

# Note méthodologique

## Preuve de concept

### Rappel du contexte

Dans le cadre de "*Place de Marché*", une marketplace de e-commerce qui utilise la classification automatique des annonces des articles mis en ligne par les vendeurs grâce aux images des produits, le modèle DenseNet121 a prouvé son efficacité pour cette tâche de computer vision. Suite à une veille technologique sur ce sujet, nous avons trouvé une nouvelle approche qui pourrait améliorer les performances sur cette classification automatique des annonces.

### Dataset retenu

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.

# Les concepts de CLIP

Les systèmes de vision par ordinateur modernes sont limités par la nécessité de données étiquetées supplémentaires pour apprendre de nouveaux concepts visuels.

CLIP est une alternative qui met en place un apprentissage à partir de texte brut associé aux images. Cette méthode multimodale utilise une tâche de pré-entraînement simple consistant à associer des légendes à des images, permettant d'apprendre des représentations d'images. Les données utilisées pour ce pré-entraînement sont issues de 400 millions de paires (image, texte) collectées sur Internet.

Après le pré-entraînement, le modèle peut être transféré à diverses tâches de vision par ordinateur sans formation spécifique. Les performances de cette approche ont été évaluées sur plus de 30 ensembles de données, montrant que le modèle est souvent compétitif avec des modèles entièrement supervisés. Par exemple, il atteint la précision du ResNet-50 sur le dataset ImageNet en *zero-shot learning*.

CLIP forme conjointement un encodeur d'image et un encodeur de texte pour prédire les paires correctes, permettant de créer un classificateur *zero-shot learning* en intégrant les noms ou descriptions des classes cibles.

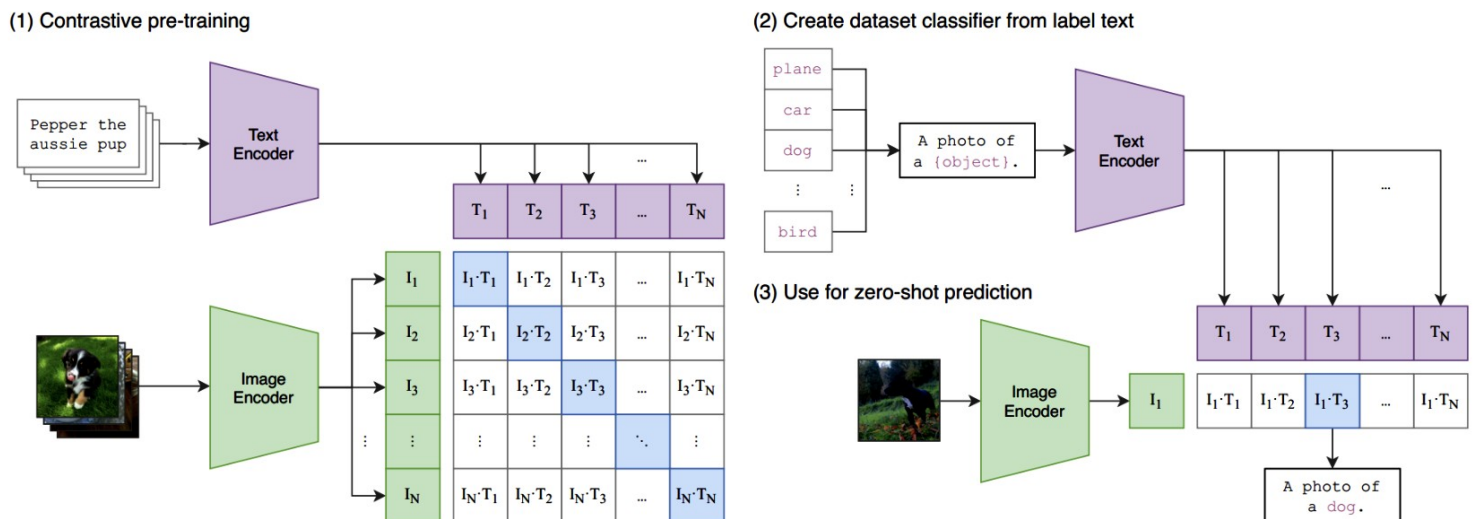


Figure 1: CLIP at train and inference time [4].

