Study of the multimodal understanding of vision-language transformer models

Emmanuelle Salin

Aix*Marseille université
Socialement engagée

NNT : 2023AIXM0425

THÈSE DE DOCTORAT

Soutenue à Aix-Marseille Université
le 29 novembre 2023, par

Emmanuelle SALIN

Étude de la compréhension multimodale des modèles transformeurs vision-langage

Discipline
Informatique

École doctorale
ED 184 MATHEMATIQUES ET INFORMATIQUE

Laboratoire/Partenaires de recherche
Laboratoire d'Informatique et Systèmes

Composition du jury

Eric GAUSSIER Professeur Université Grenoble Alpes	Rapporteur
Ewa KIJAK Maître de conférences Université Rennes 1	Rapporteure
Cecile CAPPONI Professeure Aix Marseille Université	Présidente du jury
Camille GUINAUDEAU Maître de conférences Université Paris Saclay	Examinatrice
Diane LARLUS Research Scientist Naverlabs Europe	Examinatrice
Benjamin LECOUTEUX Professeur Université Grenoble Alpes	Examinateur
Stephane AYACHE Professeur Aix Marseille Université	Co-directeur de thèse
Benoit FAVRE Professeur Aix Marseille Université	Directeur de thèse

LIS Laboratoire d'Informatique & Systèmes
cnrs
Institut Archimède Mathématiques-Informatique
Aix*Marseille Université

Affidavit

I, undersigned, Emmanuelle Salin, hereby declare that the work presented in this manuscript is my own work, carried out under the scientific supervision of Benoit Favre and Stéphane Ayache, in accordance with the principles of honesty, integrity and responsibility inherent to the research mission. The research work and the writing of this manuscript have been carried out in compliance with both the French national charter for Research Integrity and the Aix-Marseille University charter on the fight against plagiarism.

This work has not been submitted previously either in this country or in another country in the same or in a similar version to any other examination body.

Marseille, 21/09/2023

Liste de publications et participation aux conférences

Liste des publications réalisées dans le cadre du projet de thèse :

1. Salin, E., Farah, B., Ayache, S., & Favre, B. (2022, June). Are vision-language transformers learning multimodal representations? A probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 11248-11257)
2. Nikolaus, M., Salin, E., Ayache, S., Fourtassi, A., & Favre, B. (2022, December). Do Vision-and-Language Transformers Learn Grounded Predicate-Noun Dependencies? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 1538-1555).
3. Salin, E. (2022, November). 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)* (pp. 181-187). CNRS.
4. Salin, E. (2023). État des lieux des Transformers Vision-Langage : Un éclairage sur les données de pré-entraînement. In *18e Conférence en Recherche d'Information et Applications*
   16e Rencontres Jeunes Chercheurs en RI
   30e Conférence sur le Traitement Automatique des Langues Naturelles
   25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (pp. 14-29). ATALA.
5. (Forthcoming) Salin, E., Ayache, S. & Favre, B. (2023) Towards an Exhaustive Evaluation of Vision-Language Foundation Models. In *ICCV Workshops 2023*.

Participation aux conférences au cours de la période de thèse :

1. AAAI Conference on Artificial Intelligence, AAAI 2022
2. Language Resources and Evaluation Conference, LREC 2022
3. Conférence Nationale en Intelligence Artificielle, CNIA 2022
4. Conference on Empirical Methods in Natural Language Processing, EMNLP 2022

5. Conférence sur le Traitement Automatique des Langues Naturelles, TALN 2023
6. International Conference on Computer Vision, ICCV 2023

Participation aux écoles d'été au cours de la période de thèse :

1. International School on Deep Learning, DeepLearn 2022 Spring
2. École d'été en Traitement Automatique des Langues, ETAL 2023

Résumé

Les domaines du traitement du langage naturel et de la vision par ordinateur ont connu une forte croissance au cours des dernières années, ce qui a conduit à l'émergence de *modèles de fondation*. Ces modèles visent à apprendre des représentations générales en utilisant une grande quantité de données. L'architecture Transformer a joué un rôle majeur dans le développement de ces modèles. Basée sur le mécanisme de l'attention, l'architecture Transformer permet le pré-entraînement de modèles sur de très grands jeux de données, qui peuvent ensuite être adaptés à un large éventail de tâches.

Cependant, les modèles monomodaux ont des difficultés à associer le langage à d'autres modalités, telles que la vision. Cela a conduit à un intérêt croissant pour l'apprentissage automatique multimodal. Les modèles vision-langage basés sur l'architecture Transformer, en particulier, ont permis des améliorations significatives par rapport à l'état de l'art en apprentissage automatique multimodal. Ces modèles sont pré-entraînés sur des tâches prétextes textuelles, visuelles et multimodales, en utilisant des jeux de données multimodaux, généralement composés de paires *(texte, image)*. Ils apprennent des représentations multimodales qui peuvent servir de base à un large éventail d'applications vision-langage.

Cependant, le développement rapide des transformeurs vision-langage a laissé peu de temps pour une étude plus approfondie de ces modèles. En effet, nous comprenons encore mal comment les choix de conception affectent les compétences de ces modèles. Nous ne savons également pas précisément quelles informations ces modèles doivent apprendre à extraire pour être utiles aux applications vision-langage concrètes.

À travers cette thèse, notre objectif est de parvenir à une meilleure compréhension des modèles vision-langage, à travers le prisme des modèles basés sur les transformeurs. En particulier, nous étudions la capacité de ces modèles à apprendre des représentations multimodales qui peuvent servir de base à un large éventail de tâches vision-langage. Nous nous concentrons aussi sur la façon dont différents choix de pré-entraînement peuvent avoir un impact sur leurs performances. Les méthodes que nous développons ont également pour but d'aider à l'étude de futurs modèles vision-langage, indépendamment de leur architecture.

Nous commençons par donner un aperçu du domaine de la multimodalité vision-langage. En particulier, nous présentons la diversité des modèles de transformeurs vision-langage pour donner du recul sur les avancées actuelles du domaine. Ensuite, nous remettons en question les méthodes actuellement utilisées pour évaluer les transformeurs vision-langage. Nous soutenons que plutôt que d'évaluer ces modèles sur quelques tâches complexes, il serait intéressant de mieux appréhender la com-

préhension multimodale générale de ces modèles. Ainsi, nous proposons d'examiner les compétences multimodales de ces modèles. Pour cela, nous créons une taxonomie des compétences multimodales en multimodalité vision-langage. Ensuite, nous développons des tâches d'évaluation et des jeux de données pour sonder les transformeurs vision-langage sur des compétences textuelles, visuelles et multimodales. Nous constatons que ces modèles ont des difficultés à appréhender certains concepts à un niveau multimodal, comme la position des objets. Enfin, nous élaborons différents protocoles de pré-entraînement afin d'étudier comment les choix de conception influencent les performances des modèles vision-langage.

Mots clés : Apprentissage automatique, Traitement automatique du langage, Vision par ordinateur, Multimodalité vision-langage, Transformeurs, Modèles de fondation

Abstract

The fields of Natural Language Processing and Computer Vision have experienced strong growth during the last few years, leading to the emergence of *foundation models*. These models aim to learn general representations using a large amount of data. The Transformer architecture has played a major role in the development of such models. Based on the attention mechanism, the Transformer architecture enables the pre-training of models on large-scale datasets. These models can then be adapted to many tasks in different domains.

However, monomodal models lack grounding in real-world experiences. Thus, they have difficulties associating language to other modalities, such as vision. This has resulted in a growing interest in multimodal machine learning. Vision-language models based on the Transformer architecture, in particular, have enabled significant improvement over previous state-of-the-art in multimodal machine learning. Those models are pre-trained on textual, visual and multimodal pretext tasks, using multimodal datasets, usually made of *(text, image)* pairs. They learn multimodal representations that can serve as a basis in vision-language applications.

However, the fast-paced development of vision-language transformers has left little time for a deeper study of those models. Indeed, we still have a poor understanding of how design choices affect their skills and generalization abilities. We also lack insight into what information these models should be able to extract for real-world multimodal applications.

In this thesis, our goal is to reach a better understanding of vision-language models, through the lens of transformer-based models. In particular, we study the ability of those models to learn multimodal representations that can be a basis for a wide range of vision-language tasks. We also focus on how different pre-training choices can impact their performances. The methods we develop in this thesis are also aimed at the study of general-purpose vision-language models, irrespective of their architecture.

We first provide an overview of vision-language multimodality in machine learning. In particular, we introduce and compare a number of vision-language transformer models to provide hindsight on the current advances of the field. Then, we question the current methods used to evaluate vision-language transformers. We argue that rather than evaluating such models on a few complex tasks, it would be interesting to get a better apprehension of their general multimodal understanding. To that end, we propose to consider granular multimodal capabilities of vision-language models, and make a first attempt at a taxonomy of vision-language capabilities. Subsequently, we develop evaluation tasks and datasets to probe state-of-the-art vision-language transformers on specific textual, visual and multimodal capabilities. We find that models have difficulty apprehending some concepts at a multimodal level, such

as object position. Finally, we elaborate different pre-training protocols to study how design choices affect the performances of vision-language models on those capabilities.

Keywords: Machine Learning, Natural Language Processing, Computer Vision, Vision-Language Multimodality, Transformers, Foundation Models

Résumé Long

Les domaines du traitement du langage naturel et de la vision par ordinateur ont connu une forte croissance au cours des dernières années. Celle-ci a conduit à l'émergence de *modèles de fondation*, qui visent à apprendre des représentations en utilisant une grande quantité de données non étiquetées. Ces représentations peuvent ensuite être utilisées dans un grand éventail de tâches en aval.

