Note méthodologique : preuve de concept

Sources consultées et points clés :

1. **Article : "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (Google Research, 2021)**
    - **Résumé :** Cet article introduit ViT, un modèle basé sur les Transformers pour la vision par ordinateur. Il explique comment ViT divise les images en patches, utilise des embeddings et des mécanismes d'attention.
    - **Détails mathématiques :**

- Encodage des patches : $z_0 = [x_{[CLS]}; E(x_1); ...; E(x_N)] + E_{pos}$
  Où E est l'embedding linéaire et E_{pos} l'encodage positionnel.
- Auto-attention :
  $$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$
  Où Q, K, V sont respectivement les matrices de requêtes, clés et valeurs.
- Classification : Une couche dense est ajoutée après le token $[CLS]$ pour la prédiction.

2. Blog : "Hugging Face's Vision Transformer Guide" (2023)

Résumé : Présentation des implémentations open-source de ViT sur des tâches variées.

Points clés :

Utilisation d'outils pré-entraînés tels que transformers et datasets de Hugging Face.

Comparaison entre ViT et CNN sur des petits jeux de données : CNN performe mieux sans fine-tuning.

3. Conférence : CVPR 2023 – "Vision Transformers: A Survey"

Résumé : Cette présentation synthétise les dernières avancées sur ViT, incluant des variantes comme DeiT et Swin Transformer.
Points clés :

- DeiT améliore les performances de ViT sur des petits jeux de données grâce à la distillation.
- Swin Transformer introduit une attention hiérarchique pour réduire le coût computationnel.

1. Dataset retenu

Le dataset utilisé pour cette preuve de concept est constitué de 1 050 images réparties en 7 catégories distinctes, correspondant aux classes définies dans la colonne main_category. Les images ont été extraites à partir d'une base de données existante et sont stockées dans un répertoire structuré.

Chaque image est associée à des métadonnées comprenant son identifiant unique, sa catégorie réelle, et d'autres attributs. Les catégories sont équilibrées afin d’assurer une évaluation équitable des modèles testés.

Les données ont été prétraitées en redimensionnant les images à une taille de 224x224 pixels, conformément aux exigences des algorithmes modernes comme Vision Transformer (ViT).

2. Les concepts de l’algorithme récent : *Vision Transformer (ViT)*

2.1. Introduction

Vision Transformer (ViT) est un modèle basé sur l'architecture des Transformers, initialement conçue pour le traitement du langage naturel. Introduit par Dosovitskiy et al. (2020), ViT adapte le Transformer pour traiter des images en les décomposant en **patches** de taille fixe, comme des "mots" dans un texte.

2.2. Principe de fonctionnement

1. **Patchification des images :**
   Une image est divisée en patches de taille fixe (par exemple, 16x16 pixels). Ces patches sont aplatis en vecteurs unidimensionnels.
2. **Encodage des patches :**
   Chaque patch est enrichi avec un vecteur d’encodage positionnel pour conserver la structure spatiale de l'image.
3. **Passage dans le modèle Transformer :**
   Les vecteurs sont traités par plusieurs couches du Transformer, qui applique des mécanismes d’**attention multi-tête** et des couches de feed-forward pour capturer des relations globales dans l'image.
4. **Classification :**
   Un vecteur spécial appelé *class token* est utilisé pour agrégat les informations globales et est passé dans une couche dense pour produire la prédiction finale.

2.3. Avantages de ViT

- **Efficacité sur des ensembles de données volumineux** : ViT excelle lorsque de larges bases de données sont disponibles.
- **Capture de relations globales** : Grâce au mécanisme d'attention, ViT modélise efficacement les relations à longue distance entre les pixels.

La modélisation

1. Méthodologie de modélisation

1.1 Préparation des données

Nous avons commencé par charger les données depuis un fichier CSV contenant les informations des produits et les liens vers leurs images. Les images ont été extraites et prétraitées pour être adaptées aux modèles de classification.

