Note méthodologique : preuve de concept

Version	Auteur	Description
012025	Adeline Le Ray	Première émission

Table of Contents

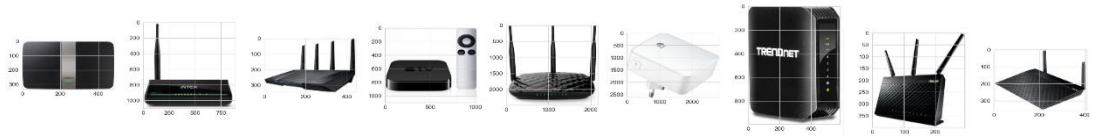
Dataset retenu

Le dataset retenu est celui utilisé pour le projet 6 – Classifiez automatiquement des biens de consommation.
Ce dataset provient du site de e-commerce Flipkart et contient des descriptions textuelles de biens de consommation et les images associées.

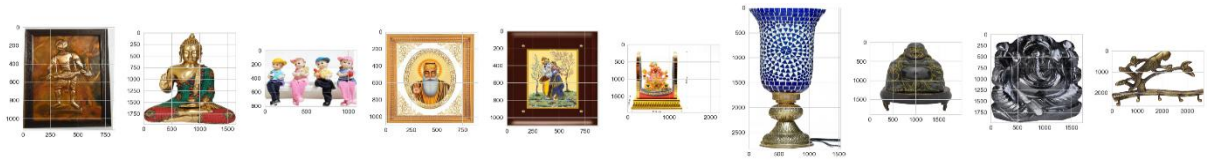
- **Baby Care :** articles pour bébés et enfants comme Infant Wear, Baby & Kids Gift, Baby bedding

- **Beauty and Personal Care :** produits capillaires, soins de la peau, maquillage comme Fragances, Combos and kits, Makeup, Body and Skin Care

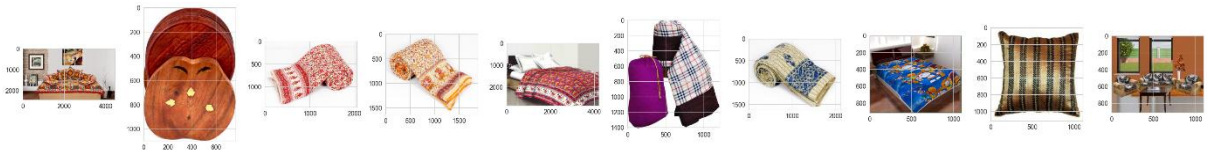
- **Computers :** ordinateurs portables, périphériques et accessoires comme Laptop Accessories, Network Components, Computer Peripherals.



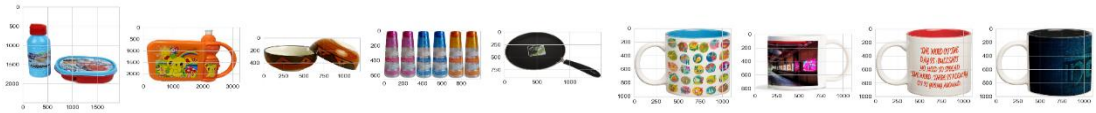
- **Home Decor & Festive Needs :** articles de décoration murale, des lampes et des accessoires pour fêtes comme Showpieces, Table Decor & Handicrafts, Wall Decor & Clocks, Candles & Fragrances.



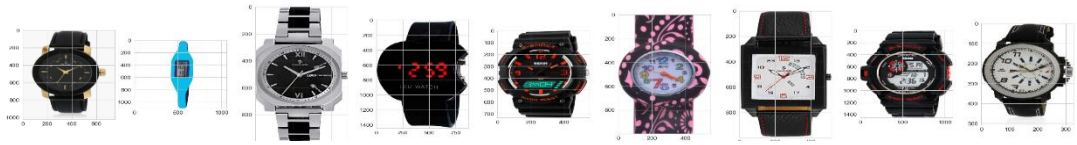
- **Home Furnishing :** articles destinés à l'ameublement et à la décoration intérieure comme Bed Linen, Bath Linen, Curtains & Accessories, Kitchen & Dining Linen.



- **Kitchen & Dining :** ustensiles de cuisine et articles pour la table (batteries de cuisine, vaisselles et couverts) comme Coffee mugs, Cookware, Kitchen Tools.



