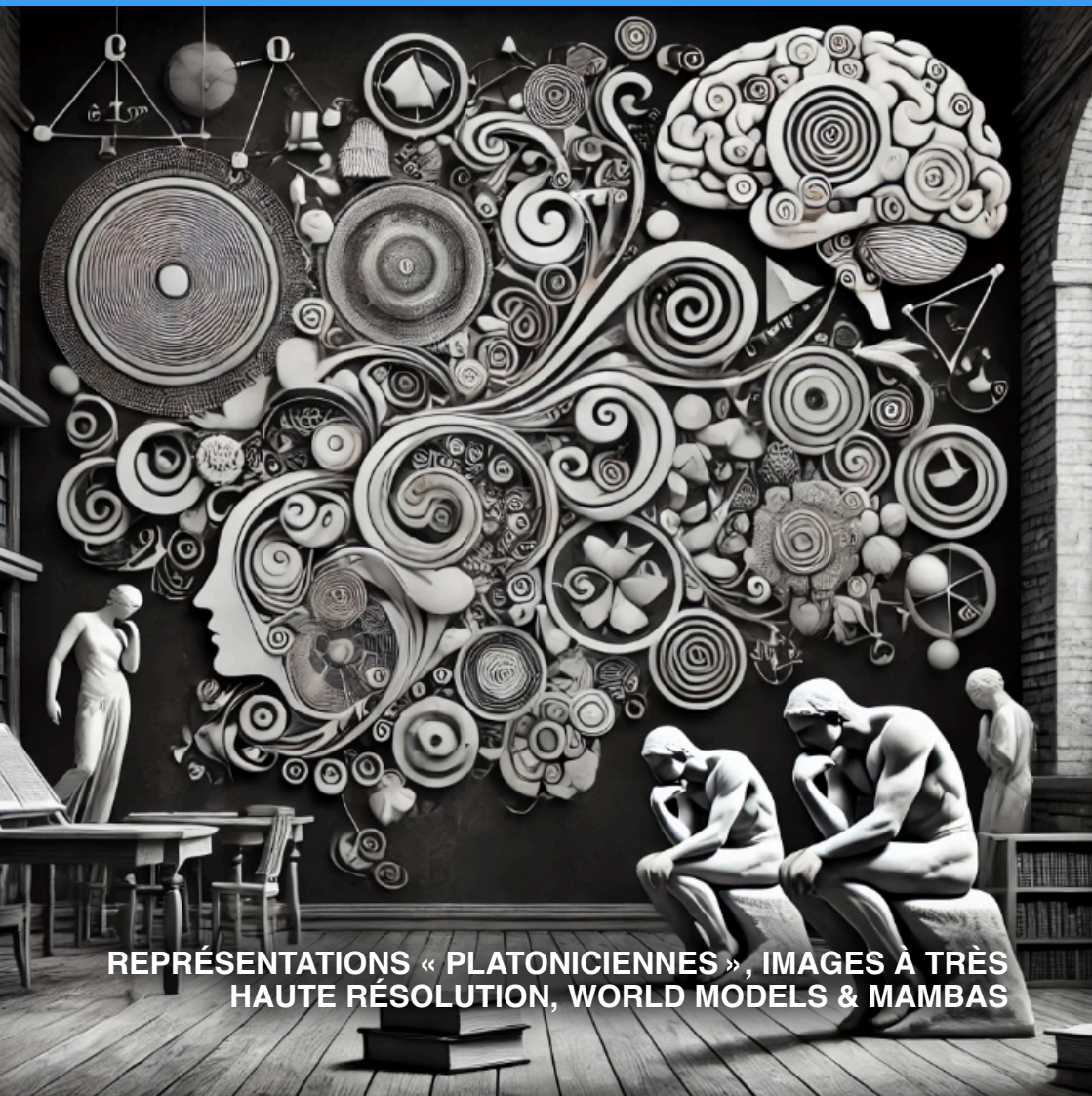
Datalchemy

ECHOS

DE LA RECHERCHE #16

JUILLET 2024



REPRÉSENTATIONS « PLATONICIENNES », IMAGES À TRÈS HAUTE RÉSOLUTION, WORLD MODELS & MAMBAS

TL;DR ?

Cinq mot-clés de ces échos

#Embeddings, #haute résolution, #diffusion, #état d'un système, #mamba

Pourquoi lire cette publi peut vous être utile concrètement ?

Mieux comprendre ce qu'apprennent les réseaux de neurones est fondamental pour appréhender notre domaine de travail, et nous avons ici une publication pertinente (quoique un peu ambitieuse) montrant que ces représentations se ressemblent à travers les architectures et les modalités. Au-delà, un nouveau travail permet de traiter des images à très haute résolution pour une consommation mémoire contrôlée, ouvrant de nouvelles applications.
Enfin, deux travaux permettent de respectivement mieux comprendre les modèles de diffusion, et la complexité des architectures, avec de nouveaux résultats sur le désormais célèbre *Mamba*

Quels process métier seront probablement modifiés sur la base de ces recherches ?

Les *Foundation Models* visent à apprendre des représentations génériques à travers une ou plusieurs modalités (texte, image, son, etc.). Que ces représentations se rapprochent naturellement permet de mieux envisager leur utilisation. Autre sujet, l'utilisation d'IA sur des images de très haute résolution est aujourd'hui bloquée, empêchant d'interpréter une image complexe correctement (images satellites par exemple). Enfin, les modèles de diffusion apparaissent de plus en plus comme un outil efficace pour modéliser un environnement et sa dynamique, permettant aux *world models* de continuer leur positionnement comme outil efficace et intéressant pour entraîner des agents autonomes.