- **Chargement des données** : Nous avons utilisé pandas pour charger les données depuis un fichier CSV.
- **Prétraitement des images** : Les images ont été redimensionnées à une taille de 224x224 pixels, puis normalisées pour être compatibles avec les modèles pré-entrainés.

1.2 Extraction des caractéristiques avec Vision Transformer (ViT)

Le modèle **Vision Transformer (ViT)** a été utilisé pour extraire les caractéristiques des images. Contrairement aux approches classiques basées sur les descripteurs de caractéristiques comme SIFT ou les couches convolutionnelles de VGG-16, ViT adopte une approche basée sur des "patches" d'image, transformant chaque patch en un vecteur de caractéristiques.

- Nous avons utilisé le modèle pré-entrainé **ViT (google/vit-base-patch16-224-in21k)** et le **ViTFeatureExtractor** pour préparer les images.
- Les caractéristiques des images ont été extraites en calculant la moyenne des activations de la dernière couche cachée du modèle ViT, produisant un vecteur de caractéristiques par image.

1.3 **Réduction de dimension**

Afin de rendre les données plus maniables et d'améliorer les performances des étapes suivantes (clustering et visualisation), une réduction de dimension a été effectuée sur les caractéristiques extraites :

- **PCA** a été utilisé pour conserver 99 % de la variance des caractéristiques.
- **t-SNE** a été utilisé pour une réduction de dimension à 2D, ce qui a permis de visualiser les données dans un espace de caractéristiques réduit.

1.4 **Clustering avec KMeans**

Une fois les caractéristiques réduites en dimension, nous avons appliqué le **clustering KMeans** pour regrouper les images en 7 clusters, en utilisant le nombre de clusters correspondant au nombre de catégories réelles dans les données.

- Le modèle KMeans a été exécuté avec un nombre de clusters égal au nombre de catégories réelles (len(data["main_category"].unique())).
- L'**ARI (Adjusted Rand Index)** a été calculé pour évaluer la qualité du clustering en comparant les clusters obtenus avec les étiquettes réelles des catégories.

1.5 **Évaluation et Matrice de confusion**

Pour évaluer la qualité du clustering, une **matrice de confusion** a été calculée entre les clusters et les catégories réelles. La matrice a été ensuite transformée pour ajuster les clusters aux catégories réelles en utilisant la correspondance des indices de la matrice de confusion.

- Un **rapport de classification** a été généré pour obtenir des métriques détaillées telles que la précision, le rappel, et la f-mesure.

2. **Comparaison avec SIFT et VGG-16**

2.1 **Méthodologie avec SIFT**

Pour la méthode basée sur **SIFT** (Scale-Invariant Feature Transform), nous avons extrait les descripteurs SIFT des images, puis avons effectué un clustering KMeans sur ces descripteurs pour créer des groupes d'images similaires. Ce processus est très différent de l'approche ViT, qui repose sur l'extraction des caractéristiques

globales à partir des images.

2.2 **Méthodologie avec VGG-16**

Dans l'approche basée sur **VGG-16**, nous avons utilisé un modèle pré-entrainé VGG-16 pour extraire les caractéristiques des images. Ces caractéristiques ont ensuite été réduites en dimension à l'aide de PCA et t-SNE, puis un clustering KMeans a été appliqué de manière similaire à l'approche ViT. Cependant, contrairement à ViT, VGG-16 est un modèle basé sur des convolutions, qui extrait des informations plus locales que ViT.

2.3 **Comparaison des performances**

Pour comparer les trois approches (ViT, SIFT, et VGG-16), nous avons utilisé les mêmes métriques d'évaluation :

- **ARI (Adjusted Rand Index)** pour mesurer la similarité entre les clusters obtenus et les étiquettes réelles.
- **Matrice de confusion** pour visualiser comment les clusters correspondent aux catégories réelles.
- **Rapport de classification** pour évaluer les performances détaillées de chaque méthode.

