Actes de CORIA-TALN 2023
Paris, France, 5 au 9 juin 2023

État des lieux des transformers Vision-Langage : un éclairage sur les données de pré-entraînement

Emmanuelle Salin[1]
(1) LIS, 163 avenue de Luminy, Marseille, France
emmanuelle.salin@lis-lab.fr

RÉSUMÉ

Après avoir été développée en traitement automatique du langage, l'architecture transformer s'est démocratisée dans d'autres domaines de l'apprentissage automatique. Elle a permis de surpasser l'état de l'art dans de nombreuses tâches. Afin d'améliorer les performances de ces modèles, de très grands jeux de données ont été créés. En multimodalité vision-langage, les résultats encourageants des transformers favorisent la collecte de données image-texte à très grande échelle. Cependant, évaluer la qualité de ces données et leur influence sur la performance de ces modèles est difficile, car notre compréhension des transformers vision-langage est encore limitée. Nous explorons les études du domaine pour mieux comprendre les processus de collecte des jeux de données image/texte, les caractéristiques de ces données et leurs impacts sur les performances des modèles.

ABSTRACT

State of the art on Vision-Language transformers : Insights on pre-training data

The transformer architecture is becoming increasingly popular in many areas of machine learning, after its development in Natural Language Processing. It has surpassed the state-of-the-art in many tasks and has led to the creation of very large datasets to improve model performances. In vision-language multimodality, the good results of transformer models have led to the gathering of large scale image-text datasets. However, it is difficult to assess the quality of these new datasets, as well as their influence on the performance of these models, as our understanding of vision-language transformers is still limited. We explore studies in the field to better understand the processes of image/text dataset collection, the characteristics of this data and its impact on model performance.

MOTS-CLÉS : Langage, Multimodalité, Vision, Jeux de données.

KEYWORDS: Language, Multimodality, Vision, Datasets.

1 Introduction

L'architecture transformer a été développée en traitement automatique du langage, permettant de surpasser l'état de l'art dans de nombreuses tâches, comme la traduction automatique ou les questions-réponses. Cependant, l'utilisation de cette architecture conjointement avec l'auto-supervision nécessite de grandes quantités de données pour améliorer les performances, conduisant au développement de modèles entraînées sur toujours plus de données, comme GPT-3 (Brown *et al.*, 2020) ou Bloom (Scao *et al.*, 2022). Cette configuration mène à l'émergence de comportements intéressant tels que la capacité de traiter de nouvelles tâches sans supervision classique (Ouyang *et al.*, 2022). Suite à ces succès, d'autres domaines applicatifs de l'apprentissage automatique ont incorporé l'architecture

Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 3 : prises de position en TAL, pages 14–29.

transformers et l'auto-supervision aux systèmes qu'ils produisent. C'est notamment le cas des modèles multimodaux vision-langage (Chen *et al.*, 2019; Tan & Bansal, 2019) qui, à partir de données textuelles et visuelles, sont capables de réaliser des tâches multimodales comme les questions-réponses visuelles ou le raisonnement sur une image et du texte.

De même que pour les modèles de langage, les modèles vision-langage plus récents sont ainsi entraînés avec un apport de données considérable, par auto-supervision, notamment sur la tâche d'association entre une image et sa légende. Par exemple, CLIP (Radford *et al.*, 2021) a montré que le passage à l'échelle en termes de données peut entraîner des gains très importants en termes de performances des modèles et de robustesse à différentes tâches. Cela a ainsi conduit à la collecte de nouveaux jeux de données appropriés pour l'apprentissage autosupervisé. Cependant, la taille d'un jeu de données vision-langage n'est pas la seule caractéristique qui peut grandement influencer le pré-entraînement. Toutefois, il est coûteux de réaliser des études d'ablations analysant précisément l'influence de la qualité des données sur les performances d'un modèle. En effet, les modèles les plus performants sont maintenant entraînés sur plus d'une dizaine de millions de paires d'image et texte. La réalisation de telles études demande donc beaucoup de ressources.

Dans cet article, nous soulevons certaines questions liées à l'utilisation de jeux de données image-texte pour le pré-entraînement de transformers vision-langage :

— Quelles sont les caractéristiques les plus importantes pour de tels jeux de données ?
— Comment évaluer la qualité d'un jeu de données vision-langage ?
— Quels problèmes éthiques peuvent être rencontrés lors de la création d'un jeu de données ?

Nous parcourons les études réalisées dans ce domaine, ainsi que dans les domaines de traitement du langage et de vision par ordinateur. Nous voulons ainsi apporter des éléments de réponse avec un impact potentiel sur ces questions, indépendamment de la modalité.

2 Vers toujours plus de données ?

La tendance actuelle des divers domaines du traitement automatique du langage et de la vision par ordinateur semble progresser vers une augmentation de la taille des modèles et jeux de données. C'est d'autant plus le cas pour les modèles basés sur l'architecture transformers. Cependant, certaines études portent un regard critique sur l'utilisation de jeux de données toujours plus gros.

En traitement automatique du langage Les itérations successives du modèle GPT (Brown *et al.*, 2020; Radford *et al.*, 2019) sont une illustration d'une règle qui s'est installée pour les nouveaux modèles de langues : des modèles constitués de plus de paramètres, pré-entraînés sur des jeux de données plus importants, aideront à obtenir de meilleures performances. De fait, (Hendrycks *et al.*, 2020) montre que la robustesse des modèles transformers semble s'améliorer quand ceux-ci sont entraînés sur plus de données, en comparant BERT (Devlin *et al.*, 2019) et RoBERTa (Liu *et al.*, 2019). Plus les données sont diversifiées, meilleure sera la capacité de généralisation du modèle.