Dans la figure ci-dessus, issue de la documentation sur l'explicabilité du modèle CLIP, nous voyons les différentes étapes de cette approche :

- 1) un pré-entraînement qui associe les paires Texte+Image avec un encoder spécifique pour chaque objet (un pour le texte, un pour l'image)
- 2) À chaque inférence, CLIP crée un dataset avec les Textes uniquement, encodés
- 3) CLIP va associer les textes et les images en se basant sur la similarité cosine des textes et des images encodés.

L'encoder de CLIP pour les textes a une architecture basée sur celle des Transformer.

L'encoder de CLIP pour les images peut être sélectionné lors du chargement du modèle parmi différentes variations de ResNet et de Vision Transformer ('RN50', 'RN101', 'RN50x4', 'RN50x16', 'ViT-B/32', 'ViT-B/16').

Ce sont ces encoder qui ont été pré-entraînés sur les 400 millions de paires Texte-Image.

## La modélisation

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) :
  - 1) soit les textes bruts sans modification avant la préparation de l'input au modèle,
  - 2) soit une tokenisation avec le RegexTokenizer de NLTK au préalable.

Il est à noter que la modélisation du DenseNet121 a bénéficié des capacités de calcul de 4 GPU. Cependant, pour CLIP, le support CUDA ne fonctionnait pas sur notre matériel. CLIP a donc été modélisé uniquement avec le CPU de notre machine.

Pour notre modélisation de CLIP, nous avons sélectionné l'encoder Vision Transformer (Base, 32x32 patches), le ViT-B/32.

### 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 basée 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.

## La synthèse des résultats

	Scenario	Description	Accuracy_Train	Accuracy_Test	Durée_totale_computation
0	Baseline_DenseNet121	AVEC poids imagenet & SANS data-augmentation	97.533631	88.607597	2 min 54 sec
1	TEST_CLIP_model_A	SANS tokenisation NLTK et troncature avec CLIP...	96.547619	96.666667	2 min 42 sec
2	TEST_CLIP_model_B	AVEC tokenisation NLTK et troncature avec CLIP	95.833333	95.714286	2 min 45 sec

Nous constatons donc que le modèle le plus performant sur nos données est le CLIP model-A, sans préparation des textes préalablement à l'envoi au modèle.

