Comment spécialiser DINOv2 pour l'astronomie ?

Matthieu LE LAIN[1] Sébastien LEFÈVRE[1,2]

[1]IRISA, UMR CNRS 6074, Université Bretagne Sud, Vannes, France

[2]UiT – The Arctic University of Norway, Tromsø, Norway

Résumé – Cette étude évalue la performance des modèles de fondation visuels existants, basés sur ViT, SwinV2, BEiT ou DINOv2, pour des applications astronomiques, en particulier la classification d'images de galaxies à l'aide de l'ensemble de données Galaxy10 DECaLS. Nous utilisons une méthodologie de spécialisation (finetuning) pour évaluer la viabilité de l'adaptation des modèles pré-entraînés aux tâches astronomiques, dans le but de contourner le coûteux processus de ré-entraînement complet des modèles de fondation. La comparaison des modèles étudiés conduit à une analyse approfondie du modèle DINOv2, dont les résultats se révèlent prometteurs. Cette étude vise à évaluer les bénéfices d'une spécialisation de l'ensemble des paramètres du modèle par rapport à une spécialisation limitée à la tête de classification. L'objectif est d'optimiser les performances du modèle tout en identifiant ses éventuelles limites. Notre recherche contribue au domaine émergent des modèles de fondation astronomiques en identifiant les limites actuelles et en définissant les orientations futures de la recherche en IA pour l'astronomie.

Abstract – This study evaluates the performance of existing visual foundation models, relying on ViT, SwinV2, BEiT or DINOv2, for astronomical applications, in particular the classification of galaxy images using the Galaxy10 DECaLS dataset. We use a finetuning methodology to assess the viability of adapting pre-trained models to astronomical tasks, with the aim of bypassing the costly process of fully re-training foundation models. Experimental comparison of the models leads to an in-depth analysis of the DINOv2 model, with promising results. The aim of this study is to assess the benefits of fine-tuning all the model parameters versus finetuning only the classification head. The aim is to optimize the model performance while identifying its potential limitations. Our research contributes to the emerging field of astronomical foundation models by identifying current limitations and defining future directions of AI research for astronomy.

1 Introduction

La dernière décennie a été marquée par des avancées majeures dans le domaine de l'intelligence artificielle, avec des résultats décisifs dans de nombreux domaines d'application, comme la vision par ordinateur ou le traitement du langage naturel. Le domaine de l'astronomie a également bénéficié de ces développements, les réseaux neuronaux profonds, et en particulier les CNN, étant conçus pour diverses tâches telles que l'identification des exoplanètes par la classification des courbes de lumière à l'aide de données réelles et synthétiques [1], ou l'identification de galaxies à faible luminosité de surface parmi les artefacts présents dans les images astronomiques du Dark Energy Survey [7], entre autres exemples.

Un nouveau changement de paradigme se produit actuellement dans l'IA, avec l'avènement des modèles de fondation. Ces modèles, qui se caractérisent par leur entraînement sur des quantités extrêmement importantes de données non étiquetées, sont capables d'apprendre des caractéristiques plus génériques, ce qui leur permet de s'adapter à une variété de tâches avales. Ces nouveaux modèles ont déjà un impact majeur dans le traitement du langage naturel, avec par exemple le succès de BERT et GPT-3 puis ChatGPT, mais aussi (dans une moindre mesure cependant) dans le domaine de la vision par ordinateur et en particulier de la génération d'images, avec Dall-E. Il est prévu que ces modèles de fondation impactent aussi fortement l'analyse des données astronomiques [6].

En effet, les ensembles de données existants générés par les télescopes et les projets en cours nécessitent des outils pour une analyse toujours plus efficace, étant donné la complexité et le volume toujours croissant des données impliquées dans la recherche astronomique moderne. Ces besoins semblent correspondre aux capacités des modèles de fondation. Des travaux récents ont démontré les avantages potentiels de la mise en œuvre de telles architectures dans le domaine de l'astronomie. Par exemple, dans le domaine du traitement du langage naturel, un modèle génératif appelé AstroLLaMA a été développé [4]. Il est basé sur l'architecture du modèle LLaMA2 et a été entraîné sur plus de 300 000 résumés d'articles en astronomie provenant de la plateforme arXiv. Son objectif est de générer des complétions de texte et d'extraire des représentations vectorielles plus pertinentes et scientifiquement exploitables.