Cependant, le coût économique et environnemental de cette tendance est non négligeable (Schwartz *et al.*, 2020; Bender *et al.*, 2021). Le coût de certains modèles est quantifié dans (Strubell *et al.*, 2019). Une manière de l'atténuer serait de prendre en compte l'efficacité du pré-entraînement d'un modèle lors de l'évaluation de celui-ci. D'autre part, une grande quantité de données n'est pas nécessairement

équivalente à une grande diversité dans les données. En effet, les méthodes de collecte et de filtrage des données peuvent engendrer des biais considérables qui pourraient être dommageables à l'utilisation de ces modèles. Par exemple, elles peuvent privilégier des points de vue hégémoniques (Bender *et al.*, 2021). La taille des jeux de données peut également faire passer inaperçu un manque de qualité de sous-ensembles de données. En étudiant des jeux de données multilingues, (Kreutzer *et al.*, 2022) font apparaître un taux d'erreurs significatif, particulièrement pour des langues à faibles ressources.

En vision par ordinateur Contrairement aux modèles convolutionnels, les modèles transformers n'ont pas l'architecture nécessaire pour apprendre à reconnaître efficacement la structure locale des images. Pour obtenir de bonnes performances, ils nécessitent de larges jeux de données (Dosovitskiy *et al.*, 2021). Bien que des modèles plus efficaces continuent à être élaborés (Touvron *et al.*, 2021), la taille des jeux de données reste l'un des principaux facteurs limitants (Zhai *et al.*, 2022).

Ces jeux de données à grande échelle présentent d'autres problèmes. Une grande partie des images disponibles sur internet contiennent des personnes. Ainsi, les jeux de données collectés à partir de ces images sans consentement préalable peuvent enfreindre la vie privée de ces personnes. Il est également difficile de collecter, filtrer et annoter des images à grande échelle. Ces données peuvent contenir des images à caractère pornographique (Prabhu & Birhane, 2020), ou présenter des biais qui nuisent à certaines catégories de la population (Kay *et al.*, 2015). Les annotations peuvent notamment être basées sur des stéréotypes, et contenir des catégories insultantes (Prabhu & Birhane, 2020). De plus, l'utilisation d'une quantité plus importante de données peut engendrer des rendements décroissants. En effet, (Sun *et al.*, 2017) observent sur des tâches de détection d'objets que la performance d'un modèle croit de façon logarithmique avec la taille des données d'entraînement.

Qualité des données En traitement automatique du langage, plusieurs méthodes de filtrages ont été développées afin d'améliorer la qualité (Wenzek *et al.*, 2019) des données de pré-entraînement. La création d'un tel protocole est souvent itérative, car il peut être difficile d'évaluer la qualité et l'impact de ces corpus. L'utilisation de jeux de données multilingues, notamment, nécessite de porter une plus grande attention au filtrage des données (Suarez *et al.*, 2019; Abadji *et al.*, 2022). En multimodalité vision-langage, de tels jeux de données n'ont été collectés que récemment, et ne sont parfois pas rendus publics pour des questions juridiques. Il est intéressant de se pencher sur le processus de collecte de ces corpus image-texte, leurs caractéristiques et leurs impacts sur le pré-entraînement des modèles.

3 Transformers vision-langage

Nous présentons dans cette section les transformers vision-langage ainsi que leurs données de pré-entraînement. Ces modèles ont été développés pour associer des concepts textuels et visuels afin de construire des représentations multimodales. Comme pour les modèles de langue, ils peuvent ensuite être affinés puis utilisés pour de nombreuses tâches liées aux modalités textuelles et visuelles. Celles-ci incluent le raisonnement visuel en langage naturel (Suhr *et al.*, 2019), la recherche multimodale d'images et de textes (Lin *et al.*, 2014), ou la génération de légendes (Agrawal *et al.*, 2019). Afin de générer des représentations multimodales, ces modèles sont pré-entraînés sur des tâches auto-supervisées textuelles, visuelles et multimodales sur une grande quantité de données.

Ces tâches se sont d'abord inspirées des tâches développées en traitement automatique du langage. Ainsi, les modèles sont généralement pré-entraînés en utilisant une tâche de 'masked language modeling', inspirée par BERT (Devlin *et al.*, 2019) et son équivalent en tâche visuelle. Les modèles utilisent également diverses tâches d'alignement entre image et texte pour obtenir une meilleure compréhension des interactions multimodales.

L'architecture des transformers multimodaux a évolué depuis les premiers modèles. Ceux-ci utilisaient d'abord des représentations issues de détecteurs d'objets pour incorporer des informations visuelles (Chen *et al.*, 2019; Tan & Bansal, 2019). Depuis le développement des transformers en vision (Dosovitskiy *et al.*, 2021), de nouvelles architectures sont apparues. Ainsi, de nouveaux modèles utilisent des transformers monomodaux pour extraire les informations de chaque modalité ainsi qu'un transformer multimodal qui combine les modalités (Li *et al.*, 2021; Yang *et al.*, 2022).

3.1 Jeux de données de pré-entraînement

Le pré-entraînement de ces modèles nécessite des corpus de données parallèles image et texte, utiles pour l'apprentissage de l'alignement multimodal. Un modèle vision-langage est ainsi pré-entraîné à partir d'un ensemble de larges jeux de données image/texte. Ces jeux de données sont généralement constitués de paires, composées d'une image et d'un court texte descriptif.