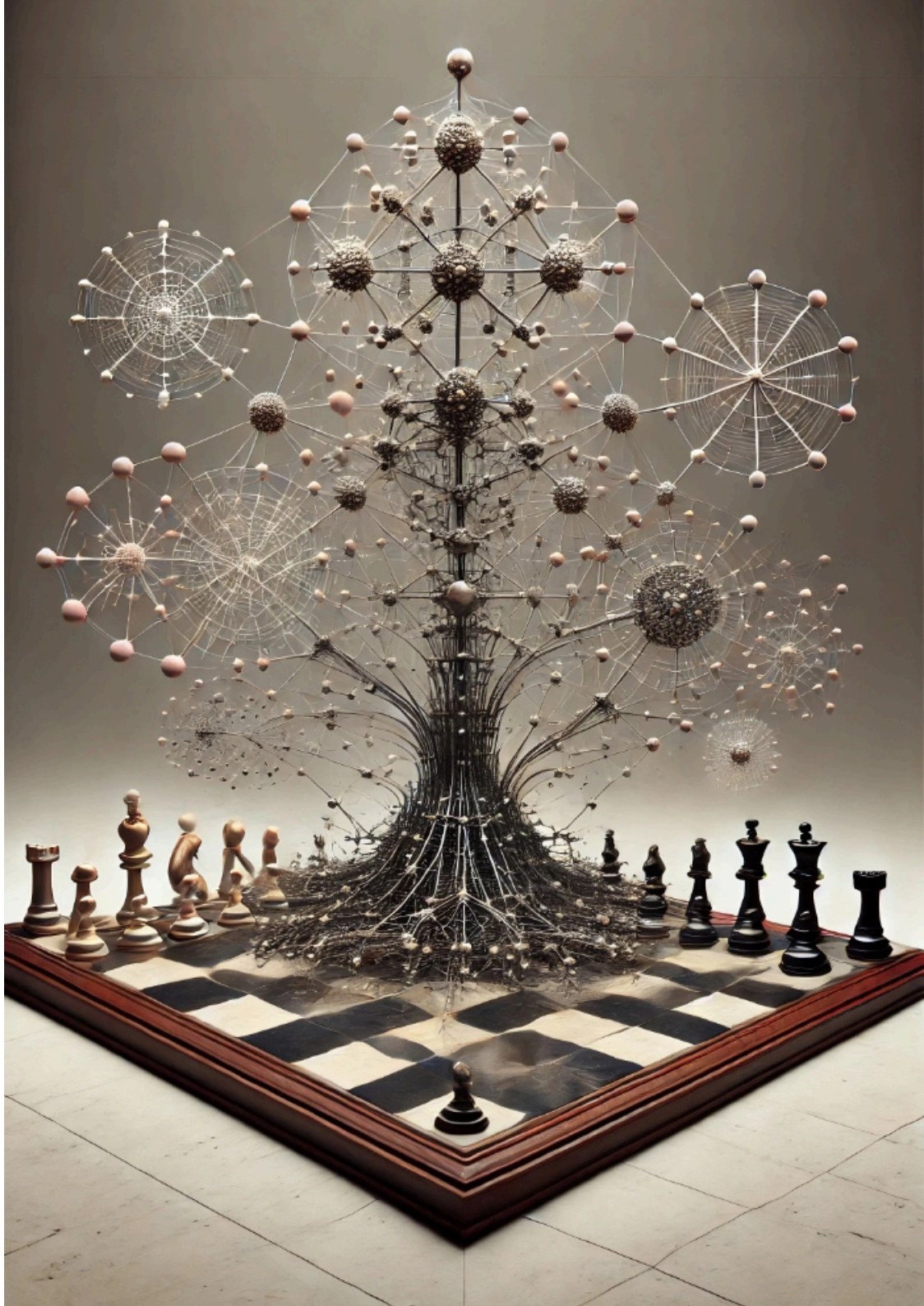
Si vous n’avez qu’une minute à consacrer à la lecture maintenant, voici le contenu essentiel en 4 points

1. Des chercheurs observent que les représentations intermédiaires apprises par différents réseaux de neurones sont beaucoup plus proches les unes des autres que ce que l’on pourrait supposer, laissant entrevoir une « convergence » de ces représentations vers une modélisation générale. Cette convergence s’observe à travers les modèles, mais aussi à travers différentes modalités : texte, image, etc. On parle d’alignement des modèles, et cet alignement est d’autant plus important que les modèles sont complexes.

2. And now something completely different : la gestion d’images de très haute résolution est encore à date une épine dans le pied des chercheurs qui nous impose de découper ou de dégrader ces images pour pouvoir travailler dessus. Hors, ces techniques détruisent potentiellement l’information de contexte global de l’image, indispensable dans certains scénarios. Ici, les auteurs proposent une approche générique intéressante qui est stable en occupation mémoire et peut gérer de très grandes images. Cette approche est aussi l’occasion de tester l’architecture Mamba sur ces problèmes.

3. Autre sujet : les world models sont une technique visant à apprendre à modéliser un environnement pour ensuite entraîner un agent autonome face à cet environnement « virtuel ». Un nouveau travail exploite les modèles de diffusion pour obtenir un nouvel état de l’art. Ce travail permet d’éclaircir notre compréhension des modèles de diffusion, notamment pour vérifier la pertinence de l’approche conseillée par NVIDIA, mais aussi pour comparer leur capacité à d’autres modèles comme les VA-VAE.

4. Enfin, un travail étudie la complexité des architectures Transformer et Mamba pour observer des limites théoriques et fondamentales dans le suivi de l’état d’un système simple (par exemple, un jeu d’échec). Il permet notamment de mieux caractériser la capacité fondamentale d’une architecture.



