RAPPORT

Projet 7 : Développer une preuve de concept

Étudiant : Zeruk Viktoriya

Mentor : Louis Willems

Sommaire

INTRODUCTION

Les réseaux de neurones convolutif, ou CNN, sont actuellement les méthodes les plus importantes dans le champ visuel de l'ordinateur.

Dans ce rapport, nous examinerons de plus près une nouvelle tendance récente : Vision Transformer (ViT). Depuis Alexey Dosovitskiy et al. appliqué avec succès un transformateur sur une variété de références de reconnaissance d'image, il y a eu une quantité incroyable de travaux de suivi montrant que les CNN pourraient ne plus être l'architecture optimale pour la vision par ordinateur.

Dans le projet précédent (Classez des images à l'aide d'algorithmes de Deep Learning), nous avons testé des réseaux de neurones convolutif pour classer les images de chiens par race.

Rappel sur la méthodologie :

CNN from scratch	Transfer learning
➢ **Convolution :** ▪ Formation d'une base de convolution ▪ Mesure des fonctions d'activations ➢ **Classification :** ▪ Mesure des optimiseurs ▪ Mesure d'une couche de Flatten ▪ Mesure d'une couche dense supplémentaire ▪ Mesure d'une couche de batch normalization ▪ Mesure d'une couche de dropout ➢ **Mesure de l'effet de la data augmentation**	➢ Phase d'expérimentation sur 5 modèles et 12 races : ▪ Entrainement d'un modèle de base ▪ Entrainement avec data augmentation ▪ Entrainement avec data augmentation et fine tuning ➢ Mise à l'échelle sur 60 races ➢ Mise à l'échelle sur 120 races.

CNN Transfer learning - Expérimentation 12 races:

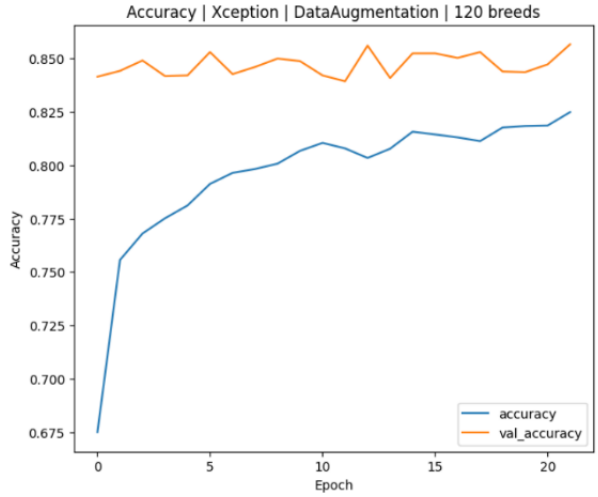
Models	Base	DataAugmentation	DataAugmentation & FineTuning
VGG16	0.67	0.65	0.09
ResNet50	0.25	0.18	0.2
MobileNetV2	0.91	0.89	0.61
InceptionV3	0.98	0.98	0.96
Xception	0.98	0.96	0.95

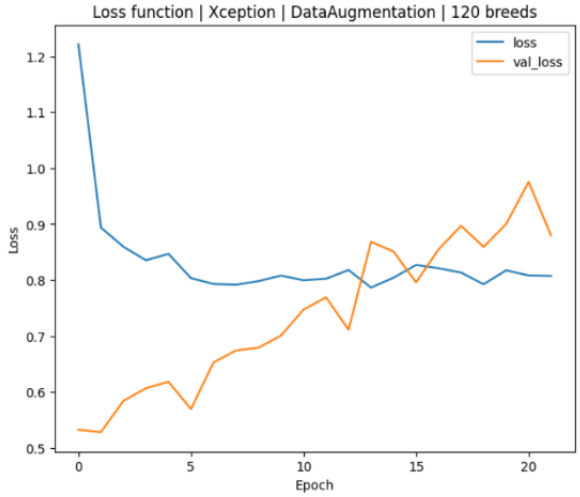
Meilleur modèle : Xception.

CNN Transfer learning - Mise à l'échelle 120 races :

```
147/147 [==============================] - 26s 176ms/step - loss: 0.5899 - accuracy: 0.8377
```

Models	DataAugmentation
Xception	0.84





L'idée est donc de comparer, en termes de précision et temps de calcul, un CNN et un ViT. Nous avons décidé de comparer le modèle Xception au modèle ViT-B / 16 de Google. Pour ce faire, nous réutiliserons le Stanford Dogs Dataset.