La disponibilité des données est l'un des principaux facteurs limitants des modèles image-texte. En effet, contrairement aux jeux de données utilisés pour le pré-entraînement des modèles de langue, les données vision-langage nécessitent une supervision supplémentaire, pour assurer la correspondance entre une image et sa description. Afin de favoriser le pré-entraînement de nouveaux modèles, des jeux de données ont été créés par diverses équipes de recherche, avec des protocoles de collecte de données différents, notamment au niveau du filtrage et de l'annotation. Nous décrivons dans cette partie ces jeux de données et leurs caractéristiques.

Nous nous intéressons aux principaux corpus publics utilisés pour le pré-entraînement des modèles d'état de l'art, sélectionnés en fonction des papiers récemment publiés dans le domaine. Ceux-ci sont de langue anglaise, ou multilingues. Une instance de corpus image/texte est généralement composée d'une image comportant au moins un objet, et d'une description courte associée à cette image.

Nom	MS COCO	Visual Genome
Nb Images	111 000	103 000
Nb Textes	558 000	5 millions
Ex. Image		
Ex. Texte	A horse carrying a large load of hay and two people sitting on it.	Park bench is made of gray weathered wood

TABLE 1 – Jeux de données image-texte annotés manuellement

Données annotées manuellement Avant l'utilisation de très grands jeux de données issus de Common Crawl, ils étaient généralement constitués à partir d'annotations manuelles. MS COCO (Lin *et al.*, 2014) et Visual Genome (Krishna *et al.*, 2016) sont deux exemples de larges corpus image-texte annotés manuellement (voir Table 1).

— MS COCO est composé d'images dites "non iconiques", c'est-à-dire comportant plusieurs objets et non centrées sur un élément visuel spécifique. Les textes associés à ces images sont écrits par différents annotateurs humains. Ces derniers ont pour instruction de décrire toutes les parties importantes de la scène en au moins 8 mots, afin d'obtenir des légendes riches.
— Visual Genome est composé d'images similaires à celles de MS COCO, mais il est spécifiquement orienté sur la description de régions d'objets. Chaque image comporte ainsi plusieurs annotations correspondant à des descriptions de régions.

Nom	SBU	Conceptual Captions	LAION
Nb Instances	1 million	3/12 millions	0.4/6 milliards
Ex. Image			
Ex. Texte	Man sits in a rusted car buried in the sand on Waitarere beach	a worker helps to clear the debris.	cat, white, and eyes image

TABLE 2 – Jeux de données image-texte annotés automatiquement

Données annotées automatiquement La création de très grands jeux de données vision-langage a commencé à se développer ces dernières années. Comme l'annotation manuelle des corpus limite la taille des jeux de données, des équipes de recherches ont décidé de recueillir et filtrer automatiquement des données disponibles sur internet : SBU (Ordonez *et al.*, 2011), Conceptual Captions (Changpinyo *et al.*, 2021; Sharma *et al.*, 2018) puis LAION (Schuhmann *et al.*, 2021, 2022) (voir Table 2). Ces jeux de données sont moins stables, car certaines données peuvent ne plus être disponibles. D'autres corpus sont en cours de développement, comme (Byeon *et al.*, 2022).

— Les images et les légendes de SBU sont collectées sur Flickr et filtrées afin que les légendes correspondent à un contenu visuel. Pour cela, les instances sont obtenues à partir de paires de requêtes formées de termes tels que des objets ou des attributs. Les descriptions collectées doivent correspondre à certains critères : taille, utilisation de préposition spatiales.
— Les deux ensembles de données de Conceptual Captions sont collectés automatiquement, en utilisant des champs 'alt_text' comme légendes, avec quelques filtres et transformations de texte. Les filtres s'assurent par exemple de la présence de parties de discours pertinentes dans la description. La correspondance entre image et texte est assurée en utilisant des modèles de vision pour assigner des labels aux images et les comparer au texte. Une version contenant 3 millions d'instances transforme le texte pour supprimer les informations relatives aux entités nommées, tandis qu'une nouvelle version ayant une taille de 12 millions de paires n'applique aucune transformation au texte.

— LAION est un corpus composé de 400 millions d'instances, également obtenu à partir du Common Crawl, avec des critères de filtrage plus souples. Il est filtré en établissant un seuil de similarité des représentations CLIP (Radford *et al.*, 2021) des deux modalités pour vérifier la correspondance entre texte et image. D'autres filtres sont apportés, notamment pour éliminer le contenu illégal. Une nouvelle version est composée de 5,85 milliards de paires image-texte.

Ainsi, les jeux de données texte-image utilisés pour le pré-entraînement des modèles transformers sont obtenus suivant des protocoles très différents. Ils sont ensuite agrégés dans un même corpus de pré-entraînement, avec différentes manières d'obtenir des échantillons. Cependant, n'y a actuellement pas de consensus sur les meilleures manières de collecter, filtrer ou transformer les données.

4 Impact des données de pré-entraînement sur la performance

L'impact des différentes caractéristiques d'un jeu de données sur le pré-entraînement des modèles vision-langage a été encore peu exploré. Cette question n'en demeure pas moins essentielle pour une meilleure compréhension des transformers multimodaux. En effet, cela permettrait d'aider à établir des protocoles de collecte, filtrage et traitement des données vision-langage. Cela pourrait également permettre de pré-entraîner les modèles de manière plus efficace, avec moins de ressources.

De nombreuses études s'accordent pour dire qu'utiliser de plus grands jeux de données améliore les performances des transformers vision-langage. C'est notamment visible à travers les études d'ablations réalisées pour différents modèles, montrant une amélioration significative de la performance des modèles sur les tâches en aval avec le passage à l'échelle des données de pré-entraînement (Li *et al.*, 2021; Yang *et al.*, 2022). Cependant, au vu des différents corpus disponibles, il est intéressant d'explorer les autres caractéristiques d'un corpus qui peuvent influencer la performance des modèles. Nous avons identifié à partir des articles du domaine comment certaines de ces caractéristiques, puis nous les avons regroupées en cinq catégories décrites ci-dessous.