REPRÉSENTATION « PLATONICIENNE » : LES RÉSEAUX DE NEURONES APPRENNENT-ILS UNE MÊME REPRÉSENTATION DE LA RÉALITÉ

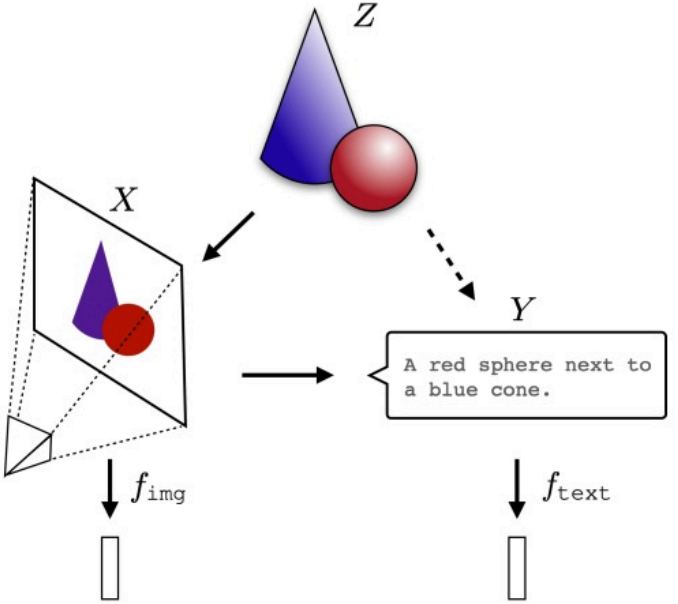
Attention, la publication *The Platonic Representation Hypothesis, Huh et al* est à manipuler avec des pincettes. Si elle pose des questions fondamentales et lève des observations passionnantes sur ce qu'apprennent nos chers réseaux de neurones, il convient de séparer les affirmations scientifiques des spéculations plus philosophiques. Mais n'allons pas trop vite, et replaçons le contexte : de quoi parlons-nous ?

Cela fait bientôt dix ans que s'est imposé un courant de recherche fondamental dans notre domaine du *Deep Learning*, celui du *Representation Learning*. À la base de ces travaux, une observation : quand un réseau de neurones apprend à adresser une tâche (par exemple de classification), il apprend implicitement à représenter la donnée d'entrée sous des formes de plus en plus simples (i.e. ayant une dimensionnalité beaucoup plus faible). On considère aujourd'hui que cet apprentissage est probablement la vraie « magie » du *Deep Learning* : apprendre à représenter une donnée extrêmement complexe sous une forme simplifiée plus facile à manipuler. C'est notamment ce qu'avaient observé *Milokov et al* qui avaient généré les premiers *embeddings* représentant des mots. Ce courant s'est ensuite poursuivi via l'entraînement de *Foundation Models* comme notre précieux *DinoV2*, capable de représenter une image par un vecteur très expressif et suffisant pour adresser un grand nombre de tâches spécifiques. Nous avions aussi fait un webinaire sur le phénomène des *embeddings* cross-modalité, où le même concept exposé sous deux formes (par exemple, image et texte), sera isolé comme un unique vecteur de représentation.

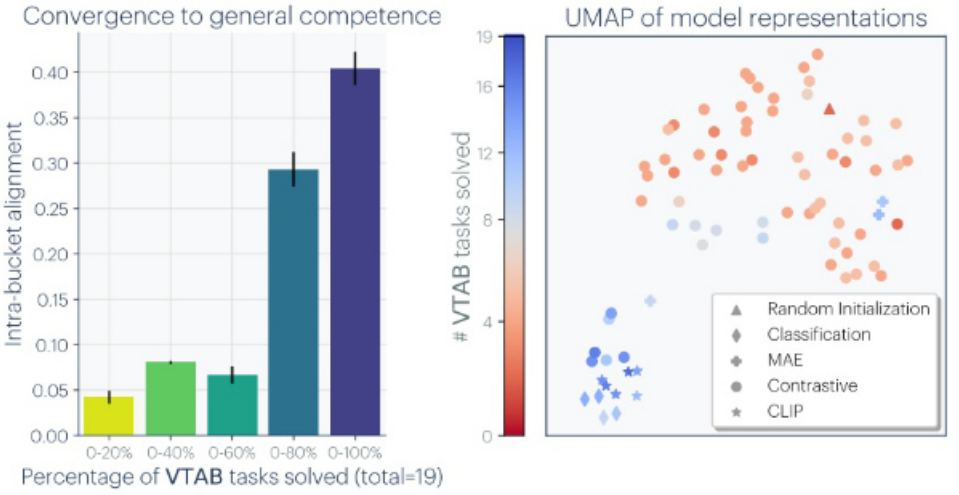
Néanmoins, si chaque réseau de neurones apprend une représentation de la donnée en entrée, une question fondamentale est de savoir à quel point deux représentations issues de deux réseaux de neurones différents seront proches. Dit autrement : chaque réseau apprend-il une représentation unique et spécifique pendant son apprentissage, où existe-t-il une « destination » de représentation vers laquelle se dirigeraient chaque réseaux soumis à apprentissage. Pour reprendre la vision des auteurs et leur hypothèse principale : **existe-t-il une représentation unique vers laquelle tendraient naturellement les représentations de réseaux de neurones ?**

The Platonic Representation Hypothesis

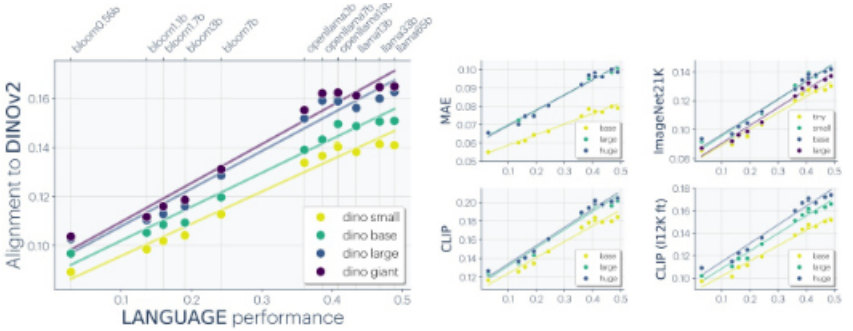
Neural networks, trained with different objectives on different data and modalities, are converging to a shared statistical model of reality in their representation spaces.



