Rapport de Veille Technologique

Développez une preuve de concept

Projet 7 du parcours Ingénieur Machine Learning

Ce rapport présente le travail de veille technologique réalisé dans le cadre de ce projet. Le sujet de veille abordé est la reconnaissance d'objet par une technique de Deep Learning relativement innovante : les Vision Transformers.

Sommaire :

I. Thématique

La thématique de veille abordée dans ce travail est celle des Vision Transformers.
Les Vision Transformers sont des algorithmes de Deep Learning appliqués à la reconnaissance d'objets sur des images. Ils sont nommés ainsi car ils représentent l'application directe; dans le domaine de la vision par ordinateur; de l'architecture phare des modèles d'apprentissage profond utilisés en Traitement Automatique du Langage (ou NLP pour *Natural Language Processing* en anglais), les *Transformers*.

Les Transformers ont été introduits en 2017 dans le papier de recherche *Attention Is All You Need (Vaswani et al.)*[1] publié par une équipe de recherche de Google. Ce sont des réseaux de neurones particulièrement efficaces pour apprendre des relations entre plusieurs données d'une séquence, par exemple d'une séquence de mots dans une phrase, grâce au mécanisme "d'attention" alors introduit par cet article.

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, *Attention Is All You Need*, 2017, in arXiv:1706.03762

Jusque là appliqués essentiellement sur les applications de Traitement Automatique du Langage pour lesquelles les Transformers ont été conçus, ils ont récemment fait l'objet d'études cherchant à transposer leur principe de fonctionnement sur les applications de vision par ordinateur sous la forme des Vision Transformers.

II. Etat de l'art

A. Mise en contexte dans l'état de l'art de la vision par ordinateur

La vision par ordinateur (*Computer vision en anglais*) est un domaine technique très dynamique depuis le début des années 2010 et l'avènement des réseaux de neurones, principales architectures de modèles permettant un apprentissage qualifié de "profond".

Parmi les principales tâches comprises dans ces applications exploitant des images ou des flux vidéos, se trouvent la reconnaissance d'objets (ou la détection), la classification, la segmentation d'images ou encore l'OCR (*Optical Character Recognition*, qui peut s'apparenter à une application spécifique de reconnaissance d'objets).

Pour reconnaître un objet présent sur une image, les techniques actuelles de l'état de l'art sont fortement basées sur la convolution. Concrètement, on conçoit et on entraîne des algorithmes de type "réseaux de neurones convolutifs" (ou CNN en anglais pour Convolutional Neural Network) dont l'architecture est globalement composée de deux parties :

- La première, les couches les plus profondes, prennent l'image en entrée et réalisent des opérations de convolution permettant d'en extraire des motifs. Les premières couches extraient des motifs très primaires comme des bords ou des coins puis les suivantes réalisent des extractions de motifs de plus en plus avancés comme des doigts puis une main etc. …
- La seconde partie de ce type de réseaux est constituée de couches de neurones complètement connectés (des couches denses) dont l'objectif est d'utiliser les motifs détectés dans les couches convolutives afin de réaliser une tâche de décision (classification d'image ou segmentation par exemple).

Le principe de la convolution et son application à travers les CNN a permis de très grands progrès dans l'ensemble des tâches de vision par ordinateur. La forte dynamique de recherche a permis de développer des réseaux convolutifs toujours plus performants et l'état de l'art actuel est principalement accaparé par les ResNet, des réseaux de neurones convolutifs utilisant un mécanisme de "mémoire" entre les couches, permettant de contrebalancer un effet négatif du grand nombre de couches des réseaux profonds que l'on nomme "le vanishing gradient".

B. Etat de l'art antérieur des Vision Transformers

Les Vision Transformers, ou "ViT", portent l'espoir d'une alternative tout aussi performante aux réseaux de neurones convolutifs.

Inspirés directement des Transformers utilisés jusque là dans les applications de Traitement Automatique du Langage évoquées plus tôt, l'idée générale est de traiter une image ou un flux d'image comme une séquence de données et d'exploiter non-pas une extraction de motifs; comme les réseaux convolutifs; mais un mécanisme d'attention pour des tâches de reconnaissance d'images.
Avant le papier de recherche qui a définitivement installé les Vision Transformers dans le panorama des algorithmes de vision par ordinateur et sur lequel nous avons fait un focus dans la partie suivante, plusieurs essais d'application des Transformers au traitement d'image ont été conduits depuis la publication de *Attention Is All You Need (Vaswani et al.)*.