Variabilité Une grande variabilité dans les données permet un transfert plus facile du modèle pré-entraîné aux tâches en aval, quand celles-ci utilisent des données similaires aux données de pré-entraînement. C'est le cas de CLIP (Radford *et al.*, 2021), qui utilise le vocabulaire de Wikipédia pour collecter les images, afin de couvrir une grande diversité d'objets. Les auteurs l'évaluent sur de nombreuses tâches et observent, sans entraînement supplémentaire, des résultats compétitifs aux modèles spécialisés sur ces tâches. Cependant, les auteurs observent aussi que CLIP montre une faible généralisation sur des données hors distribution, comme celles utilisant des images hors du domaine de pré-entraînement. Ainsi, plus le jeu de données de pré-entraînement couvre une variété d'éléments visuels importante, meilleure sera la performance du modèle. Le modèle BLIP (Li *et al.*, 2022) utilise des légendes générées automatiquement pour augmenter les données de pré-entraînement. Les auteurs constatent que générer des légendes ayant une plus grande variabilité augmente les performances du modèle, plutôt que de générer des légendes plus probables.

Exactitude Les auteurs de BLIP (Li *et al.*, 2022) constatent également que l'utilisation de données inexactes pendant le pré-entraînement a un effet négatif sur les performances, et mettent au point une technique de filtrage pour les éliminer. En étudiant les performances d'un modèle sur différents jeux de données, (Hendricks *et al.*, 2021) montrent qu'un modèle pré-entraîné sur SBU (Ordonez

et al., 2011) donne de moins bonnes performances sur les tâches en aval que ceux entraînés sur des jeux de données plus petits, comme MS COCO. Ils constatent que les données de SBU présentent moins de chevauchement entre les objets et les mots du texte que d'autres jeux de données. Cela semble cohérent, car la méthode de filtrage utilisée pour générer SBU repose peu sur la similarité entre texte et image. De plus, (Hendricks & Nematzadeh, 2021) montre que des modèles entraînés sur des données annotées manuellement comme MS COCO (Lin *et al.*, 2014), qui sont moins bruitées, sont plus sensibles aux légères différences sémantiques entre deux images que des modèles entraînés sur des données collectées automatiquement, comme Conceptual Captions (Sharma *et al.*, 2018), qui parviennent moins à les distinguer.

Compositionnalité En fonction du type de jeu de données, une certaine proportion d'images ne montrent qu'un seul objet, tandis que d'autres montrent de multiples objets avec diverses interactions. De plus, les annotations peuvent se concentrer sur le point central, tandis que d'autres peuvent décrire les relations entre les divers objets. Nous appelons la compositionnalité d'une instance le nombre d'éléments distincts décrits par une annotation et la complexité de leurs relations. Ainsi, selon (Nikolaus *et al.*, 2022), la présence pendant le pré-entraînement de légendes manuellement annotées et plus descriptives peut aider les modèles à mieux comprendre les dépendances multimodales. De même, l'utilisation de jeux de données favorisant le raisonnement spatial semble nécessaire à la compréhension multimodale des concepts de position (Salin, 2022), comme le fait LXMERT (Tan & Bansal, 2019) en utilisant pendant le pré-entraînement des jeux de données de raisonnement visuel (VQA(Antol *et al.*, 2015), GQA(Hudson & Manning, 2019), VG-QA(Zhu *et al.*, 2016)).

Biais Les modèles d'apprentissages automatique amplifient les biais présents dans leurs jeux de données (Zhao *et al.*, 2017). En effet, les données et les annotations sont deux des cinq principales sources de bais de ces modèles (Hovy & Prabhumoye, 2021). Le transformers vision-langage sont notamment sujets à des biais de genre (Hendricks *et al.*, 2018). Ces modèles se reposent aussi parfois sur des biais textuels plus que sur des informations visuelles (Goyal *et al.*, 2017). En outre, les corpus collectés sont généralement fortement biaisés en faveur de la culture occidentale, et les performances de ces modèles chutent sur des exemples hors de ce domaine (Liu *et al.*, 2021).

Similarité entre pré-entraînement et fine-tuning Dans (Singh *et al.*, 2020), les auteurs montrent que la similarité entre les données de pré-entraînement et d'évaluation peut impacter fortement les performances d'une tâche. De même, (Hendricks *et al.*, 2021) constatent qu'en prenant deux jeux de données avec les mêmes images, celui qui a une plus forte similitude de langage (calculée grâce à la perplexité) avec les données de la tâche en aval conduira à de meilleures performances dans cette tâche. Ainsi, si la méthode d'annotation des images d'évaluation varie fortement par rapport celles du pré-entraînement, les modèles peuvent observer une baisse de performance.

Nous essayons d'établir quelles caractéristiques des données influencent les performances des modèles. Certaines études insistent sur l'importance d'avoir un jeu de données de taille toujours plus grande, alors que d'autres accordent beaucoup d'importance à d'autres caractéristiques des données. La taille d'un jeu de données, qui est généralement corrélée avec la diversité de ces données, permet une utilisation dans un plus grand nombre de domaines. La compositionnalité d'un jeu de donnée semble nécessaire pour des raisonnements multimodaux plus précis. De plus, ces caractéristiques peuvent avoir différents impacts sur la compréhension textuelle, visuelle et

multimodale des modèles. En effet, une grande diversité monomodale des données, qui peut être apportée par de très larges jeux de données, peut cacher une plus faible diversité multimodale, due à un manque de compositionnalité ou d'interactions multimodales. Il est donc important d'étudier ces aspects individuellement lors de l'évaluation de la qualité d'un jeu de données de pré-entraînement.