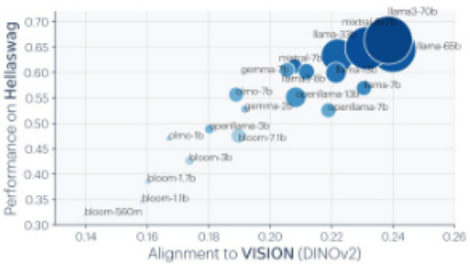
Ce travail est une occasion salutaire de répertorier ce que la communauté scientifique a découvert à ce sujet. Plusieurs observations ont ainsi été faîtes au cours des dernières années. Typiquement, via la méthode du *model stitching*, on sait déjà qu'il est possible, à partir de deux réseaux entraînés sur un même problème, d'extraire *n* première couches de l'un et *p* dernières couches de l'autre pour les accoler via une simple transformation linéaire. Cela implique déjà que deux réseaux de ce type apprennent des représentations extrêmement proches (à une transformation linéaire près). Ce type de transfert entre deux réseaux a été poursuivi jusqu'à arriver à des méthodes « zero shot » (sans ré-apprentissage spécifique) et plus particulièrement, entre différentes architectures et même différents problèmes à adresser. Cette polyvalence est déjà intéressante, surtout dans notre domaine où les déficits théoriques ne cessent de limiter notre compréhension. **Les auteurs ont donc étudié à quel point deux modèles différents, entraînés sur des problèmes différents, ont des représentations internes proches.** Pour ce faire, la logique a été de comparer, entre les deux représentations, les plus proches voisins d'un même élément et d'observer si ces groupes de voisins sont proches ou différents. Sans être absolue, cette approche a le mérite de minimiser les problèmes de métriques qui, face à des vecteurs assez complexes, peuvent vite perdre leur sens. Un premier résultat intéressant permet d'exposer, à travers 78 modèles différents de classification d'images, et donc à travers des architectures très différentes, à quel point ces modèles sont « alignés » dans leurs représentations. À gauche, on observe que plus ces modèles sont performants sur le *VTAB* (*Visual Task Adaptation Benchmark*), plus leurs représentations sont proches. À droite, une réduction de dimension (*UMAP*) projette ces modèles en deux dimensions :

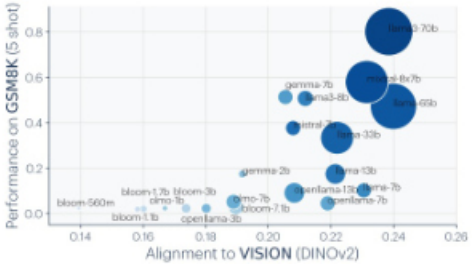


Cette idée que l'alignement inter-modèle s'améliore avec la performance est un point très intéressant. Au-delà, les auteurs affirment que ces représentations convergent à travers différentes modalités de la donnée. Nous savions déjà qu'il était possible de « coller » un modèle de vision et un modèle de langage au prix d'une transformation linéaire. Ici, les auteurs observent un alignement plus global entre modèles de vision et modèles de langage, avec un alignement d'autant plus important que le modèle de langage est performant. Les auteurs observent aussi que le *CLIP* d'*OpenAI*, entraîné aussi sur le contenu textuel, présente un alignement plus important qui se dégrade dès lors qu'on opère un *fine tuning* vers *ImageNet*.



Dernière expérience pertinente : les auteurs comparent la corrélation entre alignement d'un modèle de langage avec le modèle de vision DinoV2 et performance sur des tâches spécifiques de langage. Hellaswag présente une corrélation relativement linéaire, quand GSM8k lui présente (visuellement, attention !) une forme d'émergence.