Plusieurs approches ont ainsi permis d'avancer la possibilité d'exploiter les mécanismes d'attention sur les images en proposant divers moyens d'utiliser les données des images (les intensités des pixels) tels que sont utiliser les mots et phrases dans les applications de TAL (NLP), avec comme principal frein le volume de données bien plus conséquent des images.

III. Analyse de l'article de référence

A. Présentation de l'article

L'article principalement étudié ici, sur lequel s'est fortement basée notre expérimentation, est dénommé "*An image is worth 16x16 words : transformers for image recognition at scale*" et a été publié en 2021 à travers la conférence internationale "International Conference on Learning Representations" par une équipe de Google[2].

Ce papier de recherche suggère que ces Vision Transformers atteignent des performances remarquables sur des tâches de reconnaissance d'image; parfois même meilleures que les performances de l'état de l'art réalisées par des réseaux de neurones convolutifs. Il démontre ainsi que les Vision Transformers peuvent être de bonnes alternatives aux réseaux exploitant des convolutions.

B. Architecture proposée d'un Vision Transformer

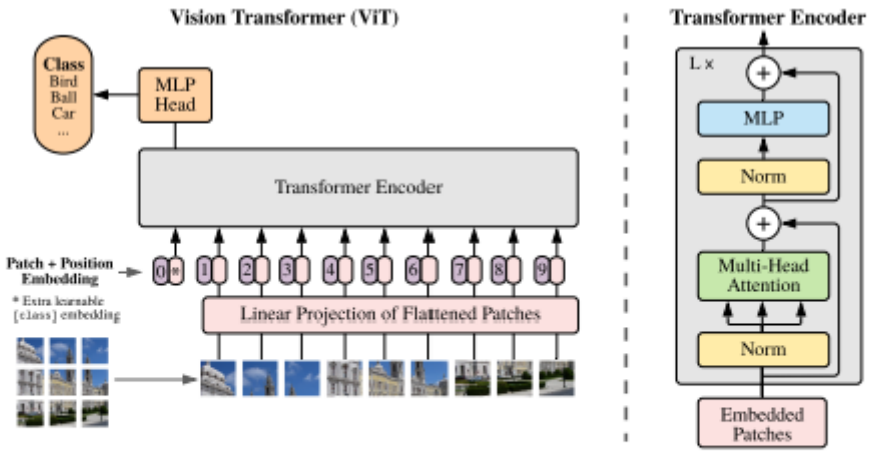
Pour réaliser l'adaptation du mécanisme d'attention à des images et avec l'objectif d'entraîner leur modèle sur des jeux de données particulièrement volumineux, l'équipe de chercheurs de Google propose une approche très similaire à celle proposée dans l'article "*On the Relationship Between Self-Attention and Convolutional Layers*" publié pour l'édition précédente de l'ICLR par une équipe de recherche de l'Ecole Polytechnique Fédérale de Lausanne[3].

---

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. *An image is worth 16x16 words: transformers for image recognition at scale*, in ICLR 2021

[3] Jean-Baptiste Cordonnier, Andreas Loukas, Martin Jaggi, *On the Relationship Between Self-Attention and Convolutional Layers*, 2020, in ICLR 2020

L'idée est ainsi de décomposer une image en un certain nombre de patches; afin de pouvoir l'interpréter comme une séquence; puis de les projeter linéairement avec un paramètre correspondant au positionnement de la partie d'image dans l'image globale, comme visible dans le schéma suivant provenant de l'article :



La séquence de vecteurs issue de ces projections est ensuite utilisée comme entrée d'un encodeur *Transformer*, décrit sur la droite du schéma, qui met en œuvre le mécanisme d'attention. La sortie du *Transformer* est ensuite exploitée par des couches denses afin de procéder à la résolution du problème de classification comme le ferait un réseau convolutif.

C. Principaux résultats obtenus

D'après cet article, les Vision Transformers sont avant tout capables d'atteindre de meilleures performances de classification que les modèles de l'état de l'art basés sur des ResNet dès lors que le ViT est suffisamment large et entraîné sur un jeu de données particulièrement important (JFT-300M contient 300 millions d'images).

Un intérêt sous-jacent des Vision Transformers est leur efficience comparée aux modèles convolutifs de l'état de l'art. En effet, comme bien explicité dans le benchmark effectué, les ViT atteignent des performances similaires aux ResNet pour des coûts de calculs jusqu'à 2 à 4 fois moindres.