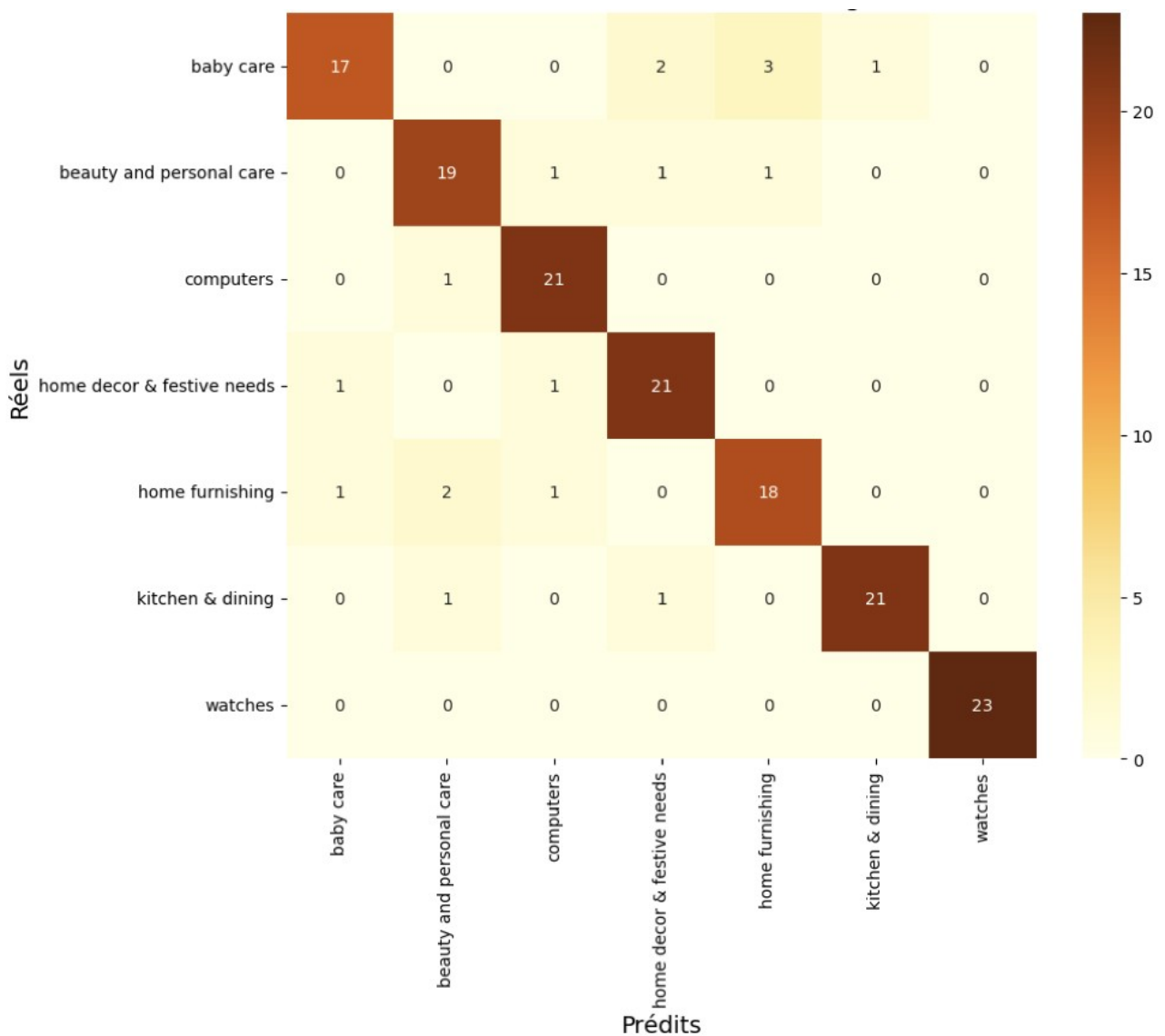
Malgré la non utilisation des GPU pour les approches CLIP, il s'avère que CLIP est beaucoup moins coûteux en terme de temps de calcul que le DenseNet avec support CUDA.

## Les Matrices de confusion sur les données de Test :

### 1. DenseNet121

Le modèle se trompe particulièrement sur la catégorie 'baby care' et la catégorie 'home furnishing'.

Ces catégories en particulier regroupent des types de produits assez hétérogènes. Pour la catégorie 'baby care' par exemple, il y a des serviettes et des langes, qui ressemblent assez à simplement du ligne de maison, mais aussi des articles de décoration de chambre.