L'architecture Transformer a joué un rôle majeur dans le développement de ces modèles. Basée sur le mécanisme de l'attention, elle est utilisée dans de nombreux domaines, comme ceux du Traitement Automatique du Langage (i.e., TAL) et de la vision par ordinateur. Dans cette thèse, je m'intéresse particulièrement à la fusion des modalités visuelles et textuelles. En effet, textes et images sont souvent complémentaires dans nos communications, et les informations présentent sous des formats de données divers peuvent être nécessaires à la réalisation d'une même tâche. Divers types de tâches vision-langage existent, basées sur la génération ou la classification, comme les tâches de dialogue multimodal ou de recherche d'image et texte. Cela a donc soulevé le besoin d'architectures fusionnant efficacement ces deux modalités. Les avancées permises à ce sujet par l'architecture Transformer ont conduit à un intérêt croissant pour l'apprentissage automatique multimodal.

Les modèles Transformers vision-langage sont pré-entraînés sur des tâches prétextes textuelles, visuelles et multimodales, en utilisant des jeux de données multimodaux de grande taille, composés de paires *(texte, image)* non étiquetées. Cependant, le développement rapide de ces modèles a laissé peu de temps pour une étude plus approfondie. En effet, nous comprenons encore mal quelles sont les informations multimodales extraites et utilisées par ces modèles, et comment les nombreux choix de conception possibles impactent les performances de ces modèles. Nous ne savons également pas précisément quelles compétences ces modèles doivent acquérir pour être utiles aux diverses applications vision-langage qui continuent à se développer.

État de l'art des modèles vision-langage

À travers cette thèse, mon objectif est de parvenir à une meilleure compréhension des modèles multimodaux, à travers le prisme des Transformers vision-langage. Cela nécessite tout d'abord une étude du domaine de la multimodalité vision-langage. Dans un premier temps, nous présentons dans le Chapitre 1 la diversité des Transformers vision-langage pour prendre du recul sur les récentes avancées. Comme ceux-ci sont en constante évolution, nous avons arrêté l'étude de ces modèles à l'année 2023.

Architecture Les premiers Transformers vision-langage conçus utilisent un détecteur d'objet pour générer des représentations visuelles de l'image. Le pré-traitement des textes est lui inspiré des modèles de langage unimodaux. Les représentations de texte et d'image sont ensuite combinées à l'aide d'une architecture Transformer, qui peut utiliser des mécanismes d'attention classique ou croisée.

Suite au développement des Transformer visuels, certains modèles vision-langage ont adopté une architecture exclusivement basée sur les Transformers, se passant des détecteurs d'objets. Les représentations visuelles sont ainsi formées à partir des pixels eux-mêmes, traités sous le format d'une grille. Par la suite, afin de profiter des grands modèles de langage (LLM) développés ces dernières années, d'autres modèles ont choisi d'incorporer la modalité visuelle à ces LLM pré-entraînés, dont les poids sont en partie gelés pendant la phase de raffinement.

Cependant, il n'y a pas encore de consensus sur l'architecture la plus efficace pour le pré-entraînement vision-langage. Il est d'autant plus difficile de tirer des conclusions concernant l'architecture que la plupart des modèles de pointe ne partagent pas le même protocole de pré-entraînement, les mêmes ensembles de données ou les mêmes tâches de pré-entraînement.

Données de pré-entraînement Les jeux de données sont une autre variable majeure de l'entraînement d'un modèle vision-langage. Les ensembles de données sur lesquels ces modèles sont pré-entraînés sont souvent un facteur limitant. En effet, même si les données utilisées pour les modèles vision-langage sont non étiquetées, la nécessité d'avoir un jeu de donnée appairé entre textes et image rend la conception de ces jeux de données plus difficile et coûteuse.

Dans un premier lieu, les jeux de données de pré-entraînement étaient conçus à l'aide d'annotations manuelles, pour obtenir des légendes descriptives d'images. Cependant, ce procédé est coûteux et passe difficilement à l'échelle. La tendance actuelle est donc d'utiliser des ensembles de données collectés automatiquement. Ceux-ci sont ensuite filtrés pour garantir une meilleure qualité des images et annotations, mais les données obtenues sont en grande partie moins complexes en termes de relation entre le texte et l'image. L'utilisation de jeux de données collectés automatiquement peut également soulever des problématiques éthiques, notamment liées au respect du consentement et de la vie privée.

Tâches de pré-entraînement De nombreuses tâches de pré-entraînement ont été développées pour les modèles vision-langage, souvent inspirées de tâches existantes dans les modalités visuelles et textuelles. Les deux tâches textuelles prédominantes sont les tâches de masquage de mot (Masked Language Modeling ou MLM) et de prédiction du mot suivant. Le but de ces tâches, appliquées aux modèles vision-langage, est d'utiliser le contexte textuel non masqué, ainsi que l'image, pour deviner les mots masqués.

Les tâches de pré-entraînement visuel utilisent un mécanisme similaire. Il s'agit, à partir d'une représentation visuelle d'une partie de l'image, de retrouver la représen-

tation de la partie masquée. Dans le cas des modèles utilisant des détecteurs d'objets, une variante de cette tâche consiste à classifier une zone de l'image en fonction de l'objet présent sur cette image. Cependant, les tâches de pré-entraînement visuelles ne font pas de consensus, et plusieurs modèles récents n'en utilisent pas pendant le pré-entraînement.

Enfin, les tâches multimodales sont utilisées pour entraîner les modèles à distinguer entre des textes qui correspondent aux images et ceux qui ne correspondent pas. Pour cela, deux méthodes sont principalement utilisées : la classification pour la tâche de correspondance image texte (i.e., image-text matching, or ITM), ou l'apprentissage contrastif.

Évaluation Plusieurs tâches ont été créées pour évaluer les performances multimodales des modèles vision-langage, allant du raisonnement multimodal à la recherche de textes et d'images (i.e., image-text retrieval). Cependant, des études ont montré que plusieurs de ces ensembles de données présentent des biais qui faussent l'évaluation. De plus, ces tâches évaluent généralement une compréhension multimodale générale, mais se concentrent rarement sur des capacités précises, ce qui rend l'analyse des faiblesses de ces modèles difficile. Cela complique également l'interprétation des résultats en termes de comparaison des capacités entre deux modèles distincts.

C'est pourquoi je propose dans une deuxième partie une nouvelle méthodologie d'évaluation des Transformers vision-langage.

Taxonomie des capacités vision-langage

En s'inspirant du travail accompli dans l'évaluation des modèles de fondation monomodaux, nous avons voulu ouvrir une discussion sur l'évaluation des Transformers vision-langage. En effet, les modèles vision-langage peuvent être adaptés à de nombreuses applications multimodales, ce qui soulève donc la question de leur évaluation. Il s'agit en effet d'avoir une évaluation représentative des applications possibles. Ces modèles sont habituellement testés sur un petit éventail de tâches complexes, comme la réponse aux questions visuelles et la recherche d'images et de textes. Cependant, des travaux ont révélé des faiblesses dans leur compréhension de certains concepts multimodaux, qui n'étaient pas identifiés à travers cet éventail de tâches. En effet, l'utilisation de tâches complexes n'aide pas à mettre en évidence des faiblesses spécifiques des modèles, car elles ne permettent généralement pas d'isoler les performances d'un modèle vis-à-vis d'un concept précis.

Ainsi, l'étude de capacités spécifiques a montré que les modèles vision-langage ont une compréhension multimodale limitée de la position RÖSCH et al. 2022a ; SALIN et al. 2022 et de l'ordre des mots THRUSH et al. 2022. Cependant, aucune tentative n'a été faite pour concevoir une méthodologie d'évaluation exhaustive de ces modèles à travers une large gamme de capacités.

L'objectif principal de la discussion entamée dans le Chapitre 2 est d'atteindre une meilleure explicabilité des performances de ces modèles dans les diverses applications

possibles. Nous nous intéressons spécifiquement à l'évaluation de la *Compréhension Multimodale Générale*, c'est-à-dire les diverses capacités reliées à la compétence d'un modèle à extraire des informations textuelles et visuelles et à les utiliser dans le cadre de concepts multimodaux.

Nous soutenons donc qu'il est essentiel d'évaluer les performances de ces modèles sur une large gamme de capacités spécifiques, représentatives des applications réelles dans lesquelles ces modèles sont utilisés. Ainsi, nous construisons une méthodologie d'évaluation en considérant d'abord les applications réelles, puis en construisant à partir d'elles une taxonomie des capacités vision-langage. Une telle méthodologie a pour but d'utiliser de pouvoir effectuer un diagnostic des capacités des modèles vision-langage.

À partir d'applications telles que la navigation vision-langage ou le diagnostic médical assisté par ordinateur, nous identifions plusieurs capacités vision-langage. Nous regroupons ensuite ces capacités en quatre catégories, en fonction de leur rôle : les capacités de dénotation, d'ancrage, de connotation et de raisonnement. Enfin, nous mettons en relation cette taxonomie avec des tâches d'évaluation existantes, pour mettre en évidence des manquements dans l'évaluation actuelle des modèles.

Cependant, l'utilisation d'une telle taxonomie présente également des limites, en raison du potentiel biais de conception dans la détermination des capacités jugées utiles. Tout d'abord, elle peut ne pas refléter toutes les applications possibles des modèles fondamentaux vision-langage et est influencée par les tâches déjà existantes. En particulier, elle peut aussi être biaisée à cause d'une certaine optique des relations vision-langage qui n'est pas nécessairement la même en fonction les cultures. De plus, il est difficile d'évaluer la taxonomie. À l'avenir, il serait intéressant de renforcer cette taxonomie à l'aide de perspectives supplémentaires et de compléter la couverture des diverses applications vision-langage.

