NOTE METHODOLOGIQUE DE VEILLE TECHNIQUE :

VISION TRANSFORMER

Nom – Prénom :	FAUGERE Armand
Date :	14/02/2024
Objet :	Note méthodologique

Objectif de la note :
A partir d'une veille technique réalisée, présenter une méthode innovante pour la classification des images :

- Un état de l'art de la méthode retenue
- Une analyse comparative avec une autre méthode
- Une preuve de concept
- Les références utilisées

Références :

resnet-50 :
Deep Residual Learning for Image Recognition : arXiv:1512.03385v1 [cs.CV] 10 Dec 2015
Transfert learning et fine tuning :
A Comprehensive Survey on Transfer Learning : arXiv:1911.02685v3 [cs.LG] 23 Jun 2020
Deep transfer learning for image classification: a survey : arXiv:2205.09904v1 [cs.CV] 20 May 2022
Transformers :
Attention Is All You Need : arXiv:1706.03762v7 [cs.CL] 2 Aug 2023
Vision transformers :
TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE : arXiv:2010.11929v2 [cs.CV] 3 Jun 2021
Vision transformers et CNN:
Do Vision Transformers See Like Convolutional Neural Networks? : arXiv:2108.08810v2 [cs.CV] 3 Mar 2022

SOMMAIRE

I) Dataset retenu

I.1) Présentation du jeu de données

Le jeu de donnée provient de la société place de marché. Il a déjà fait l'objet d'un projet classification et de plusieurs modélisations (VGG16, RESNET-50…).

Le jeu de données possède **1050 produits** avec des données textuelles et 1050 images au format jpg.

Il est composé de 7 catégories avec une répartition homogène : Home Furnishing, Baby Care, Watches, Home Decor & Festive Needs, Kitchen & Dining,Beauty and Personal Care, Computers.

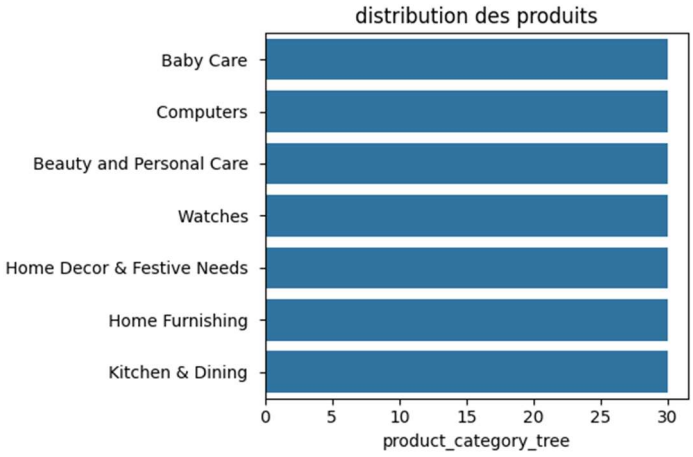
Ces données ont été compilées dans un dataframe nettoyé.

I.2) Sélection des données pour le POC

L'objectif est de présenter un POC avec une approche simple et économique. Pour cela un échantillonnage a été réalisé pour retenir **210 produits** répartis de façon homogène.

Cela permet de gagner du temps dans la réalisation du POC et suffisant pour réaliser la démonstration qui est une analyse comparative sur des périmètres identiques.

Distribution des données :

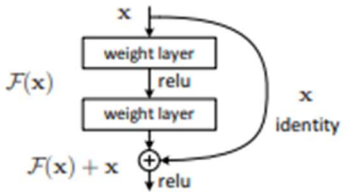


II) Les concepts de l'algorithme récent

II.1) L'algorithme étalon : le RESNET-50

Le principe est d'utiliser un réseau de neurones à convolutions profond CNN pour réaliser la prédiction.

La spécificité de cet algorithme est la mise en place d'un mécanisme de shorcut qui permet d'éviter que l'information soit bloquée dans les couches (ajout de l'entrée à la sortie) et ainsi de rendre efficace et possible l'apprentissage profond.