2 Méthode

Malgré les tentatives récentes d'affiner les modèles de fondation dans le contexte astronomique, la compréhension et le développement de ces modèles en sont encore à leurs débuts et se concentrent principalement sur les applications textuelles utilisant des grands modèles de langage (LLM).

2.1 Modèles

Notre travail se concentre sur les images astronomiques et vise à évaluer la performance des principaux modèles de fondation visuels existants, basés sur ViT, SwinV2, BEiT ou DINOv2.

Nous adoptons une méthodologie commune basée sur la spécialisation (*finetuning*) pour évaluer la pertinence de l'adaptation des modèles pré-entraînés aux tâches astronomiques. Pour chaque architecture, nous avons choisi les tailles des modèles existants *nano*, *small*, *base*, and *large*. Nous avons

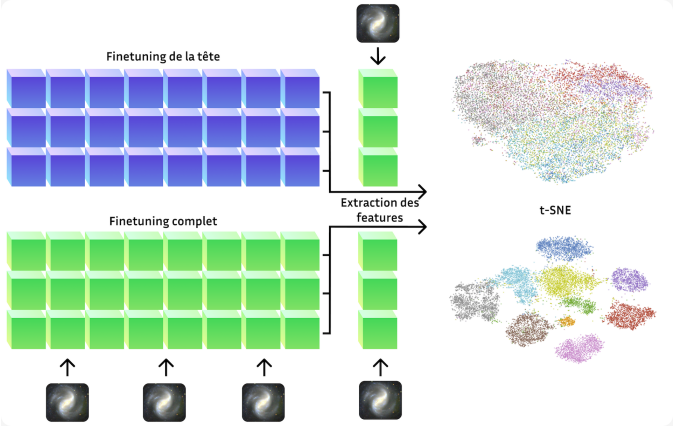

FIGURE 1 : Diagramme de la méthodologie de spécialisation supervisée utilisée dans notre étude : modèles pré-entraînés en haut ou après spécialisation complète en bas, têtes de classification en vert, puis extraction des caractéristiques générées par chaque modèle pour la représentation par t-SNE.

sélectionné des modèles pré-entraînés avec ImageNet, avec et sans tête de classification, à partir de ces architectures. Le modèle DINOv2 est pour sa part pré-entrainé avec LVD-142M et sa tête de classification spécialisée avec ImageNet-1k.

Dans un second temps, nous avons évalué la capacité de ces modèles a extraire les caractéristiques des données astronomiques à partir de représentations t-SNE, comme présenté en Figure 1, puis nous avons analysé l'impact de l'encodeur et de la complexité des données dans ces extractions.

2.2 Données

Nous considérons une tâche de classification de galaxies en utilisant le jeu de données publiques Galaxy10 DECaLS[3], qui a été préparé à partir de GalaxyZoo DECaLS [8]. Ce jeu de données contient 17 736 images de galaxies colorées (bande g, r et z) avec des dimensions de 256x256 pixels. Toutes les images sont classées en 10 classes mutuellement exclusives et étiquetées sur la base des votes de volontaires qui ont été sélectionnés et filtrés sur la base de scores élevés.

Par ailleurs, pour approfondir l'analyse de nos résultats (voir section 3.5), nous avons également utilisé le jeu de données GalaxyMNIST[8] qui contient 10 000 images de galaxies colorées de 224x224 pixels réparties en 4 classes distinctes.

Pour assurer la facilité d'utilisation et la reproductibilité, nous avons préparé deux versions disponibles sur la plateforme HuggingFace [1], basées sur ces deux jeux de données avec un découpage classique 80/20 entre entraînement et validation.

3 Résultats

Dans cette section, nous présentons les résultats obtenus avec les différentes stratégies de spécialisation. Au vu de la distribution non-équilibrée des classes (cf. matrice de confusion en Figure 4), nous évaluons les modèles avec le F1-Score. Nous présentons le F1-Score des meilleurs modèles en fonction de leur complexité, puis nous explorons plus précisément les raisons ne permettant pas d'obtenir de meilleures performances.