Dans le cadre de cette thèse, j'étudie en particulier diverses capacités de dénotation, qui peuvent être considérées comme étant le fondement de la compréhension multimodale, et sur lesquelles se reposent les autres capacités.

Étude de diverses capacités des modèles vision-langage

Les tâches de raffinement, qui sont généralement des tâches complexes, ne sont pas suffisantes pour identifier les faiblesses spécifiques des modèles vision-langage. Ainsi, nous créons des tâches pour déterminer quelles informations les modèles pré-entraînés extraient dans leurs représentations multimodales.

Dans le chapitre 3, nous utilisons deux méthodologies d'évaluation différentes : le sondage de modèle, et l'évaluation en utilisant les têtes de pré-entraînement avec leurs poids gelés.

Sondage de modèles d'état de l'art sur des capacités visuelles, textuelles et multimodales

Nous utilisons une série de tâches de sondage pour étudier des capacités monomodales et multimodales des modèles vision-langage. Nous comparons également comment différents choix en matière d'entraînement et d'architecture peuvent affecter ces représentations. Nous étudions spécifiquement les capacités suivantes :

— Compréhension de la syntaxe (Décalage de bigrammes, Étiquetage des parties du discours)
— Compréhension de la structure d'une image (Comptage d'objets)
— Compréhension des détails d'une image (Classification fine d'objets)
— Compréhension multimodale des attributs d'un objet (couleur, position, taille)
— Compréhension multimodale non spécifique à un concept (Identification de différences minimes entre image et texte liées au nombre, aux objets, aux actions)

Nous cherchons à répondre notamment aux questions suivantes : Les modèles vision-langage encodent-ils des informations pertinentes pour ces capacités textuelles, visuelles et multimodales dans leurs représentations ? Si oui, comment les modèles pré-entraînés se comparent-ils aux modèles affinés ? Comment les choix de conception impactent-ils les informations encodées par ces modèles ? Nous construisons des ensembles de données pour chacune de ces tâches et concevons aussi une méthodologie pour évaluer l'impact du biais textuel sur les performances des modèles vision-langage.

En ce qui concerne les capacités linguistiques, nos résultats montrent que les modèles vision-langage ont une compréhension syntaxique légèrement moins bonne que les modèles uniquement linguistiques comme BERT (DEVLIN et al. 2019), ce qui pourrait être dû à la structure syntaxique moins variée des jeux de données de légendes utilisés pour le pré-entraînement. Pour ce qui est de la modalité visuelle, nous constatons que UNITER (Y.-C. CHEN et al. 2019) et LXMERT (TAN et al. 2019), des modèles vision-langage qui reposent sur des détecteurs d'objets Faster R-CNN (REN et al. 2015) présentent des performances significativement inférieures à ViLT (W. KIM et al. 2021), ce qui suggère que l'utilisation de détecteurs d'objets pour extraire des représentations de l'image est un facteur limitant pour la compréhension visuelle des modèles.

De plus, nos résultats montrent que les modèles vision-langage sont capables de capturer certaines informations multimodales comme la couleur des objets, mais ils ont encore du mal à appréhender des concepts tels que la taille et la position des objets. En outre, nous constatons que le raffinement n'améliore pas les performances à ce sujet. Enfin, nos expériences évaluant les biais monomodaux mettent également en évidence le fait que les modèles vision-langage reposent significativement sur les biais textuels pour ces tâches. Cela confirme le besoin de créer des tâches d'évaluation équilibrées vis-à-vis des biais monomodaux.

Étude de la compréhension du concept de position

Dans une étude complémentaire, nous nous étudions plus en détail la compréhension de la position d'un objet, pour mieux expliquer la difficulté des modèles à appréhender ce concept. Nous cherchons à répondre aux questions suivantes : Les modèles comprennent-ils le concept de position au niveau monomodal ? Si oui, quelles sont les raisons de leur échec à comprendre la position au niveau multimodal ? Quelles mesures devraient être prises lors du pré-entraînement pour améliorer leur compréhension multimodale de la position ? Pour répondre à ces questions, nous élaborons des tâches en utilisant des données synthétiques basées sur l'ensemble de données CLEVR (JOHNSON et al. 2017).

Les résultats montrent que les modèles réussissent à extraire des informations visuelles relatives à la position d'un objet dans le cadre de l'image. Cependant, ils ont plus de difficulté à les combiner avec des informations textuelles liées à la position, ce qui entraîne de moins bonnes performances dans les évaluations multimodales. Concernant la compréhension de la distance, les modèles ne semblent pas associer les informations visuelles de position aux mots correspondants, probablement en raison du caractère plus contextuel du concept de 'distance'. En outre, nous constatons que le modèle vision-langage LXMERT (TAN et al. 2019) paraît avoir une meilleure compréhension multimodale des concepts spatiaux 'gauche' et 'droite'. Ceci pourrait être attribuable au fait que LXMERT est également pré-entraîné sur des tâches de question vision-langage, qui incluent des questions sur la position d'objet. Cependant, nos résultats sont basés sur des données synthétiques et pourraient être influencés par la nature limitée des catégories d'objets utilisées.

Étude de la compréhension des dépendances multimodales : le cas des dépendances nom-prédicat

Nous évaluons enfin comment les modèles vision-langage comprennent les dépendances multimodales, et plus précisément les dépendances nom-prédicat. Nous visons à répondre aux questions suivantes : Les modèles vision-langage comprennent-ils les dépendances multimodales fines ? Cette compréhension est-elle corrélée à leur compréhension de concepts multimodaux plus généraux ? Comment les différents choix de pré-entraînement impactent-ils leur compréhension de ce concept ? Pour répondre à ces questions, nous créons un jeu de données de correspondance image-texte, composé de triplets (image $i1$, légende correcte a, légende incorrecte b). Nous équilibrons chaque instance du jeu de données avec une instance 'opposée' ($i2$, b, a) afin de réduire liés potentiels biais monomodaux. En effet, bien qu'il a été montré que modèles vision-langage ont une compréhension multimodale de concepts tels que la couleur et les objets, il est intéressant d'évaluer leur compréhension dans le cadre d'instances plus complexes (i.e. compositionnelles).

Les résultats de nos expériences nous montrent que la compréhension des dépendances nom-prédicat est une capacité significativement plus complexe pour les modèles vision-langage que la compréhension des objets. Cependant, certains mo-

dèles (UNITER (Y.-C. CHEN et al. 2019), LXMERT (TAN et al. 2019)) montrent des résultats significativement supérieurs aux autres modèles d'état de l'art. Nous émettons des hypothèses pour expliquer ces différences de résultats entre modèles. Tout d'abord, la taille du jeu de données de pré-entraînement ne semble pas corrélée à la capacité du modèle à capturer les dépendances nom-prédicat. Cependant, les jeux de données composés de légendes descriptives semblent favoriser l'apprentissage de telles dépendances, contrairement à des jeux de données collectés automatiquement sur internet, moins précis. Les objectifs de pré-entraînement paraissent également jouer un rôle crucial dans la compréhension des dépendances multimodales. En effet, les modèles LXMERT et UNITER, utilisent des objectifs de pré-entraînement multimodaux plus précis que la tâche de correspondance image-texte, et montrent de meilleures performances que ViLT (W. KIM et al. 2021) et VinVL (P. ZHANG et al. 2021), qui sont entraînées à l'aide de cette dernière. Enfin, l'architecture du modèle a un impact sur sa performance. En effet, CLIP (RADFORD, J. W. KIM et al. 2021), qui n'a pas d'architecture permettant la fusion multimodale des représentations, montre des performances inférieures aux autres modèles.

Étude de l'impact des choix de conception sur les capacités des Transformers vision-langage

À partir des expériences réalisées dans la partie précédente, nous pouvons formuler des hypothèses concernant les facteurs limitants du pré-entraînement de modèles vision-langage. En particulier, il est intéressant d'étudier l'impact des tâches et jeux de données de pré-entraînement. Cependant, les modèles que nous évaluons dans le chapitre précédent utilisent des protocoles de pré-entraînement qui peuvent grandement varier. Ainsi, il est difficile de conclure quelles sont les tâches de pré-entraînement les plus utiles pour une capacité multimodale si les modèles que nous comparons n'ont pas les mêmes hyper-paramètres ou jeux de données de pré-entraînement.

Nous voulons mener une étude des facteurs de pré-entraînement. Afin d'avoir un meilleur contrôle, nous réalisons des études d'ablation en pré-entraînant un même modèle vision-langage et faisant varier différents facteurs de pré-entraînement individuellement. Cependant, comme le pré-entraînement des modèles vision-langage est gourmand en ressources, nous pré-entraînons des modèles plus réduits en termes de données et de durée de pré-entraînement. Nous menons les expériences sur les tâches et jeux de données conçus dans la précédente partie. En effet, nous espérons donc que les différences de performances soient ainsi plus facilement interprétables qu'avec des tâches plus complexes. Nous nous intéressons plus particulièrement à trois aspects du pré-entraînement : la durée, les tâches et les jeux de données.

Nos expériences semblent montrer que les modèles vision-langage nécessitent une longue durée de pré-entraînement, et qu'il y a peu de risque de sur-apprentissage sur les jeux de données considérés (e.g., COCO T.-Y. LIN et al. 2014). De plus, l'utilisation combinée des tâches ITM (i.e., correspondance image-texte) et MLM (masquage de

texte) améliore significativement la capacité d'un modèle à encoder des informations multimodales comparée à l'utilisation de chaque tâche indépendamment. L'utilisation de la tâche ITM sans la tâche MLM pourrait ainsi conduire à une mauvaise prise en compte des informations syntaxiques telles que l'ordre des mots.

