Note méthodologique : preuve de concept avec le modèle CLIP

Dataset retenu

Pour ce projet, le dataset retenu provient du projet 6 "Classifier automatiquement des biens de consommation". Place de Marché, une entreprise anglophone, souhaite automatiser l'attribution des catégories de produits sur sa marketplace e-commerce, une tâche actuellement réalisée manuellement par les vendeurs. Le dataset se compose de 1 fichier CSV et d'un dossier contenant 1050 images. Chaque enregistrement dans le fichier CSV correspond à un article et inclut des informations telles que le nom du produit, sa description, et un lien vers l'image de l'article. Les données sont variées, sans doublons, et ne comportent aucune contrainte de propriété intellectuelle.

Les concepts de l'algorithme récent

Présentation de CLIP

CLIP (Contrastive Language-Image Pre-Training), développé par OpenAI, est conçu pour comprendre les relations entre le texte et les images. Entraîné sur un large éventail de paires (image, texte) issus d'Internet, CLIP vise à prédire le texte le plus pertinent pour une image donnée, sans optimisation directe pour une tâche spécifique, imitant les capacités "zero-shot" des modèles GPT-2 et GPT-3. Il résout plusieurs problèmes des approches actuelles en apprentissage profond, comme le coût élevé de création pour un ensemble restreint et l'adaptation laborieuse à de nouvelles tâches (Radford et al, 2021). Le modèle utilise un encodeur ResNet50 ou un transformateur de vision, et un encodeur de texte basé sur des mécanismes d'auto-attention pour capturer les relations complexes entre les modalités texte et image. Resnet50 est un type de réseau de neurones profond de 50 couches, utilisé pour la reconnaissance d'image.

Fonctionnement de CLIP

CLIP utilise une méthode de pré-entraînement contrastif, apprenant à associer des paires (image, texte) en maximisant leur similarité. Le but est de créer des représentations vectorielles des images et des textes proches dans l'espace vectoriel lorsque les concepts se correspondent (par exemple, une image de chien et le texte "un chien"), et éloignées sinon. La fonction de perte contrastive optimise ce processus en rapprochant les vecteurs des paires correspondantes et en éloignant ceux des paires non correspondantes.

Pour utiliser CLIP dans des tâches de classification spécifiques, on commence par créer un ensemble de descriptions textuelles correspondant aux catégories de labels. CLIP encode ces descriptions textuelles en vecteurs, tout comme il encode les images du dataset.

Ce processus permet de créer un classificateur en comparant les vecteurs des images avec ceux des descriptions textuelles, facilitant ainsi l'association de chaque image à la catégorie la plus similaire.

Une des capacités principales de CLIP est la prédiction "zero-shot", où le modèle peut classer des images dans des catégories sans avoir été spécifiquement entraîné sur ces catégories. Pour cela, lorsqu'une nouvelle image est présentée, CLIP génère un vecteur pour cette image et le compare aux vecteurs des descriptions textuelles des catégories de labels créés précédemment. La catégorie dont le vecteur est le plus proche de celui de l'image est choisie comme prédiction. Cette méthode permet à CLIP de généraliser à des tâches nouvelles et arbitraires de classification d'images sans nécessiter de réentraînement, illustrant sa robustesse et sa flexibilité.

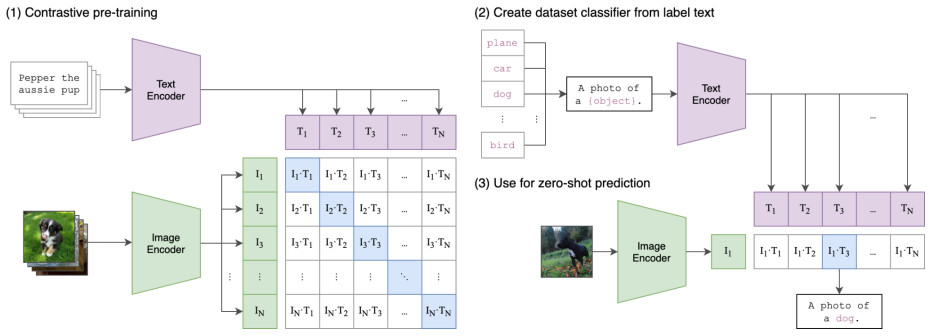

Figure 1 *- Concept de fonctionnement du modèle CLIP (Contrastive Language-Image Pre-Training)*

La modélisation

Dans ce projet, nous avons adopté deux approches distinctes pour la classification d'images de biens de consommation : une méthode basée sur VGG16 avec intégration de data augmentation et une approche basée sur CLIP. Les données comprennent un ensemble d'images associées à des étiquettes de catégorie principale, prétraitées pour la normalisation des images et l'encodage des étiquettes de catégorie, puis divisées en ensembles d'entraînement et de test.