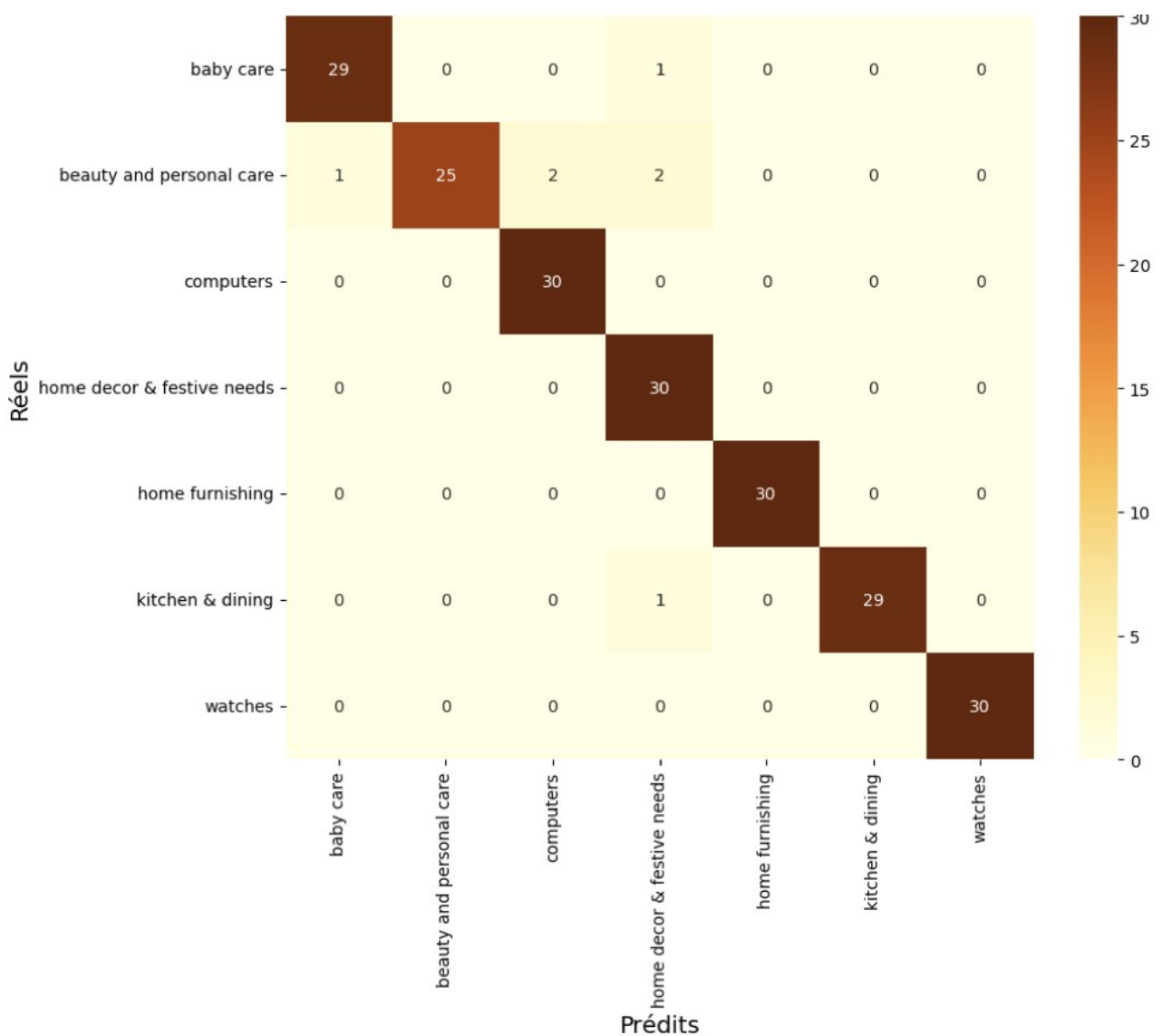
## 2. CLIP model-A

On ne retrouve pas les même catégories qui posaient problème au DenseNet121.

La confusion entre les rares classes mal retrouvées se joue principalement entre 'beauty&personnal care' et 'computers'.

Avec la visualisation des produits mal classés par CLIP nous pourrions mieux comprendre ce phénomène.


Cependant nous pouvons déjà noté que l'addition des informations issues du texte et de l'image permet au modèle de lever une grande partie des doutes légitimes soulevés par certaines images.



## Les produits mal classés avec CLIP model-A :


La majorité des produits mal classés auraient certainement susciter le doute chez un opérateur humain, comme certains produits électroniques qui sont en réalité du matériel de soin cosmétique et non du matériel informatique, ou la bande LED en rouleau qui ressemble à un rouleau de câble informatique. L'affiche de film mis en 'beauty and personal care' est moins compréhensible.

