Prospective IN2P3
Survey Synergies with Machine Learning
GT05+GT09

Éric Aubourg, Jim Bartlett, Alexandre Boucaud,
Ken Ganga, Yannick Giraud-Héraud, Maude Le Jeune,
Cécile Roucelle, Françoise Virieux (APC),
Dominique Fouchez (CPPM),
Hélène Courtois (IP2I Lyon),
Réza Ansari, Jean-Éric Campagne (LAL),
Emmanuel Gangler, Emille Ishida, Anais Möller (LPC Clermont)

October 2019

Abstract

We propose to develop new machine learning techniques, focusing on bayesian deep learning, to address the analysis challenges of future surveys like LSST, Euclid or WFIRST. Those techniques would enable multi-bandpass, multi-instrument processing of individual images, targeting science objectives like shear measurement and photometric redshift estimation on heavily blended objects, as well as time-domain measurements for supernova identification.

1 Introduction

Astronomical surveys planned for the next decade will produce data that present analysis challenges not only because of their scale (hundreds of petabytes), but also by the complexity of the measurement challenges on very deep images (for instance subpercent-level measurement of colors or shapes on blended objects). Novel machine learning techniques appear very promising to address these challenges : once trained, they are very fast, and excel at extracting features from complex images. On the other hand, astronomical surveys offer a unique environment to motivate and test the further development of some machine learning techniques. For example, preliminary results using a special type of deep neural networks (called variational auto-encoders), as galaxies deblenders on simulated images are very encouraging. As a matter of fact, machine learning techniques can be applied directly to multi-bandpass, multi-instrument individual images to address the key observation challenges without going through the traditional steps of image stacking, explicit deblending, catalog generation, that lose information at each stage. Beyond the aforementioned image treatment, analyses pipelines could be deeply changed by the inclusion of machine learning techniques. The example of gravitational lensing which is a major probe for Dark Energy studies expected to be revolutionized by the coming LSST (Ivezić et al. 2008; LSST Science Collaboration et al. 2009), Euclid (Laureijs et al. 2011) and later the NASA WFIRST project (Spergel et al. 2015) surveys will be taken here.

2 Machine learning to help analyses and survey synergies

The Euclid satellite (Laureijs et al. 2011) should be launched in 2022. It will gather about 30 petabytes of imaging data. The ground-based LSST telescope project (Ivezić et al. 2008; LSST Science Collaboration et al. 2009) will start a ten-year survey in 2022 (with commissioning data available in 2020), and will accumulate hundreds of petabytes of images. Later, the NASA WFIRST project (Spergel et al. 2015) should also reach multi-petabyte-scale. The analysis of these next generation astronomical data will nevertheless pose challenges beyond their sheer volume. Those large projects have numerous science objectives, each with specific constraints e.g.: very deep images mean that a large fraction of objects overlap, and have to be *deblended*; measuring weak gravitational shear requires a measurement of the local average shape of galaxies at the sub-percent level, and to keep optical and atmospheric systematic effects under control; measuring the distance of galaxies through their colors ("photometric redshifts" or "photo-z") is a necessary ingredient of shear analysis and requires a measurement of the objects relative flux in different filter bandpasses at the sub-percent level – LSST for instance takes data with six different optical filters, whereas Euclid combines a wide-band optical filter and three near-infrared filters. The treatment of these problematics alone considering the volume of data collected and the requirements for dark energy studies supposes an evolution of our methods.

In addition, the new analyses developments to tackle the dark energy issue rely on joint probes analyses and further down the road joint surveys analyses (Rhodes et al. 2017). Many state-of-the-art astronomical analysis software process each band independently for an individual survey, then combine the measurements once catalogs have been built and the treatment of multiple exposures even for a single survey is currently suboptimal. LSST will visit each sky location of the survey about a thousand times : it is possible to "stack" those shallow images to make deeper images, but information is lost in the process, and more annoyingly, each shallow exposure has different observing conditions, and thus a different point spread function (PSF). Because of CCD gaps and defects, the PSF of a stack is ill-defined and discontinuous.

To make most of the analyses for dark energy studies, work can be done at the pixel level, on individual frames. As a consequence, multi-band (and ultimately multi-instrument) pixel- and frame-level processing has to emerge.

