Développer une preuve de concept

Classification d'images

Lors d'un précédent projet, nous avions utilisé le réseau de neurones convolutif Xcpetion pour procéder à la classification d'images de chiens en fonction de la race. Les réseaux de neurones convolutifs, ou CNN, sont actuellement les méthodes les plus en vues dans le domaine de la vision par ordinateur. Cependant, depuis quelques mois, une nouvelle méthode vient concurrencer les CNN. Il s'agit du Vision Transformer, ou ViT, une adaptation au traitement d'images des Transformers couramment utilisés en traitement du langage.

L'idée est donc de comparer, en termes de précision et temps de calcul, un CNN et un ViT. Pour ce faire, nous réutiliserons le Stanford Dogs Dataset et le modèle Xception précédemment entrainé que l'on comparera au modèle ViT-B/16 de Google.

Sommaire:

I. Tour d'horizon de la Classification d'images 2

1. Qu'est-ce que la classification d'images ? 2
2. Fonctionnement général 2
3. Etat de l'art 3
4. Fonctionnement d'un CNN 3
5. Exemple du modèle Xception 4

II. Vision Transformers 5
1. Architecture et fonctionnement 5
2. ViT vs CNN 6

III. Méthodes de Classification 6

1. Xception 7
2. ViT-B/16 7
3. Comparaison des modèles 7

IV. Conclusion et perspectives 8

V. Sources bibliographiques 8

I. Tour d'horizon de la 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.

L'image est finalement décrite par une matrice à deux dimensions pour une image en noir et blanc, ou trois dimensions pour une image en couleur, de dimension nombre de pixels en hauteur/nombre de pixels en largeur et contenant la valeur, ou le triplet de valeurs, caractéristique de chaque pixel.

1. Qu'est-ce que la 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.

La classification d'images à un rôle important avec l'explosion de la quantité de données. De plus, les images obtenues à l'aide de caméras ou de capteurs divers sont le plus souvent non structurées, ce qui rend de plus en plus nécessaire l'utilisation de méthodes de Machine Learning ou Deep Learning pour faciliter et améliorer le traitement.

2. Fonctionnement général:

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.

3. Etat de l'art :

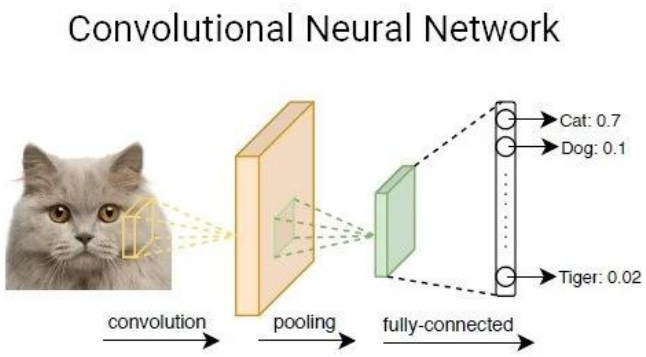
Les premières méthodes de traitement sont apparues à la fin du siècle dernier, notamment avec le célèbre algorithme SIFT, publié en 1999. Ces premières méthodes nécessitait une certaine expertise et maîtrise de la vision par ordinateur, car il fallait implémenter toutes les étapes d'extraction, de description et de regroupement des caractéristiques. Cela laissait un certain contrôle sur le processus mais était relativement technique et long à mettre en place.

Puis à partir de 2012 sont apparus les premiers algorithmes dit de Deep Learning : les réseaux de neurones convolutifs (CNN, pour Convolutional Neural Network). Il s'agit de réseaux de neurones profonds spécialement conçus pour le traitement d'images.

4. Fonctionnement d'un CNN :

Les réseaux de neurones convolutifs ont une méthodologie similaire à celle des méthodes traditionnelles d'apprentissage supervisé : ils reçoivent des images en entrée, détectent les caractéristiques de chacune d'entre elles, puis entraînent un classifieur dessus. Cependant, les caractéristiques sont apprises automatiquement. Les CNN réalisent eux-mêmes tout le boulot fastidieux d'extraction et de description: lors de la phase d'entraînement, l'erreur de classification est minimisée afin d'optimiser les paramètres du classifieur ET les caractéristiques. De plus, l'architecture spécifique du réseau permet d'extraire des caractéristiques de différentes complexités, des plus simples au plus sophistiquées.