**Catégorie réelle : baby care**  
Pred: home decor & festive needs




**Catégorie réelle : beauty and personal care**


Pred: computers




Pred: baby care




Pred: home decor & festive needs



Pred: computers




Pred: home decor & festive needs

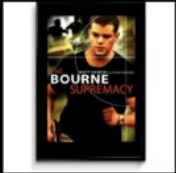


**Catégorie réelle : home decor & festive needs**


Pred: computers



Pred: beauty and personal care



**Catégorie réelle : kitchen & dining**  
Pred: home decor & festive needs



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

Ci-dessous voici quelques exemples d'articles bien et mal classés par CLIP pour lesquels nous avons récupéré les informations ayant retenu l'attention de CLIP lors de sa prise de décision, aussi sur les textes (mots en gras) que sur les images (heatmap d'attention).

Nous avons également récupéré les probabilités que CLIP a calculé avant le choix de la catégorie des articles concernés.

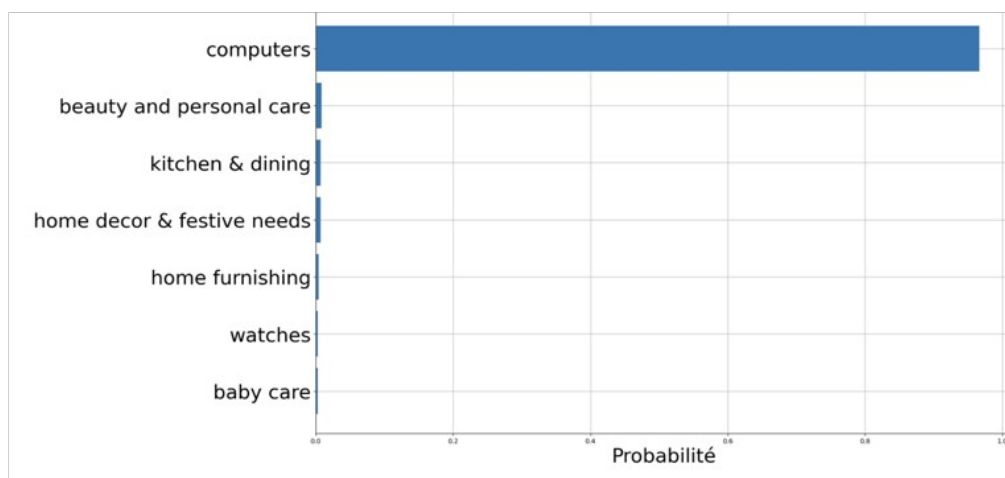
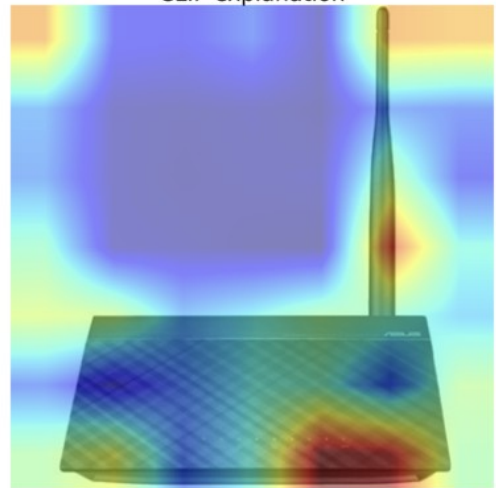
## Exemple : le modem bien classé

Produit "Asus DSL-N10E Wireless-N150 ADSL Modem Router"  
Catégorie réelle "computers"



Buy (0.04) **Asus** (0.10) **DSL – N10E** (0.04) **Wireless – N150** (0.03)  
ADSL **Modem** (0.03) Router **only** (0.02) **for** (0.02) **Rs.** (0.04) 5000  
from Flipkart.com. Only Genuine Products. 30 Day **Replacement**  
(0.03) **Guarantee.** (0.03) Free Shipping. Cash On Delivery!