[1] https://huggingface.co/matthieulel

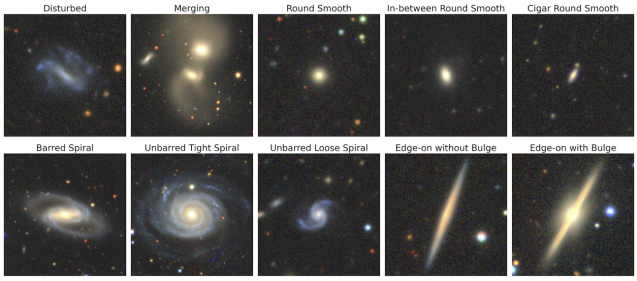

FIGURE 2 : Exemples des classes du jeu de données Galaxy10 DECaLS préparé à partir des données GalaxyZoo [3].

TABLE 1 : Meilleurs scores F1 de chaque architecture visuelle pré-entrainée sur ImageNet (LVD-142M pour DINOv2) et spécialisée sur Galaxy10 DECaLS.

Modèles	F1		Δ F1	Paramètres
(Sélection des meilleurs scores)	Complet	Tête	(C-T)	(millions)
vit-base-patch16-384	85.59	56.73	28.86	86
beit-base-patch16-224-pt22k	**85.87**	23.13	62.74	85
swinv2-base-patch4-win12to16 [2]	85.65	56.60	29.05	86
dinov2-base-imagenet1k-1-layer	85.40	**60.81**	**24.59**	86

3.1 Précision et complexité des modèles

La table 1 montre les résultats des architectures ayant obtenu les meilleurs scores F1 lors de la spécialisation de 27 variantes, issues de 4 modèles de fondation visuels. Nous pouvons observer que les modèles étudiés, tous basés sur l'architecture *Transformer* ne dépassent pas une précision de 85%, la tâche semblant trop complexe pour ces modèles génériques.

Dans le cadre de notre analyse comparative, nous introduisons une nouvelle métrique (Δ F1) pour évaluer l'apport d'une spécialisation de tous les paramètres au lieu d'une spécialisation limitée à la tête de classification. Il apparaît que le modèle DINOv2 [5] ne démontre pas une supériorité significative en termes de performance par rapport aux autres modèles évalués, bien que les écarts observés soient minimes. Il convient toutefois de souligner que DINOv2 présente des résultats optimaux lors de la spécialisation de la seule couche de classification, montrant le Δ F1 le plus bas. Par ailleurs, une étude complémentaire [2] met en lumière l'efficacité de la spécialisation de la couche de classification de DINOv2 sur le jeu de données GalaxyMNIST, ce qui suggère une capacité notable d'extraction des caractéristiques pertinentes des données astronomiques. Ces observations soulignent le potentiel de DINOv2 pour l'analyse d'images astronomiques.

3.2 Capacité d'extraction des caractéristiques

Afin d'évaluer les capacités d'extraction des caractéristiques des modèles de fondation visuels analysés, nous avons eu recours à l'algorithme t-SNE (t-distributed Stochastic Neighbor Embedding) pour réduire la dimensionnalité des vecteurs de caractéristiques et permettre leur visualisation en deux dimensions. La Figure 3 illustre la représentation des caractéristiques pour les modèles de fondation génériques pré-entraînés sur ImageNet (et LVD-142M pour DINOv2), en comparaison avec ces mêmes modèles après une spécialisation de l'ensemble de leurs paramètres sur le jeu de données Galaxy10 DECaLS.