Nous avons vu préalablement que le pré-entraînement visuel semble être un facteur limitant pour les modèles vision-langage. Les expériences réalisées dans le chapitre 4 semblent confirmer ce point, au moins pour les architectures similaires à ViLT. En effet, il est difficile de concevoir une tâche de pré-entraînement visuelle pertinente pour les Transformers vision-langage. L'utilisation de patchs de pixels, qui sont un atout pour capacités visuelles des modèles, ne semble pas propice à la représentation au niveau multimodal. Ainsi, des travaux supplémentaires sont nécessaires pour élaborer des représentations visuelles et tâches de pré-entraînement pertinentes pour favoriser à la fois l'inclusion de détails visuels et les liens entre représentation visuelle d'objets et mots correspondants.

Enfin, nous avons également montré que si la taille du jeu de données et la durée du pré-entraînement sont des facteurs qui impactent fortement les performances, la qualité de ces données est également importante. Il faut notamment prendre en compte la précision, variabilité et compositionnalité des données.

Problématiques éthiques

Au-delà des capacités des Transformers vision-langage, il est essentiel de discuter les problématiques éthiques soulevées par ces modèles :

- **Biais :** Les modèles de traitement du langage naturel et de vision par ordinateur sont sujets à des biais qui peuvent avoir des répercussions sociales importantes. Ces biais, comme ceux liés au genre, à la couleur de peau ou à l'âge, peuvent ainsi influencer les décisions prises par les modèles. Cette problématique est également présente dans le cadre de modèles multimodaux, et l'impact est souvent difficile à évaluer.
- **Collecte des données :** La plupart des modèles de vision-langage sont entraînés sur des données récupérées sur internet, ce qui soulève des questions éthiques liées au consentement des personnes dont les données sont utilisées. De plus, les annotateurs humains sont fréquemment soumis à des conditions de travail nuisibles, qui peuvent être aggravées lorsqu'il s'agit de vérifier si les données sont conformes à la loi (i.e. contenu potentiellement violent...).
- **Impact environnemental :** L'entraînement et l'utilisation de ces modèles nécessitent d'importantes ressources en énergie et en eau, ce qui a un impact environnemental significatif.
- **Impact social :** L'utilisation généralisée des modèles d'apprentissage profond dans divers domaines (médical, légal, industriel...) a un impact notable sur la société, en particulier en ce qui concerne les communications (deep fakes), l'éducation et l'emploi. Il est donc crucial de mener des recherches interdisciplinaires pour évaluer et comprendre les potentiels impacts de ces modèles.

Dans cette thèse, j'ai étudié les Transformers vision-langage et proposé une méthodologie d'évaluation qui repose sur l'évaluation de capacités vision-langage distinctes. J'ai ainsi créé des tâches d'évaluation afin de mieux identifier les capacités et faiblesses de ces modèles, comme leur compréhension multimodale de la position et les dépendances multimodales telles que les dépendances nom-prédicat. Nous avons également étudié l'impact des choix de conception des modèles sur l'apprentissage des capacités textuelles, visuelles et multimodales, et émis des hypothèses quant à l'évolution future de ces modèles. Un des buts de cette thèse est ainsi d'offrir une prise de recul sur le domaine de la multimodalité vision-langage et d'offrir des perspectives quant à l'évolution de l'apprentissage automatique multimodale face aux nouvelles architectures.

Remerciements

Je souhaiterais remercier les personnes qui m'ont accompagnée pendant ces trois années de thèse pour leur soutien, leurs nombreux conseils et leurs encouragements.

J'aimerais avant tout remercier mes directeurs de thèse, Benoit Favre et Stephane Ayache, pour leur encadrement, leur aide et leur bienveillance tout au long de cette thèse. Merci de m'avoir fait confiance et de m'avoir donné l'opportunité de découvrir le monde de la recherche académique.

Je voudrais aussi remercier chaleureusement les membres de mon jury de thèse. Tout d'abord Dr Ewa Kijak et Pr Eric Gaussier, d'avoir accepté d'être les rapporteurs de ma thèse, et également Pr Cécile Capponi, Dr Camille Guinaudeau, Dr Diane Larlus et Pr Benjamin Lecouteux d'avoir accepté de participer à ce jury. Merci à tous d'avoir bien voulu donner de votre temps pour évaluer mon travail et pour les discussions intéressantes lors de la soutenance.

Je remercie une nouvelle fois Cécile, ainsi que Laure Soulier, pour leur soutien et leurs conseils pertinents en tant que membres de mon comité de suivi de thèse.

Merci aussi à mes collègues Mitja Nikolaus, Abdellah Fourtassi, et mon ancien stagiaire Badreddine Farah, pour les collaboration que nous avons menées.

Je tiens également à remercier l'institut Archimède, qui m'a permis de mener à bien cette thèse.

Je suis aussi reconnaissante à Gregory Senay, de m'avoir fait découvrir le TAL, et à Meriem Bendris, qui m'a encouragée à me lancer dans une thèse.

Un grand merci aussi au Laboratoire d'Informatique et Systèmes de m'avoir accueillie pendant ces trois années, et en particulier aux équipes TALEP et QARMA, pour m'avoir fait découvrir tant de choses pendant ces réunions du jeudi et du vendredi.

Merci à tous mes amis des équipes TALEP et QARMA, à mes voisins de bureau et d'étage, pour les discussions, les (nombreuses) pauses café, les 'apérobots' pendant ces trois années ; et bien sûr à Elie et François, pour toutes ces parties de Dune et Hanabi qui ont beaucoup contribué à la rédaction de cette thèse !

Merci aussi vivement au personnel administratif du LIS et aux membres des commissions JC et parité, pour leur investissement auprès du labo et des doctorants.

Enfin, je tiens à remercier Maman, Papa, Annette, Nicolas et Héloïse, ma famille et mes amis à Marseille, Lyon, Paris et ailleurs. Ils m'ont soutenue dans les moments un peu difficiles. Merci d'avoir été présents autour de jeux et de bonnes bouffes, et de m'avoir aidée à rester motivée !

Contents

List of Figures

List of Tables

List of Acronyms

CL
Contrastive Learning. 59, 64

CNN
Convolutional Neural Network. 36, 37, 42, 46–48, 50, 53, 54, 82, 113, 170, 172

ITM
Image Text Matching. 61, 63, 64, 73, 106, 109, 155, 163–169, 171, 179, 183, 187

MLM
Masked Language Modeling. 44, 58–60, 63, 64, 73, 109, 150, 155, 163–169, 171, 179, 183

MPR
Masked Patch Regression. 169–172, 184

NLP
Natural Language Processing. 76, 79, 83, 84, 91

NLVR
Natural Language Visual Reasoning. 66, 110, 119, 121–123, 146

PLM
Prefix Language Modeling. 58, 59, 64, 73, 183

R-CNN
Region-based Convolutional Neural Network. 50, 53, 109, 114, 119, 123, 140, 145, 147, 172

VQA
Visual Question Answering. 29, 61, 63, 64, 67, 82, 87, 89, 98, 99, 110, 121–123, 131

WRA
Word Region Alignment. 61, 64, 110, 155, 163–169, 171, 172

Glossary

connotation
Capabilities related to the subjective analysis of a text-image instance, from symbolic interpretation to qualitative evaluation. 85

denotation
Text explicitly depicts or refers to image elements and does not require grounding, reasoning or evoke connotation. 85, 104, 154

foundation model
A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks. 28, 34, 76, 181

General Multimodal Understanding
The ability to extract and combine information from different modalities as a human would, across a wide range of domains, with good generalization ability on unseen data. 31

grounding
Capabilities requiring the use of information that is not directly accessible using the inputs (2D image and text), or the understanding of concepts that cannot be described using those modalities (e.g., time, space, knowledge, sound, mathematical documents). 85

reasoning
Capabilities requiring the application of abstract thinking or logic to the analysis of an image-text instance. 85

Introduction

Language has developed across human cultures around the world to convey meaningful information. In its written form, it has many uses. It can be used to communicate, search for information or document findings. As humans, we learn language through our interaction with the world. Our perception of the world through our various senses (e.g., sight, touch, smell, hearing) has greatly influenced what we communicate with our languages. In particular, we perceive visual information, which can be described using words representing multimodal concepts. Thus, our languages are inherently multimodal, in the way that they can represent multimodal concepts.

Towards general machine learning Since the beginnings of computer science, researchers have attempted to extract meaning from texts and images. In addition, the development of new technologies has led to an exponential growth in our ability to produce and process them. This has led to incredible improvements in the machine learning fields of Natural Language Processing and Computer Vision, that we touch upon in Chapter 1. Machine learning aims to take advantage of the massive amounts of data to learn patterns from this data. In particular, Natural Language Processing models aim to process and generate human language, while Computer Vision models learn to extract visual information or generate images.

One recent development in those fields is the emergence of foundation models. In Bommasani et al. 2021, the authors propose a definition of those models: "A foundation model is any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks." That is to say, the goal of these models is to serve as the basis of multiple tasks by learning general representations, for instance of texts or images, on a large amount of data. Those model are progressively more used in their respective fields, as they can be adapted to real-world tasks with minimal further training, and often outperform smaller task-specific models. This has shifted the research in Computer Vision and Natural Language Processing towards the learning of general-purpose representations.

