**Note méthodologique : preuve de concept**

## **Dataset retenu**

## Le jeu de données utilisé provient du projet n°6 qui s’intitule :

## « **Classifier automatiquement des biens de consommation** ».

## Ce dataset est composé de **1 050** images accompagnées de descriptions textuelles, issues du site e-commerce (place de marché) de type Consumer to Consumer (C2C), où les vendeurs proposent des articles aux acheteurs.

## Chaque article est classé, à l’aide de méthodes de traitement de données visuelles et textuelles (Computer Vision & NLP), dans l’une des **7 catégories** suivantes :

## Home Furnishing

## Baby Care

## Watches

## Home Decor & Festive Needs

## Kitchen & Dining

## Beauty and Personal Care

## Computer

## Les **images** représentent les produits mis à la vente et sont accompagnées de descriptions textuelles.

## Le dataset est équilibré : il y a 150 images reparties dans chaque catégorie.

## Dans la présente note méthodologique seules les **données visuelles sont exploitées**.

## Nous aborderons donc exclusivement l’approche **Vision par Ordinateur** souvent nommée « Computer Vision ».

## Nous division les images en trois sous-ensembles afin d’obtenir :

## **Un jeu de données d’entraînement,**

## **Un jeu de données de validation,**

## **Un jeu de données de test,**

## Les dimensions des images ont été normalisées à 224×224 pixels afin de correspondre au format d'entrée requis par les modèles de vision par ordinateur (tels que VGG16 ou ViT).

## **Les concepts de l’algorithme récent**

## L’algorithme retenu pour cette preuve de concept est le :

## **Vision Transformer (ViT**).

## Il s’agit d’une architecture de Deep learning introduite par Google Research en 2021, dans l’article intitulé : « An Image is Worth 16×16 Words : Transformers for Image Recognition at Scale ».

Traduction en Français : « **Une image vaut 16×16 mots : des Transformers pour la reconnaissance d’images à grande échelle** »

## Une image contenant texte, capture d’écran, Police Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Police

Le contenu généré par l’IA peut être incorrect.

## Contrairement aux CNN classiques (comme VGG16), le ViT adopte une approche inspirée des Transformers utilisés en NLP (Natural Language Processing)

## Les images sont découpées en petits patches (par exemple 16x16 pixels), chacun est ensuite aplati et transformé en un vecteur d'embedding, analogue à un "mot" dans une phrase.

## Ces vecteurs sont enrichis avec un « positional encoding » avant d’être traités par une série de blocs Transformer (self-attention + MLP).

## Le Vision Transformer « ViT » repose principalement sur le mécanisme de self-attention, qui permet à chaque patch de l’image de "regarder" tous les autres patches, pour identifier ainsi **les relations globales**.

## Cela offre une meilleure contextualisation spatiale que les convolutions locales.

Une image contenant texte, Appareils électroniques, capture d’écran, logiciel

Le contenu généré par l’IA peut être incorrect.

## **Les avantages du « ViT » sont nombreux :**

## Meilleure scalabilité avec de grandes quantités de données.

## Très bonne capacité de généralisation.

## Possibilité d'interprétation via les poids d’attention, permettant de visualiser les zones d’attention du modèle.

*Présentez, en 2 pages maximum, les principes de fonctionnement du nouvel algorithme.*

**La modélisation**

**Méthodologie de modélisation**

La démarche de modélisation repose sur l’utilisation du modèle **Vision Transformer (ViT)**, pré-entraîné sur ImageNet et accessible via la bibliothèque timm.  
ViT est une déclinaison des Transformers initialement conçus pour le NLP (comme BERT ou GPT-2), appliquée ici à la vision par ordinateur.

Le modèle utilisé a été **fine-tuné** sur un jeu de données de 1 050 images réparties de manière équilibrée entre 7 catégories de biens de consommation.  
Le fine-tuning consiste à adapter les **poids d’un modèle pré-entraîné** à un nouveau jeu de données en modifiant uniquement les dernières couches (ici la tête de classification).

Le dataset a été découpé en trois sous-ensembles :

* **70 %** pour l’entraînement
* **15 %** pour la validation
* **15 %** pour le test

Les images ont été redimensionnées à **224×224 pixels**, puis normalisées, conformément aux prérequis du modèle.

**Métrique d’évaluation retenue**

Pour évaluer les performances du modèle, nous avons retenu plusieurs métriques classiques de classification multi-classes :

* **Accuracy** : proportion d’étiquettes correctement prédites sur l’ensemble du jeu de test
* **F1-score macro** : moyenne harmonique entre précision et rappel sur toutes les classes
* **Matrice de confusion** : pour analyser les erreurs par classe

Ces métriques permettent de mesurer non seulement les performances globales, mais aussi la **capacité du modèle à bien traiter chaque catégorie individuellement**.