1. CLASSIFICATION D'IMAGES

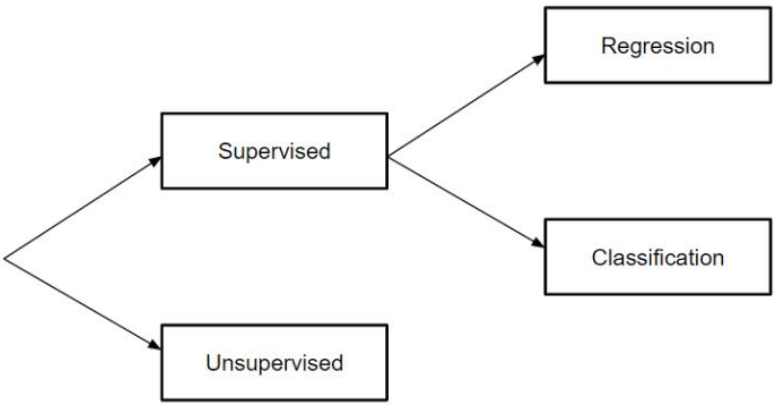
Tout d'abord, rappelons comment une image est représentée numériquement. L'ordinateur ne pouvant lire des objets continus, il faut discrétiser l'image. Pour cela, elle est découpée en petites unités de bases appelées pixels. Chaque pixel est ensuite caractérisé par :

- Une valeur entière entre 0 et 255 décrivant le niveau d'intensité de gris pour une image en noir et blanc.
- Un triplet d'entiers entre 0 et 255 décrivant les niveaux d'intensité de rouge, vert et bleu pour une image en couleur.

Classification d'images.

La classification d'images consiste à construire un système capable d'assigner correctement une catégorie à n'importe quelle image en entrée en fonction de règles particulières. La loi de catégorisation peut être appliquée par une ou plusieurs caractérisations spectrales ou texturales. Les techniques de classification d'images sont principalement divisées en deux catégories : Les techniques de classification d'images supervisées et non supervisées.

- La classification non supervisée est une méthode entièrement automatisée. Cela signifie que des algorithmes d'apprentissage automatique sont utilisés pour analyser et regrouper des ensembles de données non étiquetées en découvrant des modèles cachés ou des groupes de données sans nécessiter d'intervention humaine. Les caractéristiques particulières d'une image sont reconnues systématiquement au cours de l'étape de traitement de l'image.
- La classification supervisée utilise elle des échantillons de référence préalablement classés (la vérité du terrain) afin d'entraîner le classifieur et de classer ensuite de nouvelles données inconnues.

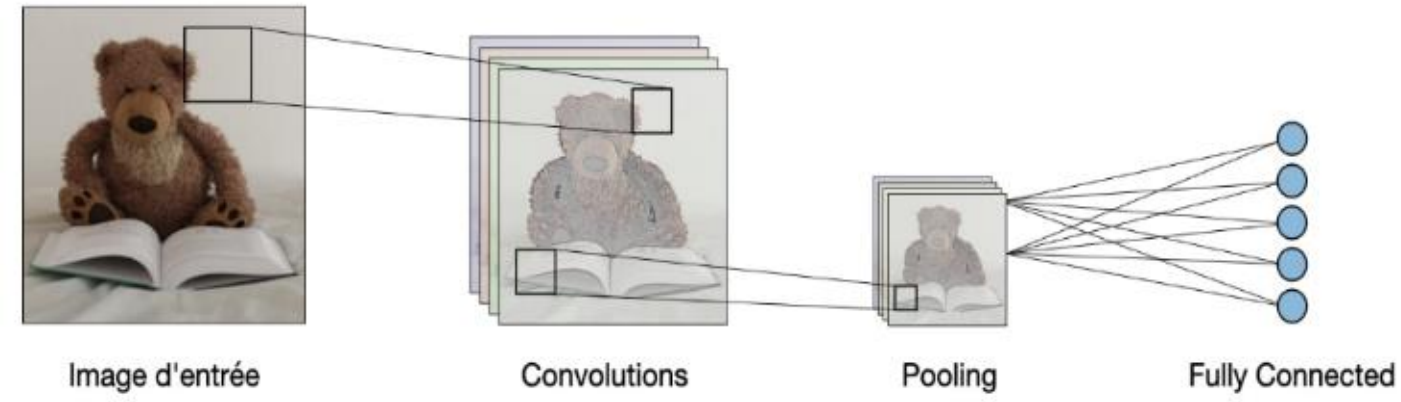


Fonctionnement :