5 Évaluer la qualité d'un jeu de données texte-image

Dans cette section, nous voulons évaluer un jeu de données de pré-entraînement pour un modèle qui serait utilisé sur une grande variété de domaines et de tâches. Les relations vision-langage peuvent différer en fonction de leur *information mutuelle*, *statut* et *corrélation sémantique* comme le décrit (Otto *et al.*, 2019). Dans les jeux de données usuels des tâches vision-langage, le texte est subordonné à l'image, et sert d'ancrage (*anchorage*) en offrant une manière de décrire l'image. Nous voulons donc évaluer la qualité d'un tel jeu de données en nous concentrant sur les images et les textes de manière indépendante, ainsi que la relation d'ancrage de ces images et textes. Nous étudions des méthodes permettant de vérifier que les données répondent à certains critères de la variabilité, d'exactitude, de compositionnalité et de réduction des biais, en réponse aux analyses de la partie 4.

5.1 Évaluation des textes

Différentes méthodes ont été développées pour évaluer la qualité d'un corpus textuel. Les auteurs de (Mishra *et al.*, 2020) étudient en détail des mesures de qualité pour les corpus de traitement automatique du langage. Celles-ci qui peuvent être appliquées à l'évaluation d'un jeu de données image-texte.

Le vocabulaire du jeu de données, en prenant en compte les parties de discours, peut donner des indications quant à la diversité de ces données. On peut par exemple utiliser comme métrique le rapport entre la taille du vocabulaire et celle du corpus ou la présence de mots de divers domaines. Une autre manière d'assurer la variabilité est d'évaluer la similarité entre les textes du corpus, notamment au niveau de leur structure syntaxique. L'étude de la diversité des structures, notamment au niveau des parties de discours et de la taille des textes, permet aussi de mieux appréhender la compositionnalité des textes. Dans (Mishra *et al.*, 2020), des méthodes de réduction de biais sont également proposées, comme la pseudonymisation des entités nommées. Le biais lié aux stéréotypes peut également être étudié en examinant la fréquence de N-grammes se rapportant à certains mots. Cela peut également permettre d'éviter une présence trop forte de biais textuels dans le corpus. Une manière de contrôler l'exactitude des textes est de limiter le bruit des données. Dans (Baldwin *et al.*, 2013), les auteurs comparent des textes issus de réseaux sociaux, pour évaluer le bruit des données. En analysant les mots hors vocabulaire et la grammaticalité des textes, ils évaluent d'une certaine manière l'exactitude des données collectées. De plus, ces auteurs proposent des méthodes pour nettoyer les textes afin de les rendre moins bruités.

5.2 Évaluation des images

La qualité d'un jeu de données de pré-entraînement pour la vision est d'abord déterminée par la qualité des images elles-mêmes, ce qui est compliqué de faire automatiquement. La taille des images

et les métadonnées associées peuvent fournir une aide pour cette évaluation.

Il est également important d'avoir une variabilité des images. On peut rechercher une couverture importante des possibles catégories, tels que des objets (Deng *et al.*, 2009), des scènes (Zhou *et al.*, 2017) ou des actions (Poppe, 2010). Pour cela, des ressources telles que Wikipédia ou WordNet (Miller, 1995) sont disponibles et peuvent être utilisées en lien avec des modèles détecteurs d'objets. En plus des objets représentés dans les images, il est intéressant d'obtenir diverses configurations d'images, comme des images iconiques et non iconiques, avec des dispositions et des nombres d'objets différents. Ceci peut être important pour favoriser la compréhension de la compositionnalité et une meilleure adaptation à divers domaines. Quant au bruit lié aux images, il peut être évalué en étudiant les catégories. En effet, une partie du biais d'un jeu de données peut provenir de l'utilisation de catégories bruitées, mal équilibrées ou contenant des stéréotypes (Prabhu & Birhane, 2020).

5.3 Évaluation de l'ancrage

La vérification de l'exactitude de l'ancrage requiert soit des annotations manuelles, soit l'utilisation d'un modèle vision-langage dont la performance n'est pas assurée. Ainsi, sur de grands jeux de données comme LAION, le modèle CLIP est utilisé pour mesurer la similarité entre texte et image. Cependant, le seuil de similarité requis semble fixé arbitrairement à partir d'observations d'annotateurs humains. Cela ne garantit pas la qualité de l'ancrage, notamment pour des différences texte-image fines qui peuvent ne pas être observées par CLIP. L'utilisation de modèles supplémentaires, comme les détecteurs d'objets, utilisés par les auteurs de Conceptual Captions, pourrait rendre l'évaluation plus robuste, et moins sensible aux spécificités d'un seul modèle. D'autre part, ces méthodes d'évaluation peuvent nuire à la diversité des données. En effet, CLIP est lui-même entraîné à partir de données collectées automatiquement, et peut reproduire les biais de ses données d'entraînement pendant l'évaluation. C'est pourquoi il serait intéressant d'utiliser conjointement plusieurs modèles entraînés sur des domaines ou tâches différents pour avoir différentes manières de juger l'ancrage texte-image des données.

Une autre problématique liée à l'évaluation de l'ancrage texte-image est le biais de l'annotation, qui se retrouve aussi en traitement automatique du langage (Geva *et al.*, 2019). En effet, dans de nombreux exemples de relation texte-image, celle-ci est subjective et dépend des annotateurs. Ils peuvent choisir de décrire différentes parties d'une image, en employant un langage varié, influencé par leurs diverses cultures. Il n'y a donc pas un unique ancrage textuel possible pour chaque image. Pour limiter ce biais, il est important d'obtenir des données provenant de sources variées, car un seul annotateur ne pourra être représentatif de cette subjectivité (Aroyo & Welty, 2015). Dans le cas d'annotations manuelles, il serait intéressant de donner différentes consignes d'annotation pour varier le type d'ancrage, avec diverses approches de description. Bien qu'il ne soit pas possible d'enlever tous les biais, la diversification des sources et la variabilité des textes et images peut aider à les atténuer.