Parmi les autres constats réalisés par l'équipe de recherche, l'intérêt d'utiliser des architectures de Vision Transformers plus larges semble directement dépendre de l'importance du jeu de données sur lequel le modèle va être (pré-)entraîné. Ainsi, la version "large" obtient des résultats similaires à la version "huge" si les deux sont pré-entraînés sur des jeux de données comme ImageNet-21k (contenant déjà 21k classes et 14M d'images). La donne change pour un pré-entraînement sur JFT-300M (plus de 300 millions d'images).

De la même façon, les Vision Transformers nécessitent un pré-entraînement sur des jeux de données plus volumineux que les ResNet : les ViT ont ainsi tendance à moins bien performer que les ResNet sur les "petits" jeux de données (ils auraient notamment une

tendance plus importante à sur-ajuster ces jeux de données). Les auteurs de la publication expliquent cette différence par l'intérêt du "biais convolutif induit" inhérent aux réseaux convolutifs (invariance en translation par exemple) sur des jeux de données plus petits.

Les Vision Transformers sont donc moins pertinents que les ResNet dès lors qu'ils ne peuvent pas être (pré-)entraînés sur des jeux de données très volumineux. Cette nécessité est un frein à son application dans certains domaines où il peut être difficile de rassembler des jeux de données aussi imposants (notamment dans le domaine médical) et où les techniques de transfer learning depuis des datasets volumineux ne permettent pas une bonne transposition des performances.

D. Un article pertinent qui révèle définitivement les Vision Transformers

De manière générale, ce travail de recherche semble avoir bien tenu compte de l'état de l'art et des variétés des possibilités de transposer le principe des transformers à la reconnaissance d'image.

Ainsi, l'équipe de Google évoque plusieurs possibilités de mettre en œuvre un plongement (embedding) à partir d'une image en faisant référence à la littérature existante et en expliquant leur choix de réaliser ce plongement par la décomposition d'une image en patches. Ce principe n'est pas complètement original, comme ils le soulignent, mais le benchmarking des performances réalisé dans ces travaux apporte une véritable valeur opérationnelle ; démontrant l'intérêt des Vision Transformers par rapport aux réseaux convolutifs.

D'autre part, la mise en perspective des résultats obtenus est particulièrement pertinente avec l'essai de plusieurs variantes architecturales (simple, large, très large), l'hybridation en réutilisant les sorties d'un ResNet (cartographie de caractéristiques / feature mapping) ainsi que la comparaison directe avec les performances obtenues avec des ResNet.

Enfin, le benchmark réalisé semble d'autant plus pertinent qu'il s'est appuyé sur des jeux de données reconnus : pré-entraînement sur ImageNet, ImageNet-21k et JFT-300M et évaluation sur ImageNet, CIFAR-100 et VTAB.

Globalement, si cet article est devenu très populaire dans la communauté de recherche en Deep Learning, c'est bien parce qu'il intronise définitivement les Vision Transformers comme des alternatives efficaces aux sacro-saints réseaux convolutifs en proposant une architecture facilement implémentable et une comparaison robuste et prometteuse de leurs performances avec les ResNet, l'état de l'art des réseaux convolutifs.

IV. Expérimentations & conclusion

A. Jeu de données utilisé

Afin de mener des expérimentations sur un Vision Transformer, nous choisissons de réutiliser le jeu de données du projet précédent de classification d'images afin d'obtenir rapidement une bonne idée des performances d'entraînement et de résultats, notamment en le comparant avec un réseau convolutif que l'on a déjà optimisé.

Le jeu de données est donc un ensemble de 20580 photos de chiens, classées en fonction de leur race parmi 120 races différentes. Ce dataset est mis à disposition par l'Université de Stanford et disponible à l'adresse http://vision.stanford.edu/aditya86/ImageNetDogs/ .

L'objectif sur lequel nous avons ici testé notre Vision Transformer est la classification d'une image de chien en fonction de sa race.

B. Modèle utilisé et analyse comparative des résultats

Pour réaliser le test d'un Vision Transformer sur ce dataset, étant donné la nécessité d'être pré-entraîné sur de larges datasets, nous avons choisi d'opérer par transfer learning en récupérant sur le dépôt Tensorflow Hub un modèle ViT pré-entraîné.

Nous avons ainsi trouvé un dépôt particulièrement pertinent car proposant différentes implémentations Tensorflow de Vision Transformers qui sont en fait les modèles pré-entraînés et mis à disposition par l'équipe de recherche de Google sur GitHub et convertis depuis la librairie JAX.