Les algorithmes commencent par séparer l'image en une série de ses caractéristiques les plus marquantes, notamment en recherchant des bords, des coins ou des variations importantes des valeurs des pixels. Ces caractéristiques sont ensuite représentées sous forme de vecteurs pour pouvoir être utilisables par la suite. Certaines caractéristiques peuvent représenter un même objet ou détail sous différents angles, on les regroupe alors en classes. Par exemple, toutes les caractéristiques décrivant une roue de voiture seront regroupées dans une même et unique classe. Ces groupes de caractéristiques sont enfin utilisés par un classifieur pour obtenir une idée de ce que l'image représente et de la classe dans laquelle elle peut être considérée.

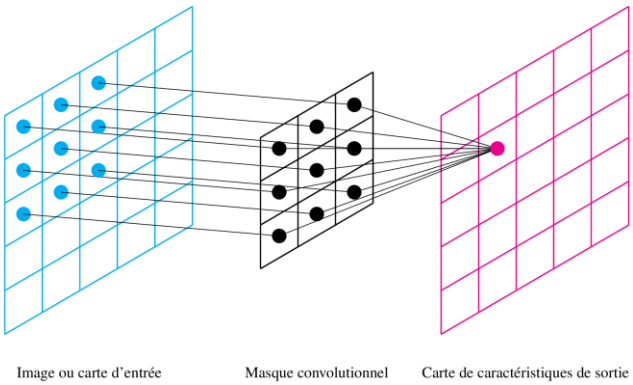
La classification d'images, en particulier la classification supervisée, dépend aussi énormément des données fournies à l'algorithme. Un jeu de données de classification bien optimisé et étiqueté fonctionne très bien par rapport à un mauvais jeu de données présentant un déséquilibre des données en fonction de la classe et une mauvaise qualité des images et des annotations.

Architecture d'un CNN traditionnel.

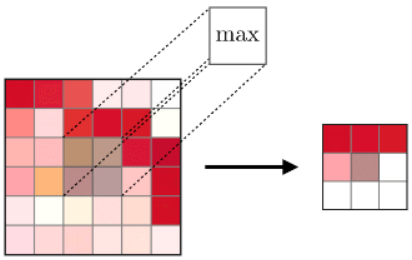
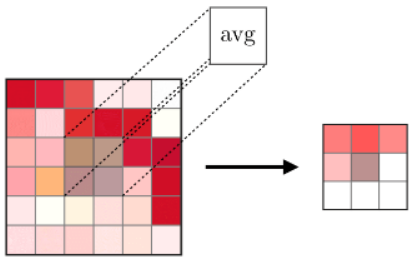


Les réseaux de neurones convolutionnels (en anglais Convolutional neural networks), aussi connus sous le nom de CNNs, sont un type spécifique de réseaux de neurones qui sont généralement composés des couches suivantes :

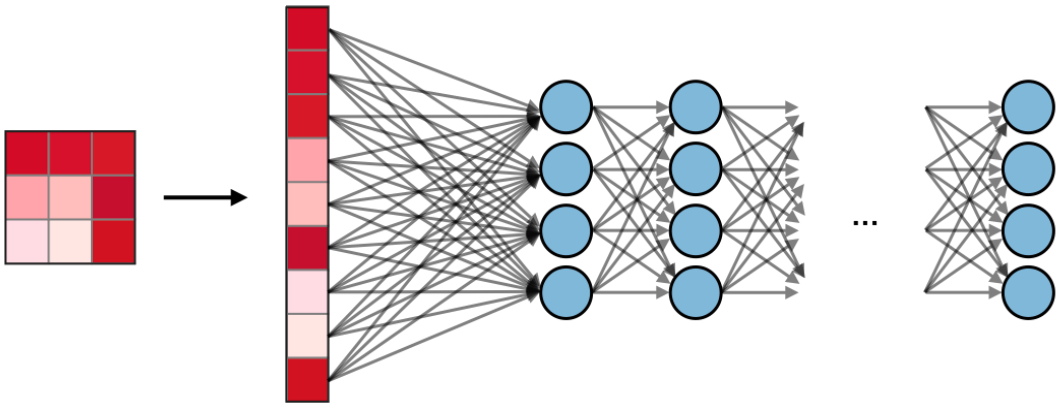
- **Couche de convolution (CONV) :** son but est de repérer la présence de caractéristiques dans les images reçues en entrée. Elle utilise des filtres, des petites fenêtres représentant une caractéristique et apprise au fur et à mesure de l'entraînement, qu'elle fait glisser sur l'image pour détecter la présence de caractéristiques particulières. Il s'agit de la principale couche utilisée dans un CNN.