Pour l'approche avec VGG16, nous utilisons un modèle pré-entraîné sur ImageNet pour l'extraction des caractéristiques visuelles des images. En parallèle, des techniques de data augmentation telles que la rotation, le zoom et le renversement horizontal sont appliquées aux images afin d'accroître la diversité des données d'entraînement. L'architecture du modèle a été enrichie par l'ajout de couches supplémentaires adaptées à la classification des biens de consommation. L'entraînement a été réalisé avec l'optimiseur Adam et une fonction de perte de catégorisation croisée, avec validation croisée pour évaluer les

performances tout au long du processus. Des ajustements de hyperparamètres comme le taux d'apprentissage et la taille du batch sont effectués à l'aide de techniques comme la validation croisée, tandis que des callbacks tels que Early Stopping et Model Checkpoint sont utilisés pour prévenir le surapprentissage et sauvegarder le meilleur modèle basé sur la perte de validation.

La seconde approche avec CLIP utilise un modèle basé sur le Transformer (ViT-B/32) pré-entraîné pour apprendre les correspondances entre le texte et les images. Les descriptions textuelles des catégories ont été converties en tokens à l'aide de CLIPProcessor. Les images ont été chargées et prétraitées pour être compatibles avec CLIP. Le modèle CLIP est pré-entraîné sur un ensemble de données volumineux incluant texte et images pour générer des représentations textuelles et visuelles cohérentes, évaluées ensuite par leur exactitude sur l'ensemble de test. Contrairement à VGG16, CLIP ne nécessite pas de techniques de data augmentation étant donné sa capacité à généraliser à partir d'un ensemble d'entraînement plus large et diversifié.

Les métriques d'évaluation retenues sont l'accuracy qui représente le pourcentage d'images correctement classifiées par le modèle parmi toutes les images dans l'ensemble de données de test. Nous utilisons aussi la précision (mesure de la proportion d'éléments correctement identifiés parmi ceux qui sont prédits comme appartenant à une classe spécifique), le rappel (mesure la proportion d'éléments appartenant à une classe spécifique qui sont correctement identifiés parmi tous les éléments qui appartiennent réellement à cette classe) et le F1-score (mesure de la précision globale du modèle qui est calculé comme la moyenne harmonique de la précision et du rappel). De plus, nous observons aussi la matrice de confusion, qui permet de visualiser la performance du modèle en détaillant le nombre de prédictions correctes et incorrectes pour chaque classe.

En résumé, cette méthodologie vise à évaluer et comparer deux approches de classification d'images en utilisant des modèles pré-entraînés et à déterminer celle qui offre les meilleures performances sur notre jeu de données spécifique de biens de consommation.

Une synthèse des résultats

Les résultats obtenus par deux approches distinctes, VGG16 avec data augmentation et CLIP pour la classification d'images, ont été comparés en termes de performances sur les ensembles d'entraînement et de test.

Pour VGG16, l'accuracy sur l'ensemble d'entraînement atteint 84.44%, illustrant sa capacité à apprendre efficacement les caractéristiques des données fournies. L'accuracy sur l'ensemble de test est de 74.67%, ce qui confirme la robustesse du modèle face à de nouvelles données, bien que légèrement inférieure à celle des ensembles d'entraînement :

Performances d'Entraînement :

- Précision : 91.33%
- Rappel : 78.07%
- F1-score : 84.23%

Performances de Test :