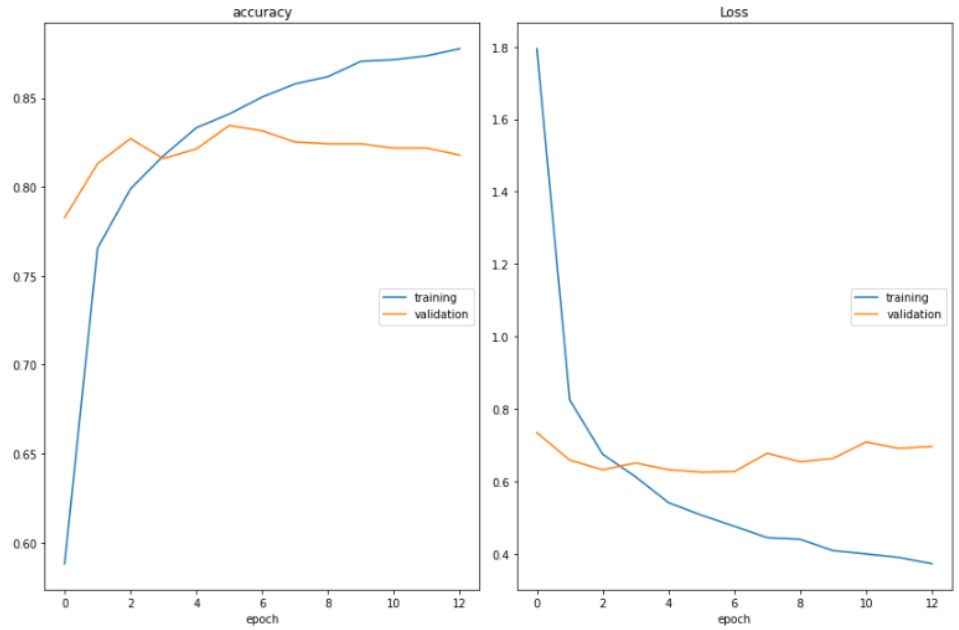
Le modèle que nous avons choisi est un Vision Transformer "S/16" qui est un modèle composé de 22 millions de paramètres et performant une décomposition des images en 16x16 patches. Il a été pré-entraîné sur ImageNet-21k et optimisé sur ImageNet-1k.

On en récupère une version ne contenant que la partie de "feature extraction" (projection des patches + transformer) et on lui ajoute une couche de décision softmax (une couche neurones de 120 unités correspondantes à chacune des races de chien).

Pour le transfer learning, les paramètres correspondants aux premières couches sont figés, seuls les 46200 paramètres correspondants à la dernière couche softmax que l'on a rajoutée sont entraînés.

La figure ci-après présente les courbes d'apprentissage obtenues pour l'entraînement sur notre dataset (avec data augmentation). On y aperçoit un apprentissage relativement efficace sur les toutes premières époques mais qui surajuste très rapidement le jeu de données d'entraînement :

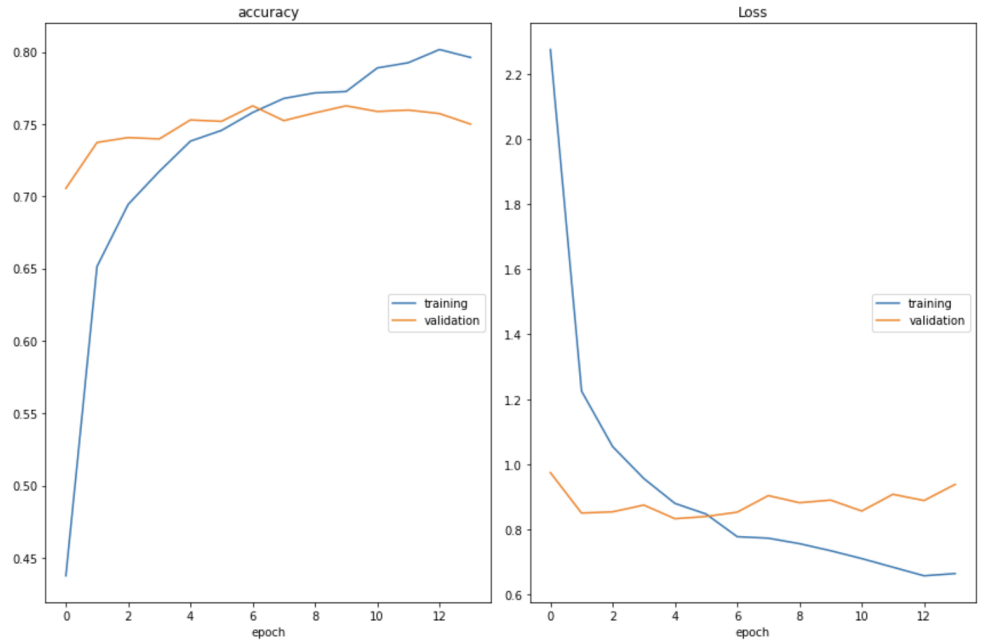
Courbes d'apprentissage du Vision Transformer pré-entraîné :




```
accuracy
        training                (min:    0.588, max:    0.878, cur:    0.878)
        validation              (min:    0.783, max:    0.834, cur:    0.818)
Loss
        training                (min:    0.372, max:    1.795, cur:    0.372)
        validation              (min:    0.625, max:    0.734, cur:    0.696)
```

