

Sujet de stage Ecole d'ingénieur 3ème année/Master 2

Etude de l'impact du self-supervised learning et des modèles de fondation sur la tâche « downstream » de prédiction du « redshift » avec un fort « covariate-shift » (i.e. décalage de « redshift »)

Mots-clefs : Apprentissage profond, Apprentissage auto-supervisé, Modèles de fondations, Images télescope multi-bandes, Astrophysique, Redshift, Covariate-Shift, Big-data

Encadrement :

Marc Chaumont (LIRMM, ICAR, CNRS/Université de Montpellier) - marc.chaumont@lirmm.fr,
Jérôme Pasquet (TETIS, Université Paul Valérie Montpellier) – jerome.pasquet@univ-montp3.fr,
Sébastien Lefèvre (IRISA, OBELIX, CNRS/Université de Bretagne Sud) - sebastien.lefevre@irisa.fr,
Frédéric Comby (LIRMM, ICAR, CNRS/Université de Montpellier) - frederic.comby@lirmm.fr
Matthieu Le-Lain (IRISA, OBELIX, CNRS/Université de Bretagne Sud) - matthieu.le-lain@univ-ubs.fr,

Contexte du stage :

Afin de déterminer les constantes astrophysiques et de mieux comprendre l'univers, les astrophysiciens ont notamment besoin de déterminer la distance aux galaxies (via le redshift).

Pour cela, de nombreuses campagnes d'observation ont été menées par le passé. Nous nous intéresserons plus particulièrement à l'étude COSMOS.

« L'étude COSMOS est l'une des plus anciennes études de galaxies. Pendant plus de deux décennies, nous avons recueilli des observations de rayons X, de lumière optique, infrarouge et radio pour obtenir les distances des galaxies et comprendre leur évolution au cours de 13,7 milliards d'années d'histoire cosmique. L'édition 2020 du catalogue principal des galaxies, utilise des techniques de pointe d'ajustement des modèles de galaxies de The Farmer pour mesurer les propriétés des galaxies. »

Extrait de <https://astroweaver.github.io/project/cosmos2020-galaxy-catalog/>
[Cosmos2020]

Dans le cadre de ce stage, le laboratoire d'astrophysique de Marseille, et en particulier Stéphane Arnouts et Marie Treyer (stephane.arnouts@lam.fr et marie.treyer@lam.fr) avec lesquels nous travaillons via le projet ANR DEEPDIP¹ [Bairouk 2023], nous fournirons des données de COSMOS2020. Ces données sont fournies et partiellement labellisées. Il y a plus de 40 millions de sources avec des images composés de 5 à 11 bandes.

Objectif du stage :

Un des objectifs du stage est de mesurer l'impact qu'il y a à entraîner de manière préliminaire sur de gros volume d'images (+ de 40 millions de sources) en auto-supervisé un gros modèle pour ensuite le spécialiser sur la tâche de prédiction du « redshift ». On comparera donc cette approche

¹ <https://anr.fr/Projet-ANR-19-CE31-0023>

d'apprentissage en deux étapes à l'approche directe (apprentissage de manière supervisé sans passer par un pré-apprentissage) [Vaswani 2017], à l'approche proposée par [Ait Ouahmed 2023] [Treyer 2024] [Ait Ouahmed 2024], et également à une approche utilisant un modèle de fondation pré-entraîné comme par exemple DINO [DINO v1][DINO v2] ou CLIP [CLIP].

Par ailleurs, il faudra gérer aux mieux le « covariate-shift » entre les données labellisées, qui ont un « redshift » globalement plus faible que celui des données non labellisées. En effet, les algorithmes supervisés classiques sont fortement *biaisés* par ce « décalage ».

D'un point de vue astrophysique, l'objectif est d'étudier si le biais avec les algorithmes fondation sont aussi important qu'avec les algorithmes classiques [Lin 2024].

A noter qu'il y aura peut-être des problèmes d'équilibrage de classes à gérer (si l'on découpe en intervalles l'intervalle de définition du redshift); problèmes également présent si l'on considère une régression plutôt qu'une classification.

Il faudra également utiliser le calculateur Jean Zay.

Pré requis :

Programmation, apprentissage profond, sciences des données, connaissances en traitement des images, anglais écrit scientifique, intérêt pour les applications en astrophysique.

Langages : Python et PyTorch, etc.

Conditions de stage :

Durée : 5 à 6 mois avec un début entre septembre et octobre 2024.

Indemnités : ~650 € / mois

Le stage se déroulera au sein de l'UMR TETIS (500 Rue Jean François Breton, 34000 Montpellier France) à Montpellier.