The idea of generalization abilities in Artificial Intelligence has already been raised, through the concept of Artificial General Intelligence (AGI). Goertzel 2014 define a 'core AGI hypothesis' that we can use as basis for the concept: "the creation and study of synthetic intelligences with sufficiently broad (e.g. human-level) scope and strong generalization capability[...]" With this definition in mind, the development of foundation models trained on huge and diverse amounts of data can seem like a step towards AGI. As a result, public discussion has seen a growing focus on this concept. Indeed, those models can be applied to a diverse range of Natural Language Processing and Computer Vision tasks, with results that sometimes outperform human capabil-

ities on those tasks (T. Brown et al. 2020). However, evaluating a model on specific tasks is not sufficient to characterize the generalization ability of a model, especially in a domain as complex as understanding human languages. In a paper taking stock of the advances of foundation models in Natural Language Processing, Emily M. Bender et al. 2021 argue that language-only models cannot, by design, understand meaning at a human level. Indeed, there are not sufficiently grounded in real-world experiences to be able to grasp the meaning of expressions that represent multimodal concepts.

Multimodality To overcome this issue, one solution could be the use of multimodal models able to associate words with information from other modalities. Through this thesis, we will often refer to *multimodality*, a broad term which has different interpretations across domains. From the perspective of a human, modality can be equivalent to a human sense, such as vision or hearing. From the field of machine learning, modality refers to a type of input data of a machine learning algorithm. In the field of *vision-language multimodality*, this term specifically refers to the combined use of both the textual and visual modalities as input data. Consequently, vision-language models use multimodal instances composed of both texts and images as input. In this specific case, texts and images are complementary modalities. While some information between a text and an image may be redundant, the way information is encoded in those modalities is completely different. Images can be incredibly detailed, and can help us observe and analyze the world. They are often deemed to provide information at the object-level, but also offer significant information through finer details or interactions between various objects. Written language, on the other hand, often provides interpretation, and helps communication of information between humans. It is mostly studied at a word, sub-word or sentence level.

Until a few years ago, vision-language models were trained and optimized from scratch with a specific task in mind, in order to extract information from the two modalities relevant for this single task. For example, such models have been created for the tasks of image-text retrieval (Frome et al. 2013; Faghri et al. 2018) or Visual Question Answering (VQA) (Q. Wu et al. 2017). This paradigm changed a few years ago with the development of the Transformer architecture (Vaswani et al. 2017), a deep learning model based on self-attention, first used in Natural Language Processing tasks. By transferring this architecture to the field of vision-language multimodality, researchers started the trend of pre-trained vision-language models. Such models can be applied to many tasks involving both textual and visual information:

— Visual Dialog (Das, Kottur, et al. 2017): In this task, a model and a user dialog while referencing to visual content. For instance, a user may ask a question about an image. These questions mostly relate to image descriptions. They can be a tool to help visually impaired persons.

— Medical Assisted Diagnosis tasks: These regroup vision-language tasks aimed at medical data. Among them, medical visual question answering (Bazi et al. 2023), aims at answering questions about radiology or MRI images. Other tasks are image segmentation (Z. Li et al. 2023), or radiology report generation (Zhongzhen Huang et al. 2023).

— Vision-Language Navigation (Jing Gu et al. 2022): Vision-language navigation tasks evaluate the ability of a model to understand natural language instructions, while considering their environment through visual data. It is a branch of vision-language multimodality closely related to robotics.

This list is non-exhaustive, and progress in vision-language multimodality encourages the development of new applications.

Vision-language transformer models In this work, we study models aimed at learning general-purpose representations, which are adaptable to many tasks and domains. In particular, we study multimodal models through the lens of pre-trained vision-language transformers (Y.-C. Chen et al. 2019; Tan et al. 2019; J. Lu, Batra, et al. 2019b), and design appropriate methods. Such methods could also be applicable for future models intended for vision-language applications, irrespective of their architecture. Vision-language transformers take advantage of existing text-image datasets, usually composed of *(text, image)* pairs, to create multimodal representations from both visual and textual information. This method is based on self-supervised pre-training and removes the need for labeled data. The multimodal representations can then be adapted to specific tasks or domains through further training. The goal of pre-trained models is to develop general-purpose representations that can be used in a wide variety of tasks with minimal change. At the start of this thesis, only a few of such models existed. Since then, numerous other methods have been developed, taking advantage of advances in the fields of Natural Language Processing and Computer Vision. The various models that have emerged differ in multiple ways:

— Unlabeled data used for self-supervised pre-training;
— Text and image processing;
— Vision-language modality fusion;
— Losses to optimize during training;
— Adaptation methods to specific tasks.

They have enabled better performances and significant improvement over previous state-of-the-art vision-language models. However, our understanding of those successive models, in particular their strengths and weaknesses, remains limited. Indeed, their fast-paced development has left little time for a deeper study of those models. In terms of model pre-training, we have a poor understanding of how our model design choices affect specific capabilities of those models. We also lack insight in what kind of information those representations should contain in order to be applied to real-world tasks.

Vision-language model evaluation Following this observation, we take a step back on the last few years of progress in the field of vision-language machine learning. Vision-language models are usually evaluated on selected vision-language tasks, which are designed to be difficult to achieve. As a result, significant improvement over previous state-of-the-art models is often correlated with a better understanding of multimodal concepts. However, those models are optimized for those selected

multimodal tasks, which can skew the direction of research chosen for those models. This can lead to suboptimal model development, depending on the purpose of the model. In particular, one of the main goals of machine learning model evaluation is to anticipate the performance of a model in real-world tasks. In the case of vision-language multimodality, the possible domains of applications are very diverse, from E-commerce recommendation W. Shin et al. 2022 to biomedical data interpretation Boecking et al. 2022. Evaluating the accuracy of a model on a few complex tasks offers limited insight into its performances for those applications. Indeed, design choices made to pre-train vision-language models may negatively affect some of their capabilities, in favor of a better performance on others.

This work aims to reach a better explainability of the performances pre-trained vision-language transformer models. Specifically, we aim to study whether those models learn multimodal representations that can serve as good basis for a wide range of vision-language tasks. Accordingly, we introduce the concept of General Multimodal Understanding. Using the previous AGI definition, we propose a definition of *General Multimodal Understanding* as:

Definition 1 (General Multimodal Understanding) *The ability to extract and combine information from different modalities as a human would, across a wide range of domains, with good generalization ability on unseen data.*

As this definition is very broad, it becomes clear that a few fine-tuning tasks are insufficient to characterize the performance of a vision-language model. As such, through this work, we propose alternative methods to apprehend the general multimodal understanding of those models. In particular, we introduce new methods and tasks aimed at vision-language model evaluation.

Important questions Through this thesis, we study pre-trained vision-language transformers, with the aim of reaching a better understanding of their performances and limitations. Indeed, vision-language multimodality is a growing field of study, and the use of pre-trained models has considerably shifted research in the domain. Thus, we think it is important to consider how vision-language transformers have impacted the domain, what issues are still unanswered despite those advances and potential ways to further improve those models. Beyond this, we aim to further develop methods to study multimodal models, irrespective of the precise architecture. Indeed, the current focus on vision-language models seems to be a precursor for the development of other multimodal models with other architectures and modalities.

The main contributions of this thesis are:

— An overview of current vision-language transformer research to provide hindsight on the current advances of the field;
— A first attempt at theorizing the concept of general multimodal understanding through a taxonomy specific vision-language capabilities;
— Tasks and datasets developed for the study of specific capabilities of state-of-the-art vision-language models;

— A study of the impact of pre-training choices on specific capabilities of a vision-language model;

Through these contributions, we aim to resolve the following questions:

1. What are vision-language transformers? What sets them apart from previous multimodal machine learning models? How have they evolved since their development?
2. What methodology should be used to evaluate such models, or in a broader sense, general-purpose multimodal models? What are the difficulties?
3. How should vision-language tasks and datasets be created? What are the risks encountered when building such tasks?
4. What are the strengths and weaknesses of vision-language transformer models? Are some aspects of multimodal understanding easier to understand for them than others?
5. How does the pre-training of a model affect its performances? What is the role of architecture, pre-training protocol, dataset and fine-tuning?
6. What are the potential impacts of vision-language transformer models?

Thesis structure In Chapter 1, we introduce key concepts of vision and language machine learning, relating to transformer models. We provide an overview of vision-language transformers, and in particular the methods used for their pre-training. We specifically focus on the different architectures and pre-training protocols. Finally, we also analyze the current trends to better understand how the field is expected to evolve.

In Chapter 2, we propose a taxonomy of vision-language capabilities geared towards the evaluation of vision-language models. Indeed, with recent development, vision-language models can be described as multimodal foundation models. As their monomodal counterparts, one key difficulty with vision-language foundation models is how to evaluate them. In this chapter, we detail a method to evaluate the *General Multimodal Understanding* of a model. In particular, by studying precise capabilities of such models, we hope to gain a better understanding of their potential blind spots, and their generalization abilities. We encourage a different structuration of the evaluation of vision-language models, geared towards specific capabilities rather than complex tasks. We also identify how existing vision-language datasets relate to the proposed taxonomy, to identify potential blind spots in vision-language research. Associated publication:

— (Forthcoming) Salin, E., Ayache, S. & Favre, B. (2023) Towards an Exhaustive Evaluation of Vision-Language Foundation Models, In *ICCV Workshops 2023*.

In Chapter 3, we study state-of-the-art transformer models on specific capabilities. To that end, we draw inspiration from methods developed in the field of Natural Language Processing. We provide different tasks and datasets designed for the study of monomodal and multimodal capabilities of vision-language models. The first study focuses on probing vision-language models on a range of linguistic, visual and

multimodal capabilities. In the second study, we specifically evaluate the ability of models to understand position. Finally, in the third study, we evaluate how vision-language models understand multimodal dependencies through the lens of noun-predicate dependencies. Those results enable us to draw hypotheses on the choices of architecture, datasets, and pre-training of state-of-the-art models. Associated publications:

— Salin, E., Farah, B., Ayache, S., & Favre, B. (2022, June). Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 10, pp. 11248-11257)

— Nikolaus, M., Salin, E., Ayache, S., Fourtassi, A., & Favre, B. (2022, December). Do Vision-and-Language Transformers Learn Grounded Predicate-Noun Dependencies?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 1538-1555).