Voici ci-dessous les mêmes courbes d'apprentissage pour un modèle ResNet-50 pré-entraîné lui aussi sur ImageNet et dont on ne ré-entraîne là aussi que les dernières couches :

Courbes d'apprentissage du ResNet-50 pré-entraîné et optimisé :




```
accuracy
        training                (min:    0.438, max:    0.802, cur:    0.796)
        validation              (min:    0.706, max:    0.763, cur:    0.750)
Loss
        training                (min:    0.658, max:    2.276, cur:    0.665)
        validation              (min:    0.834, max:    0.975, cur:    0.939)
```

La comparaison des deux modèles et de leurs courbes d'apprentissage nous permet de faire les constats suivants :

- Les phases d'apprentissage sont assez similaires mais la généralisation de l'apprentissage semble légèrement meilleure sur le ResNet-50 que le Vision Transformer.
- Les performances atteintes sur le jeu de validation sont supérieures pour le Vision Transformer (accuracy à 83.4% contre 76.3%) alors même que les dernières couches du ResNet ont été légèrement optimisées sur ce dataset lors du projet précédent (ajout d'un GlobalAveragePooling et d'un dropout de 15% avant la couche softmax).

Ces deux modèles étant pré-entraînés sur le même jeu de données (ImageNet), cette différence de performances pourrait s'interpréter par le fait que le mécanisme d'attention permet de sélectionner plus efficacement les formes utiles pour la classification que les convolutions sur ce jeu de données dans lequel un certain nombre de photos peuvent limiter les capacités d'extraction de motifs pendant l'apprentissage : en effet, au-delà de la variabilité de cadrages et d'expositions, certaines photos comportent plusieurs chiens voire des humains, certains chiens sont habillés etc. …

C. Perspectives de recherche et conclusion

L'une des principales interrogations soulevées par les auteurs de l'article de référence concerne la capacité des méthodologies d'auto-apprentissage à améliorer les performances ou l'efficience d'apprentissage des Vision Transformers.

On peut également se poser la question de l'adaptation des Vision Transformers à l'exploitation non pas d'une image mais d'une séquence d'images (flux vidéo, images 3D). On pourrait ainsi imaginer recycler le principe de décomposition en patches + embedding en décomposant non pas une image mais une série d'images : chaque patch étant lui-même une image de la séquence d'image. Quelles performances pourraient alors atteindre des Vision Transformers de ce type ?

D'autre part, le sujet étant couvert que partiellement dans cet article, on peut se poser diverses questions sur les possibilités d'hybridation de Vision Transformers avec des réseaux de neurones convolutifs (notamment autres que des ResNet).

Pour conclure, les Vision Transformers constituent ainsi une alternative concrète aux réseaux convolutifs, même s'ils manquent de pertinence par rapport à ceux-ci s'ils ne sont pas pré-entraînés sur de très larges jeux de données, ils peuvent dans le cas contraire atteindre des performances similaires pour un coût de calcul bien moindre.

V. Bibliographie

- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, *Attention Is All You Need*, 2017, in arXiv:1706.03762

- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. *An image is worth 16x16 words: transformers for image recognition at scale*, 2021, in ICLR 2021.

- Jean-Baptiste Cordonnier, Andreas Loukas, Martin Jaggi, *On the Relationship Between Self-Attention and Convolutional Layers*, 2020, in ICLR 2020

- Maithra Raghu, Thomas Unterhiner, Simon Kornblith, Chiyuan Zhang, Alexey Dosovitskiy. *Do Vision Transformers See Like Convolutional Neural Networks ?* In arXiv:2108.08810