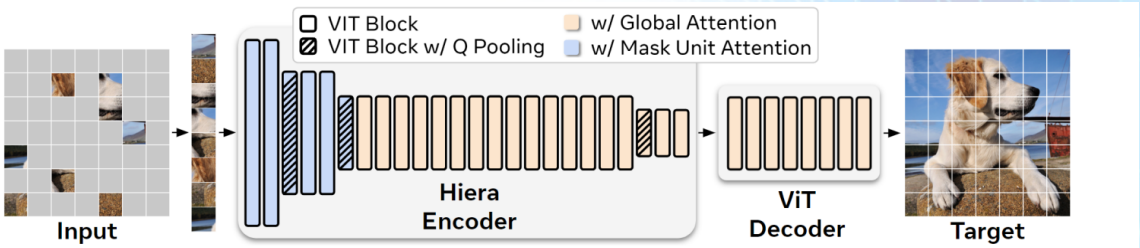
Un nouvel opus de Transformers, cinéma ou réalité ?

Ne cachons pas notre joie, voilà une revue de recherche qui sera assez éloignée du fracas des modèles de langage et autres GPT-like (qui ont fait l'objet d'un webinaire dédié en juillet), pour nous intéresser à trois travaux passionnants et fondamentaux sortis ces deux dernières semaines. Au programme : une nouvelle architecture très simple et diablement efficace pour la vision, un modèle génératif convainquant pour la génération de musique, et un travail scientifique prometteur pour l'interprétabilité du Deep Learning et la génération de représentations de la donnée. En route !

Un Vision Transformer de référence
[https://arxiv.org/abs/2306.00989]

Que se passe-t-il ?
Quoi qu'on pense de Meta en tant qu'entreprise, force est de reconnaître que le laboratoire *Meta AI* reste un des acteurs les plus intéressants du domaine. Un de leurs avantages reste de chercher des architectures efficaces, à contre-courant des approches académiques classiques.
Ici, avec *Hiera*, Meta AI nous propose une nouvelle architecture de base autour du *Vision Transformer*, plus simple et plus efficace. Si le *ViT* est en effet une architecture qui s'impose doucement depuis trois ans dans le monde académique, sa grande jeunesse a conduit à de nombreux travaux qui tentaient d'améliorer ses résultats via différentes astuces plus ou moins contrôlées.
Comme l'observent les auteurs, les évolutions récentes du *ViT* ont certes amélioré ses résultats, mais au prix d'une plus grande lourdeur. Hors, dès que l'on s'extrait du monde

académique pour s'intéresser à l'ingénierie, la rapidité d'entraînement d'un modèle est un facteur clé permettant d'itérer plus rapidement sur un problème et donc d'avoir des outils mieux adaptés.
Ici, *Hiera* est donc un modèle beaucoup plus simple où les auteurs ont retiré un certain nombre d'évolutions proposées ces deux ou trois dernières années, au profit d'un pré-entraînement beaucoup plus efficaces basé sur le *MAE* (Masked AutoEncoder) où le modèle apprend à appréhender le biais spatial dans les images, biais qui n'est pas présent structurellement dans l'architecture du *Transformer*. Notons que cette absence de biais spatial dans le *Transformer* est un vieux sujet sur lequel de nombreuses approches ont été tentées, mais au prix d'une lourdeur supplémentaire. Les auteurs ici argumentent sur le fait que le *MAE* est un pretraining qui introduit ce biais spatial, mais pour un sur-coût d'entraînement minimal.



Le modèle est qui plus est un modèle hiérarchique avec donc un traitement de l'image en entrée à différents niveaux de résolution. Notons que l'approche *MAE (*consistant à masquer une partie de l'image et à entraîner le modèle à retrouver cette partie masquée*)* est fondamentalement mal adaptée pour un modèle hiérarchique, les auteurs proposent ici une adaptation spécifique.

Pourquoi c'est intéressant

Le *Deep Learning* est un domaine de recherche, et cela pose souvent problème en application. En effet, de nombreuses améliorations sont proposées, souvent validées selon une approche spécifique, mais sans recul sur la combinaison et la pertinence réelle de ces améliorations. Qui plus est, ces améliorations sont souvent des sources de complexité et de lenteur supplémentaires, qui freinent l'utilisation de ces modèles en application directe.
Ici, une nouvelle architecture plus simple et plus contrôlée est une excellente nouvelle pour nous, car nous avons ici un point de départ plus stable et plus rapide pour approcher de nouveaux sujets.
Au delà, rappelons que ces modèles peuvent jouer un rôle de "*backbone*" dans des modèles plus spécialisés (notamment en segmentation, détection, etc.) Il y a donc fort à parier que ce modèle pourra faire directement évoluer de nombreux outils plus spécialisés.