- **Couche de pooling (POOL) :** son but est de réduire la taille de l'image en entrée tout en conservant les caractéristiques importantes détectées. Elle est souvent placée entre deux couches de convolution.

Type	Max pooling	Average pooling
But	Chaque opération de pooling sélectionne la valeur maximale de la surface	Chaque opération de pooling sélectionne la valeur moyenne de la surface
Illustration		
Commentaires	• Garde les caractéristiques détectées • Plus communément utilisé	• Sous-échantillonne la *feature map* • Utilisé dans LeNet

- **Couche fully-connected (FC) :** toujours la dernière couche du CNN, sont but est de classer l'image. Elle renvoie un vecteur indiquant la probabilité d'appartenance à chaque classe.



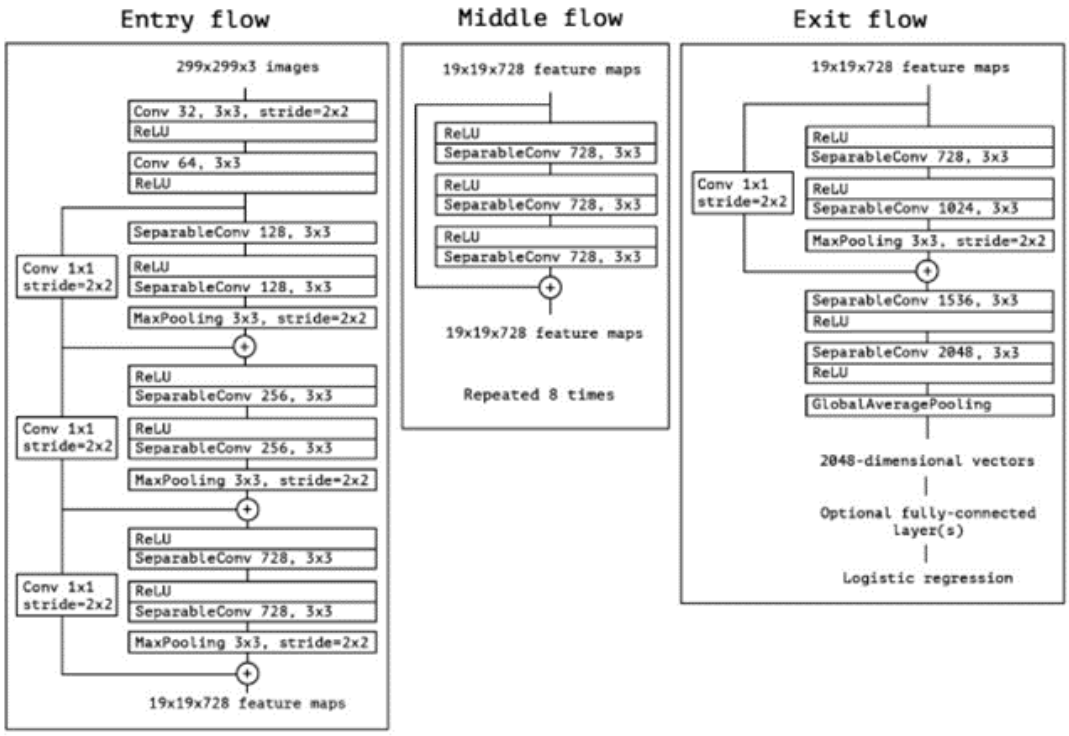
2. LES MODELES

2.1 Xception

Le modèle Xception est un réseau de neurones convolutif profond développé par Google en 2017. Sa principale caractéristique est qu'il contient des couches de convolution profondes séparables. Ce sont des alternatives aux couches de convolution classiques qui ont pour but de réduire les temps de calcul. Une couche de convolution classique va appliquer les filtres sur tous les canaux (3 canaux pour une image en couleur) en même temps, alors qu'une couche de convolution profonde séparable va les appliquer sur un seul canal à la fois puis appliquer une combinaison linéaire des sorties. L'idée principale est donc de diviser le travail de recherche de caractéristiques en tâches distinctes.