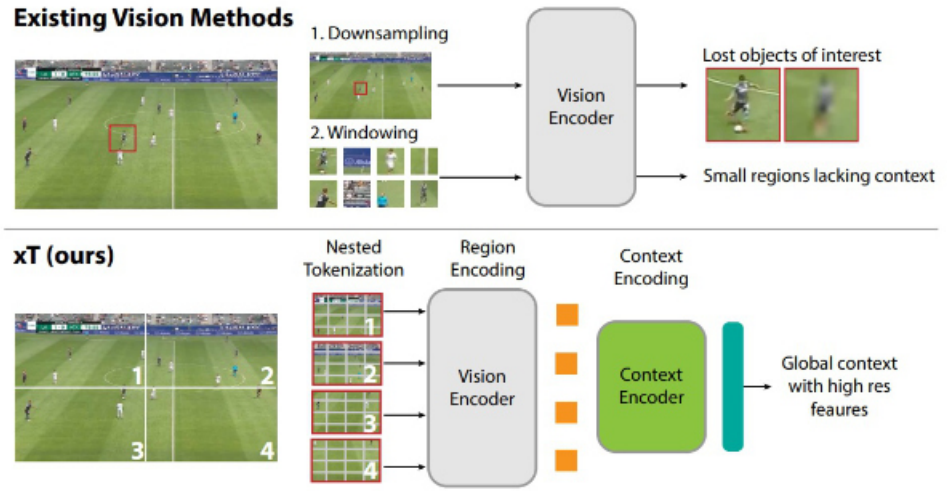


Ces travaux présentent un intérêt remarquable car la question des représentations apprises par un réseau de neurones est une question fondamentale liée à l'interprétabilité et notre compréhension de ces modèles. Ici, observer que ces représentations « se rapprochent » en fonction de la complexité d'un modèle a quelque chose de très rassurant. Face au chaos intellectuel que représente le Deep Learning, l'idée que nos modèles apprennent plus ou moins bien une représentation « universelle » de l'information, ce à travers le texte ou l'image, est une très bonne nouvelle. Néanmoins, il convient aussi de se méfier d'une précipitation trop grande. L'hypothèse platonicienne des auteurs est, précisément, une hypothèse. Et si les auteurs parlent de convergence des représentations (un terme qui chatouille le vocabulaire mathématique), nous observons juste une corrélation. Qui plus est, ces vecteurs de représentation ont beau être une version simplifiée de la donnée d'entrée, ils restent très complexes (des vecteurs de dimension 500, 1000…) et chaque méthode de comparaison (ici les plus proches voisins) présente ses qualités et défauts sans s'imposer définitivement face aux (nombreuses) autres méthodes. Ceci dit, cette direction de recherche est particulièrement importante, surtout à l'heure des Foundation Models, donc un sujet à surveiller de très près.

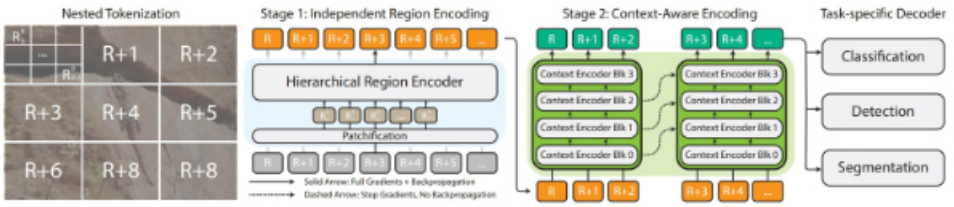
LES IMAGES À TRÈS HAUTES RÉSOLUTIONS : UN TALON D'ACHILLE DU DEEP LEARNING

Ce point est connu de tous les praticiens ayant développé des architectures Deep Learning en Computer Vision : l'immense majorité des modèles existants sont, fondamentalement, incapables de gérer une image à haute résolution, par exemple une image de qualité 4K (environ 8 millions de pixels) ne sera pas naturellement traité en un bloc par un réseau de neurones classique. Cette image sera soit redimensionnée, soit découpée, de manière à pouvoir rentrer dans les résolutions d'entrée des modèles.

C'est le sujet qu'adresse xT: Nested Tokenization for Larger Context in Large Images de Gupta et al. L'enjeu n'est pas anodin. Pouvoir traiter une image de grande résolution permet de qualifier correctement un sujet en fonction de l'intégralité du contexte. Par exemple, j'ai peu de chance de comprendre le comportement d'un footballeur si je ne voie pas l'intégralité des autres joueurs sur une image suffisamment précise pour observer leurs poses. De même, si on travaille en image satellite, détecter un élément de petite taille peut être déjà complexe, mais le contexte de cet élément joue un rôle important. C'est cette problématique qu'illustre le schéma ci-après :



Le problème fondamental est, évidemment, un problème de ressources disponibles. La complexité inhérente aux traitements internes d'un réseau de neurones (quelle que soit l'architecture) est telle que la majorité de ces outils prennent en entrée des résolutions autour de 500 pixels et occupent déjà un espace mémoire conséquent. L'enjeu est donc un enjeu d'architecture, et nous allons voir qu'ici, l'approche pose des questions intéressantes. Représentée dans le schéma ci-dessous, l'approche xT va déjà découper l'image en patchs et en sous-patchs (Nested Tokenization). Chaque patch va être encodé en utilisant un réseau de neurones pour représenter le patch (nous parlions de représentations précédemment 😊). Cet encodage sera indépendant pour chaque patch, et permettra de simplifier la représentation de l'image. Afin que le modèle puisse utiliser chaque patch et donc l'intégralité du contexte, la troisième partie va consister en un Context-Aware Encoding qui recevra l'encodage de chaque patch pour, enfin, donner une prédiction finale en fonction du problème adressé via un decoder ad-hoc.



Les résultats sont intéressants à double titre. Déjà, les auteurs s'attaquent au dataset iNaturalist xView3-SAR, composé d'images satellite à très haute résolution (29400x24400 pixels), où les éléments sont fortement dépendants d'un contexte global de l'image. Mais au-delà, les auteurs testent le dataset iNaturalist ainsi que différentes types de Context-Aware Encoding. Un des types testés étant ce bon vieux Mamba. Si vous ne voyez pas de quoi nous parlons, n'hésitez pas à aller consulter la revue de recherche dédiée à cette architecture qui fait parler d'elle depuis plusieurs mois, en proposant un mécanisme de sélection efficace et plus économique en mémoire. Cette architecture est ici un choix pertinent, et sans se distinguer totalement, elle présente souvent des scores finaux compétitifs ou supérieurs aux approches Transformer.

Un autre argument porte sur l'efficacité de cette méthode. Les auteurs mesurent en fonction de la rapidité d'exécution (ici en nombre de régions traitées par seconde) la précision finale obtenue. On observe ici une efficacité beaucoup plus intéressante à qualité égale de cette approche (à gauche). Deuxième schéma à droite, l'approche occupe une ressource mémoire constante malgré la résolution d'entrée, là où les modèles classiques bloquaient et ne pouvaient plus répondre.