Rappelons néanmoins que *MetaAI* proposait il y a un an un nouveau modèle convolutif de référence (*A Convnet for the 2020s [https://arxiv.org/abs/2201.03545]*) contre l'évolution des *Vision Transfomers*. Cette nouvelle étude semble acter que les convolutifs rentrent peu à peu dans l'histoire 🙂

IA Générative et musique, une nouvelle étape

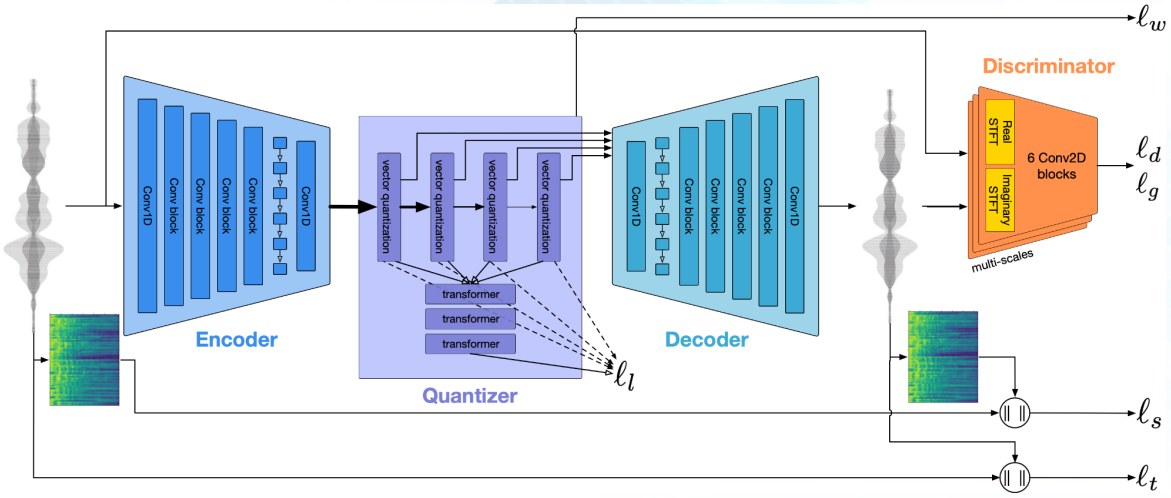
[https://arxiv.org/abs/2306.05284]
L'IA générative est un sujet très actif depuis les explosions de *Stable Diffusion* et des *Large Language Models*, adaptée donc à l'image ou au texte. La musique est un sujet très intéressant mais assez frustrant, car beaucoup moins efficace que ce qu'on peut observer ailleurs. Plusieurs raisons à ce retard :

- Le fait que nous sommes beaucoup plus tolérants à des erreurs dans une image qu'à des erreurs dans du son. Quelques pixels erronés ne seront même pas captés par notre cerveau. Des erreurs sur la phase d'un signal généré, en revanche, seront une expérience auditive désagréable
- Le constat que la musique est un domaine moins intéressant que l'image ou le texte en application, avec donc des enjeux moins forts sur le plan économique
- Le fait que nous ayons relativement moins de données en musique qu'en image, dans un domaine où la quantité de données reste un facteur important pour entraîner un modèle de réseaux de neurones.

Si l'on laisse de côté les approches très récents où l'on mélangeait modèles de diffusion et son qui étaient assez décevants, la dernière approche convaincante était le *Jukebox* d'*OpenAI* en 2020 [https://openai.com/research/jukebox]

Dès lors, le récent travail de *Meta AI* (encore eux, promis, nous essayons d'équilibrer nos veilles) sur la musique est un travail passionnant qui pose de nouveaux résultats bluffants en termes de résultats. Le sujet n'est pas encore "conclu", mais nous vous encourageons déjà à aller écouter sur la page du projet les différents exemples :
https://ai.honu.io/papers/musicgen/