CLIP explanation





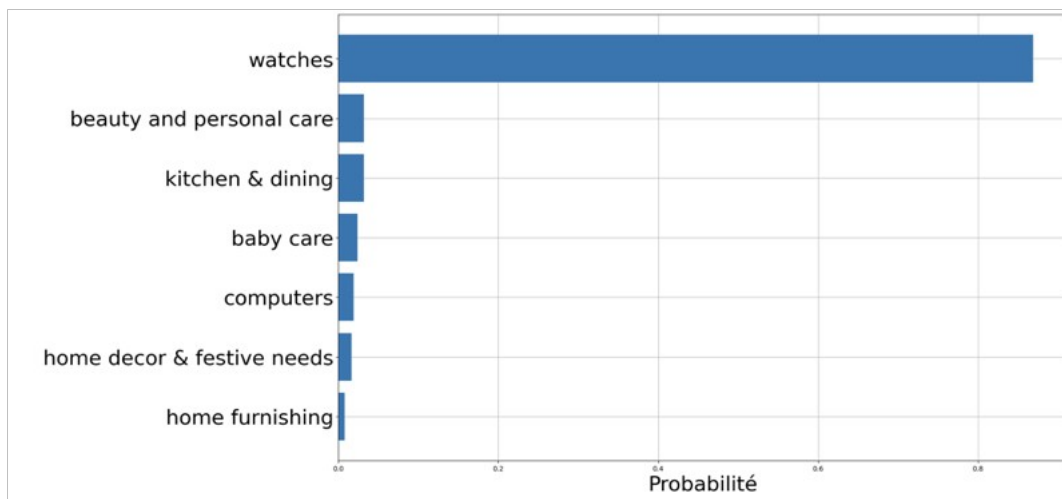
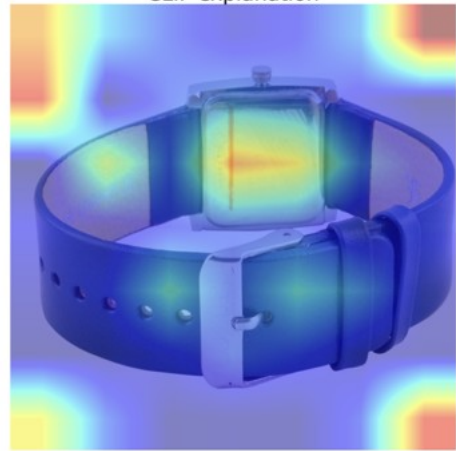
## Exemple : la montre sur l'envers bien classée

Produit "Logues LGSWATCHES760SL Analog Watch - For Women"  
Catégorie réelle "watches"




Logues (0.04) LGSWATCHES760SL (0.07) Analog (0.03) Watch (0.03)  
- (0.04) For Women (0.02) - (0.04) Buy (0.05) Logues (0.04)  
LGSWATCHES760SL (0.07) Analog (0.03) Watch (0.03) - (0.04) For  
Women (0.02) LGSWATCHES760SL (0.07) Online at Rs.725 in India Only  
at Flipkart.com. - (0.04) Great Discounts, Only Genuine Products, 30 Day  
Replacement Guarantee, Free Shipping. Cash On Delivery!

CLIP explanation




## Exemple : la gourde d'apprentissage mal classée

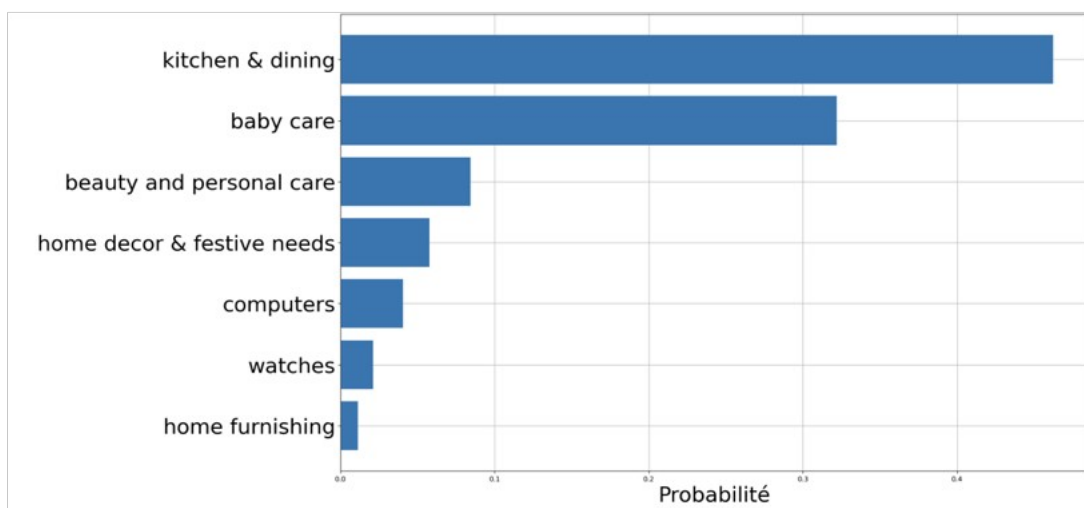
Produit "Munchkin Mighty Grip Trainer Cup"  
Catégorie réelle "baby care"



