Mémoire

Projet 7 Machine Learning

Preuve de concept : Vision Transformers

Rédigé par : Arslane Lahmer, apprenti ingénieur Machine Learning

Sommaire

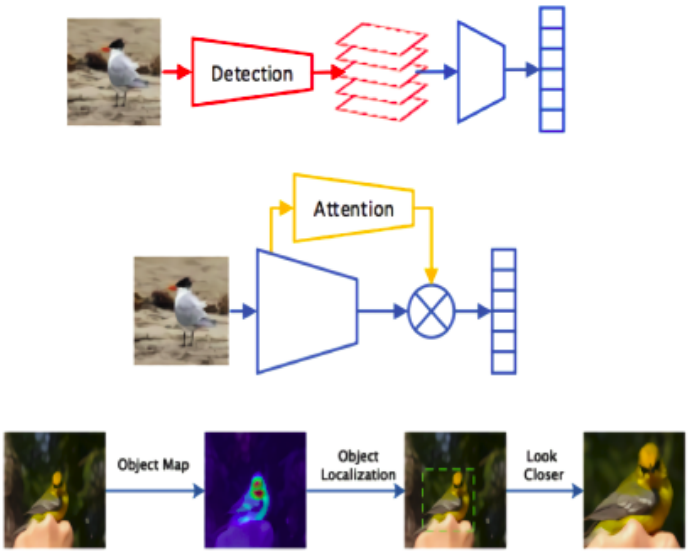
1. Introduction

La classification des images/objets appartenant à des sous-catégories (comme la détection de race de chien sur une image) est une tâche pouvant s'avérer particulièrement complexe en raison des subtiles différences inter-classes ainsi que des importantes différences intra-classes.

Les méthodes de classification les plus avancées ont recours, pour traiter ce problème, à la délimitation des zones les plus discriminantes de l'image pour effectuer leur classification. Les Vision Transformers se sont montrés particulièrement efficaces dans cette tâche, en réalisant des performances outrepassant celles des méthodes traditionnelles telles que les réseaux convolutifs.

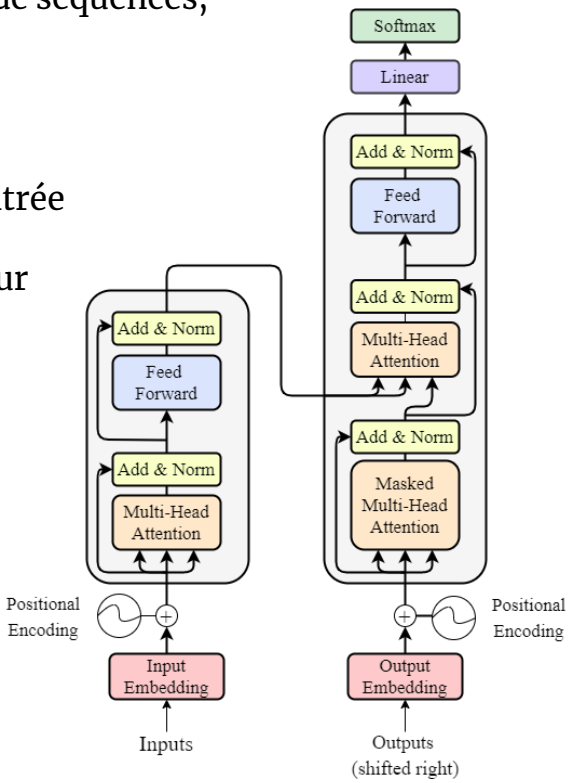
Dans ce projet, nous appliquerons un framework *ViT* à plusieurs blocs pour notre classification, qui localise les zones les plus importantes de l'image à l'aide du mécanisme d'attention tout en gardant l'architecture originelle du transformer. Notre choix de jeu de données s'est porté sur Stanford Cars.

Les performances seront interprétées et comparées à celles réalisées par d'autres modèles.



2. Transformers

Le transformer est un modèle d'apprentissage profond qui exploite le mécanisme de self-attention consistant à déterminer l'importance de chaque partie des données en entrée. Il est principalement utilisé dans les domaines du traitement du langage et de la détection d'images.
A l'instar des réseaux de neurones récurrents, les transformers sont conçus pour traiter des données sous forme de séquences, telles que le langage naturel, en l'appliquant dans des processus tels que la traduction et l'analyse des sentiments.
Néanmoins, contrairement aux RNNs, les transformers traitent toutes les données en entrée en une seule fois.
Le mécanisme d'attention fournit un contexte pour chaque position en entrée.



Si les données en entrée représentent une phrase de langage naturel, le transformer ne traite pas un mot à la fois mais considère la séquence dans son entièreté.
Cela a pour conséquence, notamment, de réduire le temps d'apprentissage nécessaire et de solliciter moins de ressources matérielles

Les Transformers ont été lancés en 2017 par une équipe d'ingénieurs de **Google Brain** .
Depuis lors, c'est devenu un modèle de choix pour les problématiques liées au traitement du langage, supplantant même, en terme d'efficacité les modèles traditionnels tels que les réseaux de neurones récurrents ainsi que les LSTMs (long-short memory service).