- **Watches :** montres pour hommes, femmes et enfants comme Wrist Watches, Clocks.



Les concepts de l'algorithme récent

L'algorithme **CLIP** (Contrastive Language-Image Pretraining) est un modèle **multimodal** développé par OpenAI et publié en 2021 [1].
Multimodal signifie qu'il peut traiter à la fois des données visuelles (images) et des données textuelles.

- **Encodage multimodal :**

CLIP utilise 2 types d'encodeurs *(voir Figure 1)* :

- **Encodeur d'images** : CLIP utilise un réseau de neurones, comme un **Vision Transformer (ViT)** ou un réseau convolutif, pour encoder les images en vecteurs dans un espace de caractéristiques.
- **Encodeur de texte** : CLIP utilise un modèle de type **Transformer**, similaire à GPT ou BERT, pour transformer les descriptions textuelles en vecteurs dans le même espace de caractéristiques que les images.

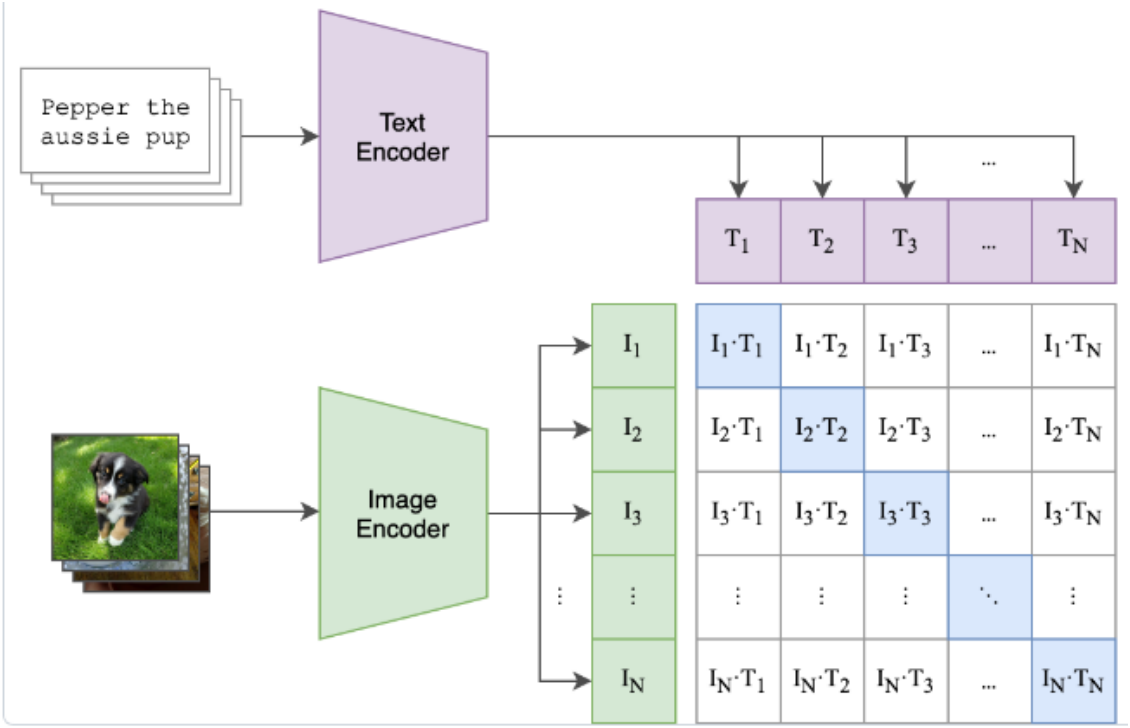


Figure 1 - Architecture de CLIP issue de la publication officielle d'OpenAI (source : https://openai.com/index/clip/)

- **Apprentissage contrastif et espace d'embedding partagé**

CLIP utilise un **apprentissage contrastif** pour relier les images et les textes. Le modèle est entraîné à reconnaître les correspondances entre des paires d'images et de descriptions textuelles, et à distinguer celles qui ne correspondent pas. Son objectif est de maximiser la **similarité cosine** entre les images et les textes corrects tout en minimisant celle des mauvaises paires, créant ainsi un espace d'embedding commun où les représentations visuelles et textuelles coexistent. La paire avec la plus grande similarité est prédite comme étant correspondante. Pour cela, CLIP utilise une fonction de perte contrastive, qui rapproche les représentations des paires image-texte correctes et éloigne celles des paires incorrectes dans cet espace vectoriel (voir Figure 2).