Principe de shortcut

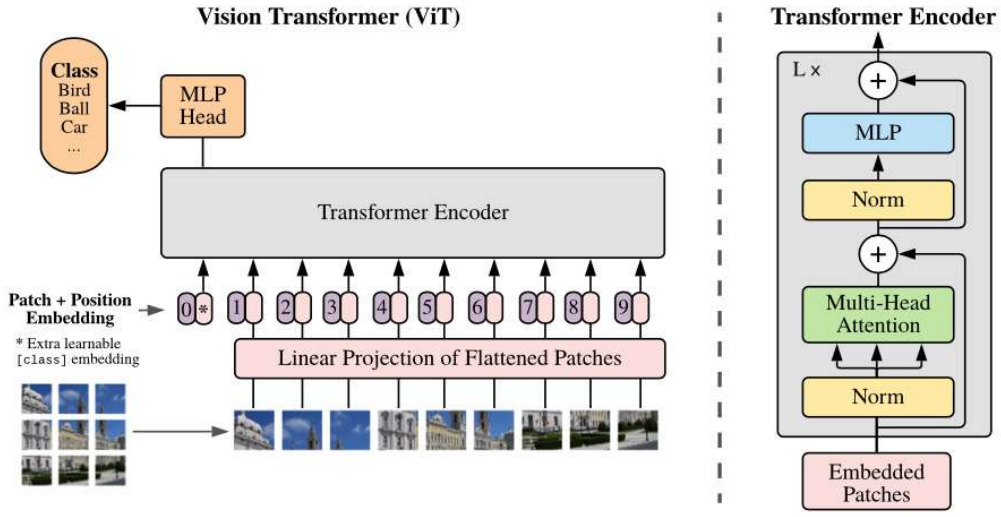
Principe des shortcus sur 34 layers

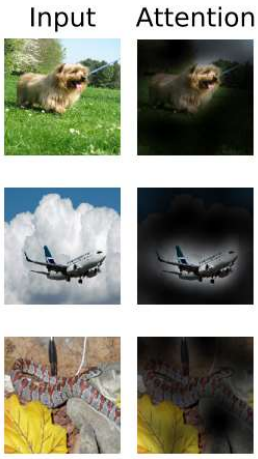
layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
		3×3 max pool, stride 2				
conv2_x	56×56	[3×3, 64; 3×3, 64] ×2	[3×3, 64; 3×3, 64] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3	[1×1, 64; 3×3, 64; 1×1, 256] ×3
conv3_x	28×28	[3×3, 128; 3×3, 128] ×2	[3×3, 128; 3×3, 128] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×4	[1×1, 128; 3×3, 128; 1×1, 512] ×8
conv4_x	14×14	[3×3, 256; 3×3, 256] ×2	[3×3, 256; 3×3, 256] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×6	[1×1, 256; 3×3, 256; 1×1, 1024] ×23	[1×1, 256; 3×3, 256; 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512; 3×3, 512] ×2	[3×3, 512; 3×3, 512] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3	[1×1, 512; 3×3, 512; 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Architectures et resnet-50

II.2) L'algorithme du POC : le VIT

Architecture VIT

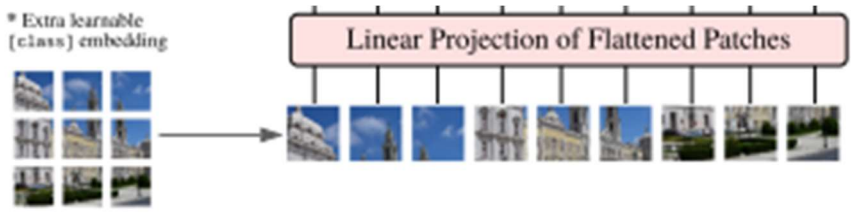




Le Principe du Vision Transformer est une des dernières ruptures technologiques en termes d'algorithme, et des améliorations ne cessent d'y être apportées (prepossessing, transformer encoder, layers…)

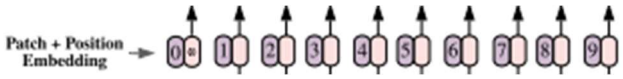
Basé sur le principe de mécanisme d'attention permettant de cerner les informations pertinentes il est multimodal et peut travailler sur différents types de data (vidéos, images, sons…), le principe étant d'avoir une approche par patches en réduisant les dimensions, permettant de conserver les informations pertinentes et le sens, de passer par un Transformer Encoder avec différentes couches qui va extraire les features pertinentes, et enfin de réaliser une classification avec un module MLP.