CLIP explanation



**Buy** (0.03) **Munchkin** (0.09) **Mighty** (0.07) **Grip** (0.05) **Trainer** (0.06) **Cup** (0.04) **for** (0.04) **Rs. 349** (0.02) online. **Munchkin** (0.09) **Mighty** (0.07) **Grip** (0.05) **Trainer** (0.06) **Cup** (0.04) **at** (0.03) **best** (0.02) prices with FREE shipping & cash on delivery. Only Genuine Products. 30 Day Replacement Guarantee.



Quand nous regardons les features importances pour les textes et les images, voici les observations que l'on peut faire :

1. certainement grâce à la présence de mots-clés très précis dans le texte (exemple : le modem), on peut raisonnablement supposer que cela permet à CLIP d'être certain de la catégorie, même dans le cas où l'image est peut conventionnelle pour ce type de produit (exemple : la montre montrée sur l'envers)
2. Lorsqu'il y a erreur, il y a eu une forte hésitation, et c'est probablement l'absence de mots-clés pertinents de la catégorie qui a fait défaut. Ainsi dans l'exemple de la gourde pour enfant, malgré la présence des mots « *munchkin* » (nabot/liliputien) et « *trainer cup* » (tasse d'entraînement), cela n'a pas été suffisant pour comprendre précisément le contexte.

## Les limites et les améliorations possibles

En dehors d'augmenter le volume des données d'entraînement, nous pourrions aussi chercher à comprendre les raisons des erreurs de CLIP sur les paires Texte+Image mal classées.

Nous pourrions donc envisager de faire une **étude avec un opérateur humain** sur les erreurs de CLIP.

En lui présentant les images et les légendes associées des articles mal classés par CLIP, nous demanderons à cet opérateur humain de les classer parmi les 7 catégories à disposition et nous lui demanderons d'indiquer si il a douté ou non.

**En cas de doute**, il serait peut-être opportun de prévoir un travail de révision des catégories de produits, car si même un opérateur humain doute de la catégorie des produits, cela signifie qu'il y a des ambiguïtés à lever.

Dans l'interface de dépôt d'annonce, peut-être pourrions nous indiquer de courtes recommandations aux vendeurs quant à l'usage de certains mot-clés dans les descriptions des produits pour les catégories principales.

**En cas de certitude** pour l'opérateur humain, il faudrait réentraîner le modèle et dans ce cas de figure plusieurs pistes s'offrent à nous :

1. Réentraîner le modèle en modifiant certains paramètres d'entraînement (batchsize, pourcentage de données dans le train et dans le test etc..),
2. Travailler les images en les passant en niveau de gris, pour concentrer l'information sur la forme de l'objet,
3. Mettre plus de poids sur la présence de certains mots dans les textes en faveur des catégories auxquelles l'article pourrait correspondre. Par exemple, « *kid* », « *baby* », et le fameux « *munchkin* », si nos analyses sur les textes le confirme, pourraient faire partie de mots-clés avec plus de poids que les autres pour indiquer une probabilité plus importante d'appartenance à la catégorie 'Baby'.