Références :

[Cosmos2020] "COSMOS2020: A Panchromatic View of the Universe to $z \sim 10$ from Two Complementary Catalogs", J. R. Weaver, O. B. Kauffmann, O. Ilbert, H. J. McCracken, A. Moneti, S. Toft, G. Brammer, M. Shuntov, I. Davidzon, B. C. Hsieh, Published 2022 January 10, 2022. The Author(s). Published by the American Astronomical Society. The Astrophysical Journal Supplement Series, Volume 258, Number 1 Citation J. R. Weaver et al 2022 ApJS 258 11 DOI 10.3847/1538-4365/ac3078 <https://iopscience.iop.org/article/10.3847/1538-4365/ac3078>

[DINO v1] "Emerging Properties in Self-Supervised Vision Transformers," Caron, Mathilde and Touvron, Hugo and Misra, Ishan and Jegou, Hervé and Mairal, Julien and Bojanowski, Piotr and Joulin, Armand, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9630-9640, doi: 10.1109/ICCV48922.2021.00951. https://openaccess.thecvf.com/content/ICCV2021/papers/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.pdf

[DINO v2] "DINO v2: Learning Robust Visual Features without Supervision", Maxime Oquab and Timothée Darcet and Théo Moutakanni and Huy V. Vo and Marc Szafraniec and Vasil Khalidov and Pierre Fernandez and Daniel HAZIZA and Francisco Massa and Alaaeldin El-Nouby and Mido Assran and Nicolas Ballas and Wojciech Galuba and Russell Howes and Po-Yao Huang and Shang-Wen Li and Ishan Misra and Michael Rabbat and Vasu Sharma and Gabriel Synnaeve and Hu Xu and Herve Jegou and Julien Mairal and Patrick Labatut and Armand Joulin and Piotr Bojanowski, Transactions on Machine Learning Research, issn=2835-8856, 2024. <https://openreview.net/forum?id=a68SUt6zFt>

[CLIP] “Learning Transferable Visual Models From Natural Language Supervision”, Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh 1 Gabriel Goh 1 Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, *Proceedings of the 38 th International Conference on Machine Learning*, PMLR 139, 2021.

<https://cdn.openai.com/papers/Learning Transferable Visual Models From Natural Language Supervision.pdf>

[Ait Ouahmed 2023] « Estimation du redshift des galaxies à partir d'images multibandes avec l'apprentissage profond : performances et obstacles », Mohammed Reda AIT OUAHMED, 18 Décembre 2023, Directeur de these M. Stéphane ARNOUTS Aix-Marseille Université – LAM, CoDirecteur de these M. Jérôme PASQUET Université Paul-Valéry Montpellier 3, Rapporteur M. Dino IENCO TETIS/LIRMM, INRAE, Rapporteur M. Marc HUERTAS-COMPANY Université Paris Diderot, Président Mme Laurence TRESSE Aix-Marseille Université – LAM, Examineur M. Pascal PONCELET Université Montpellier, Examineur Mme Sandra BRINGAY Université Paul-Valéry Montpellier 3, Examineur M. Alexandre BENOIT Polytech Annecy-Chambéry.

[Ait Ouahmed 2024] “Multimodality for improved CNN photometric redshifts”, R. Ait-Ouahmed and S. Arnouts and J. Pasquet and M. Treyer and E. Bertin, 2023, *A&A*, Volume 683, March 2024, A26, 17 pages, Extragalactic astronomy, <https://doi.org/10.1051/0004-6361/202347395>, <https://www.aanda.org/articles/aa/pdf/2024/03/aa47395-23.pdf>

[Treyer 2024] “Photometric redshifts in the SDSS at $r \leq 20$ ”, M Treyer, R Ait Ouahmed, J Pasquet, S Arnouts, E Bertin, D Fouchez, *CNN*, *Monthly Notices of the Royal Astronomical Society*, Volume 527, Issue 1, January 2024, Pages 651–671, <https://doi.org/10.1093/mnras/stad3171> <https://academic.oup.com/mnras/article/527/1/651/7320325?login=false>

[Lin 2024] “Photometric redshift estimation with convolutional neural networks and galaxy images: Case study of resolving biases in data-driven methods”, Q. Lin, D. Fouchez, J. Pasquet, M. Treyer, R. Ait Ouahmed, S. Arnouts and O. Ilbert, *A&A*, Volume 662, June 2022, A36, 28 pages, Numerical methods and codes, <https://doi.org/10.1051/0004-6361/202142751>, 14 June 2022. <https://www.aanda.org/articles/aa/pdf/2022/06/aa42751-21.pdf>

[Bairouk 2023] “ConvEntion: Astronomical Image Time Series Classification Using Convolutional attEntion”, Anass Bairouk, Marc Chaumont, Dominique Fouchez, Jerome Pasquet, Frédéric Comby, Julian Bautista. In section 15, “Numerical methods and codes” of *Astronomy & Astrophysics (Impact Factor = 6,240)*, March 2023. https://www.lirmm.fr/~chaumont/publications/ConvEntion_Bairouk_Chaumont_Fouchez_Pasquet_Comby_Bautista_2023.pdf

[Vaswani 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010. <https://dl.acm.org/doi/pdf/10.5555/3295222.3295349>