Nous utilisons les méthodes développées dans cette section pour évaluer la qualité de deux jeux de données en Annexe A. Nous observons ainsi que LAION (Schuhmann *et al.*, 2021), qui est collecté automatiquement, montre une plus forte diversité de vocabulaire, mais une moindre compositionnalité que MS-COCO (Lin *et al.*, 2014), qui est annoté manuellement. Les annotations sont également moins descriptives et présentent plus d'information sur les métadonnées. Il serait

intéressant d'ajouter une étape d'analyse syntaxique lors du filtrage automatique des données afin de sélectionner des légendes plus descriptives et moins biaisées (moins de métadonnées et de noms propres). De plus, l'utilisation de détecteurs d'objets permettrait de sélectionner une plus grande diversité d'images en termes de nombre et catégories d'objets. Ainsi, une évaluation de la syntaxe et du vocabulaire des textes, ainsi que des objets contenus dans les images, permettrait d'améliorer la diversité et la compositionnalité des instances et de réduire l'impact du biais.

6 Discussions sur les problématiques éthiques

La collection d'un ensemble de données pour le pré-entraînement de modèles vision-langage peut soulever des questions éthiques, selon la manière dont les données sont obtenues et filtrées.

— L'utilisation de données provenant d'internet soulève la question du consentement lors de l'acquisition de ces données. Par exemple, Conceptual Captions (Sharma *et al.*, 2018) ou LAION (Schuhmann *et al.*, 2021) contiennent des images qui ne font pas l'objet d'une licence explicite. Certains collectent également des images et des textes contenant des informations personnelles sans demander le consentement des personnes concernées. Cela peut inclure des images provenant de sources problématiques et des données illégales.(Birhane *et al.*, 2021).

— Dans le cas d'ensembles de données annotées par des humains ou qui nécessitent des évaluations humaines, il est important de tenir compte de la rémunération des travailleurs et de leurs conditions de travail, par exemple de l'impact psychologique des contenus préjudiciables (Díaz *et al.*, 2022).

— Les données peuvent faire l'objet de biais nuisibles, en particulier dans le cas de données peu filtrées. L'une des formes de ces biais les plus courantes est l'absence de représentation de certains groupes sociaux (Zhao *et al.*, 2021), que l'on appelle 'biais de représentation'. Il peut avoir un impact important dans les applications en aval (Birhane *et al.*, 2021). Pour atténuer ce biais, il est important de tenir compte des sources de données visuelles et, lorsqu'il y a des annotateurs humains, de la diversité des expériences de ces annotateurs. (Díaz *et al.*, 2022).

7 Conclusion

Comme en traitement automatique des langues, l'utilisation de très grands jeux de données améliore très fortement le pré-entraînement des transformers vision-langage. Cependant, l'utilisation de plus grandes quantités de données conduit à un coût environnemental et économique non négligeable, ce qui rend leur démocratisation difficile. D'autres problèmes éthiques sont également soulevés, comme l'absence de consentement explicite lors de la collecte de ces données. Outre la taille des données, d'autres caractéristiques des jeux de pré-entraînement peuvent avoir un impact majeur sur la performance des modèles. En étudiant diverses analyses des transformers vision-langage, nous regroupons ces caractéristiques en variabilité, exactitude, compositionnalité et biais. Nous défendons un filtrage plus précis des grands jeux de données texte-image, pour mieux répondre à ces critères. Nous espérons qu'en accentuant l'importance de la qualité des données plutôt que leur quantité, le pré-entraînement des modèles vision-langage devienne plus efficace, et moins coûteux. Des études supplémentaires permettraient d'affiner les règles proposées.

8 Remerciements

Je voudrais remercier mes encadrants Benoit Favre et Stéphane Ayache, ainsi que les relecteurs pour leurs commentaires et suggestions.

Ces travaux ont bénéficié d'un accès aux ressources en IA de l'IDRIS au travers de l'allocation de ressources 2023-AD011013880 attribuée par GENCI. Ce travail a bénéficié d'une aide du gouvernement français au titre du Programme Investissements d'Avenir Initiative d'Excellence d'Aix-Marseille Université - A*MIDEX (Institut Archimède AMX-19-IET-009)

Références

ABADJI J., SUAREZ P. O., ROMARY L. & SAGOT B. (2022). Towards a cleaner document-oriented multilingual crawled corpus. In *International Conference on Language Resources and Evaluation*.

AGRAWAL H., DESAI K., WANG Y., CHEN X., JAIN R., JOHNSON M., BATRA D., PARIKH D., LEE S. & ANDERSON P. (2019). Nocaps : Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, p. 8948–8957.

ANTOL S., AGRAWAL A., LU J., MITCHELL M., BATRA D., ZITNICK C. L. & PARIKH D. (2015). Vqa : Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, p. 2425–2433.

AROYO L. & WELTY C. (2015). Truth is a lie : Crowd truth and the seven myths of human annotation. *AI Magazine*, **36**(1), 15–24.

BALDWIN T., COOK P., LUI M., MACKINLAY A. & WANG L. (2013). How noisy social media text, how diffrnt social media sources ? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, p. 356–364.

BENDER E. M., GEBRU T., MCMILLAN-MAJOR A. & SHMITCHELL S. (2021). On the dangers of stochastic parrots : Can language models be too big ? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, p. 610–623.