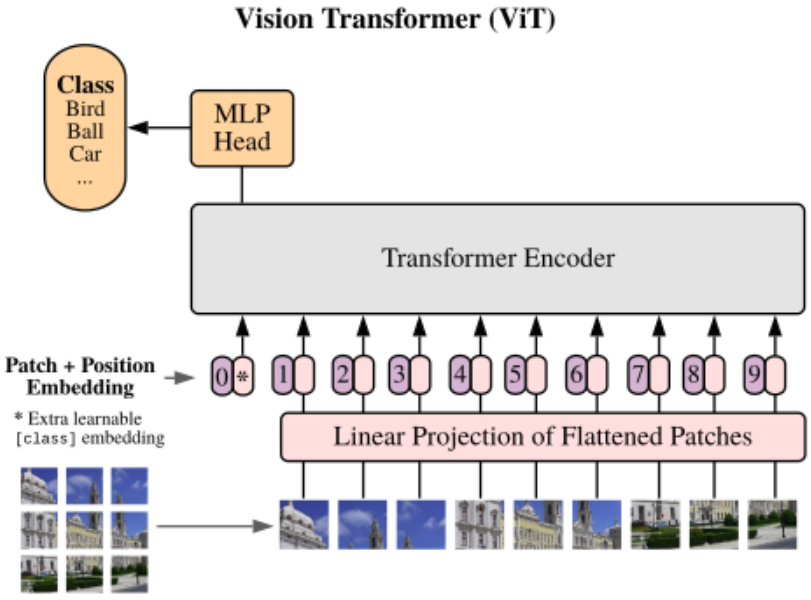
3. Vision transformers

Le concept de Vision Transformer (ViT) est une application des Transformers à la vision par ordinateur. Leur implémentation a subi de légères modifications afin de pouvoir traiter des données visuelles. L'architecture du modèle reste la même cependant tout en utilisant différentes méthodes pour la tokénization ainsi que l'embedding. En effet, l'image en entrée est divisée en patchs qu'on pourrait appeler *tokens visuels*. Les tokens sont transformés en vecteurs d'une certaine dimension. La position de chaque patch de l'image est intégré au vecteur et donnée à un transformer encoder dont l'architecture est fondamentalement la même que celui conçu pour traiter des données textuelles. Comme on peut le voir dans l'exemple ci-dessous.

L'encoder ViT est constitué de trois blocs: *Layer Norm*, *Multi-head Attention Network (MSP)* et *Multi-Layer Perceptrons (MLP)*. *Layer Norm* assure le processus d'apprentissage du modèle et l'adapte aux variations sur le training set. *MSP* génère des attention maps à partir des tokens visuels. Ces attention maps servent à se concentrer sur les zones les plus discriminantes de l'image.

MLP consiste en deux layers de classification avec la fonction d'activation GELU (*Gaussian Error Linear Unit*) à sa fin. Le dernier bloc MLP, représente la terminaison du transformer. On y applique la fonction *softmax* pour l'adapter à la classification.

En raison de sa nature générique, les applications des *ViTs* englobent quasiment tous les aspects de la vision par ordinateur. Cela inclut la classification d'images, la génération de phrases en langage naturel à partir de texte,

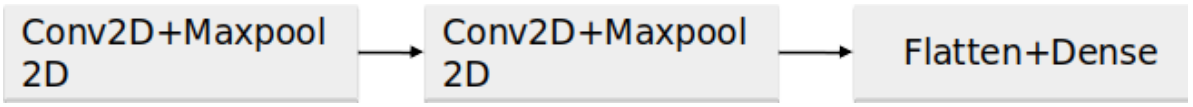


4. Choix du dataset

Il s'agit du dataset Stanford Cars qui contient 16.815 images divisées en 196 classes correspondant aux labels. Le jeu d'entraînement comporte 8.144 images, et le jeu de test 8.041, chaque label ayant un nombre d'images quasi équivalent dans les deux jeux de données.
La classe fournit des informations sur la marque, le nom du modèle ainsi que l'année de production. Exemple : *Audi V8 Sedan 1994*.

4. Méthode baseline

Le tableau ci-dessous montre les résultats obtenus par le modèle baseline. Il s'agit d'un réseau convolutif de 10 layers comprenant des layers de convolution, de max pooling ainsi que des layers de classification.



batch size	learning rate	max test accuracy
16	1e4	0.0055
16	5e4	0.0051
32	1e4	0.0056
32	5e4	0.0054

5. Méthode état de l'art

Il s'agit d'un modèle ViT préalablement entraîné sur la banque ImageNet. En le déployant, on obtient les résultats suivants

batch size	learning rate	max val accuracy	max test accuracy
16	1e4	0.67	0.83
16	5e4	0.65	0.85
32	1e4	0.66	0.87
32	5e4	0.65	0.84

6. Veille technologique

Il s'agit des modèles les plus performants sur les jeux de données CUB (jeu de données d'images correspondant à 200 espèces d'oiseaux) ainsi que Stanford Cars.

Method	Backbone	CUB	Stanford Cars
VGG-16	VGG-16	77.8	85.7
Inception-V3	Inception-V3	83.0	86.8
WS-DAN	Inception v3	89.4	94.3
MMAL-Net	Res-Net-50	89.6	95.0
ViT	ViT-B/16	91.0	94.3

7. Conclusion

Le ViT pré entraîné permet d'atteindre des résultats de niveau état de l'art en classification d'images grâce au mécanisme de l'attention qui permet de déterminer les zones les plus discriminante de l'image ce qui lui procure un avantage considérable en comparaison aux réseaux convolutionnels classiques qui se révèlent incapables de détecter les patterns propres à chaque classe en raison principalement des trop subtiles différences intra-classes.

Bibliographie

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei.
3d object representations for fine-grained categorization. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov,Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby.
An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

https://en.wikipedia.org/wiki/Vision_transformer

https://towardsdatascience.com/getting-started-with-pytorch-image-models-timm-a-practitioners-guide-4e77b4bf9055#5e2d