Le modèle a été entraîné pendant **10 epochs (ittération)**, avec un **batch size de 32** et une fonction de perte CrossEntropyLoss, adaptée à une tâche de classification.  
L’optimisation a été assurée avec l’algorithme **AdamW**, avec un taux d’apprentissage initial de 2e-5.

Nous avons suivi les pratiques standards d’optimisation :

* Suivi de la **loss d’entraînement**
* Validation croisée pour **surveiller le surapprentissage**
* Évaluation finale sur un jeu de test **non vu pendant l’entraînement**

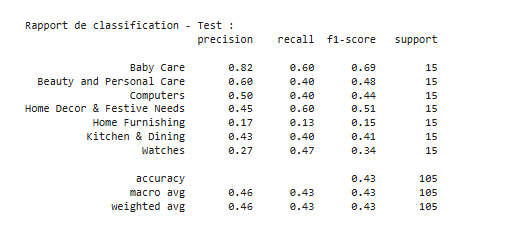
Enfin, une analyse d’**attention moyenne** a été réalisée en sortie de ViT, permettant d’identifier visuellement les zones d’image les plus discriminantes selon le modèle, contribuant ainsi à l’**interprétabilité** du réseau.

**Une synthèse des résultats**

L’objectif de cette expérimentation était de comparer les performances d’un modèle classique basé sur CNN (VGG16) à celles d’un modèle récent de type Transformer, le Vision Transformer (ViT), dans une mission de classification de biens de consommation (réf projet #6 Classifiez automatiquement des biens de consommations).

Les performances des deux modèles

***Rapport de classification CNN VGG16***



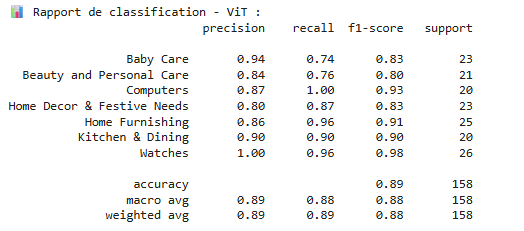
**1. Précision globale :**

Le modèle **ViT** surpasse largement le CNN en termes d’accuracy :  
**89 % contre 46 %**, soit plus du double en performance de prédiction globale.

**2. F1-score macro :**

Le **f1-score macro** reflète la capacité du modèle à bien classer **toutes les classes, même les moins représentées**. ViT obtient un **f1-macro de 0.88**, contre **0.43** pour le CNN.

**Rapport de classification ViT**



**3. Robustesse inter-catégories :**

L’analyse des matrices de confusion montre que :

* Le **CNN** confond plusieurs classes proches visuellement (ex : "Home Decor" avec "Home Furnishing").
* **ViT** présente **des prédictions beaucoup plus nettes et précises**, avec une diagonale dominante dans la matrice.

CNN

Une image contenant texte, capture d’écran, diagramme, Rectangle

Le contenu généré par l’IA peut être incorrect.

Vit

Une image contenant texte, capture d’écran, diagramme, Parallèle

Le contenu généré par l’IA peut être incorrect.

**4. Classes les mieux prédites :**

* Avec CNN : seule la classe "Baby Care" dépasse 0.6 de f1-score.
* Avec ViT : **toutes les classes dépassent 0.80**, avec une précision parfaite sur "Watches" (F1 = 1.00).

**Interprétabilité du modèle ViT**

Grâce à l’analyse des poids d’attention du Transformer, ci-dessous une carte d’attention moyenne a été générée. Elle révèle que le modèle se concentre naturellement sur les zones centrales des images, là où se trouvent souvent les objets clés, renforçant la lisibilité et la transparence du modèle.

Une image contenant Caractère coloré, capture d’écran, motif, carré

Le contenu généré par l’IA peut être incorrect.

**CONCLUSION**

Cette preuve de concept démontre clairement l’intérêt d’adopter des architectures récentes comme **ViT** dans les tâches de classification d’images :

* Meilleures performances globales
* Robustesse inter classes accrue
* Interprétabilité intégrée via les mécanismes d’attention
* Meilleure scalabilité pour des datasets complexes

En conclusion, **ViT constitue une avancée majeure** pour les problématiques de Computer Vision, et son intégration dans des pipelines de traitement d’images peut significativement améliorer la qualité des prédictions.

**1. Feature Importance locale (au niveau d'une image)**

Pour comprendre comment le modèle ViT prend ses décisions sur **une image individuelle**, nous avons utilisé une technique appelée **Attention Rollout**, qui exploite les **poids d’attention des couches internes** du modèle.

Cette méthode permet de visualiser **où le modèle "regarde" réellement** lorsqu’il effectue une prédiction. Chaque image est découpée en **patches de 16×16 pixels**, et le modèle attribue un niveau d’attention à chacun d’eux.