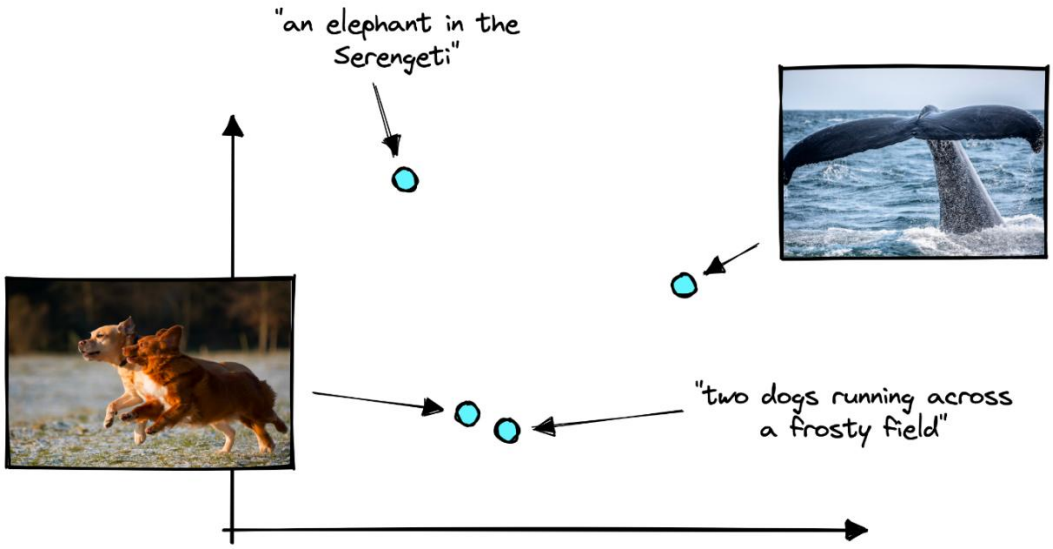


Figure 2 - Un texte et une image similaires seront encodés dans un espace vectoriel similaire. Un texte et une image dissimilaires ne partagent pas un espace vectoriel similaire (source: https://www.pinecone.io/learn/series/image-search/clip/)

- **Cas d'utilisation : Zero-shot Learning**

L'un des aspects les plus puissants de CLIP est sa capacité à effectuer du **zero-shot learning**. Cela signifie que le modèle peut classifier des objets ou accomplir des tâches sans avoir été explicitement entraîné sur ces classes ou tâches spécifiques.

- **Fonctionnement** : Lorsque CLIP est confronté à une nouvelle image, il compare sa représentation à une série de descriptions textuelles (par exemple, "une photo d'un chat", "une photo d'un chien"). La classe avec la plus grande similarité dans l'espace vectoriel sera celle choisie par le modèle.
- **Avantage** : Cela permet à CLIP de généraliser à de nouvelles tâches sans avoir besoin de données d'entraînement supplémentaires pour chaque classe, contrairement aux modèles supervisés classiques.

La modélisation

Les techniques de modélisation d'images sélectionnées pour l'analyse comparative sont les suivantes :

- **Approche récente** : CLIP en Zero-Shot Learning.
- **Approche plus classique** : DenseNet avec Data Augmentation.

La variante de CLIP choisie pour cette étude repose sur la **Vision Transformer (ViT)**, accessible via Hugging Face, afin d'effectuer des prédictions en zero-shot sur les catégories du dataset analysé.

Le modèle **DenseNet** a été pré-entraîné sur **ImageNet**, puis affiné via un entraînement supervisé sur notre dataset spécifique. Pour optimiser ses performances, nous avons intégré une stratégie de **data augmentation**, appliquant diverses transformations sur les images, telles que rotation, zoom et inversions horizontales.

Pour évaluer les performances de DenseNet et les comparer à un modèle de Zero-Shot Learning, plusieurs métriques ont été retenues:

- **Accuracy** : Proportion d'images correctement classées par rapport à l'ensemble des images testées.

