LES VITS : L'AVENIR DE LA VISION PAR ORDINATEUR ?

Ali AHMAD
Consultant Coperneec

Les récentes avancées dans le domaine de la vision par ordinateur ont vu le passage des réseaux de neurones convolutifs (CNN) vers des modèles plus récents : les transformers de vision (ViT). Les ViT sont une nouvelle architecture de réseau neuronal qui a démontré des performances supérieures à celles des CNN dans une variété de tâches de vision par ordinateur. Au fur et à mesure que l'utilisation des ViT se répand, pourraient-ils devenir l'architecture de référence pour les applications de vision par ordinateur ?

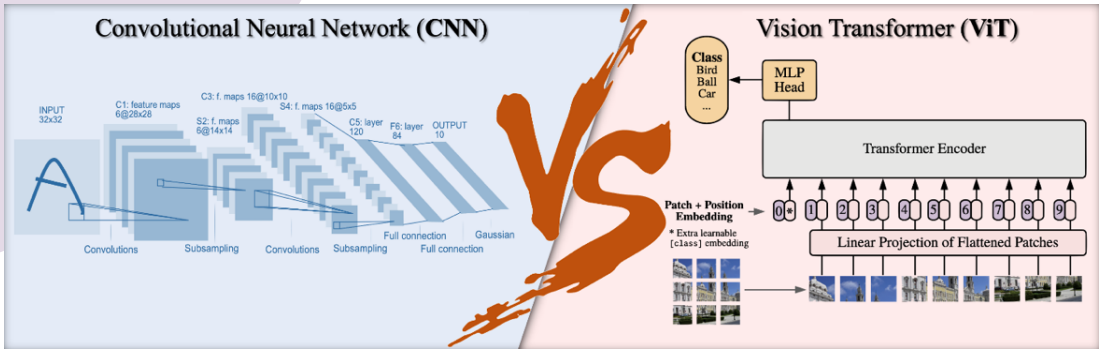

Figure 1: Le concept de CNN Vs le concept de ViT. Source: Bai et al, Are Transformers More Robust Than CNNs? NeurIPS 2021.

I. LES CNN EN VISION PAR ORDINATEUR

Le CNN est un type d'algorithme d'apprentissage profond qui analyse les données visuelles. Il est constitué de plusieurs couches de neurones interconnectées hiérarchiquement. Chaque couche est conçue pour extraire des caractéristiques telles que les bords, les formes et les textures des données d'entrée telles que les images, les vidéos, etc.

Le début de la CNN n'a pas connu en grand succès. En effet, l'utilisation des premiers CNN, développés par Yann LeCun dans les années 1990, a été limitée en raison du besoin de données et de ressources informatiques considérables. Quelques décennies plus tard, l'algorithme CNN a fait ses premiers pas dans le domaine applicatif de la vision par ordinateur grâce au succès d'AlexNet. C'est le nom donné à une architecture de réseau neuronal convolutif (CNN) qui a participé et remporté une compétition de classification d'images (le défi 2012 ImageNet) avec des performances spectaculaires.

Depuis lors, le CNN est resté le réseau le plus dominant dans le domaine, évoluant vers différentes variantes d'architectures, telles que les architectures profondes InceptionNet (GoogleNet) en 2014 et le VGG en 2015, les architectures avec des connexions résiduelles telles que ResNet pour résoudre le problème des gradients évanescents, et les réseaux plus légers optimisés pour les applications mobiles, telles que EfficientNet en 2020. Ces réseaux sont actuellement utilisés comme backbones (blocs de base) pour de nombreuses applications de vision par ordinateur.

II. LES TRANSFORMERS DE VISION (VIT) : UNE ALTERNATIVE COMPÉTITIVE ET INNOVANTE

En 2022, le ViT est apparu comme une alternative compétitive aux réseaux de neurones convolutifs (CNN). Grâce à son mécanisme d'auto-attention (self-attention mechanism) et les performances des modèles de Transformers dans le traitement du langage naturel (NLP) depuis son introduction en 2017, la recherche s'est orientée vers l'application des mêmes principes à la vision par ordinateur.

Le premier ViT appliqué à la classification d'images est développé par Google Research Brain Team en 2020. Depuis cette date, plusieurs variantes du ViT ont vu la lumière dans le domaine de la vision par ordinateur. Par exemple, le DeiT (Data-Efficient Image Transformer) est développé en 2021 pour améliorer la performance de ViT en utilisant une nouvelle technique de distillation spécifique au Transformers. Durant la même année, un groupe de recherche a développé le Swin Transformers (Swin pour shifted windows) pour surmonter la perte partielle d'informations spatiales aux bordures des patches ainsi que la complexité informatique quadratique de ViT.

Le ViT et ses variantes sont ainsi utilisés, seul ou en architecture hybrid (CNN + Transfromers), non seulement pour la classification d'images mais aussi dans plusieurs autres domaines de la vision par ordinateur, tels que la segmentation (e.g, UNETR, Swin-Unet, SegFormer,etc.), la détection (e.g, DETR), la génération d'images (e.g, TransGan) , restauration d'images (e.g, IPT, TTSR) et le traitement des séquences de vidéos (e.g, TimeSformer).

III. COMPARAISON DES PERFORMANCES

Le point fort du ViT est d'extraire le contexte global et les dépendances à long terme entre les éléments d'entrée par rapport aux méthodes CNN qui ont tendance à se concentrer sur les détails locaux. Cela permet à ViT d'apprendre des caractéristiques plus complexes que les CNN, ce qui se traduit par une meilleure précision et des temps d'apprentissage plus rapides ainsi qu'une efficacité en termes de mémoire et de ressources de calcul. Cependant, le ViT nécessite une très grande base de données pour apprendre un modèle à partir de zéro et pour surpasser les CNN.
Les CNN, en revanche, sont plus performants dans les régimes de données faibles grâce à leur fort biais inductif. Cependant, quand un grand nombre de données est disponible, les biais inductifs (fournis par les CNN) limitent la capacité globale du modèle CNN.

IV. CONCLUSION

Les ViT représentent une réelle avancée dans le domaine de la vision par ordinateur. Ils sont plus performants que les CNN dans les régimes de données élevées et peuvent apprendre des caractéristiques plus complexes. Actuellement, de nombreux efforts de recherche sont déployés pour tenter de réduire la complexité et les exigences en matière de matériel et de données nécessaires pour entrainer des Transformers de vision. Est-ce que l'évolution rapide des ViTs dans le domaine de la vision par ordinateur est un signe de la fin de l'ère des CNN?... Seul le temps nous le dira.

ALI AHMAD
Consultant Coperneec

Titulaire d'un doctorat et Data Scientist, Ali s'est spécialisé en intelligence artificielle et vision par ordinateur. Ali possède également de solides compétences en traitement d'images et de données, apprentissage automatique (ML/DL), systèmes de vision et simulation numérique de données.