— Salin, E. (2022, November). 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)* (pp. 181-187). CNRS.

In Chapter 4, we evaluate how different pre-training protocols impact the performances of a vision-language model. Indeed, while comparing state-of-the-art models may help us better understand their inner workings, their pre-training are too different to precisely pinpoint how each choice impacts their performances. As a result, we use the hypotheses drawn in the previous chapter to devise experiments and test several pre-training protocols. We base our experiments on a single state-of-the-art vision-language models, for better comparison purposes. Associated publication:

— Salin, E. (2023). État des lieux des Transformers Vision-Langage: Un éclairage sur les données de pré-entraînement. In *18e Conférence en Recherche d'Information et Applications*
16e Rencontres Jeunes Chercheurs en RI
30e Conférence sur le Traitement Automatique des Langues Naturelles
25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (pp. 14-29). ATALA.

Finally, through discussion on vision-language transformer models, we provide guidelines for the design of evaluation tasks and datasets aimed at the study of vision-language models. We also draw our conclusions on the generalization abilities of transformer models, and attempt to provide insight on the future trends of the domain.

1. State of the Art: Vision-Language Transformers

Table of Contents

1.1. Introduction

One major goal of machine learning research has always been to use data to emulate human capabilities. For instance, this is the case of Natural Language Processing models, which aim to understand human language. The past few years have brought considerable advances in the field, specifically due to the Transformer architecture (Vaswani et al. 2017). This architecture, that we will study in detail in Section 1.1.2, is based on the attention mechanism to learn how different words of a sentence relate to each other. It has furthered the development of contextual language models. Contextual models learn representations of a word by considering the context of this word. In particular, this has led to the emergence of linguistic foundation models. Those foundation models are pre-trained on a large amount of textual data, available at scale. This is done using pretext tasks that serve as proxy to learn appropriate

representations of the data. Foundation models can then be adapted to many tasks using these representations as a basis. The concept of pre-training is further detailed in section 1.1.3.

However, while language models have come close to human performances on a significant number of tasks, researchers are reluctant to liken the human understanding of language, to the one of a machine learning model. Indeed, a large part of human language is used to help their interaction with the world, and uses their other senses. For instance, an important number of words represent a concept that can be apprehended with diverse senses, such as vision or hearing. Yet, this information is rarely represented by language-only models. This has led to a growing interest in multimodal machine learning (i.e., machine learning concerned with the fusion of multiple modalities).

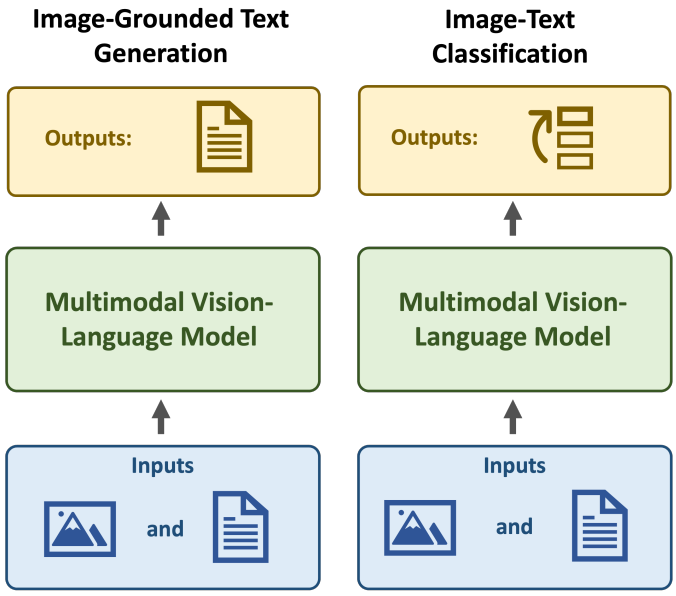


Figure 1.1. – Illustration of image grounded text generation and image-text classification

Vision-language multimodality In this thesis, we focus on vision-language multimodality, which associates Natural Language Processing with Computer Vision. The latter aims to extract information from images. By combining information from texts and images, vision-language machine learning aims to ground our language in visual modality. To that end, they are pre-trained using the large amount of text-image data available on the web. More precisely, the goal of vision-language models is to associate visual and textual information to learn multimodal text-image representations. By encoding information from both text and images in a same multimodal space, these representations can be used in a variety of tasks. The most common vision-language tasks are a variation of image-grounded text generation and image-text classification, illustrated in Figure 1.1. For instance, the goal of image-text retrieval is to associate a

text or image with its most appropriate counterpart, by classifying a text-image pair as *matching* or *non-matching*. Visual question answering aims at answering a question using visual clues given in an image. These tasks can be applied to many domains, from natural images to scientific data, leading to a number of real-world applications, such as:

— Assistance for visually impaired people (Zongming Yang et al. 2022);
— Crisis analysis (M. Li et al. 2022);
— Satellite images understanding (Wen et al. 2023).

In this chapter, we provide context on research in vision-language machine learning. We specifically focus on pre-trained vision-language models based on the Transformer architecture. Indeed, in this thesis, our goal is to provide further insight on those models. In recent years, most state-of-the-art vision-language models have been based on this architecture. There has been an increasing interest in using the transformer architecture in Computer Vision and Natural Language Processing, that has translated to vision-language multimodality. Figure 1.2 shows the evolution of keywords related to Recurrent Neural Networks (RNNs) and in particular LSTMs (Hochreiter et al. 1997), Convolutional Neural Network (CNN)s (LeCun et al. 2015) and the Transformer architecture in titles and abstracts of vision-language papers accepted to the *ACM International Conference on Multimedia*, the *Conference on Empirical Methods in Natural Language Processing* and the *Conference on Computer Vision and Pattern Recognition*. This figure shows that vision-language multimodality as a whole has generated increasing interest, and a growing part of those papers use or study the transformer architecture.

Most state-of-the-art vision-language models are pre-trained using the Transformer architecture as a basis. However, they vary in the specifics of modality preprocessing, fusion and pre-training protocols. In this first section, we introduce several notions and models that can be helpful for the understanding of vision-language transformers. In section 1.2, we give a survey of the architecture transformer-based vision-language models. Then, in section 1.3, we focus on the data and tasks used to pre-train those models. Finally, we explain in section 1.4 how these models have been compared and evaluated.

1.1.1. Vision-Language Models: Background

Vision-language machine learning rests on associating textual representations to visual representations. Texts and images differ at a local and structural level, in the ways they encode information. As a result, both those modalities have been historically separated into different subfields of machine learning, though methods from each subfield have been used to inspire the other. The processing of text has been tackled through Natural Language Processing algorithms. Bag of words (Sebastiani 2002) help take into account the most relevant words of a text, thereby extracting features that could represent textual information. Similar methods have been developed in computer vision, to extract a visual bag of words (T. Li et al. 2010). For instance, SIFT

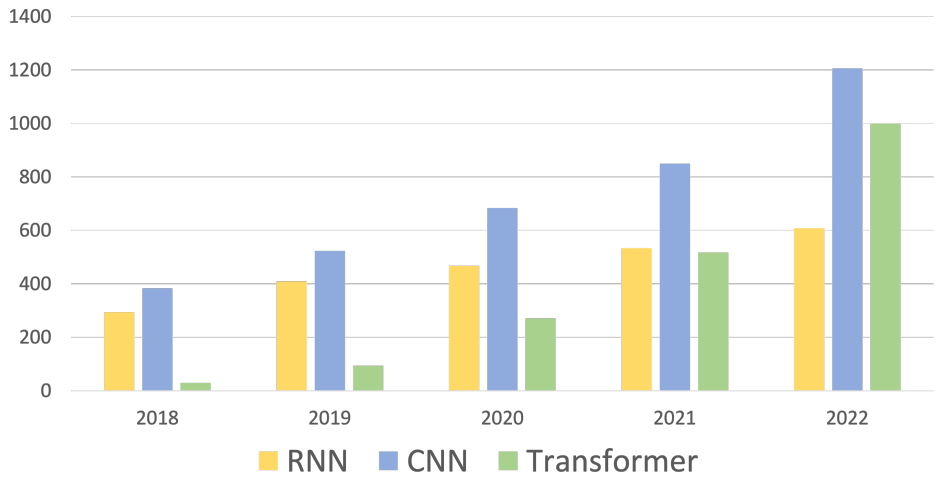


Figure 1.2. – Evolution of vision-language papers accepted to ACM Multimedia, CVPR and EMNLP using RNN, CNN or Transformer keywords in titles and abstracts

(Lowe 2004) extracts features with semantic relevance that enable matching with semantically related images. More recently, the growth of deep learning has incited the use of convolutional neural networks to extract features from images (Jiuxiang Gu et al. 2018). In Natural Language Processing, contextual models have also been developed to consider contextual information, with recurrent neural networks (Y. Yu et al. 2019).

In parallel with the extraction of textual and visual features, vision-language machine learning hopes to combine the information from both modalities. The goal is to create textual and visual representations that are semantically related. The main idea of text-image representation learning is first to use recurrent neural networks and convolutional networks to extract monomodal features. Then, multimodal models learn visual semantic embeddings by projecting visual and textual embeddings in the same multimodal space. The use of deep learning models to extract monomodal features makes it possible to adapt the extraction of monomodal features to the structure of the multimodal data. Vision-language multimodal machine learning caters to several applications, among which historical ones are:

— Image description generation: Karpathy et al. 2015 use an object detector to extract image features and an RNN to compute word embeddings. To evaluate multimodal similarity, they compute pairwise similarity between image and text tokens.