Sur les exemples testés, la carte d’attention locale montre une **forte concentration de l’attention sur les zones centrales** de l’image, là où se situent souvent les produits. Cela confirme que le modèle apprend à focaliser son attention sur les régions visuellement discriminantes.

Une **heatmap est superposée à l’image** pour illustrer clairement les régions d’intérêt du modèle. Ces visualisations permettent une **interprétation fine des décisions** et renforcent la confiance dans le modèle.

*Carte d’attention locale : le modèle se concentre sur les régions centrales de l’image testée*

**2. Feature Importance globale (au niveau de l’ensemble du dataset)**

En complément de l’analyse locale, nous avons généré une **carte d’attention moyenne** sur un ensemble de plusieurs images issues du jeu de test.

Cette **feature importance globale** permet de dégager des tendances sur le comportement du modèle. La moyenne des poids d’attention du token [CLS] vers tous les patches révèle :

* Une **attention systématiquement concentrée au centre de l’image**
* Une **faible importance accordée aux bords**, ce qui correspond à une stratégie d’attention efficace pour la classification d’objets centrés

Cette carte moyenne confirme que le modèle **extrait des régularités visuelles à grande échelle** et qu’il a appris **une priorisation spatiale cohérente** avec le contenu des images du dataset.

L’analyse des poids d’attention fournit une forme d’explicabilité intégrée au modèle ViT.  
Elle permet de vérifier que le modèle se concentre sur les bonnes zones, d’identifier d’éventuelles erreurs de prédiction injustifiées, et de justifier ses décisions auprès d’un public non technique.

En résumé, l’approche par Transformer ne se contente pas d’être performante : elle est aussi plus interprétable que les architectures CNN classiques.

*Nous avons utilisé 3 images pour générer la carte d’attention.*

*L’image affichée ci-dessus montre une attention* ***systématiquement centrée***

**Les limites et les améliorations possibles**

1. **Dépendance à un modèle pré-entraîné**  
   Le Vision Transformer (ViT) utilisé repose sur un entraînement préalable sur le dataset ImageNet. Cela peut engendrer un **biais de généralisation** si les caractéristiques visuelles des produits diffèrent fortement de celles du corpus d’origine.
2. **Absence d’exploitation des données textuelles**  
   Dans cette preuve de concept, seules les **images** ont été utilisées. Or, chaque produit est aussi accompagné d’une **description textuelle riche**, qui pourrait être exploitée pour améliorer la précision du classement.
3. **Temps de calcul et mémoire**  
   Les modèles Transformer appliqués aux images, bien que performants, sont **plus gourmands en ressources** que les CNN classiques. Cela peut poser des contraintes en production ou sur de gros volumes.
4. **Patch uniformes**  
   Le découpage de l’image en patches 16×16 est fixe, ce qui peut parfois **perdre des détails fins** pour des objets petits ou mal centrés.

**Pistes d’amélioration envisagées**

1. **Fusion Multimodale (Image + Texte)**  
   Intégrer une modélisation conjointe des images et des descriptions via des architectures **multimodales**, comme **CLIP** ou des Transformers bimodaux (ex. : ViLT), permettrait de **capturer des signaux complémentaires**.
2. **Data Augmentation ciblée**  
   Appliquer des techniques comme **MixUp**, **CutMix**, ou des **transformations de style** permettrait d’augmenter la diversité du jeu d’entraînement, réduisant le surapprentissage et améliorant la robustesse.
3. **Fine-tuning sur plus d’epochs avec plan d’apprentissage**  
   Optimiser le taux d’apprentissage avec des **schedulers**, entraîner sur plus d’epochs avec early stopping, ou utiliser la validation croisée pour affiner les hyperparamètres.
4. **Explicabilité avancée**  
   En complément des cartes d’attention, intégrer des outils d’explicabilité visuelle comme **Grad-CAM pour ViT** ou **LIME sur les sorties du modèle** permettrait une **interprétation plus fine** des décisions, même en cas d’erreur.

**En résumé :**

L’approche basée sur ViT démontre des performances élevées, mais elle peut être renforcée par des optimisations techniques et l’intégration d’autres modalités (texte).  
L’objectif serait de bâtir une **chaîne de classification plus robuste, multimodale et interprétable**.

**Source**

**Technique** : Vision Transformer (ViT)

**Date** : Introduit par Google Research en **2021**

* **Article de référence** :

Dosovitskiy et al., *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, publié à ICLR 2021

Disponible sur **arXiv** : <https://arxiv.org/abs/2010.11929>

* **Mentionné sur les site** :

ArXiv, PapersWithCode, HuggingFace blog, MIT Technology Review, Towards Data Science

**Référence bibliographique**

* Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929.
* Google AI Blog. (2021). Exploring ViT for Image Classification. <https://ai.googleblog.com/2020/12/transformers-for-image-recognition-at.html>
* PapersWithCode: Vision Transformer. <https://paperswithcode.com/method/vision-transforme>