To address those various challenges, ML algorithms able to process petabyte-scale astronomical data will be developed. Once trained, neural networks are very fast, so are well adapted to the scale of the forthcoming surveys. This type of techniques will moreover make use of all the available information in the original dataset, avoiding loss in stack or catalog steps. It will also use in an optimal way the color information, whereas many techniques process separately each channel.

ML algorithms are extremely well suited to address the main challenges of the astronomical surveys planned for the next decade: object identification and separation (deblending), shape measurement, relative photometry measurement, time series identification, all of this with a consistent treatment of uncertainties. With deep learning, they can also be designed to make most of those analyses at the pixel level, on individual frames allowing to integrate multi-band and multi-instrument information. If the development of ML is mature enough to advocate for their use in dark energy studies at this time, this approach will go much beyond the current state-of-the-art in ML and will create novel synergies in particular with pure advances in computer science[1].

To elaborate further on this point: input astronomical images have carefully characterized noise properties (photon noise, CCD readout noise, electronic noise, etc.). Traditional algorithms propagate those uncertainties to the final measurement. This capability is a very strong requirement for any astronomical image processing software. For instance, when deblending overlapping galaxies, the input noise will end up being split between the images of the individual galaxies and a sky background image. Pixels of the output images might have correlated noise, and this noise correlation must also be characterized. Noise properties of the final measurement have to be clearly established to allow their use in dark energy oriented analyses.

[1] The APC team has started collaborations on this topic, it has received an ANR grant for the period 2019-2023 with Inria as one of the partners.

A paradigm shift has thus to be made in research for quantifying uncertainty in deep learning used to identify when deep models are *'guessing at random'* and the development of a new class of probabilistic methods for practical yet principled estimation of deep learning uncertainty. Recent advances in deep learning performance have been spectacular, but reasoning about uncertainty becomes crucial when deploying deep models in real-world applications, and especially for astrophysical applications where imaging datasets suffer multiple types of noises. The field of bayesian deep learning (BDL, Gal 2016), a frontier in machine learning, offers a principled framework to quantify deep models' uncertainty. In BDL, probability distributions are placed over deep models' parameters, giving bayesian neural networks (BNNs) the possibility to infer statistically meaningful uncertainties. But a chasm separates the capabilities of current tools and capabilities needed in applications. The problems are inherent to current BDL theory, requiring heavy theoretical and engineering machinery to solve. Practical and scalable BDL tools for astronomical imaging datasets have to be developed.

Other research work can be motivated by the recent developments in machine learning and especially in deep learning where neural networks achieves very good results in image classification, speech recognition and natural language processing. Many methods are developed to analyze and explore these data in a supervised setting. However, most of the produced data does not come with metadata and/or labeling. This raises two possible directions of investigation: (i) learning to classify using a few labeled examples, and (ii) transferring what was learned in a given domain into another domain. Actually, active learning (AL) tries to overcome the difficulty of having a few examples available, and may involve interaction (Ishida et al. 2019b; Ishida et al. 2019a). Deep learning poses several difficulties when used in an active learning setting, as, by contrast, active learning methods generally rely on the capability to learn and update learning models from small amounts of data. Accordingly, one can adapt the idea of "bayesian active learning" based on recent methodologies about an active learning framework for high dimensional complex data combining bayesian deep learning and active learning (Gal 2016; Gal and Ghahramani 2016; Gal et al. 2017). In addition, bayesian active learning is well suited to deal with uncertainty as this is the case for standard bayesian networks (Ghahramani 2015).

3 Introducing machine learning techniques in analysis pipelines

One of the key observables brought by the LSST and Euclid surveys and that will be crucial for dark energy studies in the next decade will be the shear induced by weak gravitationnal lensing (Kilbinger 2015; Mandelbaum 2018.

If galaxy orientations are randomly distributed, their ellipticity averages out intrinsically. An observed non-vanishing orientation is a local estimate of shear induced by gravitational lensing of the intervening large-scale structure. Measuring shear gives then unbiased information about the dark-matter distribution in the Universe.