- **Precision et Recall par catégorie** : La précision mesure la capacité du modèle à éviter les faux positifs tandis que le rappel évalue sa capacité à identifier les vrais positifs.
- **F1-score par catégorie** : Moyenne harmonique de la précision et du rappel, utilisée pour offrir un équilibre entre les deux et donner une vision plus complète de la performance du modèle.

Pour **CLIP**, l'optimisation des performances a été réalisée en testant différentes formulations textuelles comme source pour les descriptions de catégories :

- **Catégories simples** : Les intitules des catégorie comme Baby Care, Beauty and Personal Care, Computers, etc. …
- **Descriptions détaillées des catégories** : Des formulations plus riches et contextuelles, incluant des attributs spécifiques de chaque catégorie comme
  - Baby Care : Clothes, toys, bed, furniture for infant, baby boys and girls,
  - Beauty and Personal Care : 'Fragances, Products and accessories for hair, body, skin, eye care for women and men',
  - Computers : 'Components and accessories for computer, laptop and tablet',

Ces deux approches ont été comparées afin de mesurer leur impact sur la qualité des prédictions zero-shot de CLIP. L'objectif était de déterminer si des descriptions textuelles plus riches permettaient une meilleure correspondance avec les images, ou si des intitulés plus simples étaient suffisants pour maximiser la performance.

Une synthèse des résultats

Les résultats des performances sur le jeu de test des modèles DenseNet avec data augmentation, CLIP zero-shot learning avec texte simple et détaillé sont synthétisés dans les Table 1, Table 2 et la Figure 3 .

Table 1 – Accuray et F1-score des modèles DenseNet, CLIP avec texte simple, CLIP avec texte détaillé sur le jeu de données test et CLIP sur l'ensemble du dataset

	Model	Accuracy	f1score - Baby Care	f1score - Beauty and Personal Care	f1score - Computers	f1score - Home Decor & Festive Needs	f1score - Home Furnishing	f1score - Kitchen & Dining	f1score - Watches
0	DenseNet avec data augmentation	0.857	0.77	0.83	0.92	0.79	0.86	0.89	0.93
1	CLIP zero-shot learning, short text	0.571	0.61	0.69	0.76	0.21	0.45	0.06	0.92
2	CLIP zero-shot learning, detailled text	0.852	0.71	0.82	0.84	0.83	0.87	0.89	1.00
3	CLIP zero-shot learning, all dataset	0.849	0.77	0.82	0.86	0.80	0.84	0.87	0.97

Table 2 – Precison et Recall des modèles DenseNet, CLIP avec texte simple, CLIP avec texte détaillé sur le jeu de données test et CLIP sur l'ensemble du dataset

	Model	Precision / recall - Baby Care	Precision / recall - Beauty and Personal Care	Precision / recall - Computers	Precision / recall - Home Decor & Festive Needs	Precision / recall - Home Furnishing	Precision / recall - Kitchen & Dining	Precision / recall - Watches
0	DenseNet avec data augmentation	0.75 / 0.80	0.96 / 0.73	0.90 / 0.93	0.82 / 0.77	0.80 / 0.93	0.85 / 0.93	0.96 / 0.90
1	CLIP zero-shot learning, short text	0.79 / 0.50	0.58 / 0.87	0.95 / 0.63	0.44 / 0.13	0.31 / 0.83	1.00 / 0.03	0.86 / 1.00
2	CLIP zero-shot learning, detailled text	0.77 / 0.67	0.81 / 0.83	0.89 / 0.80	0.79 / 0.87	0.84 / 0.90	0.87 / 0.90	1.00 / 1.00
3	CLIP zero-shot learning, all dataset	0.76 / 0.77	0.79 / 0.86	0.88 / 0.85	0.83 / 0.77	0.86 / 0.83	0.89 / 0.86	0.94 / 1.00