Le modèle Xception dispose également de connexions récurrentes, ce sont des boucles de rétropropagation qui permettent de garder en mémoire des informations obtenues lors 'étapes précédentes et de les utiliser au moment de prendre une décision dans les étapes suivantes.

L'architecture Xception:



2.2 Vision Transformers (ViT)

Le modèle Vision Transformer (ViT) a été proposé dans ***An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*** par **Alexey Dosovitskiy**, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold , Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby.

C'est le premier article qui entraîne avec succès un encodeur Transformer sur ImageNet, obtenant de très bons résultats par rapport aux architectures convolutives familières.

Cet article explore la manière dont on peut utiliser des Transformers pour tokeniser des images, tout comme on tokenise des phrases, afin de les transmettre à des modèles de transformation pour l'entraînement.

Le résumé de l'article est le suivant :

« Alors que l'architecture Transformer est devenue la norme de facto pour les tâches de traitement du langage naturel, ses applications à la vision par ordinateur restent limitées. En vision, l'attention est soit appliquée en conjonction avec des réseaux convolutifs, soit utilisée pour remplacer certains composants des réseaux convolutifs tout en maintenant leur structure globale en place. Nous montrons que cette dépendance aux CNN n'est pas nécessaire et qu'un transformateur pur appliqué directement à des séquences de patchs d'images peut très bien fonctionner sur des tâches de classification d'images. Lorsqu'il est pré-formé sur de grandes quantités de données et transféré vers plusieurs benchmarks de reconnaissance d'image de taille moyenne ou petite (ImageNet, CIFAR-100, VTAB, etc.), Vision Transformer (ViT) atteint d'excellents résultats par rapport à l'état de la réseaux convolutifs d'art tout en nécessitant beaucoup moins de ressources de calcul pour la formation. »

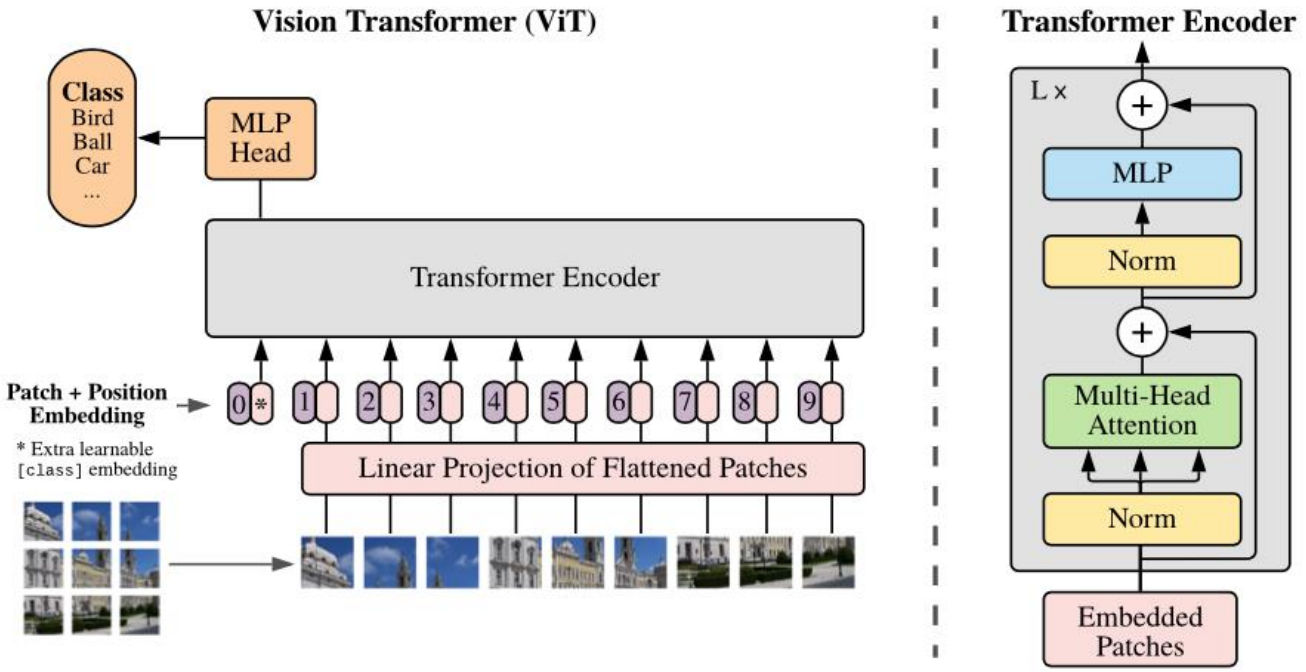
Source : https://huggingface.co/docs/transformers/model_doc/vit