- Précision : 86.92%
- Rappel : 62.00%
- F1-score : 72.37%

CLIP se distingue par une méthode de pré-entraînement contrastif qui intègre des descriptions textuelles et des images pour la classification. Sur l'ensemble d'entraînement, CLIP atteint une accuracy de 82.33%, démontrant sa capacité à apprendre à partir d'un ensemble de données non étiquetées. L'accuracy sur l'ensemble de test est de 77.33%, ce qui montre une performance légèrement supérieure à celle de VGG16 sur cet ensemble spécifique, suggérant une meilleure capacité de généralisation

Performances d'Entraînement :

- Précision : 82.90%
- Rappel : 82.33%
- F1-score : 82.28%

Performances de Test :

- Précision : 78.40%
- Rappel : 77.33%
- F1-score : 77.63%

En analysant les résultats, CLIP semble surpasser légèrement VGG16 en termes de précision et de F1-score sur l'ensemble de test, indiquant sa capacité à généraliser efficacement à de nouvelles données sans nécessiter de techniques de data augmentation spécifiques. Cette capacité est attribuée à l'utilisation conjointe de représentations textuelles et visuelles, ce qui permet à CLIP de capturer des informations complexes et des nuances dans les images.

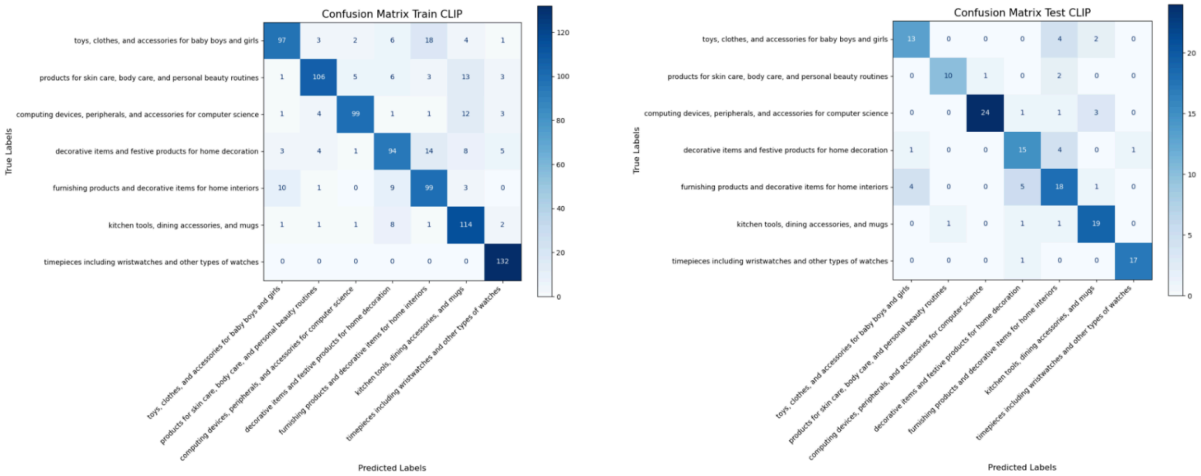

Figure 2 *- Matrice de confusion d'entraînement et de test du modèle CLIP (Contrastive Language-Image Pre-Training)*

Les matrices de confusion fournissent une compréhension plus détaillée des performances de la méthode CLIP. Ces mesures montrent comment chaque modèle se comporte dans la classification de différentes catégories, offrant ainsi des insights précieux pour le choix du modèle dans des applications réelles.

Dans cette étude comparative, CLIP se positionne comme une approche prometteuse pour la classification d'images, avec une précision légèrement supérieure à celle de VGG16 sur notre jeu de données spécifique. Cependant, le choix entre les deux méthodes dépendra des exigences spécifiques de précision, de flexibilité et des ressources disponibles. Ainsi, bien que VGG16 avec data augmentation montre des résultats impressionnants en termes de précision sur l'ensemble d'entraînement, sa capacité de généralisation est légèrement inférieure à celle de CLIP. CLIP, avec sa robustesse et ses performances équilibrées, constitue une alternative viable et efficace, particulièrement adaptée aux tâches nécessitant une bonne généralisation à de nouvelles données.

