second axis will address the problem of low numbers of observations in certain areas of the training set, e.g. low redshift galaxies with high magnitudes. The idea is to populate the training dataset with simulated images of galaxies as realistic as possible. To achieve this goal, the student will explore a deep generative approach based on the Cyclic GAN (Zhu et al., 2018). This kind of architecture learns to translate an image from one representation to another with the same semantic. In this thesis, the student will extend this kind of network by changing one or several physical parameters of the galaxies. The model will have to estimate these parameters and then change theirs values according to a parameter given by the user. The automatic extraction of the properties on an image and their modification to generate a new image are major steps in generative DL research. This technique will have many different applications for the generation of missing data in machine learning.

1.3 WORK PLAN

The first part of the PhD (1-1.5 yr) will be dedicated to the improvement of the photometric redshifts via the DL techniques discussed above. It will be supervised by Jerome Pasquet (TETIS) in Montpellier. The student being based at LAM, regular visits and meetings will be organized to follow the progress in these critical steps.

The second part of the PhD will be dedicated to the cosmic web analysis with the large imaging surveys CFHTLS and HSC-CLAUDS. In parallel, dark matter simulations will be used to first explore the impact of photometric redshift accuracy to reveal the cosmic web in thin redshift slices and to confront the simulations with the observed connectivity as a function of peak properties and redshift, and perform cosmological forecasts for the LSST.

4. SUPERVISOR AND RESEARCH GROUP DESCRIPTION

The student will be under the supervision of **Stéphane Arnouts** at LAM (**stephane.arnouts@lam.fr**), who has a long expertise on photometric redshifts and imaging surveys and co-supervised by Jérome Pasquet at TETIS (Montpellier), who is an expert in deep learning techniques. The PhD student will interact with researchers and engineers at LAM (M. Treyer, D. Vibert, S de La Torre, O. Ilbert) and IAP (S. Codis, C. Pichon, C. Laigle) covering a large panel of expertises in astronomy (observations, simulations and theory). For the Deep learning part, in addition to the supervision by Jerome Pasquet, s/he will have the opportunity to interact with the LIS namely the Qarma research team and Thierry Artières (Marseille) to benefit from their advice and expertise in generative GAN. S/he will also collaborate with researchers from CPPM (D. Fouchez) and IAP (E. Bertin, S. Codis) in the framework of the ongoing ANR DEEPDIP directly related to the topic of this PhD subject, providing a rich and dynamic environment for the student.

2. RECENT PUBLICATIONS

Pasquet, Bertin,Treyer, Arnouts, Fouchez, 2019, A&A 621,26
Pasquet, Bringay, Chaumont, EUSIPCO 2014
Malavasi, Arnouts et al., 2017, MNRAS 465, 3817
Kraljic, Arnouts, Pichon et al., 2018, MNRAS 474, 547
Laigle, Pichon, Arnouts et al., 2018, MNRAS 474, 5437
Sawicki, Arnouts, Huang et al., 2019, MNRAS 489, 5202
Shuntov, Pasquet, Arnouts et al.,2020, A&A 636, 90

3. EXPECTED PROFILE OF THE CANDIDATE

The candidate is expected to have - a good academic background with a master in physics, astronomy or data science - a background in statistics - a strong interest in machine learning and/or deep learning techniques with a motivation for their applications in an astrophysical context - a good programming skills in python - autonomy, ability to learn by him/herself and to lead his/her own investigations.

4. SUPERVISORS' PROFILE

Arnouts, Stéphane Laboratoire d'Astrophysique de Marseille **stephane.arnouts@lam.fr**

Education & career

2003 —	permanent research position at LAM
1999-2002	ESO fellowship at Garching (Germany)
1997-1999	Individual Marie Curie Fellow at Padova (Italy)
1992-1996	PhD thesis at IAP, university of Paris VII

Responsabilities

2019 —	lead of the GECO team at LAM (~50 persons)

Recent funding projects

2020 - 2023	co-PI ANR DEEPDIP	(CPPM / LAM / IAP / LIRMM)
2013 - 2018	co-PI ANR SPINe	(IAP / LAM)

Recent observation campaigns

2010 - 2015	PI VIPERS-MLS: UV (GALEX~100h) & NIR (CFHT~120h) obs.
2016 - 2019	PI CLAUDS : U band obs. (CFHT~350h) of HSC Deep fields

Developer

Le Phare	SED fitting code for photometric redshift and galaxy properties
EM Phot	Flux measurement in crowded fields with priors (for GALEX satellite)

Publications

Number of articles in refereed journals : 178 (total citation: 23,000; h-index: 62)

Supervisions

Post-doc	Blaizot (2003-04); Pollo (04); Kraljic (2014-17)
PhD thesis	Heinis (2002-05); Moutard (2013-15); Malavasi (2016); Picouet (2018-21)
Masters	Heinis (2002); Magnelli (06); Tourneboeuf (11); Pernot-Borras (16); Comte (17); Shuntov (18)

VISA DU RESPONSABLE DE L'INSTITUT ET DU DIRECTEUR DE LABORATOIRE CONCERNÉS

Visa du responsable de l'institut,
NOM Prénom

Fait à Marseille, le 14/05/2020

Signature

Eric KAJFASZ
Directeur - IΦU
Institut de Physique de l'Univers

Visa du directeur du laboratoire,
NOM Prénom Jean-Luc BEUZIT
Directeur du LAM

Fait à Marseille, le 14/5/2020

Signature