— Cross-modal retrieval: In VSE++ (Faghri et al. 2018), the model uses a convolutional model and an RNN to extract monomodal features, and project them in the same multimodal space using a triplet ranking loss.

— Visual Question Answering: Stacked attention networks (Zichao Yang et al. 2016) also employ both CNN and RNN to extract monomodal features. Additionally, the attention layers help them to focus on relevant objects of the image.

There exist several types of models producing visual semantic embeddings. In Baltrušaitis et al. 2018, authors differentiate between:

— 'Joint' embeddings, which combine mono-modal inputs into the same representation space.

— 'Coordinated' embeddings, which encode inputs of different modalities separately and ensure that their representations form a 'coordinated' space using similarity metrics.

In order to create appropriate representations for a specific vision-language task, multimodal models based on recurrent and convolutional neural networks require specific architecture and training protocol. Thus, they are difficult to adapt to multi-task learning, or to use with transfer learning. The development of transformer models (Vaswani et al. 2017) in natural language processing, and their ability to learn textual representations adaptable to many tasks (Devlin et al. 2019) has inspired the development of vision-language transformer models. While previous visual semantic models mostly form coordinated embeddings, most Transformer-based models learn joint text-image representations, and can be adapted to a wide range of tasks by adding task-specific layers to the model.

1.1.2. The Transformer Architecture

The Transformer (Vaswani et al. 2017) is a deep learning architecture which has replaced RNNs such as LSTMs (Long Short-Term Memory — Hochreiter et al. 1997) as the state-of-the-art architecture in many modeling tasks. Transformers can be used with data ranging from sequences to images and graphs.

The novel idea of this architecture is that it is only based on self-attention mechanisms, while previous models usually combined attention mechanisms with other architectures. Thus, it removes the need for any recurrent architecture.

The attention mechanism Previously used in machine translation models, the concept of attention was introduced as a mechanism that learns to make use of the most relevant tokens in a sequence. For instance, it can be used to benefit from the most appropriate context words in a sentence to translate a specific word (Bahdanau et al. 2014).

Attention is a function applied to sequences of embeddings of size n using query Q, value V and key K input vectors. It computes a weighted sum of V of dimension $n * d_v$, using Q and K of dimension $n * d_k$ to compute the attention weights. In the Transformer, authors use Scaled Dot Product Attention, which adds a $\frac{1}{\sqrt{(d_k)}}$ factor to the standard dot product attention (Equation 1.1).

$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{(d_k)}})V \tag{1.1}$$

Most attention mechanisms use weight matrices with learnable parameters W_Q, W_K, W_V to compute the Q, K, V vectors. Equation 1.2 shows the mechanism for Q,

with an input sequence x of size n.

$$q^i = W_Q x^i \text{ for } i \in [1, n] \tag{1.2}$$

There are multiple forms of attention:

— Self-attention is an attention mechanism that captures how embeddings in a sequence relate to other embeddings of the same sequence. To that end, it outputs an attention-weighted version of the input sequence. In this case, Q, V and K are all computed from the same input sequence x.

— Cross-attention is an attention mechanism that relates two sequences to each other. They can be of different sizes. In this case, Q of size n corresponds to one sequence x_1 and K and V correspond to another sequence x_2. The computed output is of size n.

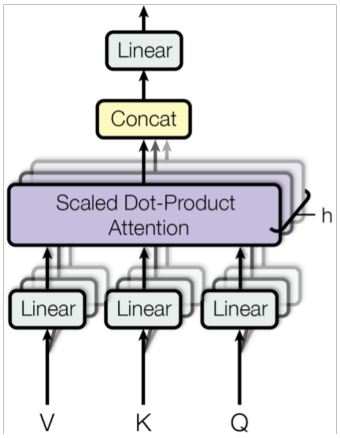


Figure 1.3. – Multi-Head Attention (Vaswani et al. 2017): The attention mechanism is applied to a number h of projections of the V, K and Q matrices. The attention weights are then obtained by concatenating the output of each head.

Multi-head attention Multi-head attention (Figure 1.3) allows the models to learn multiple ways to compute attention for a given input. The model uses a number h of attention heads. Each attention head computes attention on a given projection of K, Q, and V, using different W_Q^j, W_K^j and W_V^j matrices, for $j \in [1, h]$. The outputs computed by those attention heads are then concatenated and projected in the original dimension of V. This allows the transformer model to learn a less averaged representation of the input. The richer representations created by the multiple attention heads help the transformer models reach better results. Indeed, they capture different forms of interaction between the input embeddings with each head.

Position encoding The attention mechanism is computed on one or several sequences of input embeddings. However, the order of the embeddings in those sequences has no impact on the values of the Q, K, V matrices or the attention weights.

As a consequence, prior to the computation of the attention mechanism, the relevant positional information must be encoded in the input embeddings. To that end, a position embedding is usually added to the input embedding. This positional embedding can depend on the structure of the original input (e.g., text, image, graph). It can either be a learnable parameter or a fixed parameter.

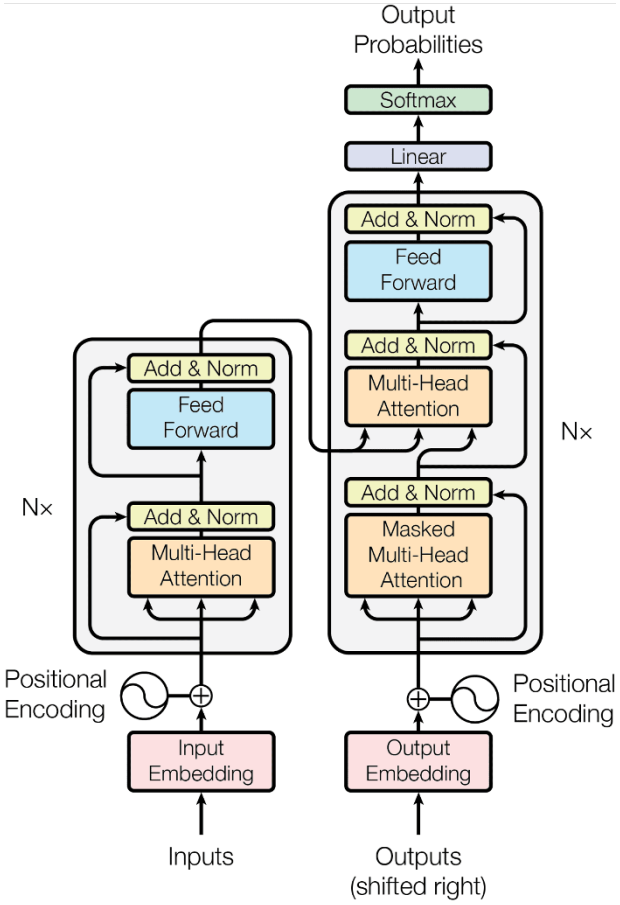


Figure 1.4. – The Transformer architecture (Vaswani et al. 2017) The original transformer architecture is composed of an encoder and a decoder. The encoder is made of N stacked multi-head self-attention layers, that encode the information from the input embeddings. The output of encoder layers is fed to the decoder, which is made of stacks of self-attention and cross-attention layers. It predicts the next token in a sequence from the input and the output, which is shifted right to avoid cheating.

The original Transformer architecture The original Transformer architecture, shown in Figure 1.4, is an encoder/decoder model with three kinds of multi-head attention mechanisms. There is first an attention mechanism specific to the encoder, then one specific to the decoder and one relating the encoder to the decoder. Contrary to recurrent models, where each hidden state h_t of a sequence is computed using the

value of the previous state h_{t-1}, the transformer relies on the attention mechanism. Thus, it draws dependencies between tokens of an input sequence in a non-sequential way. To take advantage of the sequential information necessary for most sequence to sequence tasks, the authors introduce a positional encoding, which is added to the embeddings. The authors choose to use sine and cosine functions to encode positional information, to help the model take into account relative position.

This architecture was first developed for sequence-to-sequence Natural Language Processing tasks, such as machine translation (Raganato et al. 2018). For this application, input and output tokens are text tokens.

Strengths and weaknesses of the Transformer architecture The transformer architecture has established itself both in the fields of Natural Language Processing and Machine Learning.

In Natural Language Processing, the attention mechanism allows for more parallel computing, and thus usually more efficient models than recurrent neural networks. Additionally, while recurrent models are subject to vanishing and exploding gradient problems when dealing with long input sequences, Transformers have no such issue to catch long-term dependencies.

However, one main limit of the Transformer model is the need for large datasets and resources due to its large number of parameters. While some researchers have developed less computational intensive models, one main trend of research is scaling up the data and model architecture, such as GPT-3 (T. Brown et al. 2020), which are very resource- and time-consuming. Indeed, as has been shown in the past few years, scaling Transformer models in data, training time and number of parameters enables better performances with minimal other improvements (T. Brown et al. 2020; Koubaa 2023). This scalability comes at the cost of considerable computing resources, especially when scaling the size of input sequences. Indeed, the Transformer architecture has a quadratic complexity with respect to the size of the input. However, through its collection of language models, Llama (Touvron, Lavril, et al. 2023) has shown that smaller models pre-trained can be competitive with larger models.
Since the development of the Transformer model, more researchers use self-attention-based models for their applications beyond the original transformer architecture, with variations in architecture, data preprocessing, and losses used during pre-training Devlin et al. 2019; Y. Liu et al. 2019. We call models mainly composed of stacked multi-head self-attention blocks *Transformer* or *Transformer-based* models. They have been expanded to several fields of Machine Learning research, such as Natural Language Processing (section 1.1.4) and Computer Vision (section 1.1.5), leading to new state-of-the-art performance.