L’analyse de la feature importance globale et locale du nouveau modèle

L'analyse de la feature importance globale permet de comprendre l'influence relative de chaque classe sur les prédictions du modèle. Voici les scores obtenus suite à la mesure de similarité entre les embeddings d'images et les embeddings de texte générés par le modèle CLIP. Ces scores indiquent à quel point une image est proche, en termes d'embeddings, de la description textuelle correspondante :

```
Classe : toys, clothes, and accessories for baby boys and girls, Score moyen : 0.2610509991645813
Classe : products for skin care, body care, and personal beauty routines, Score moyen : 0.23423995077610016
Classe : computing devices, peripherals, and accessories for computer science, Score moyen : 0.24053122103214264
Classe : decorative items and festive products for home decoration, Score moyen : 0.2508518695831299
Classe : furnishing products and decorative items for home interiors, Score moyen : 0.26195669174194336
Classe : kitchen tools, dining accessories, and mugs, Score moyen : 0.2521899342536926
Classe : timepieces including wristwatches and other types of watches, Score moyen : 0.27515164017677307
```

Ces scores moyens indiquent à quel point chaque classe a contribué aux prédictions globales du modèle sur l'ensemble de tests. Une valeur plus élevée suggère une plus grande importance de cette classe dans les prédictions du modèle. Dans notre cas, c’est la classe "timepieces including wristwatches and other types of watches" et "furnishing products and decorative items for home interiors" qui obtiennent les scores de similarité moyens les plus élevés, ce qui indique que le modèle CLIP parvient à bien associer les images et les descriptions pour ces catégories.

L'analyse de la feature importance locale identifie les images individuelles qui ont le plus influencé les prédictions pour chaque classe. Pour chaque classe, les images suivantes ont été identifiées comme ayant les scores de similarité les plus élevés avec la classe correspondante :

Figure 3 *- Feature importance la plus élevée pour chaque catégorie avec CLIP (Contrastive Language-Image Pre-Training)*

Ces exemples illustrent visuellement quelles caractéristiques visuelles spécifiques ont été associées à chaque classe lors des prédictions du modèle. Le modèle CLIP montre une bonne capacité à associer des images et des descriptions textuelles à travers différentes catégories de produits. Les scores de similarité obtenus sont cohérents et montrent que certaines catégories, telles que les montres et les articles d'ameublement, sont mieux associées que d'autres.

Les limites et les améliorations possibles

En terme de limite dans l'utilisation de CLIP sur ce dataset, le dataset utilisé dans ce projet, bien que varié, reste relativement limité en taille avec seulement 1050 images. Cette quantité restreinte de données peut limiter la capacité du modèle à généraliser sur des ensembles de données plus vastes et variés. De plus, un déséquilibre éventuel entre les classes peut biaiser les performances du modèle en faveur des classes majoritaires. De plus, certaines catégories de produits peuvent présenter des caractéristiques visuelles et textuelles similaires,

rendant leur classification plus complexe. CLIP peut éprouver des difficultés à distinguer ces catégories étroitement liées sans une séparation claire. Enfin, l'utilisation de modèles de grande envergure comme CLIP peut être gourmande en ressources computationnelles. Le pré-traitement et l'entraînement nécessitent des infrastructures matérielles puissantes, ce qui peut constituer une barrière pour des implémentations à plus grande échelle.

Une des améliorations majeures consisterait à augmenter la taille du dataset en collectant davantage d'images et de descriptions textuelles. Un dataset plus grand et plus diversifié améliorerait la capacité de généralisation du modèle. De plus, bien que CLIP ne nécessite pas de techniques de data augmentation classiques, des augmentations spécifiques comme la génération de variations textuelles ou l'utilisation de techniques de vision par ordinateur pour générer des variations visuelles réalistes pourraient enrichir le dataset et améliorer les performances du modèle. Enfin, l'amélioration des descriptions textuelles associées aux images, par exemple en utilisant des techniques de génération de texte pour créer des descriptions plus détaillées et spécifiques, pourrait améliorer la qualité des correspondances entre images et textes, augmentant ainsi les performances de classification.

En conclusion, bien que CLIP démontre des capacités prometteuses pour la classification d'images de biens de consommation, des améliorations dans les domaines des données, de l'interprétabilité et de l'optimisation des modèles peuvent encore être apportées pour en maximiser les performances et l'applicabilité dans des contextes réels.

Références

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. arXiv preprint arXiv:2103.00020.
2. OpenAI. (2021). CLIP: Connecting Vision and Language through Contrastive Learning. OpenAI Blog.