Pour rappel, un Transformer est un modèle de Deep Learning qui utilise les mécanismes d'attention en pondérant de manière différentielle l'importance de chaque partie des données d'entrée.

Architecture et fonctionnement :

Un ViT n'est autre que le réseau d'encodage d'un Transformer auquel on a apporté quelques modifications dans le prétraitement pour le rendre adapté à la vision par ordinateur.

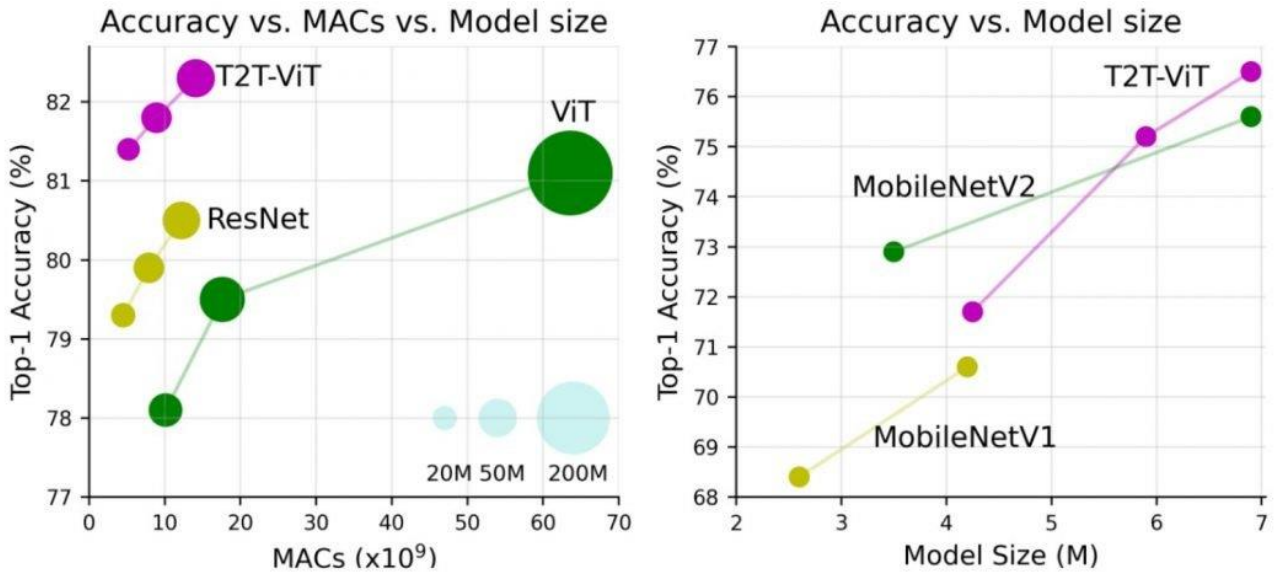
- Patching : on commence par diviser l'image en une séquence de petites parcelles. Le modèle ViT-B/16 que l'on va utiliser divise d'abord l'image en sections de 16x16 pixels qui ne se chevauchent pas. En considérant chaque canal de couleur (RVB), cela donne une matrice de dimensions [16,16,3].
- Aplatissement : la matrice obtenue est ensuite aplatie pour former un vecteur de taille 768 (16*16*3).
- Projection linéaire : un perceptron multicouches sans fonction d'activation est utilisée sur chaque patch aplati pour réduire la taille.
- Position des patchs : on ajoute un paramètre de position pour indiquer au modèle où se trouve chaque patch dans l'image originale.
- Transformer : Il ne reste plus qu'à introduire la séquence comme entrée dans Transformer



2.3 ViT vs CNN

Lorsqu'ils sont entraînés sur des ensembles de données de taille moyenne tels qu'ImageNet sans forte régularisation, les modèles ViT donnent des précisions modestes, moins bonnes que les CNN. Ce résultat apparemment décourageant peut être expliqué: les Transformers ne disposent pas de certaines caractéristiques inhérentes aux CNN, telle que l'invariance par rotation, et par conséquent ne généralisent pas aussi bien lorsqu'ils sont entraînés sur des quantités insuffisantes de données.