L'unité de base d'un CNN est le perceptron, l'équivalent informatique d'un neurone, c'est-à-dire une fonction qui applique une moyenne pondérée aux données d'entrée puis applique une fonction d'activation au résultat. On regroupe ensuite les perceptrons pour former une couche, puis on empile les couches pour former un CNN. On retrouve quatre types de couches différentes :

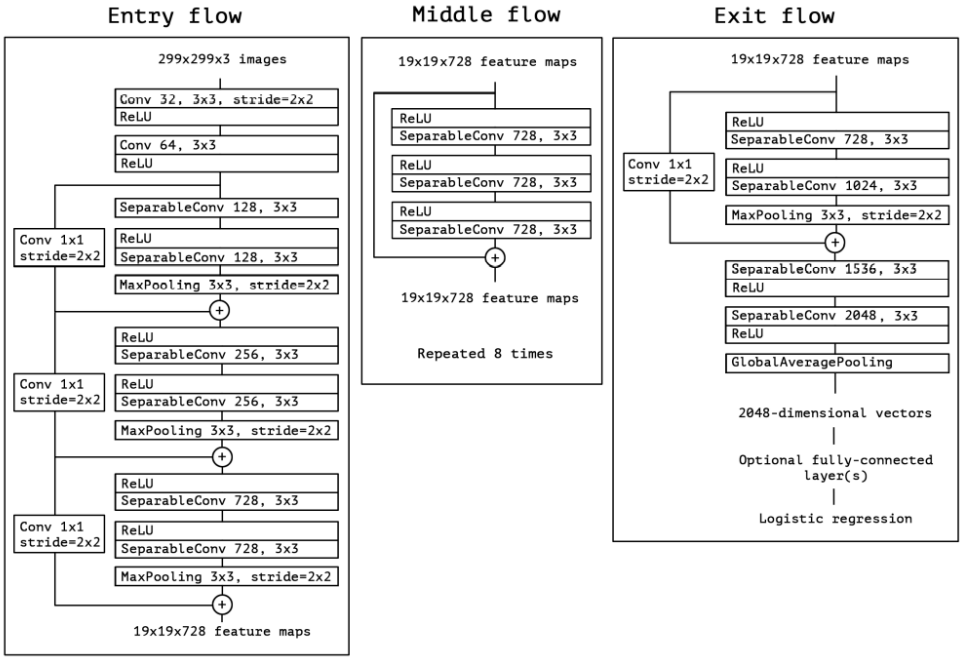


- Couche de convolution : son but est de repérer la présence de caractéristiques dans les images reçues en entrée. Elle utilise des filtres, des petites fenêtres représentant une caractéristique et apprise au fur et à mesure de l'entraînement, qu'elle fait glisser sur l'image pour détecter la présence de caractéristiques particulières. Il s'agit de la principale couche utilisée dans un CNN.
- Couche de pooling : 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.
- Couche de correction Relu : son but est de remplacer les valeurs négatives des pixels par zéro afin de limiter le sur-apprentissage et améliorer la précision et l'efficacité du CNN.
- Couche fully-connected : 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.

5. Exemple du modèle 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 d'étapes précédentes et de les utiliser au moment de prendre une décision dans les étapes suivantes.



II. Vision Transformers :

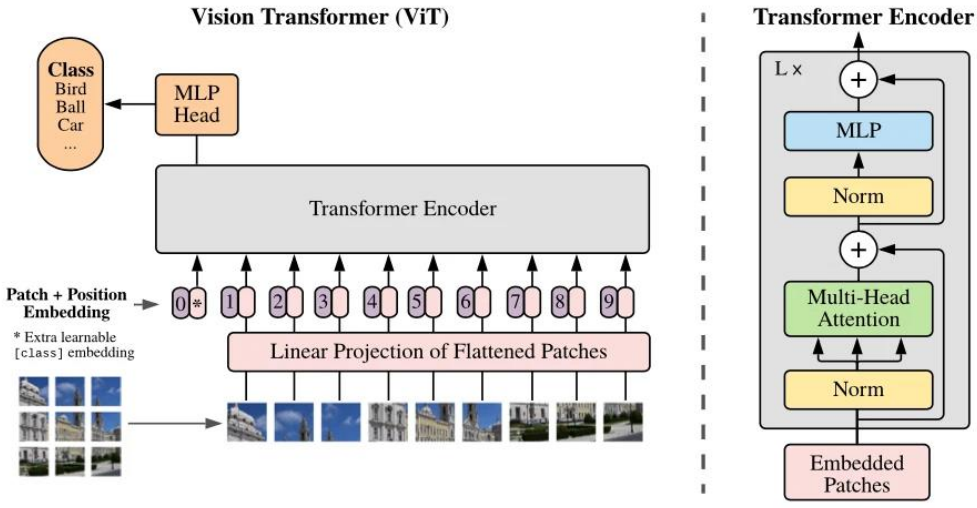
Le modèle Vision Transformer (ViT) a été présenté dans un article de recherche publié comme document de conférence à l'ICLR 2021 intitulé "An Image is Worth 16*16 Words : Transformers for Image Recognition at Scale". Il a été développé par une équipe de chercheurs de Google. Cet article explore la manière 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.