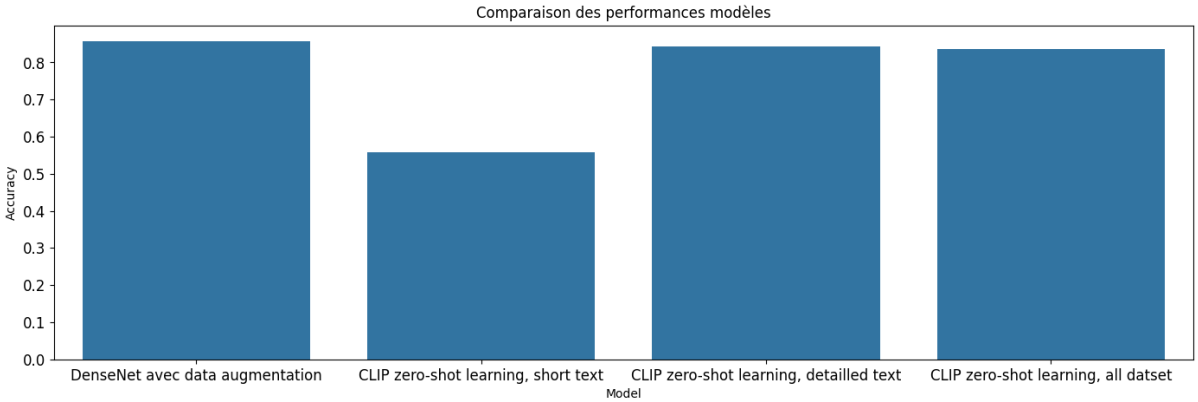


Figure 3 - Accuracy pour les modèles DenseNet, CLIP avec texte simple, CLIP avec texte détaillé sur le jeu de données test et CLIP sur l'ensemble du dataset

- **DenseNet avec Data Augmentation :**

Le modèle DenseNet, entraîné de manière supervisée sur le dataset avec des techniques de data augmentation, obtient une **accuracy globale de 85,7 %**. Il se distingue par des performances globalement solides dans presque toutes les catégories, avec des **F1-scores supérieurs à 0.75**, notamment pour les catégories "Watches" (0.93), "Computers" (0.92) et "Kitchen & Dining" (0.89). La data augmentation semble particulièrement efficace pour améliorer la robustesse du modèle face à la diversité des images, lui permettant ainsi de bien généraliser.

- **CLIP en Zero-Shot Learning :**

Deux formulations textuelles ont été testées avec CLIP : des **catégories simples** et des **descriptions détaillées**. Les résultats montrent que la formulation textuelle joue un rôle crucial dans les performances du modèle.

- **CLIP avec catégories simples** (ex. : "Beauty", "Home Decor") obtient une **accuracy globale de 57,1 %**, avec des résultats mitigés selon les catégories. Les meilleures performances sont observées dans des catégories comme **Watches** (F1-score de 0.92) et **Computers** (F1-score de 0.76), où les descriptions textuelles simples semblent bien correspondre aux images. Toutefois, certaines catégories, telles que **Kitchen & Dining** et **Home Decor**, affichent des F1-scores extrêmement bas (0.06 et 0.21 respectivement), indiquant que les simples étiquettes de catégorie ne suffisent pas à capturer les subtilités visuelles de ces produits.
- **CLIP avec descriptions détaillées** améliore considérablement les résultats **(+49%),** atteignant une **accuracy globale de 85,2 %**, proche de celle de DenseNet. Dans cette configuration, CLIP dépasse même DenseNet pour certaines catégories, notamment **Home Decor & Festive Needs** (F1-score de 0.83) et **Watches** (F1-score de 1.00). Ces résultats montrent que des descriptions textuelles plus riches, incluant des détails spécifiques aux produits, permettent une meilleure correspondance image-texte.

Les résultats de CLIP en zero-shot learning sur l'ensemble du jeu de données montrent une **accuracy globale de 84.9 %**, ce qui démontre une bonne capacité de généralisation du modèle sans entraînement spécifique.

- Les **F1-scores** par catégorie sont globalement élevés, avec des scores allant de **0.81 à 0.97**. Les meilleures performances sont observées dans des catégories comme **Watches** (0.97) et **Kitchen & Dining** (0.87).

- Les performances sont également très bonnes dans des catégories plus variées comme **Beauty and Personal Care** (0.82) et **Home Decor & Festive Needs** (0.80), soulignant la robustesse de CLIP pour capturer des nuances dans les descriptions textuelles.
- Cependant, certaines catégories comme **Baby Care** (0.77) sont légèrement en retrait, suggérant que des descriptions textuelles plus détaillées ou spécifiques pourraient améliorer ces résultats. La forme de l'objet influe sur la classification et entraîne des erreurs pour la catégorie Baby Care.