A central task in the use of this probe is obviously the galaxy shape measurement and in this exercise the calibration, since generally the estimated shear is biased. The science requirements for LSST, Euclid, or WFIRST are very challenging, demanding calibration at the 0.1% level. A major source of bias are blended objects (Euclid Collaboration et al. 2019). Going down to this exquisite precision in the calibration level requires to think out of the standard techniques, hence to look for an optimal treatment of the exposures to address these problems as developed in the previous part. The advances in ML and cosmology clearly appear to be bound. Further, bayesian inference techniques that do not require the measurement of individual galaxy shapes like in Bernstein and Armstrong 2014 and Bernstein et al. 2016 could be developed and tested, changing the classical pipelines for shear measurements in galaxy surveys.

There has been early attempts at using neural networks to address bias in shear measurements (Gruen et al. 2010), or on building a low-bias shear estimator on individual galaxies using measured features (Tewes et al. 2019). Although, if using deep learning at the pixel level is starting to be used for photometric redshifts, or for measuring galaxy features (Tuccillo et al. 2018; Huertas-Company et al. 2018), this is still mostly uncharted territories for weak lensing.

In the context of weak lensing, an important aspect to explore is the bayesian measurement of photometric redshifts from multi-band images, extracting for each galaxy a p(z) distribution, directly from the blended images (see Jones and Heavens 2019). Compared to the state of the art, which uses template fitting or neural networks applied to "colors" (difference between the measured magnitudes in two bandpass) to derive a value of the redshift with a (Gaussian) uncertainty, and is prone to "catastrophic errors" where degeneracies yield to selecting a bad solution, we plan to work directly at the pixel level, on multi-band image cubes, and output a p(z) probability distribution, using bayesian networks. Some work has already been done on that topic, within LSST-France (Pasquet et al. 2019), but using deep convolutional networks that output probability distribution functions, considering the problem as a classification problem, rather than bayesian networks. An efficient synergy between LSST and Euclid could be to use deep learning techniques, including BDL, on multicolor image cubes, using both LSST 6-band images, and LSST+Euclid 10-band images, making again the most of the next major surveys for dark energy studies by developing sustainable and innovative analyses framework through joint advances in ML.

Finally, in preparation for next generation surveys such as LSST, state-of-the-art ML frameworks are being developed for dark energy studies using supernovae. In particular, identification of supernova types using ML, has become a strong research area. Recently, in order to incorporate model uncertainties in classification of supernovae, a framework has been developed using bayesian recurrent neural networks (Möller and de Boissière 2019). Bayesian

NNs show promising results providing robust classification probabilities which express epistemic uncertainties on supernova classification. Furthermore, within LSST-France, an initiative is being developed to process the LSST alert stream with a broker called Fink [2]. Fink will combine active learning and bayesian NNs to provide increasingly more accurate classifications of alerts.

4 Conclusion

A further development of machine learning techniques is crucial in the next decade to maximize the scientific return and the synergies of the large photometric surveys supported by the in2p3. The use of those techniques will allow first to address the huge volume of these survey but more importantly will bring a new approach allowing to combine more efficiently multi-band and multi-instrument pixel- and frame-level information in more efficient analyses. As the statistical treatment of data and the mastering of uncertainties is essential for cosmological studies, development of ML techniques like bayesian deep learning should be pursued.

5 References

Bernstein, G. M. and R. Armstrong (Feb. 2014). "Bayesian lensing shear measurement". In: MNRAS 438, pp. 1880–1893. arXiv: `1304.1843`.

Bernstein, G. M. et al. (July 2016). "An accurate and practical method for inference of weak gravitational lensing from galaxy images". In: MNRAS 459, pp. 4467–4484. arXiv: `1508.05655 [astro-ph.IM]`.

Euclid Collaboration et al. (Jan. 2019). "Euclid Preparation IV. Impact of undetected galaxies on weak lensing shear measurements". In: *arXiv e-prints*, arXiv:1902.00044, arXiv:1902.00044. arXiv: `1902.00044 [astro-ph.CO]`.

Gal, Yarin (2016). "Uncertainty in Deep Learning". PhD thesis. University of Cambridge, UK.

Gal, Yarin and Zoubin Ghahramani (2016). "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning". In: *Proceedings of the 33nd International Conference on Machine Learning, (ICML)*. Ed. by Maria-Florina Balcan and Kilian Q. Weinberger, pp. 1050–1059.