Résultats attendus :

- ViT devrait avoir un avantage grâce à sa capacité à extraire des caractéristiques globales plus robustes, mais l'absence de fine-tuning pourrait limiter ses performances par rapport à une approche entièrement optimisée.
- SIFT pourrait offrir des résultats compétitifs pour des images avec des caractéristiques distinctes, mais pourrait être moins performant sur des images complexes ou variées.
- VGG-16 devrait fournir de bons résultats grâce à son entraînement sur un large ensemble de données, mais son approche convolutionnelle peut être moins efficace pour des images ayant des variations complexes.

3. **Conclusion**

Cette approche a permis de tester ViT dans le cadre d'une tâche de classification

d'images, en le comparant avec des méthodes classiques comme SIFT et VGG-16, en utilisant des techniques de réduction de dimension et de clustering pour évaluer la qualité du modèle. Les résultats montrent les avantages et les limitations de chaque approche, et la nécessité d'optimiser les modèles pour obtenir des performances optimales, notamment en fine-tunant ViT et VGG-16 pour la tâche spécifique.

Synthèse des Résultats et Comparaison des Méthodes

La comparaison des résultats des trois modèles — **SIFT**, **VGG-16**, et **ViT (Vision Transformer)** — a été réalisée en utilisant des métriques de qualité telles que l'ARI (Adjusted Rand Index), l'accuracy, et le F1-score moyen, ainsi qu'une analyse des matrices de confusion. Voici une synthèse des performances :

Tableau de Résultats Comparatifs

Métrique	SIFT	VGG-16	ViT
ARI	0.0019	0.439	0.054
Accuracy	0.14	0.65	0.30
F1-Score Moyen	0.14	0.66	0.27

Analyse des Résultats

1. **SIFT (Méthode Basique)** :
    - **ARI :** Très faible (0.0019), indiquant un faible regroupement entre les catégories réelles et prédictions.
    - **Accuracy :** Faible (14%) avec des scores F1 uniformément bas, ce qui montre que la méthode SIFT n’est pas adaptée pour capturer la complexité visuelle des images.
    - **Matrice de Confusion :** Les confusions sont élevées entre toutes les catégories. Les prédictions sont réparties de manière presque aléatoire.
2. **VGG-16 (Technique Avancée)** :
    - **ARI :** Significativement plus élevé (0.439), indiquant une meilleure

capacité de clustering par rapport à SIFT.

- **Accuracy :** Bonne performance (65%), ce qui montre une capacité notable à distinguer les catégories.
- **F1-Score :** Le score le plus élevé (0.66), avec des performances robustes sur des catégories comme **Watches** et **Beauty and Personal Care**.
- **Matrice de Confusion :** Les catégories comme **Watches** (classe 6) et **Beauty and Personal Care** (classe 1) sont bien prédites. Cependant, certaines confusions subsistent entre les catégories visuellement similaires.

3. **ViT (Vision Transformer)** :

    - **ARI :** Faible (0.054), ce qui suggère que le modèle a du mal à regrouper les images selon les catégories réelles.
    - **Accuracy :** Modérée (30%), mais meilleure que SIFT.
    - **F1-Score :** Moyenne (0.27), montrant que certaines catégories comme **Watches** et **Baby Care** ont des performances acceptables, tandis que d'autres, comme **Computers**, sont mal classées.
    - **Matrice de Confusion :** Bien que le modèle capte certaines relations, des confusions importantes persistent, en particulier dans les catégories complexes.

Conclusion

Les résultats montrent que **VGG-16**, basé sur des réseaux de neurones convolutifs pré-entraînés, surpasse de manière significative les deux autres méthodes. **SIFT**, une méthode classique de traitement d'images, est insuffisante pour des données complexes et des tâches multi-catégories. Bien que **ViT** soit une technique récente et prometteuse, elle nécessite davantage d'ajustements et d'optimisations pour obtenir des résultats compétitifs dans ce contexte.