Cette analyse montre que CLIP, même en zero-shot, offre des performances compétitives, particulièrement lorsqu'il est appliqué à des catégories avec des correspondances visuelles claires.

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

Les **cartes d'attention** montrent comment le modèle utilise les **caractéristiques visuelles** des images pour effectuer ses prédictions, en **se focalisant sur des zones spécifiques qui influencent le plus la décision**. Ces visualisations permettent d'examiner si le modèle accorde de l'importance aux bons éléments dans chaque catégorie.

Pour la catégorie **"Baby Care"** (*voir Figure 4*), lorsque les prédictions sont correctes, l'attention du modèle se porte sur des aspects critiques comme les motifs ou la coupe des vêtements. Cela inclut les formes, les couleurs et les logos caractéristiques des habits pour enfants.

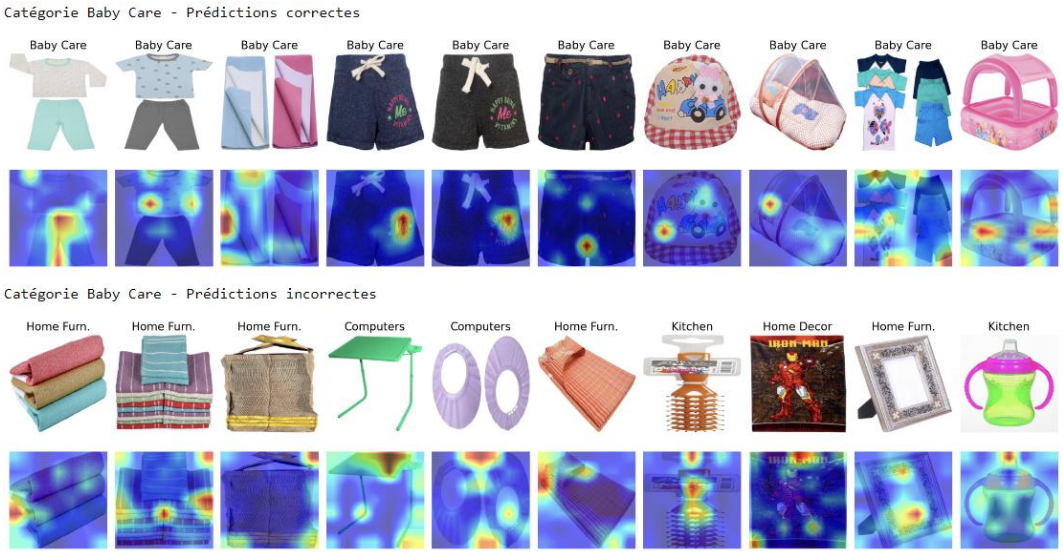


Figure 4 - Catégorie Baby Care : Prédictions correctes et incorrectes avec cartes d'attention

Dans les autres catégories, les zones d'attention les plus marquées lorsque les prédictions sont correctes incluent :

- **Beauty and Personal Care :** l'attention se focalise sur les formes des flacons et des bouteilles, les couleurs des rouges à lèvres, et des détails spécifiques comme les étiquettes ou logos sur les emballages.

- **Computers :** le modèle met l'accent sur les ports de connexion comme l'USB, la couleur noire des boîtiers d'ordinateurs et les étiquettes présentes sur les boîtes.
- **Home Decor and Festive Needs :** des détails tels que les motifs des objets de décoration sont pris en compte, avec une concentration sur les parties blanches entourant les objets.
- **Home Furnishing :** le modèle se focalise sur les motifs des tissus, élément clé pour cette catégorie.
- **Kitchen and Dining :** l'attention se porte sur des détails comme les anses des ustensiles et les dessins sur les mugs.
- **Watches :** l'accent est mis sur la forme ronde des cadrans, les chiffres sur le cadran et les attaches du bracelet.

Ces zones d'attention permettent également d'expliquer les prédictions incorrectes du modèle en zero-shot learning. En effet, dans ces cas, l'attention du modèle est souvent dirigée vers des parties non pertinentes de l'image, ce qui conduit à des erreurs de classification.
Par exemple, pour la catégorie "Baby Care", au lieu de se concentrer sur des motifs ou des formes spécifiques aux vêtements pour enfants, le modèle peut accorder de l'importance à des éléments de fond ou à des détails qui n'ont aucune valeur discriminante pour cette catégorie.