BIRHANE A., PRABHU V. U. & KAHEMBWE E. (2021). Multimodal datasets : misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv :2110.01963*.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A. *et al.* (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

BYEON M., PARK B., KIM H., LEE S., BAEK W. & KIM S. (2022). Coyo-700m : Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

CHANGPINYO S., SHARMA P., DING N. & SORICUT R. (2021). Conceptual 12M : Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*.

CHEN Y.-C., LI L., YU L., EL KHOLY A., AHMED F., GAN Z., CHENG Y. & LIU J. (2019). Uniter : Learning universal image-text representations.

DENG J., DONG W., SOCHER R., LI L.-J., LI K. & FEI-FEI L. (2009). Imagenet : A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, p. 248–255 : Ieee.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

DÍAZ M., KIVLICHAN I. D., ROSEN R., BAKER D., AMIRONESEI R., PRABHAKARAN V. & DENTON E. L. (2022). Crowdworksheets : Accounting for individual and collective identities underlying crowdsourced dataset annotation. *2022 ACM Conference on Fairness, Accountability, and Transparency*.

DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J. & HOULSBY N. (2021). An image is worth 16x16 words : Transformers for image recognition at scale. In *International Conference on Learning Representations*.

GEVA M., GOLDBERG Y. & BERANT J. (2019). Are we modeling the task or the annotator ? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv :1908.07898*.

GOYAL Y., KHOT T., SUMMERS-STAY D., BATRA D. & PARIKH D. (2017). Making the v in VQA Matter : Elevating the Role of Image Understanding in Visual Question Answering. p. 6904–6913.

HENDRICKS L. A., BURNS K., SAENKO K., DARRELL T. & ROHRBACH A. (2018). Women also Snowboard : Overcoming Bias in Captioning Models. p. 771–787.

HENDRICKS L. A., MELLOR J. F. J., SCHNEIDER R., ALAYRAC J.-B. & NEMATZADEH A. (2021). Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, **9**, 570–585.

HENDRICKS L. A. & NEMATZADEH A. (2021). Probing Image-Language Transformers for Verb Understanding. In *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, p. 3635–3644, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.findings-acl.318.

HENDRYCKS D., LIU X., WALLACE E., DZIEDZIC A., KRISHNAN R. & SONG D. X. (2020). Pretrained transformers improve out-of-distribution robustness. In *Annual Meeting of the Association for Computational Linguistics*.

HONNIBAL M. & MONTANI I. (2017). spaCy 2 : Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

HOVY D. & PRABHUMOYE S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, **15**(8), e12432.

HUDSON D. A. & MANNING C. D. (2019). Gqa : a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv :1902.09506*, **3**(8), 1.

KAY M., MATUSZEK C. & MUNSON S. A. (2015). Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, p. 3819–3828.

KREUTZER J., CASWELL I., WANG L., WAHAB A., VAN ESCH D., ULZII-ORSHIKH N., TAPO A., SUBRAMANI N., SOKOLOV A., SIKASOTE C. *et al.* (2022). Quality at a glance : An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, **10**, 50–72.

KRISHNA R., ZHU Y., GROTH O., JOHNSON J., HATA K., KRAVITZ J., CHEN S., KALANTIDIS Y., LI L.-J., SHAMMA D. A., BERNSTEIN M. & FEI-FEI L. (2016). Visual genome : Connecting language and vision using crowdsourced dense image annotations.

LI J., LI D., XIONG C. & HOI S. C. H. (2022). Blip : Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.

LI J., SELVARAJU R., GOTMARE A., JOTY S., XIONG C. & HOI S. C. H. (2021). Align before fuse : Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, **34**, 9694–9705.

LIN T.-Y., MAIRE M., BELONGIE S., HAYS J., PERONA P., RAMANAN D., DOLLÁR P. & ZITNICK C. L. (2014). Microsoft coco : Common objects in context. In *Computer Vision–ECCV 2014 : 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, p. 740–755 : Springer.

LIU F., BUGLIARELLO E., PONTI E. M., REDDY S., COLLIER N. & ELLIOTT D. (2021). Visually Grounded Reasoning across Languages and Cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 10467–10485, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics. DOI : 10.18653/v1/2021.emnlp-main.818.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019). Roberta : A robustly optimized bert pretraining approach. *arXiv preprint arXiv :1907.11692*.

MILLER G. A. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 39–41.

MISHRA S., ARUNKUMAR A., SACHDEVA B., BRYAN C. & BARAL C. (2020). Dqi : Measuring data quality in nlp. *arXiv preprint arXiv :2005.00816*.

NIKOLAUS M., SALIN E., AYACHE S., FOURTASSI A. & FAVRE B. (2022). Do vision-and-language transformers learn grounded predicate-noun dependencies ? *arXiv preprint arXiv :2210.12079*.

ORDONEZ V., KULKARNI G. & BERG T. (2011). Im2text : Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, **24**.

OTTO C., SPRINGSTEIN M., ANAND A. & EWERTH R. (2019). Understanding, categorizing and predicting semantic image-text relations. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, p. 168–176.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A. *et al.* (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv :2203.02155*.

POPPE R. (2010). A survey on vision-based human action recognition. *Image and vision computing*, **28**(6), 976–990.

PRABHU V. U. & BIRHANE A. (2020). Large image datasets : A pyrrhic win for computer vision ? *arXiv preprint arXiv :2006.16923*.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J. *et al.* (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, p. 8748–8763 : PMLR.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D., SUTSKEVER I. *et al.* (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1**(8), 9.

REN S., HE K., GIRSHICK R. & SUN J. (2015). Faster r-cnn : Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, **28**.