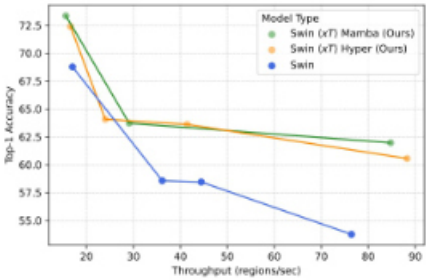


Figure 5. ***xT* offers greatly increased accuracy per throughput.** On iNaturalist classification, we find that our models only slightly diminished throughput (with the exception of Swin-T (xT) XL/Mamba) but achieved greater accuracies at each throughput threshold.

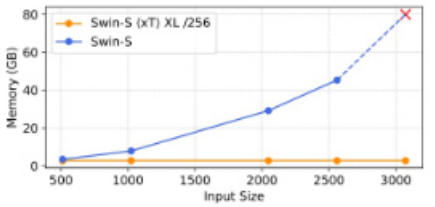


Figure 6. **Swin rapidly goes out of memory (indicated by the red X) as images grow in size whereas *xT* retains near-constant memory cost.** *xT* can scale to much larger images than the naïve usage of a vision backbone.

Ces approches restent très importantes puisque le problème soulevé (impossibilité de gérer une image de trop grande résolution) est un challenge constant pour appliquer le Deep Learning à la conception d'outils efficaces. À date, ces sujets sont adressés en découpant l'image et en appliquant ensuite des heuristiques spécifiques, et si ces approches risquent de s'imposer encore un temps, la capacité d'un modèle à adresser de hautes résolution est un sujet majeur pour aller vers des outils plus efficaces, notamment lors de leur apprentissage.

WORLD MODELS & DIFFUSION : UN RAPIDE UPDATE

Les World Models sont un sujet de recherche créé en 2018 par les célèbres Ha et Schmidhuber qui a révolutionné les approches par renforcement, soit, celles où nous voulons entraîner un agent autonome à maximiser une récompense arbitraire dans un environnement. Ces approches sont notamment très utilisées en robotique, mais plus généralement pour résoudre des problèmes d'optimisation. À l'époque, les World Models avaient été une immense bouffée d'air frais. Ce domaine cherchait auparavant à entraîner un réseau de neurones unique dans des conditions horribles (toute personne ayant entraîné ces modèles, que l'on parle de Q Learning ou de Policy Gradient, en a tiré un fort syndrome post traumatique face à des courbes de reward illisibles et une facture GPU, elle, parfaitement lisible mais déprimante). Un problème venait de cette volonté d'entraîner un unique réseau. Les World Models ont proposé une approche beaucoup plus structurée, basée sur trois étapes successives :

- L'entraînement d'un modèle chargé de compresser et de cartographier les observations de l'agent dans son environnement (à l'époque, un bon vieux Variational Autoencoder)

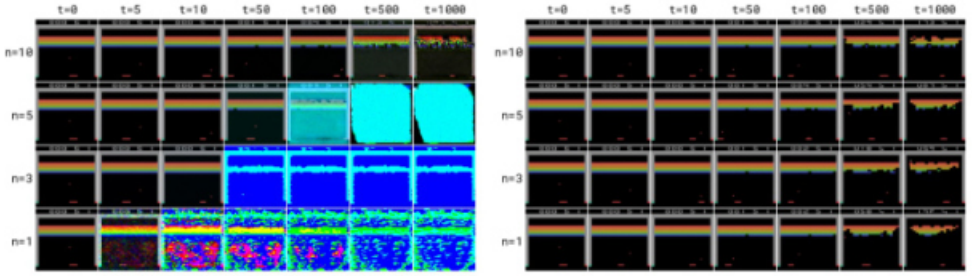
- L'entraînement d'un modèle chargé de modéliser la dynamique de l'environnement (que se passe-t-il si on exécute telle action dans tel état), ce modèle travaillant uniquement sur les vecteurs issus du premier modèle.

- Enfin, la mise au point d'un contrôleur qui utilisera les deux premiers modèles pour résoudre le problème global.

Cette division permet de poser les bonnes questions d'une manière itérative et donc de mieux répartir les sujets. Et les auteurs, à l'époque, avaient réussi l'exploit d'utiliser les deux premiers modèles comme un environnement virtuel, d'entraîner un agent uniquement dans cet environnement virtuel, pour enfin l'utiliser dans l'environnement réel.

Ces World Models ont donc vocation à « simuler le monde », tout du moins l'environnement de travail. Ils sont ainsi très proches du travail de « simulateur universel » présenté le mois dernier dans la revue. Nous suivons cette approche de très près, et avons observé avec beaucoup d'intérêt le récent Diffusion for World Modeling: Visual Details Matter in Atari d'Alonso et al. Poursuivant les travaux de ce domaine, cette approche a questionné l'architecture utilisée pour « modéliser » l'environnement, et nous retrouvons aujourd'hui nos bons vieux modèles de diffusion ! Ceux-là même qui ont révolutionné l'IA générative en image et donc nous parlions encore le mois dernier. Car ces modèles, s'ils sont effectivement très puissants, sont encore très mal compris, et tout retour d'expérience est d'un grand intérêt.