En résumé, pour la tâche de classification d’images multi-catégories :

- **VGG-16** est la méthode recommandée, offrant un bon compromis entre précision et efficacité.
- **SIFT** peut être utilisé pour des analyses exploratoires simples, mais n’est pas

viable pour des tâches complexes.

- **ViT** pourrait être exploré davantage avec des hyperparamètres ajustés ou un fine-tuning pour des résultats améliorés.

Ces résultats renforcent l'importance des modèles pré-entraînés et des réseaux de neurones convolutifs pour des tâches complexes de vision par ordinateur.

Limites et Améliorations Envisageables de l'Approche ViT (Vision Transformer)

Limites Identifiées

1. **Données d'Entraînement Limitées :**
    - Les Vision Transformers nécessitent de grandes quantités de données annotées pour une performance optimale. Avec un dataset limité comme le nôtre (1050 images), le modèle risque de sous-apprendre et de ne pas capturer les nuances entre les catégories.
2. **Complexité Computationnelle :**
    - Les ViT ont une architecture basée sur des self-attentions qui deviennent coûteuses en temps et en ressources pour des images à haute résolution.
3. **Confusions Inter-Catégories :**
    - Les résultats montrent des confusions significatives dans certaines catégories, notamment pour des classes visuellement similaires ou mal représentées dans les données.
4. **Faible Interprétabilité :**
    - Contrairement aux CNN (comme VGG-16), où les cartes de convolution permettent de visualiser les régions d'intérêt, ViT manque d'outils intégrés pour interpréter facilement les décisions du modèle.
5. **Sensibilité aux Hyperparamètres :**
    - L'efficacité des ViT dépend fortement du choix des hyperparamètres (taille des patches, nombre de couches, etc.), ce qui peut nécessiter des expérimentations coûteuses.

Améliorations Possibles

1. **Augmentation des Données :**
    - Mettre en œuvre des techniques avancées de data augmentation (e.g., mixup, augmentation par GANs) pour enrichir le dataset sans collecter manuellement davantage de données.
2. **Fine-Tuning du Modèle Pré-Entrraîné :**
    - Utiliser des modèles ViT pré-entraînés sur des datasets massifs comme ImageNet et les adapter à notre tâche spécifique grâce au fine-tuning. Cela permet de profiter des représentations déjà apprises.
3. **Réduction de la Résolution et des Patches :**
    - Réduire la taille des images ou des patches pour diminuer la complexité computationnelle tout en conservant les caractéristiques discriminantes.
4. **Exploration de ViT Hybrides :**
    - Tester des architectures hybrides combinant les avantages des CNN (extraction locale de caractéristiques) et des Transformers (modélisation des dépendances globales).
5. **Meilleure Sélection des Hyperparamètres :**
    - Mettre en œuvre des techniques d'optimisation comme la recherche bayésienne ou des algorithmes génétiques pour sélectionner les hyperparamètres de manière efficace.
6. **Explications Locales avec LIME ou SHAP :**
    - Intégrer des méthodes d'explicabilité comme LIME (Local Interpretable Model-Agnostic Explanations) ou SHAP (SHapley Additive ExPlanations) pour rendre les décisions du modèle plus interprétables.
7. **Enrichissement des Caractéristiques Contextuelles :**
    - Ajouter des métadonnées (e.g., description textuelle ou marque) pour fournir un contexte supplémentaire au modèle et améliorer les performances globales.
8. **Optimisation Matériel et Logiciel :**

- Utiliser des frameworks optimisés comme Hugging Face Transformers ou ONNX Runtime pour réduire les temps d'exécution.

Conclusion :

Les ViT représentent une avancée significative pour les tâches de vision par ordinateur, mais leur performance dépend fortement de la qualité des données et des configurations. En intégrant les améliorations proposées, il serait possible d'accroître à la fois la performance et l'interprétabilité du modèle, rendant l'approche plus robuste et adaptée à des contextes industriels.

Référence bibliographique:

Vision Transformer : Dosovitskiy et al. (2020), An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. https://arxiv.org/abs/2010.11929.