[2] swinv2-base-patch4-win12to16-192to256-22kto1k-ft

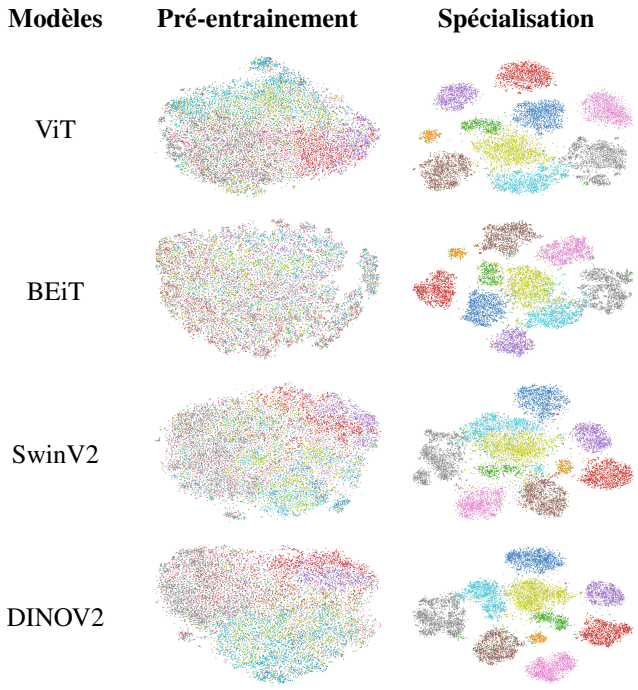

FIGURE 3 : Comparaison des représentations t-SNE après pré-entrainement ImageNet (et LVD-142M pour DINOv2) et spécialisation sur Galaxy10 DECaLS.

L'analyse des résultats met en évidence l'importance cruciale de la spécialisation pour optimiser l'extraction des caractéristiques spécifiques aux données astronomiques par les différents modèles. Cette étape s'avère indispensable pour permettre aux architectures d'adapter leurs représentations internes aux particularités du domaine étudié (ici les galaxies).

3.3 Erreurs de classification

Les résultats exposés précédemment ont conduit à approfondir notre étude en orientant spécifiquement l'analyse vers le modèle DINOv2. La matrice de confusion des résultats obtenus par le modèle DINOv2, en Figure 4, révèle que les erreurs sont principalement concentrées entre des classes similaires, en particulier entre les classes *Unbarred Tight Spiral* et *Unbarred Loose Spiral*. En outre, la classe *Disturb* présente une tendance notable à être mal classée par rapport à toutes les autres classes. La Figure 5 illustre la distribution dispersée des caractéristiques des images appartenant à la catégorie *Disturb galaxies*. Cette dispersion met en évidence la complexité rencontrée par le modèle dans l'extraction et l'identification des traits distinctifs propres à cette classe spécifique.

La complexité de cette classe, caractérisée par sa morphologie atypique (illustrée en Figure 2) plutôt qu'une apparence plus conventionnelle, semble poser un défi important aux capacités de généralisation des modèles. Ces observations soulignent la nécessité d'améliorer la précision de la caractérisation et d'intégrer les connaissances astronomiques préalables.

3.4 Performance de la tête de classification

Pour évaluer l'efficacité de la tête de classification du modèle DINOv2, qui inclut une seule couche linéaire, nous l'avons comparée à une tête de classification simple, un perceptron

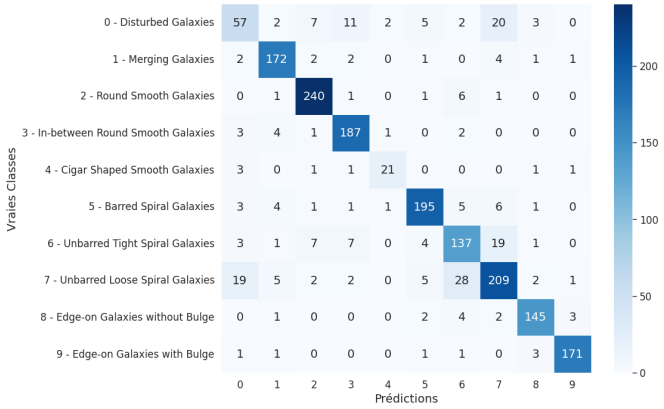

FIGURE 4 : Matrice de confusion du modèle DINOv2 après spécialisation sur les données Galaxy10 DECaLS.