Gal, Yarin, Riashat Islam, and Zoubin Ghahramani (2017). "Deep Bayesian Active Learning with Image Data". In: *Proceedings of the 34th International Conference on Machine Learning, (ICML)*. Ed. by Doina Precup and Yee Whye Teh, pp. 1183–1192.

Ghahramani, Zoubin (2015). "Probabilistic machine learning and artificial intelligence". In: *Nature* 521.7553, pp. 452–459.

Gruen, D. et al. (Sept. 2010). "Bias-free Shear Estimation Using Artificial Neural Networks". In: ApJ 720, pp. 639–651. arXiv: `1002.0838 [astro-ph.CO]`.

[2] https://owncloud.lal.in2p3.fr/index.php/s/XdQnCWcvcjbQ6Vr#pdfviewer

Huertas-Company, M. et al. (May 2018). "Deep Learning Identifies High-z Galaxies in a Central Blue Nugget Phase in a Characteristic Mass Range". In: ApJ 858, 114, p. 114. arXiv: 1804.07307 [astro-ph.GA].

Ishida, E. et al. (Sept. 2019a). "Active Anomaly Detection for time-domain discoveries". In: *arXiv e-prints*, arXiv:1909.13260, arXiv:1909.13260. arXiv: 1909.13260 [astro-ph.IM].

Ishida, E. et al. (Feb. 2019b). "Optimizing spectroscopic follow-up strategies for supernova photometric classification with active learning". In: MNRAS 483.1, pp. 2–18. arXiv: 1804.03765 [astro-ph.IM].

Ivezić, Ž. et al. (May 2008). "LSST: from Science Drivers to Reference Design and Anticipated Data Products". In: *arXiv e-prints*. arXiv: 0805.2366.

Jones, D. M. and A. F. Heavens (Feb. 2019). "Bayesian photometric redshifts of blended sources". In: MNRAS 483, pp. 2487–2505. arXiv: 1808.02846.

Kilbinger, M. (July 2015). "Cosmology with cosmic shear observations: a review". In: *Reports on Progress in Physics* 78.8, 086901, p. 086901. arXiv: 1411.0115.

Laureijs, R. et al. (Oct. 2011). "Euclid Definition Study Report". In: *arXiv e-prints*. arXiv: 1110.3193 [astro-ph.CO].

LSST Science Collaboration et al. (Dec. 2009). "LSST Science Book, Version 2.0". In: *arXiv e-prints*. arXiv: 0912.0201 [astro-ph.IM].

Mandelbaum, Rachel (Sept. 2018). "Weak Lensing for Precision Cosmology". In: *Annual Review of Astronomy and Astrophysics* 56.1, pp. 393–433. ISSN: 1545-4282. arXiv: 1710.03235.

Möller, Anais and Thibault de Boissière (Jan. 2019). "SuperNNova: an open-source framework for Bayesian, Neural Network based supernova classification". In: *arXiv e-prints*, arXiv:1901.06384, arXiv:1901.06384. arXiv: 1901.06384 [astro-ph.IM].

Pasquet, J. et al. (Jan. 2019). "Photometric redshifts from SDSS images using a convolutional neural network". In: A&A 621, A26, A26. arXiv: 1806.06607 [astro-ph.IM].

Rhodes, J. et al. (Dec. 2017). "Scientific Synergy between LSST and Euclid". In: ApJS 233, 21, p. 21. arXiv: 1710.08489 [astro-ph.IM].

Spergel, D. et al. (Mar. 2015). "Wide-Field InfrarRed Survey Telescope-Astrophysics Focused Telescope Assets WFIRST-AFTA 2015 Report". In: *arXiv e-prints*. arXiv: 1503.03757 [astro-ph.IM].

Tewes, M. et al. (Jan. 2019). "Weak-lensing shear measurement with machine learning. Teaching artificial neural networks about feature noise". In: A&A 621, A36, A36. arXiv: 1807.02120.

Tuccillo, D. et al. (Mar. 2018). "Deep learning for galaxy surface brightness profile fitting". In: MNRAS 475, pp. 894–909. arXiv: 1711.03108 [astro-ph.GA].