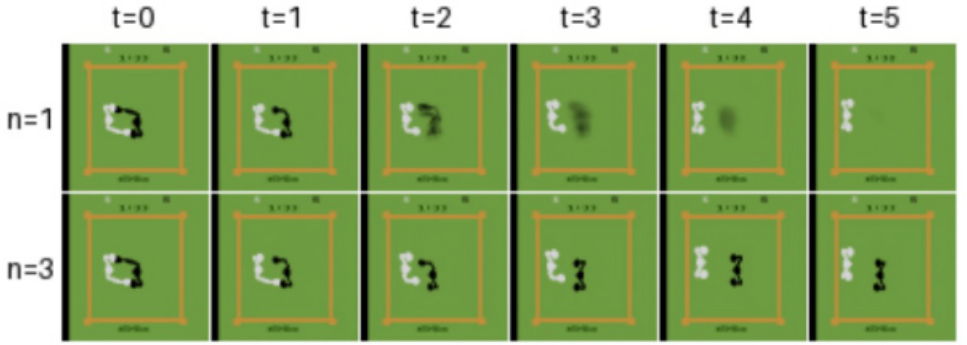
Un point notable ici est que les auteurs s'interrogent sur la méthode la plus efficace pour entraîner un modèle de diffusion. Ils observent notamment que l'approche EDM, popularisée par les équipes de NVIDIA, est beaucoup plus efficace pour simuler avec précision l'évolution visuelle d'un jeu Atari. La valeur n ci-dessous correspond au nombre d'étapes de denoising du modèle de diffusion. On retrouve une évolution de la qualité en fonction du nombre d'étapes.



(a) DDPM-based world model trajectories.

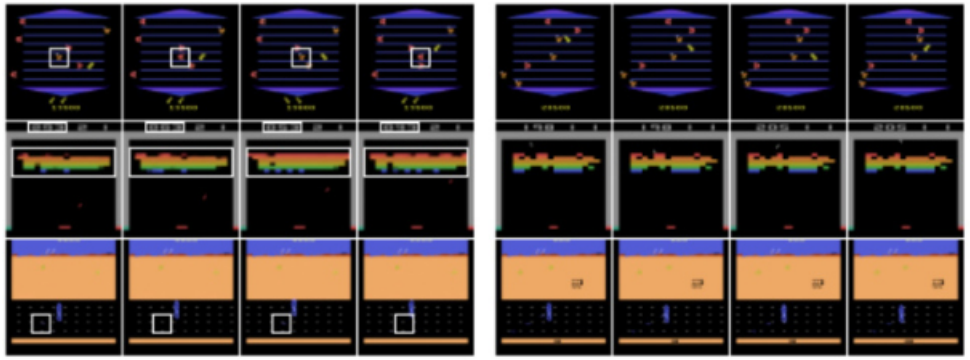
(b) EDM-based world model trajectories.

Ici, le modèle de diffusion joue un rôle de « prédiction de l'avenir » de l'environnement. Cette publication est ainsi remarquable pour évaluer la précision de ces prédictions, et notamment l'importance du nombre d'étapes sur la qualité de la génération (ci-dessous)



Autre point d'intérêt, cet article est l'occasion de comparer les modèles de diffusion à une autre architecture très utilisée pour ces problèmes de modélisation visuelle : les VQ-VAE qui, depuis la version hiérarchique de 2019, ont repris le flambeau en modélisation d'images (le Stable Diffusion de Rombach et al se base d'ailleurs sur un VQ-VAE pour transformer l'image en vecteur qui, ensuite, était soumis au processus de diffusion). Ci-dessous, à gauche un world model basé sur le VQ-VAE et à droite la version en modèle de diffusion. La version VQ-VAE (ci-dessous : IRIS) présente des incohérences (dans les carrés blancs) non présentes dans la version diffusion :

- Ligne du haut : jeu Asterix : un ennemi (orange) devient une récompense (rouge), puis alterne entre ces deux états.
- Ligne du milieu : jeu Breakout : inconsistance entre le score estimé par le modèle et les briques détruites.
- Ligne du bas : jeu Road Runner : les récompenses (petits points bleus) apparaissent et disparaissent



(a) IRIS (b) DIAMOND

Cela montre que pour modéliser une dynamique interne, le modèle de diffusion semble beaucoup plus précis que le VQ-VAE, ce qui pousse encore plus l'intérêt de ces modèles pour généraliser des mécaniques ou des comportements. Rappelons que dans une récente revue de la recherche consacrée à l'Imitation Learning en robotique, les modèles de diffusion (que nous avons pu tester et valider) étaient d'une manière surprenante excellents pour modéliser le contrôle d'un bras robotique …

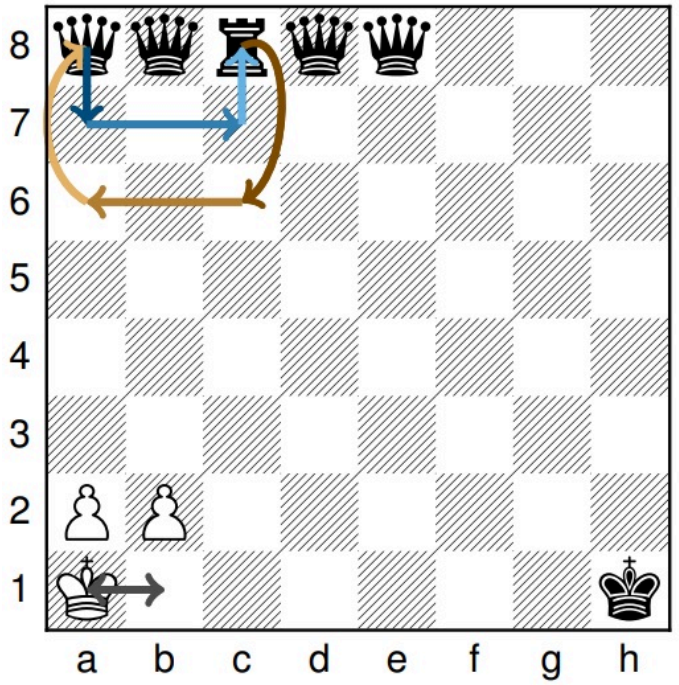
REPARLONS DU MAMBA !