FIGURE 5 : Distribution des caractéristiques de la classe *Disturb* (en vert) extraites par DINOv2, modèle pré-entrainé générique à gauche, et spécialisé sur Galaxy10 DECaLS à droite.

multicouche (MLP) dont nous avons fait varié la complexité (1, 2 et 3 couches), connecté à la sortie du modèle de fondation spécialisé avec le jeu de données Galaxy10 DECaLS.

Les résultats présentés dans la Table 2 montrent des performances équivalentes avec une tête DINOv2 plus simple par rapports aux différents MLP, suggérant une très bonne capacité d'extraction des caractéristiques par le modèle de fondation. On remarque par ailleurs que l'augmentation de la complexité du MLP n'apporte pas d'amélioration significative. Cette approche confirme l'importance d'une optimisation globale des paramètres des modèles pour améliorer leurs performances.

3.5 Complexité des données

L'étude récente de Lastufka *et al.* [2] mentionnée dans la section 3.1 évalue l'application des modèles de fondation visuels aux images astronomiques, en se concentrant particulièrement sur les galaxies dans les images optiques et radio. Elle explore le potentiel d'utilisation de modèles prêts à l'emploi avec l'incorporation d'un classificateur linéaire à une seule couche, comparable à notre étude, avec le jeu de données GalaxyMNIST. Afin d'évaluer le niveau de complexité du jeu de données Galaxy10 DECaLS par rapport à GalaxyMNIST, nous avons procédé à une spécialisation de tous les paramètres du même modèle DINOv2 sur ce jeu de données plus restreint et moins complexe. Les résultats sont présentés dans la Table 3.

De plus, afin d'évaluer l'importance d'utiliser un modèle ayant déjà des connaissances astronomiques, nous avons comparé les représentations t-SNE produite par un modèle DINOv2 générique face au même modèle après spécialisation

[3]Notons que l'étude de Lastufka *et al.* [2] rapporte un score de 88% que nous n'avons pas réussi à reproduire.

TABLE 2 : Comparaison des scores F1 pour DINOv2 après spécialisation complète sur Galaxy10 DECaLS, avec une tête MLP (1 à 3 couches) ou la tête originale DINOv2.

Classes	1 C.	2 C.	3 C.	DINOv2
Disturbed	0.56	0.54	0.55	**0.57**
Merging	0.91	**0.93**	0.92	0.91
Round Smooth	0.93	**0.94**	0.93	**0.94**
In-between Round Smooth	0.90	0.89	0.89	**0.91**
Cigar Shaped Smooth	0.78	**0.80**	0.71	0.79
Barred Spiral	0.88	**0.90**	0.89	**0.90**
Unbarred Tight Spiral	**0.76**	0.73	**0.76**	0.75
Unbarred Loose Spiral	**0.79**	**0.79**	**0.79**	0.78
Edge-on without Bulge	**0.93**	0.92	0.92	0.92
Edge-on with Bulge	**0.96**	**0.96**	0.95	**0.96**
Paramètres	394K	525K	558K	7.7K

TABLE 3 : Comparaison du score F1 des modèles DINOv2 génériques ou préalablement spécialisés sur Galaxy10 DECaLS, après spécialisation sur GalaxyMNIST de la tête uniquement ou du modèle complet.

Modèles	Tête (%)	Complet (%)
DINOv2 générique	83.67[3]	92.79
DINOv2 GZ10DECaLS	90.85	93.79

complète sur Galaxy10 DECaLS. Nous comparons dans la Figure 6 les représentations obtenues avec celles produites par un modèle ayant été spécialisé sur les données cibles GalaxyMNIST. Les résultats montrent une nette amélioration de la performance sur le jeu de données GalaxyMNIST à partir d'une spécialisation complète d'un modèle générique, ou d'un modèle déjà spécialisé complet avec des données astronomiques. Ces résultats confirment une nouvelle fois l'importance d'une optimisation globale des paramètres des modèles pour améliorer leurs performances, d'une bonne capacité du modèle DINOv2 à extraire les caractéristiques visuelles des données astronomiques et de l'intérêt d'insufler des connaissances astronomiques aux modèles en vue de tâches avales.

