Tackling the lack of annotated data in machine learning based computer vision

Suppose you want to teach a concept to someone: a possible way, instead of trying to provide a formal definition (which can be hard to provide or ineffective to communicate), is to give examples and counterexamples. Basically, this is the approach adopted by supervised machine learning; based on the complexity of the objects to be classified, the number of examples can be very high and, in general, the choice of the number of examples from the different classes that must be managed is not simple. So producing a good dataset for training supervised machine learning models is difficult and time consuming.

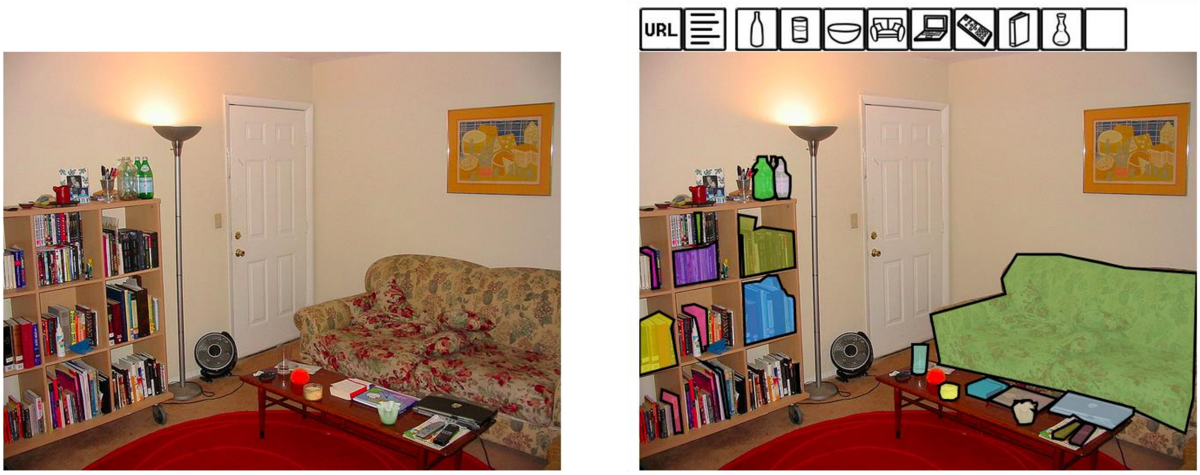

Fig.1: Example picture in Common Object in COntext [5] (COCO) with annotations

Trying to describe the amount of data provided by the astronomical imaging world, adjectives like huge, titanic, gargantuan are not exaggerations. Sadly, we cannot describe in the same way the amount of *annotated* images, which are the fuel of supervised machine learning. Manual labelling is, in the best case scenario, trivial but time consuming: imagine defining the borders of a ball or a window in an everyday picture (crowdsourcing approaches to annotate large amounts of data for training was instrumental in the production of several relevant datasets, and tools like Amazon's Mechanical Turk were instrumental in the success of this approach). But in the worst case scenario, we might not even be completely certain about how to define borders of entities in the frame, and how to classify the defined areas. This is the case in several situations of astronomical research.

Several known computer vision techniques might be of help. Some of them aim at increasing the amount of annotated data, starting from existing ones or working directly from scratch (augmentation or generation). Others aim to reuse models learned from different data and adapt them (or start from them, and do some fine tuning) on the new task. Next we'll briefly talk about some of these techniques that could be promising for astronomical objects detection tasks.

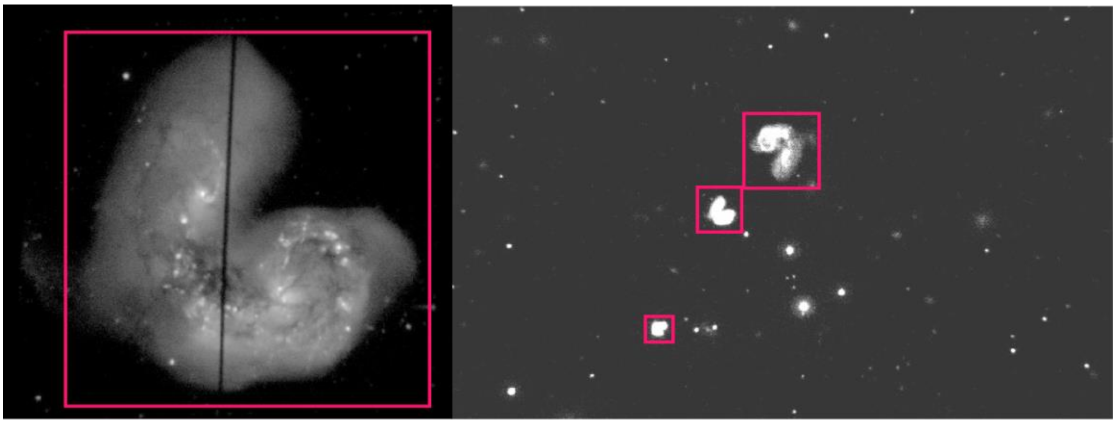

Fig.2: Image generated with Photon Simulator using a truth source of a galaxy resized and pasted in a synthetic star field (From the official docs)

First we'll talk about generation of synthetic images. In the complex real world, with objects like cars, dogs and people, a recent data driven approach to this form of simulation adopts GANs (Generative Adversarial Networks). The approach is data driven, but it is essentially unsupervised: although internally a supervised component is used, to train a GAN only examples with no labels are needed. GANs, however, are essentially large deep neural network architectures, with a huge number of parameters to be learned, and in some cases they can generate unwanted artifacts. So, an experimental evaluation of the approach is tricky.