Il y a quelques mois, nous vous proposions une revue de la recherche dédiée à une nouvelle architecture qui avait fait beaucoup de bruit dans la communauté scientifique : le Mamba, issu des structured space models. Cette architecture disposait d'un mécanisme de sélection singulièrement intéressant, potentiellement robuste à des séquences très longues, constituant un talon d'Achille gigantesque des Transformers qui soutiennent nos chers modèles de langage. Déjà à l'époque, si nous attirions l'attention sur ces travaux, nous recommandions en toute urgence d'attendre la suite pour observer comment la communauté scientifique reproduirait et critiquerait ces travaux.

Nous avons vu ci-dessous une approche exploitant le Mamba pour gérer des images à très haute résolution. Le travail que nous mettons ici en lumière pose la question de ces architectures à calculer l'état d'un système à travers une séquence longue, avec un titre assez provocateur : The Illusion of State in State-Space Models, de Merrill et al.

Une illusion ? Le Mamba part-il déjà à la poubelle ? Non, non, non 😊. Cette publication soulève des questions très intéressantes qui vont au-delà de cette simple architecture, sans invalider l'utilisation de ce modèle.

La question centrale est ici de savoir si l'architecture Mamba sont plus efficaces que les Transformers pour le suivi de l'état d'un système. Avant de rentrer dans le détail, proposons quelques exemples. Un problème de suivi d'état (state tracking si vous avez des sunglasses) est un problème où le modèle reçoit une liste de modifications de l'état d'un système, et doit à la fin fournir l'état final. Typiquement, imaginons que j'ai 5 balles posées de 1 à 5, et que j'ai ensuite une série d'instructions du type : intervertir la balle 3 et la balle 2, intervertir 1 et 4, intervertir 2 et 5, etc. L'enjeu est, à la fin de la séquence, de pouvoir donner la position finale de chaque balle. Les auteurs travaillent ici à un problème similaire de modification de l'état d'un échiquier, où le modèle reçoit une série de déplacement de pièces et doit à la fin donner la position de chaque pièce :

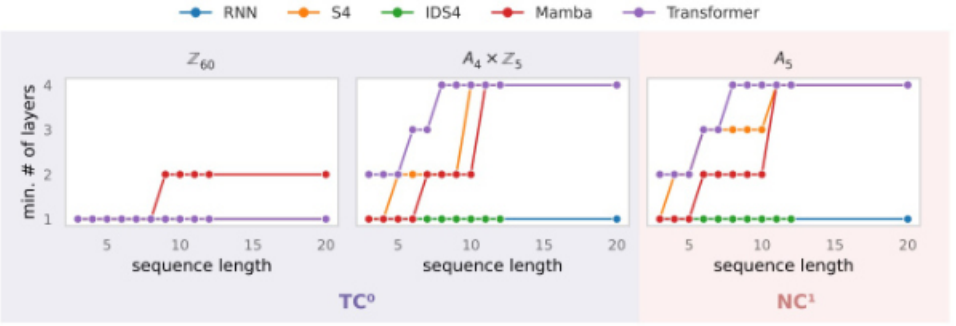


Il a déjà été démontré que les Transformers sont incapables de résoudre d'une manière générale ce type de problème (retenez le « générale », nous en reparlerons). Dans le cas présent, les auteurs utilisent une théorie mathématique assez récente et passionnante, la Circuit-Complexity, visant à exprimer, pour une architecture donnée, la complexité des problèmes qui peuvent (ou non) être résolus. Les Transformers étaient déjà condamnés à une classe de problèmes assez simple, la classe TC-0 (nous évitons ici de rentrer dans les détails, mais réfléchissons à une revue de recherche dédiée à ce sujet). Le Mamba était un candidat idéal pour espérer adresser des sujets plus complexes. Hélas (ou tant mieux, après tout, en recherche, tout résultat est bon à prendre), il est ici démontré que les structured space models et le Mamba appartiennent à la même classe de complexité. Ils sont donc incapables d'adresser un sujet de state tracking.

Mais alors, demanderait le data-scientist un peu désespéré de ce chaos intellectuel, faut-il mettre le Mamba à la poubelle ? Absolument pas. C'est dans ce contexte que nous pouvons juger le titre de la publication comme étant un peu « provocateur ».

En effet, l'impossibilité est ici démontrée théoriquement pour une classe de problème, ce qui en soit est très intéressant. Cela ne veut pas dire que ces modèles sont incapables de gérer ces sujets, seulement que pour une complexité donnée (un nombre de couche du modèle), au-delà d'une certaine longueur de séquence ou complexité, le modèle commencera à être en erreur. Il est donc possible d'utiliser ces outils pour des problèmes plus restreints. Au-delà, rappelons que même si une architecture est considérée comme pouvant adresser théoriquement ces problèmes, nous n'aurons aucune garantie de réussite totale, car sommes en Deep Learning, domaine de l'empirisme où un modèle peut, à tout moment, échouer (halluciner). Et les auteurs observent ici, empiriquement, qu'à nombre de couches égale, le Mamba est plus efficace que le Transformer classique pour ces problèmes de state tracking.

Ci-dessous des schémas donnant en fonction de la longueur de séquence le nombre de couches minimales à avoir pour obtenir plus de 90% de précision. Les deux premiers schémas portent sur des problèmes simples (TC-0), le troisième sur un problème plus complexe.



Notons enfin que les auteurs proposent une modification du Mamba (ci-dessus : IDS4) qui devrait améliorer la qualité de l'architecture. Considérant qu'une nouvelle version du Mamba est sortie le mois dernier, ces travaux sont plutôt des signaux positifs pour continuer de surveiller les progrès de cette architecture et préparer son arrivée dans notre boîte à outils 😊

Datalchemy

contact@datalchemy.net

Images extraites des articles respectifs ou générées par IA