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.

1. 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.

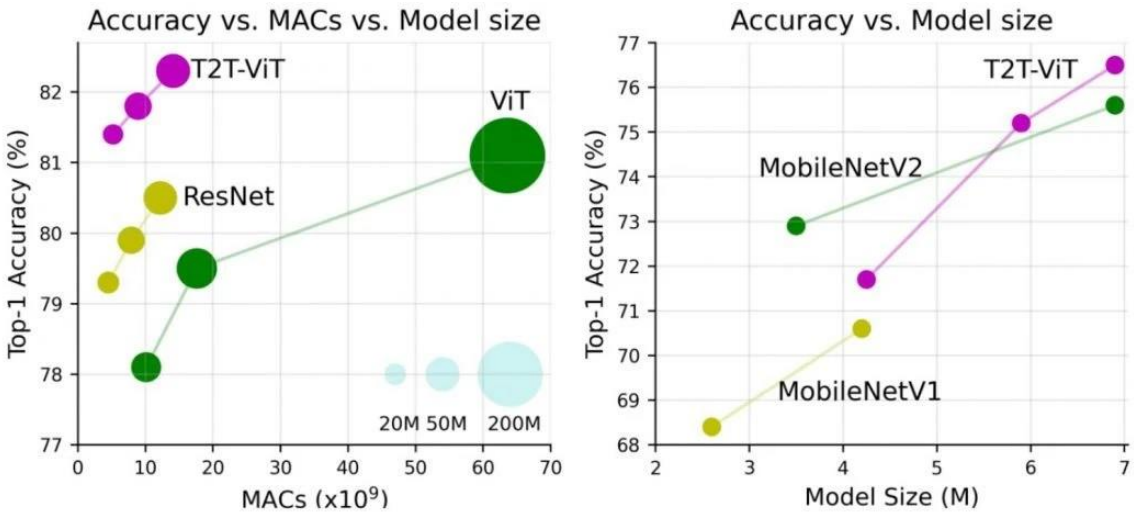
- Patching : on commence par diviser l'image en une séquence de petites parcelles. Le modèle ViT-B/16 que l'on va utiliser divise d'abord l'image en sections de 16x16 pixels qui ne se chevauchent pas. En considérant chaque canal de couleur (RVB), cela donne une matrice de dimensions [16,16,3].
- Aplatissement : 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. 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.



III. 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.

1. 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

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

	Temps entraînement (s)	Accuracy train	Accuracy test
Xception	4463	0.9878	0.7272

2. 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.

Le GPU a également été utilisé pour entraîner le modèle et l'on a obtenu les résultats suivants :

	Temps entraînement (s)	Accuracy train	Accuracy test
ViT	4566	0.9922	0.8243

3. 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.

Dans notre cas, l'utilisation d'un ViT est concluante et c'est la méthode que l'on retiendra pour la mise en production.

IV. Conclusion et perspectives :

Le principal frein à l'utilisation d'un Vision Transformer est sa complexité d'implémentation. Elle n'est pas aussi rapide et intuitive que pour un CNN. De plus, dans une optique de mise en production, il faut tenir compte des librairies proposant ces modèles pour des questions des maintenance et compatibilité. Toutefois, cette méthode étant récente et prometteuse, il y a de fortes chances pour qu'elle se développe et se démocratise dans le domaine de la vision par ordinateur. Par conséquent, les librairies d'implémentation vont devenir de plus en plus nombreuses et faciles d'utilisation, notamment pour l'optimisation des hyperparamètres.

L'inconvénient d'un ViT est qu'il nécessite plus de données d'entraînement qu'un CNN, et sera donc moins intéressant dans le cas d'une programmation complète. En revanche, dans la pratique, on utilise majoritairement le transfer learning, et dans ce cas, un ViT sera une alternative au CNN qu'il sera intéressant de tester.

V. Sources bibliographiques :

- Article de recherche : 2010.11929v2.pdf (arxiv.org)
- 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)