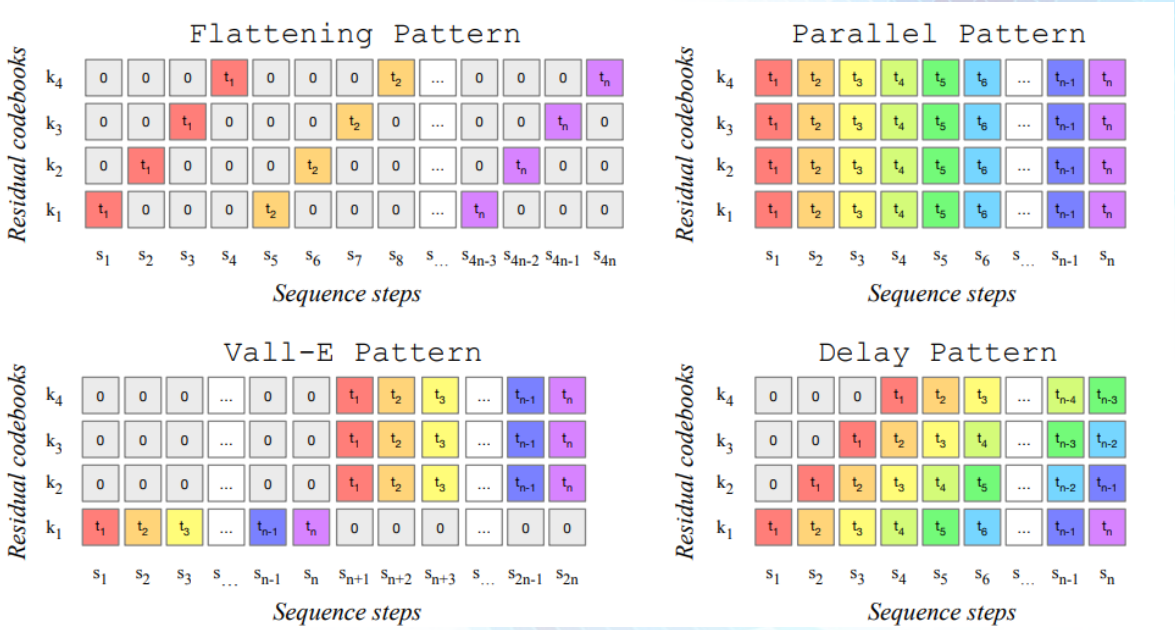
Que se passe-t-il ?
Sans rentrer trop dans la technique, les auteurs ont ici approché le problème un peu à la manière des célèbres *Latent Diffusion Models*, en entraînant un modèle génératif non pas sur la musique directement, mais sur la musique encodée via le projet *EnCodec* de *Meta* [https://github.com/facebookresearch/encodec]. Ce projet permettait d'apprendre une compression de l'information via un modèle encodeur/décodeur relativement classique.



Le nouveau projet génératif s'appuie donc sur ces représentations de haut niveau pour apprendre à générer de nouvelles musiques Ces morceaux de trente secondes peuvent être générés depuis un *prompt* textuel (qui reste assez décevant) ou depuis une mélodie d'un autre morceau (où là, la condition semble être appliquée d'une manière beaucoup plus efficace).

Le point technique important ici est sur la manière de modéliser la musique et de la générer.

Déjà, la modélisation de la musique porte sur plusieurs vecteurs quantifiés parallèles, et non pas un unique vecteur comme on le retrouve souvent en texte ou en image. Ces vecteurs complémentaires sont nécessaires pour retrouver un son de qualité. Hors, cette multiplication pose un problème d'optimisation en utilisation. Les auteurs ici travaillent plusieurs approches pour générer ces vecteurs à différents niveaux en cherchant à paralléliser les calculs. Une contrainte est que ces vecteurs sont générés séquentiellement avec donc une dépendance à l'historique. Les auteurs proposent donc un *framework* générique pour gérer ce cas de figure de *tokens* parallèles à calculer.



Deuxième point d'intérêt, le modèle génératif ici n'est pas basé sur les modèles de diffusion! A contre courant de la vague de fond démarrée en juin dernier, les auteurs ici utilisent un bon "vieux" *Transformer* autorégressif. SI les travaux ultérieurs confirment que cette approche est mieux adaptée à la génération de musique, peut-être pourrons-nous à un moment observer comment le type d'une donnée générée se corrèle avec l'architecture *Deep Learning* utilisée.

Pourquoi c'est intéressant ?

Les IAs génératives sont en même temps une apocalypse et une révolution pour les domaines concernés. Suivre ce tsunami est déjà intéressant sur le plan strictement social et économique, et ce que nous observons depuis quelques mois sur l'image (changements de paradigme économique, et apparition de nouveaux outils de création) se produira tôt ou tard en musique. Ce travail est clairement une nouvelle étape fondamentale vers le "moment *Stable Diffusion* de la musique".

Notons néanmoins ici que la musique reste souvent uniforme, et que la dépendance au *prompt* textuel est ici beaucoup moins intéressante. Il y a fort à parier que la faible taille du *dataset* (ici, 10K morceaux) y soit pour beaucoup.