SALIN E. (2022). Etude de la compréhension spatiale multimodale des modèles transformers vision-langage. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, p. 181–187 : CNRS.

SCAO T. L., FAN A., AKIKI C., PAVLICK E., ILIĆ S., HESSLOW D., CASTAGNÉ R., LUCCIONI A. S., YVON F., GALLÉ M. *et al.* (2022). Bloom : A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv :2211.05100*.

SCHUHMANN C., BEAUMONT R., VENCU R., GORDON C., WIGHTMAN R., CHERTI M., COOMBES T., KATTA A., MULLIS C., WORTSMAN M. *et al.* (2022). Laion-5b : An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv :2210.08402*.

SCHUHMANN C., VENCU R., BEAUMONT R., KACZMARCZYK R., MULLIS C., KATTA A., COOMBES T., JITSEV J. & KOMATSUZAKI A. (2021). Laion-400m : Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv :2111.02114*.

SCHWARTZ R., DODGE J., SMITH N. A. & ETZIONI O. (2020). Green ai. *Communications of the ACM*, **63**(12), 54–63.

SHARMA P., DING N., GOODMAN S. & SORICUT R. (2018). Conceptual captions : A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

SINGH A., GOSWAMI V. & PARIKH D. (2020). Are we pretraining it right ? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv :2004.08744*.

STRUBELL E., GANESH A. & MCCALLUM A. (2019). Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv :1906.02243*.

SUAREZ P. O., SAGOT B. & ROMARY L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures.

SUHR A., ZHOU S., ZHANG A., ZHANG I., BAI H. & ARTZI Y. (2019). A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 6418–6428, Florence, Italy : Association for Computational Linguistics. DOI : 10.18653/v1/P19-1644.

SUN C., SHRIVASTAVA A., SINGH S. & GUPTA A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, p. 843–852.

TAN H. & BANSAL M. (2019). LXMERT : Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 5100–5111, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1514.

TOUVRON H., CORD M., DOUZE M., MASSA F., SABLAYROLLES A. & JÉGOU H. (2021). Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, p. 10347–10357 : PMLR.

WENZEK G., LACHAUX M.-A., CONNEAU A., CHAUDHARY V., GUZMÁN F., JOULIN A. & GRAVE E. (2019). Ccnet : Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv :1911.00359*.

YANG J., DUAN J., TRAN S., XU Y., CHANDA S., CHEN L., ZENG B., CHILIMBI T. & HUANG J. (2022). Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 15671–15680.

ZHAI X., KOLESNIKOV A., HOULSBY N. & BEYER L. (2022). Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 12104–12113.

ZHAO D., WANG A. & RUSSAKOVSKY O. (2021). Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, p. 14830–14840.

ZHAO J., WANG T., YATSKAR M., ORDONEZ V. & CHANG K.-W. (2017). Men also like shopping : Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv :1707.09457*.

ZHOU B., LAPEDRIZA A., KHOSLA A., OLIVA A. & TORRALBA A. (2017). Places : A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, **40**(6), 1452–1464.

ZHU Y., GROTH O., BERNSTEIN M. & FEI-FEI L. (2016). Visual7w : Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 4995–5004.

A Comparaison de MS COCO et LAION-400

En utilisant les méthodes d'évaluation développées dans la section 5, nous comparons deux corpus collectés manuellement et automatiquement : MS COCO (Lin *et al.*, 2014) et LAION-400 (Schuhmann *et al.*, 2021). Nous étudions un sous-ensemble S de 1000 instances de chaque jeu de données et comparons les jeux de données sur plusieurs mesures :

— Variabilité du vocabulaire : En appelant d la taille du dictionnaire de S et n le nombre de mots de S, nous calculons $V = d/n$.
On obtient $V_{COCO} = 0.04$ et $V_{Laion} = 0.10$.
LAION montre ainsi une plus grande diversité de vocabulaire que MS COCO.

— Biais du vocabulaire : Afin d'obtenir plus d'information sur le vocabulaire utilisé par ces jeux de données, nous étudions quels groupes de mots sont les plus utilisés. Cela peut permettre d'étudier le biais de chaque sous-ensemble S. En particulier, après avoir éliminé les mots vide, nous regardons quel groupe G de deux mots se retrouve le plus fréquemment dans une même instance S.
On obtient $G_{COCO} =$ (group, people) et $G_{Laion} =$ (stock, photo).
Cela semble montrer que COCO contient des textes descriptifs, alors que les textes de Laion contiennent également beaucoup de métadonnées.

— Nombre d'objets : Le nombre d'objets dans une image permet d'avoir plus d'information sur la complexité de la composition de cette image. Nous calculons à laide d'un détecteur d'objet Faster RCNN (Ren *et al.*, 2015). le nombre d'objets par image de S, et on rapporte la médiane M et le troisième quartile Q.
On obtient $M_{COCO} = 4$, $Q_{COCO} = 8$ et $M_{Laion} = 1$, $Q_{Laion} = 2$.
On observe que les images de COCO contiennent plus d'objets que celles de LAION. Cette méthode a cependant des limites, notamment parce que les images du jeu de données LAION sont peu semblables à celles utilisées pour l'entraînement du détecteur d'objet.

— Syntaxe : Étudier la syntaxe d'un texte peut également nous donner plus d'information sur la qualité de celui-ci. Pour ce faire, nous utilisons les outils de Spacy (Honnibal & Montani, 2017). Nous observons les deux plus fréquentes étiquettes de partie de discours $P1$ et $P2$. On obtient $P1_{COCO}$ = Nom et $P2_{COCO}$ = Déterminant, $P1_{Laion}$ = Nom et $P2_{Laion}$ = Nom propre.
La prévalence de noms propres dans le jeu de données LAION peut augmenter le risque de biais.