1) Le preprocessing



Il s'agit de splitter l'image en patches, de les convertir en feature de dimension 1D avec une projection linaire : on produit ainsi des vecteurs z.

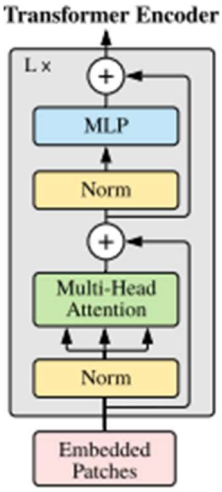
2) Le position embedding



Le vision Transformer connait l'arrangement des séquences pour le training, il existe plusieurs fonctions à appliquer pour cela (sinusoïdale, learnable embedding, rotary embedding…)

3) Le Transformer encoder

Il est composé de 2 blocs clefs : self-attention et MLP



a) Le bloc sel-attention

Basé sur le mécanisme d'attention :

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Il fait travailler plusieurs mécanismes d'attention en parallèle avant de les concaténer

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

b) Le bloc MLP

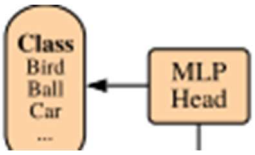
Il contient 2 couches GLU d'activation qui retournent Q, K, V.

c) Les couches de normalisation

Situées après la couche de self attention et MLP, elle permet de réduire le temps d'entrainement et de stabiliser le modèle.