4 Discussion

Nos résultats démontrent que, dans le contexte d'un modèle DINOv2 générique pré-entraîné, il est plus avantageux d'affiner tous les paramètres plutôt que de se focaliser uniquement sur la tête. Il convient toutefois de noter que la présente étude est majoritairement limitée à deux jeux de données créés à partir des mêmes données intiales [8], mais présentant des complexités et des classes différentes. Des recherches supplémentaires sont donc nécessaires pour remédier aux limites identifiées, que ce soit par l'utilisation d'autres ensembles de données, la prise en compte de contraintes physiques ou l'exploration d'approches différentes, telles que les modèles hybrides.

5 Conclusion

Notre étude de l'application des modèles de fondation visuels en astronomie révèle des résultats prometteurs mais limités. L'approche de spécialisation (*finetuning*) sur le modèle DINOv2 démontre le potentiel de l'utilisation de modèles préentraînés pour les tâches astronomiques, en particulier pour la classification d'images de galaxies. Cette approche méthodologique, combinant la spécialisation des modèles et l'utilisation de t-SNE pour la visualisation des caractéristiques extraites, offre un aperçu précieux de l'adaptabilité des modèles de fondation visuels à des tâches astronomiques spécifiques, soulignant l'importance de la spécialisation complète pour obtenir des performances optimales dans des domaines spécialisés.

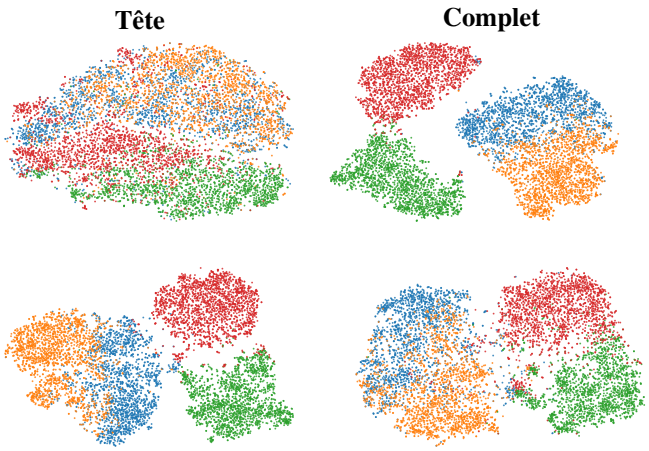

FIGURE 6 : Distribution des caractéristiques extraites par DINOv2 sur GalaxyMNIST à partir du modèle DINOv2 générique (en haut) ou spécialisé sur Galaxy10 DECaLS (en bas), avant ou après spécialisation finale sur GalaxyMNIST.

Références

[1] S. CUÉLLAR *et al.* : Deep learning exoplanets detection by combining real and synthetic data. *Plos one*, 17(5): e0268199, 2022.

[2] E. LASTUFKA *et al.* : Bridging the Gap : Examining Vision Foundation Models for Optical and Radio Astronomy Applications. *arXiv :2409.11175*, septembre 2024.

[3] H.W. LEUNG et J. BOVY : Deep learning of multi-element abundances from high-resolution spectroscopic data. *Monthly Notices of the Royal Astron. Soc.*, 2018.

[4] T.D. NGUYEN *et al.* : Astrollama : Towards specialized foundation models in astronomy. *arXiv preprint arXiv :2309.06126*, 2023.

[5] Maxime OQUAB *et al.* : DINOv2 : Learning Robust Visual Features without Supervision. *arXiv e-prints*, page arXiv :2304.07193, avril 2023.

[6] M.J. SMITH et J.E. GEACH : Astronomia ex machina : a history, primer and outlook on neural networks in astronomy. *Royal Society Open Science*, 10(5):221454, 2023.

[7] D. TANOGLIDIS *et al.* : Deepshadows : Separating low surface brightness galaxies from artifacts using deep learning. *Astronomy and Computing*, 35:100469, 2021.

[8] M. WALMSLEY *et al.* : Galaxy zoo decals : Detailed visual morphology measurements from volunteers and deep learning for 314 000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 509(3):3966–3988, 2021.