En revanche, si les modèles sont entraînés sur des ensembles de données plus importants (entre 14 millions et 300 millions d'images), on constate que les biais propres aux Transformers tendent à s'atténuer, et la précision devient semblable, voir même meilleure, que celle des CNN.



Source : [Vision Transformers (ViT) in Image Recognition – 2022 Guide]

3. MÉTHODES DE CLASSIFICATION

Le jeu de données utilisé est le Stanford Dogs Dataset. Il est composé de 20580 photos de chiens représentant 119 races différentes. Les photos sont en couleur, au format RGB. Une photo toutefois dispose en plus d'une couche alpha, couche utilisée pour coder le taux de transparence de l'image. Les méthodes de traitement d'images ne prenant souvent pas en compte cette couche, l'image a été supprimée du jeu de données.

Dans un second temps, une séparation jeu d'entraînement / jeu de test a été effectuée afin de pouvoir comparer les deux méthodes entraînées sur un même jeu de données. Les métriques utilisées pour la comparaison sont le score de précision, ou Accuracy, et le temps de calcul.

- *Xception*

  Le modèle a été entraîné via transfer learning en utilisant une méthode de fine-tuning partiel. Les couches fully-connected ont été remplacées par une couche GlobalAveragePooling2D pour réduire la dimension puis deux couches denses avec correction Relu et activation Softmax pour la classification. 10% des couches hautes de convolutions ont été réentraînées. Les hyperparamètres avaient été optimisés dans le précédent projet et sont les suivants :

  - La fonction d'activation utilisée est la fonction Relu
  - La fonction d'optimisation est la fonction Adam
  - Le nombre d'Epochs est égal à 50

- *ViT-B/16*

Le modèle provenant de la plateforme HuggingFace, il a d'abord fallu adapter les données d'entraînement et de test et les représenter sous forme de dictionnaire. Ensuite, pour préparer les données, il suffit d'utiliser le Feature Extractor associé au modèle. Après avoir choisis la métrique d'évaluation, il ne reste plus qu'à charger le modèle en précisant le nombre de classes souhaitées afin de procéder au fine-tuning. L'optimisation des hyperparamètres étant bien plus complexe que pour les CNN, nous avons conservé les valeurs par défaut.

L’entraînement du modèles à ensuite nécessité l’utilisation d’un GPU et l’on a obtenu les résultats suivants :

	Temps entraînement	accuracy train	accuracy test
Xception	4462.854993	0.987791	0.727162
Vit	4566.369979	0.9922	0.8243

Comparaison des modèles

Les deux modèles ont des temps de calcul similaires mais un score de précision significativement meilleur pour le ViT. Le modèle ViT est toutefois plus compliqué à mettre en œuvre, notamment si l'on souhaite optimiser les hyperparamètres ou ajouter de la Data Augmentation.

4. CONCLUSION

Le modèle ViT utilise l'auto-attention multi-têtes en vision par ordinateur sans nécessiter les biais spécifiques à l'image. Le modèle divise les images en une série de patchs d'intégration positionnels, qui sont traités par l'encodeur du transformateur. Il le fait pour comprendre les caractéristiques locales et globales que possède l'image. Enfin, le ViT a un taux de précision plus élevé sur un grand ensemble de données avec un temps de formation réduit.

L’inconvénient d’un ViT est qu’il nécessite plus de données d’entraînement qu’un CNN.

Sources bibliographiques

Article de recherche :

2010.11929v2.pdf (arxiv.org)

https://openreview.net/pdf?id=LtKcMgGOeLt

Article de vulgarisation :

Vision Transformers (ViT) in Image Recognition - 2022 Guide - viso.ai

Code source :

google-research/vision_transformer (github.com)

Tutoriel :

Fine-Tune ViT for Image Classification with Transformers (huggingface.co)

https://analyticsindiamag.com/complete-guide-to-t2t-vit-training-vision-transformers-efficiently-with-minimal-data/