Au delà, de nombreux sujets industriels supposent d'analyser du son (par exemple de la voix, ou le bruit d'un système), voire même plus généralement un signal continu dans le temps. Toute approche marquant des points dans le domaine de la musique et donc potentiellement pour nous un nouvel outil pour adresser ces sujets.

White-box Transformers, un peu d'intelligence dans le Deep Learning

[https://arxiv.org/abs/2306.01129]
Ce sujet est mathématiquement beaucoup plus complexe et "bas niveau" que les sujets que nous abordons usuellement dans ce blog, mais il tombe dans un sujet que nous suivons depuis des années avec attention, celui de l'interprétabilité des modèles et d'une meilleure compréhension de ce qui se passe quand nous entraînons et utilisons des réseaux de neurones.
Rappelons déjà la base : le *Deep Learning* est toujours aujourd'hui un sujet largement empirique, en déficit de compréhension via une théorie mathématique complète qui justifierait et orienterait les travaux. L'interprétabilité d'un modèle reste ainsi un sujet académique complexe où, si des solutions partielles existent, aucune approche ne peut prétendre complètement adresser le sujet.

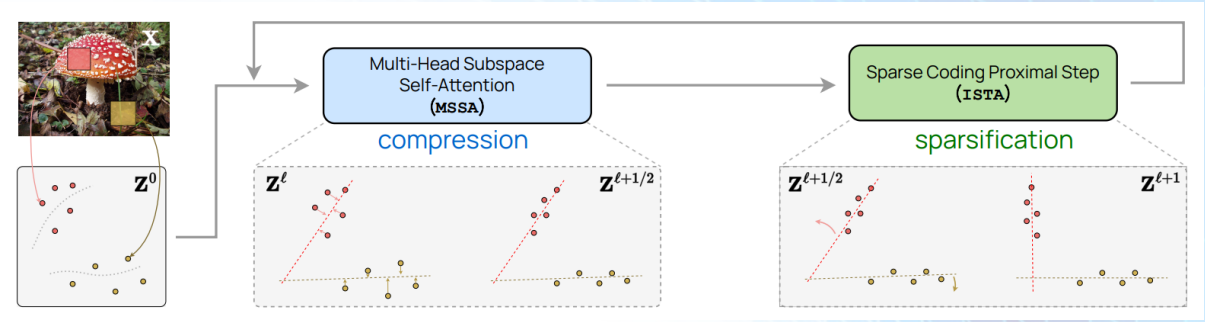
Aussi, tout travail sérieux allant dans cette direction est à surveiller particulièrement.

Que se passe-t-il ?
Une théorie forte dans le monde du *Deep Learning* est celle du *Representation Learning*. Caricaturalement : un réseau de neurones, quand il est entraîné à adresser un sujet, apprend implicitement à compresser l'information en entrée. Cette compression n'est pas juste une réduction de l'espace, mais conserve des informations haut niveau définissant la donnée en entrée.

Cette théorie est un des axes de travail fondamental en *Deep Learning*. C'est notamment dans cette optique que l'on parle des vecteurs intermédiaires dans un réseau de neurones comme de vecteurs latents ou d'*embedding*. Elle est importante, car elle semble régulièrement justifiée (par du *clustering*, de la détection d'anomalie, le *style transfert*, etc.) et nous offre un axe de compréhension fort sur ce qu'apprend un modèle IA

Ici, les auteurs proposent d'adapter l'architecture du *Transformer* et son entraînement, en se concentrant sur la compression opérée sur le modèle. Le point central est que cette adaptation, le *white box Transformer*, est mathématiquement interprétable pour chaque couche du réseau, ce qui ouvre la porte à de nombreux travaux passionnants pour mieux

analyser ce qui se passe dans ces réseaux. Le modèle est entraîné via un objectif de compression et un objectif de raréfaction (*sparsification*) de l'information, pour ensuite être adapté à de la classification classique.



Pourquoi c'est important ?

Ce travail ne donnera pas lieu dès demain à de nouveaux outils d'interprétabilité, mais le fait que les opérations des couches puissent être interprétées mathématiquement ouvrent la porte à l'apparition de nouvelles méthodes pour l'interprétabilité et donc la robustesse de nos modèles IA. Considérant que ces deux points sont, aujourd'hui, deux des plus grandes difficultés que l'on puisse rencontrer, nous ne pouvons pas ignorer ce type de publication.

Rédigé par Eric DEBEIR – Directeur scientifique de Datalchemy – eric@datalchemy.net
www.datalchemy.net