De même, pour la catégorie "Beauty and Personal Care", les erreurs peuvent survenir lorsque l'attention se porte sur des zones du flacon ou du packaging sans pertinence, comme une étiquette trop petite ou un détail de la forme qui ne permet pas d'identifier correctement le produit.

Dans "Computers", les prédictions incorrectes surviennent souvent lorsque l'attention se détourne des ports de connexion ou des éléments distinctifs et se fixe sur des parties communes à plusieurs objets, comme des formes rectangulaires ou rondes ou des zones sans texte explicatif.

Ces observations mettent en évidence les limitations du zero-shot learning dans l'interprétation des images : le modèle peut mal interpréter les zones clés qui différencient les catégories, ce qui explique en partie ses erreurs de prédiction.

Les limites et les améliorations possibles

- **Limites du Zero-shot learning :**
  Les performances du modèle CLIP en zero-shot dépendent fortement de la qualité des descriptions textuelles utilisées pour les catégories. La similarité entre certains produits appartenant à différentes catégories (par exemple, des linges de lit dans **Home Furnishing** et **Baby Care**) et le regroupement de biens très différents au sein d'une même catégorie (par exemple, **Home Decor & Festive Needs**, **Baby Care**) peuvent rendre difficile la définition de descriptions textuelles suffisamment nuancées pour bien capturer ces distinctions.
  De plus, les descriptions textuelles sont limitées à **77 tokens**, ce qui restreint la capacité à fournir des informations détaillées pour chaque catégorie, notamment pour les catégories complexes.
  Sans entraînement spécifique, CLIP peut également avoir des performances limitées.

- **Améliorations envisageables en Zero-shot learning :**
  - **Utilisation de descriptions textuelles enrichies** : Pour améliorer les performances, il serait pertinent d'optimiser encore les descriptions textuelles utilisées pour les catégories. Des descriptions plus précises, incluant des caractéristiques spécifiques aux produits, peuvent mieux capturer la diversité des images.
  - **Interprétabilité via les cartes d'attention** : L'analyse des **cartes d'attention** permet d'interpréter la manière dont CLIP et d'autres modèles focalisent leur attention sur des parties spécifiques des images et du texte. Pour améliorer l'interprétabilité du zero-shot learning, ces cartes d'attention peuvent être utilisées pour ajuster les descriptions textuelles et observer si le modèle se concentre bien sur les éléments pertinents. Cela permettrait non seulement d'optimiser les performances mais aussi de rendre les décisions du modèle plus transparentes.
- **Few-shot Learning :** Bien que CLIP ait de très bonnes perfomances en zero-shot, un entraînement en few-shot (avec quelques exemples annotés par catégorie/sous-catégories) pourrait améliorer les performances sans nécessiter un entraînement intensif sur l'ensemble du jeu de données.
  En se basant sur l'article *« Multimodal Multilabel Classification by CLIP »* [2], nous pouvons également envisager :
  - **Fusion multimodale :** Tester des méthodes de fusion comme la fusion par somme ou fusion mixte pour renforcer la combinaison des informations texte-image et améliorer la précision des prédictions.
  - **Fonctions de perte :** Utiliser des fonctions comme la focal loss ou l'asymmetric loss serait particulièrement utile en few-shot pour mieux gérer les sous-catégories déséquilibrées.
  - **Data augmentation multimodale :** Appliquer des augmentations synchronisées sur les images et les textes peut aider à maximiser l'efficacité de l'apprentissage.

Bibliographie

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning transferable visual models from natural language supervision." (2021). <https://arxiv.org/abs/2103.00020>.

[2] Guo, Yanming. "Multimodal Multilabel Classification by CLIP." (2024). <https://arxiv.org/abs/2406.16141>.

[3] Peng Xu, Xiatian Zhu, David A. Clifton. "Multimodal Learning with Transformers - A Survey." (2023). <https://arxiv.org/abs/2206.06488>.

[4] Wonjae Kim, Bokyung Son, Ildoo Kim. "ViLT - Vision-and-Language Transformer Without Convolution or Region Supervision." (2021). <https://arxiv.org/abs/2102.03334>.