A model driven alternative to the above approach might employ tools like Photon Simulator [2], derived by LSST project (Large Synoptic Survey Telescope). The Photon Simulator can be used to generate a synthetic dataset of astronomical images: it can be useful to generate images even before a real telescope is functional, provided that there is sufficient knowledge about the observed area of space, and it can pretrain models to be ready when real images come. Provided that the area of space used to generate the images is known, labeling the generated images can be time consuming, but there is no uncertainty about the labels to be assigned to the segmented areas. Photon Simulator can simulate different source types (stars or galaxies) or use truth images and "paste" it in the galactic field with different sizes. Moreover, it can simulate even a background that fits well for optical images.

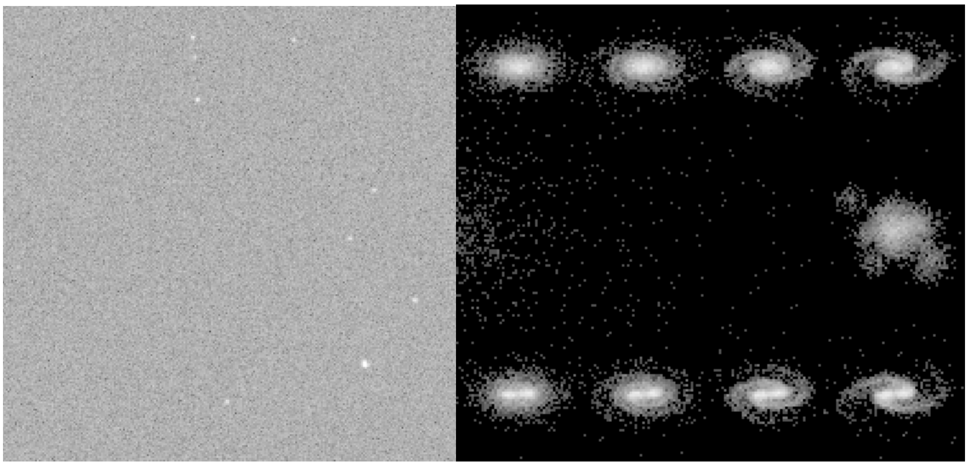

Fig.3: From left, simulated background and simulated galaxy

This approach works well with optical images and some evidence can be found in scientific papers like one of Burke, Colin J., et al. [1], but for radio astronomy the "style" of the image is slightly different, with wave-like background and with new artifacts. For this reason, it could be useful to take into account other options.

Viable and generally adopted techniques that do not require a simulator are data augmentation and transfer learning. The former represents a way to employ slightly altered versions of a single image, applying basically standard transformations. The latter, instead, adopts (and adapts) models trained in different contexts as a starting point, performing a more compact training phase compared to starting from scratch.

It's not trivial to decide which type of data augmentation technique to apply to the dataset, because some operations that distort the data beyond their nature could lead to a drop in the achieved performance. At the same time, data augmentation needs to be numerically significant (e.g. flipping or rotating a circle is ineffective and pointless). Other more complex techniques that also exploit unlabeled data have been proposed in the literature in the astronomical field [3] and they fall within the sphere of semi-supervised learning: we will consider them, and hopefully we will go deeper in the discussion and evaluation of these techniques in following posts.

About transfer learning, we want to point out that building methods that tackle lack of annotated data could benefit from taking an interdisciplinary approach, being open to potential contributions from other scientific researches. In fact, we could find similar tasks and challenges in underwater object detection [4]. Using the technique of transfer learning between similar domains could represent a promising way to increase the overall performance of machine learning approaches. Curiously, an underwater domain could have much more in common with the astronomical one than with images of everyday life, such as those in the COCO [5] dataset (it's no coincidence that octopuses are thought to come from space). Jokes aside, within WP4 we will further investigate the above techniques and we hope to achieve interesting results to be communicated, and not just through this blog but also as submissions to scientific venues, in the upcoming months.

Thomas Cecconello, Giuseppe Vizzari
Università degli studi di Milano Bicocca (UNIMIB)

[1] Burke, C. J., Aleo, P. D., Chen, Y. C., Liu, X., Peterson, J. R., Sembroski, G. H., & Lin, J. Y. Y. (2019). Deblending and classifying astronomical sources with Mask R-CNN deep learning. *Monthly Notices of the Royal Astronomical Society*, *490*(3), 3952-3965.

[2] Peterson, J. R., Jernigan, J. G., Kahn, S. M., Rasmussen, A. P., Peng, E., Ahmad, Z., ... & Grace, E. (2015). Simulation of astronomical images from optical survey telescopes using a comprehensive photon Monte Carlo approach. *The Astrophysical Journal Supplement Series*, *218*(1), 14.

[3] Ma, Z., Zhu, J., Zhu, Y., & Xu, H. (2019, July). Classification of radio galaxy images with semi-supervised learning. In *International Conference on Data Mining and Big Data* (pp. 191-200). Springer, Singapore.

[4] Zurowietz, M., & Nattkemper, T. W. (2020). Unsupervised Knowledge Transfer for Object Detection in Marine Environmental Monitoring and Exploration. *IEEE Access*.

[5] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740-755). Springer, Cham.