DataSpace

Plan prévisionnel

Contexte & Dataset retenu

L'entreprise « place du Marché » est une marketplace de e-commerce. Elle permet à des vendeurs de proposer leurs articles à des acheteurs. Les annonces des produits vendus comprennent une description et une photographie du produit.
Le vendeur met en ligne lui-même les annonces et attribue la catégorie du produit manuellement, un processus source d'erreurs notamment dû à une certaine subjectivité quant au choix des catégories. Pour améliorer l'expérience des vendeurs et des clients, il a été souhaité de développer une classification automatique des annonces lors de leur création, basée soit sur les textes des descriptions, soit sur les images des produits. Après les études de faisabilité, il ressort que c'est la classification automatique par image qui est la plus efficace sur ces données.

Le dataset se présente sous la forme d'un fichier *.csv*, comprenant les données suivantes :

- 1050 lignes, chaque ligne correspond à un produit;
- 15 colonnes d'information (précisées ci-après);
- 7 catégories de produits dites « principales », retenues pour la classification;
- Jusqu'à 5 niveaux de sous-catégories par catégorie principale.

Pour chaque ligne-article, nous avons disposé des informations suivantes :

- *uniq_id* : numéro identifiant unique
- *product_name* : nom du produit
- *product_category_tree* : arborescence complète de l'annonce du produit sur le site
- *image* : chemin d'accès vers l'image
- *description* : texte de la description du produit (en anglais)

D'autres informations sont disponibles dans le dataset, mais n'ont pas été utiles au développement du projet (l'URL de l'article, le prix de vente, les spécifications etc.)
Le dataset réceptionné est déjà nettoyé et dépourvu de NaN.
Après nos différents tests, c'est le DenseNet121 qui a été retenu pour la classification automatique des annonces avec les images des produits. Sur l'échantillon de test, il obtenait le score de justesse (‘*accuracy*’) de 89%. Ce modèle sera notre baseline.

Modèle envisagé

Bien que les performances du DenseNet121 soient plutôt bonnes sur notre jeu de données, l'article du 5 janvier 2021 intitulé **« *Learning Transferable Visual Models From Natural Language Supervision* »** issu du travail d'une équipe d'Open AI, propose un état de l'art de la computer vision et démontre les performances élevées de l'approche multimodale de **CLIP**.

En effet, dans le but de passer outre la nécessité de données étiquetées supplémentaires pour apprendre de nouveaux concepts visuels, cette équipe a développé une alternative basée sur l'apprentissage à partir des légendes associées aux images.
Ainsi ont-ils utilisé en pré-entraînement 400 millions de paires de données composées de textes (légende de l'image) et d'images, collectées sur Internet.
Après ce pré-entraînement spécifique, le modèle a pu s'adapter à diverses tâches de vision par ordinateur sans apprentissage spécifique.
Cette nouvelle approche de la computer vision, **CLIP**, ***Contrastive Language-Image Pre-training***, a démontré son efficacité en égalant la précision du ResNet-50 sur le dataset ImageNet tout en pratiquant du *zero-shot learning*. Il s'agit d'un type d'apprentissage sans coup d'essai, c'est-à-dire sans avoir eu d'exemples étiquetés des catégories ou des concepts à reconnaître lors de son entraînement.

Dans notre projet d'amélioration de la qualité de la classification existante, nous allons donc tester l'approche CLIP, en utilisant les caractéristiques combinées du texte et de l'image de chaque article et en les envoyant à un simple modèle de Régression Logistique qui prédira les catégories des produits.

Dans le cas de résultats concluants, l'intérêt principal de cette démarche dans notre contexte est in fine d'améliorer l'expérience utilisateur (vendeur et acheteur) en assurant une classification automatique des annonces qui respecte bien les catégories des produits, le tout en valorisant les données textuelles existantes.

Explication de votre démarche de test du nouvel algorithme (votre preuve de concept)

Pour notre démarche de test, nous allons procéder en reprenant les résultats du modèle DenseNet121 issus de l'entraînement spécifique sur notre jeu de données et les comparer aux résultats de l'approche avec CLIP.
Nous allons également tester 2 versions de l'approche CLIP, qui se différencieront par le traitement fait aux données textuelles.
En effet, la majorité des textes qui ont servis au pré-entraînement de CLIP étaient des légendes relativement courtes par rapport aux textes issus des descriptions de notre dataset et CLIP intègre un tokeniser propre, qui applique une troncature à 77 tokens sur les textes reçus, correspondant à la longueur maximale de contexte attendue par le modèle pré-entraîné.
Pour certaines catégories de produits dans nos données, la moyenne du nombre de tokens est plus élevées (99 token pour la moyenne la plus élevée) et certaines descriptions dépassent les 500 tokens. Bien qu'il ne soit pas nécessaire de tokeniser les textes en inputs du modèle car il intègre déjà un tokeniser, nous tester également une approche avec tokenisation préalable. Nous allons préparé une version des textes déjà tokenisés avec NLTK, dans le but de voir si il y a une restitution d'information plus importante dans ce cas, et donc une meilleure performance sur la classification finale.

Préparation des données prévues :

- Images : mise au carré des images pour éviter leur déformation lors du redimensionnement nécessaire pour l'input des modèles.
- Textes (uniquement pour l'approche CLIP) :
- soit les textes bruts sans modification avant la préparation de l'input au modèle,
- soit une tokenisation avec avec le RegexTokenizer de NLTK au préalable.

Nous comparerons les résultats sur les 3 modèles suivants :

- La baseline de référence avec le DenseNet121 uniquement sur les images,
- CLIP 'model A' sur les images et les textes sans transformation,
- CLIP 'model B sur les images et les textes tokenisés avant l'envoi au modèle.

L'évaluation sera principalement accès sur le taux de classification, l'accuracy, le temps de calcul nécessaire à l'obtention des résultats, et nous porterons également notre attention sur les erreurs de classification.

Références bibliographiques

Article développant l'approche CLIP

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger and Ilya Sutskever. "*Learning Transferable Visual Models From Natural Language Supervision.*" International Conference on Machine Learning (2021).
https://arxiv.org/abs/2103.00020

Article sur l'explicabilité de CLIP

Sepideh Mamooler - 2021 :
https://github.com/sMamooler/CLIP_Explainability/blob/main/CLIP_Explainability.pdf

Sources GIT pour l'explicabilité de CLIP

Sepideh Mamooler - 2021 :
https://github.com/sMamooler/CLIP_Explainability
Shashwat Trivedi :
https://github.com/shashwattrivedi/Attention_visualizer?tab=readme-ov-file#readme
Hila Chefer :
https://github.com/hila-chefer/Transformer-MM-Explainability/tree/main

Site Open.AI

https://openai.com/index/clip/