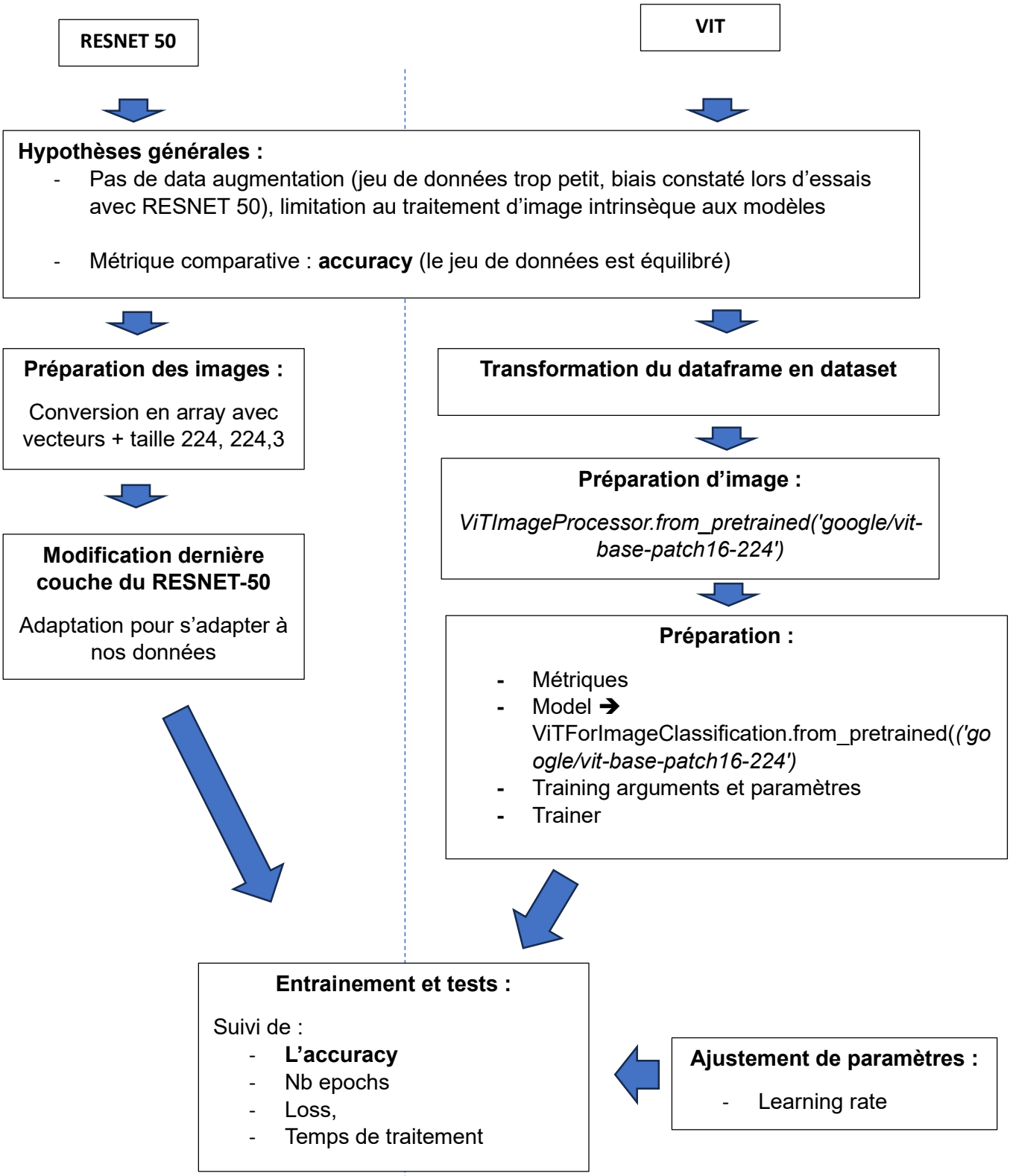
4) La prédiction finale

Il s'agit d'un dernier bloc MLP qui permet de classer l'image



III) La modélisation

III.1) Méthodologie de modélisation et optimisation



IV) Synthèse des résultats

IV.1) Résultats sur l'accuracy

	RESNET-50		VIT	
EPOCH	**Accuraccy val**	**Loss val**	**Accuraccy val**	**Loss val**
1	**0.539**	1.374	**0.714**	1.583
2	**0.666**	1.048	**0.785**	1.344
3	**0.714**	1.006	**0.809**	1.153
4	**0.698**	1.069	**0.809**	1.046
5	**0.682**	1.059	**0.785**	1.013

Malgré le jeu de donnée qui est très petits (210 images), les 2 algorithmes atteignent des résultats acceptables, notamment le VIT qui arrive à atteindre un accuracy à **0.809**.

D'un point de vue performance de classification, le VIT est plus performant et démontre son réel intérêt pour de la classification.

IV.2) Résultats temps de traitement

	RESNET-50	VIT
TEMPS TOTAL	**00 : 01 : 29**	**00 : 12 : 32**

Pour les temps de traitement, le modèle RESNET-50 est dix fois plus rapide que le modèle VIT

IV.3) Conclusion

Ces 2 algorithmes sont très performants et leur utilisation dépendra de l'usage que l'on veut en faire.

En termes de réactivité on préfèrera le RESNET, et en termes de précision on préfèrera le VIT.

V) Feature importance

Après une recherche approfondie, il s'avère qu'il n'existe pas vraiment de méthode pour mesurer l'impact des features importance globale et locale dans le domaine de la vision.

En effet ce qui pratique consiste plutôt à récupérer une couche d'attention et d'afficher les images avec une prédiction, permettant de visualiser les effets induits par le transformer.

VI) Les limites et les améliorations possibles

Les résultats du nouvel algorithme sont très encourageant, cependant il y a certaines limites et améliorations possibles.

Limites :

- Le temps de traitement est très long, et cela a forcément un impact financier.

Améliorations possibles :

- Elargir le jeu de données
- Affiner le Learning rate
- Réaliser le traitement en parallèle sur plusieurs machines