1.1.3. Pre-training and Fine-tuning

The Transformer architecture requires large training datasets to converge. However, as supervised training datasets are difficult and expensive to create, self-supervised learning has been implemented to pre-train those architectures. Indeed, self-supervised

pre-training can be used to learn representations from large unlabeled datasets. These representations can then be fine-tuned on task- or domain-specific labeled data. Thus, this method enables models to reach better performances than when only training on labeled task-specific data.

Self-supervised pre-training The idea of self-supervision is to remove the need for labeled data by using part of the input in place of a label. It is less intensive than supervised pre-training in terms of manual annotations. Indeed, such a method can take advantage of the massive amounts of unlabeled text and image data available on the web. As a result, it is used for representation learning in different branches of machine learning research, such as Natural Language Processing and Computer Vision. Self-supervised tasks can be separated into three main categories (X. Liu et al. 2021).

— Generative tasks hide or scramble part of the input and reconstruct the encoded input using a decoder. Many language models such as are pre-trained on such tasks. For instance, GPT (Radford, Narasimhan, et al. 2018) pre-training consists in predicting the next word in a sequence, while BERT (Devlin et al. 2019) consists in recovering a portion of masked words in a sequence.
— Contrastive tasks measure similarity between two parts of an encoded input. Contrastive learning can be used to learn visual representations using data augmentation, as in SimCLR (T. Chen et al. 2020). It differentiates between positive pairs (i.e., an image on which two different augmentation methods have been applied) and negative pairs (i.e., two different images).
— Adversarial tasks discriminate between real samples and fake samples created with a generator. This method is used to train Generative Adversarial Networks (Goodfellow et al. 2020).

Fine-tuning Fine-tuning is a method of transfer learning, where a model is pre-trained on a large dataset and then trained on a specific task. The pre-training task is not necessarily useful for the downstream application, but is used as a pretext task to learn general representations. The weights of the model are updated to fit this task. For more efficient fine-tuning, some weights of the pre-trained models can be frozen, i.e., not updated during fine-tuning. As lower layers usually learn low-level features, especially in computer vision architectures such as CNNs, those can be frozen, and protected, to avoid 'catastrophic forgetting' (Kirkpatrick et al. 2017). During fine-tuning, layers can also be added to those models to fit a specific task. These layers are usually added on top of the model, but new models have introduced adapter layers. These layers can be added between attention layers (Houlsby et al. 2019). Their weights are updated in instead of the attention layers, which are frozen. They use fewer parameters to enable efficient fine-tuning.

The process of fine-tuning builds upon what is already learned, and requires smaller datasets and less computing resources if the pre-trained models have already been trained to extract all relevant information. This makes it important for the pre-trained

models to learn general-purpose representations, in order to be useful in a wider range of domains and downstream tasks. This method is commonly used in Natural Language Processing, where language models are pre-trained on large datasets and then fine-tuned on specific tasks. For instance, sentiment analysis (Socher et al. 2013) consists in attributing a positive or negative sentiment to a text such as a review in a binary classification task. Another example is natural language inference (Williams et al. 2017), which uses two sentences as input and predicts whether one can be inferred from the other, or if there is a contradiction. In computer vision, fine-tuning is often used to transfer the representations from a general-purpose dataset with diverse images to a dataset in a specific domain. For instance, computer vision models can be fine-tuned on fine-grained flower (Nilsback et al. 2008) or painting (Saleh et al. 2015) classification.

1.1.4. Transformers in Natural Language Processing

In Natural Language Processing, variations of the Transformer architecture have been used to build language models. Language models are models that assign a probability distribution to sequences of words. More specifically, those models compute the probability of a text sequence being drawn from the training dataset. Language models help extract syntactic and semantic information for Natural Language Processing tasks. Indeed, they can be used to learn representations of words or text tokens. Similarly, the goal of vision-language pre-training is to extract structural and semantic information from text and image data.

Language models Training a Natural Language Processing model for a specific task, such as natural language inference (Williams et al. 2017), can require a large amount of task-specific supervised data and specific feature engineering. To overcome this, language models are pre-trained to learn textual representations that extract syntactic and semantic information from words, as a base model for later applications. With the development of deep learning, and in particular RNNs, self-supervised language models have been trained to learn probability distributions on a huge amount of unsupervised text (Peters et al. 2018; Devlin et al. 2019; Radford, J. Wu, et al. 2019). There are two main self-supervised tasks to train language models:

— Auto-regressive models use next-word prediction (Radford, Narasimhan, et al. 2018). It consists in asking the model to predict the next word in a sequence using the representation of the current word. This task is especially helpful to learn representations for generative language models. Bidirectional variations of the standard autoregressive task have also been implemented (Zhilin Yang, Z. Dai, et al. 2019).

— Auto-encoding models corrupt the input by masking some tokens or replacing them with another word (Devlin et al. 2019). The masking strategy is usually random, and the goal is to reconstruct the original input.

A language model can then be fine-tuned on a specific downstream task, using a smaller amount of supervised data. These tasks usually belong to two categories:

— Text generation: It groups tasks such as summarization (Allahyari et al. 2017) and translation (J. Zhang et al. 2015).
— Text classification: It groups tasks such as sentiment analysis (Socher et al. 2013) or natural language inference (Williams et al. 2017).

Transformer-based language models GPT (Radford, Narasimhan, et al. 2018) has first combined the pre-training of large self-supervised language models with the Transformer architecture. Since then, most state-of-the-art models on multitask benchmarks such as GLUE (A. Wang, Singh, et al. 2019 and A. Wang, Pruksachatkun, et al. 2019) have been based on variations of the transformer architecture. GPT is composed of a Transformer decoder pre-trained in an autoregressive manner by using masked self-attention. It is a unidirectional model, meaning that the model uses only previous context to predict a word. As a result, the representations learned by GPT do not take advantage of the context after a word.

As Bidirectional RNNs such as ELMo (Peters et al. 2018) outperform uni-directional RNNs, BERT (Devlin et al. 2019) authors decided to use an auto-encoding model to learn a deeper bidirectional model. Contrary to GPT, BERT learns representations of words that use both the previous and future context. To accomplish this, BERT is

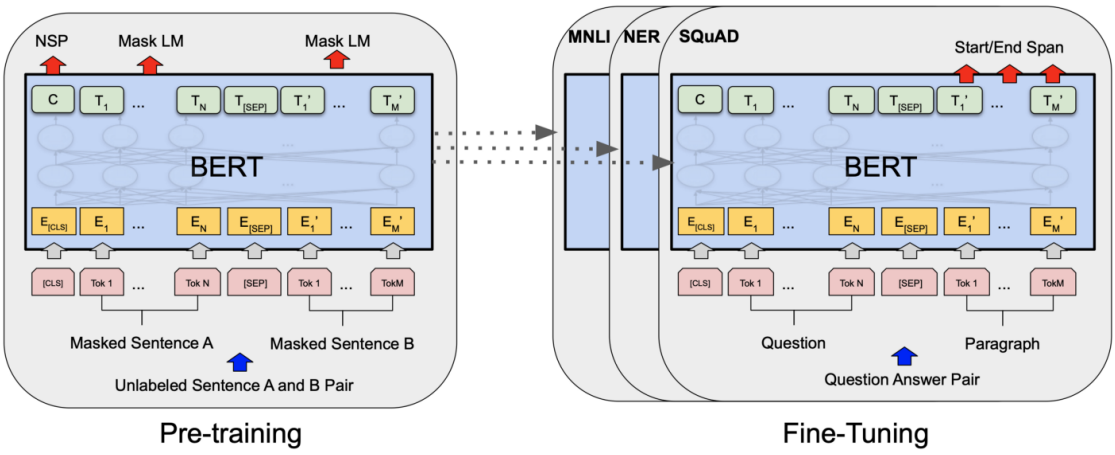


Figure 1.5. – BERT pre-training and fine-tuning (Devlin et al. 2019)

built as a Transformer encoder pre-trained with two self-supervised tasks, as shown in Figure 1.5:

— Masked Language Modeling (Masked Language Modeling (MLM)): The idea of this task, taken from the cloze task (Taylor 1953), is to randomly mask 15% of the input tokens. Most of those tokens (80%) are replaced with a [MASK] token, while some (10%) are unchanged and the rest (10%) are replaced randomly. The Transformer model will encode each token, and the task consists in predicting a masked token using the last layer representation of this token.
— Next sentence prediction: The goal of this task is to use this token to predict whether two token sequences are consecutive or not. As Natural Language

Processing tasks do not only use token-based representations, but also sequence-based representations, authors build a task to learn to model the relationship between two text sequences. When choosing the two input sequences, the second input sequence is half of the time contiguous to the previous sequence, and the rest of the time a random sequence of the corpus with no relationship to it. A classification token [CLS] is introduced at the beginning of the two-sequence input, and a separation token [SEP] marks the separation between the sentences, as shown on the figure. The [CLS] token is used as a representation of a text sequence, in particular for classification tasks. To that end, it is also used in later variations of the BERT model that do not use the next-sentence prediction task.

BERT introduces a specific token encoder, which encodes textual data into sub-word tokens. This limits the amount of out of vocabulary tokens in the data. This pre-training enables BERT to outperform GPT and GPT-2 (Radford, J. Wu, et al. 2019) on several Natural Language Processing tasks. The size of the pre-training dataset also evolved between the three models. It went from the use of the BookCorpus dataset (Yukun Zhu et al. 2015) for GPT, to the addition of English-language Wikipedia for BERT, and the use of Common Crawl data for GPT-2. Further Transformer-based models have been developed by changing their size, architecture, datasets, or pre-training tasks, such as XLNet (Zhilin Yang, Z. Dai, et al. 2019), RoBERTa (Y. Liu et al. 2019), GPT-3 (T. Brown et al. 2020), PaLM (Chowdhery et al. 2022) and Llama (Touvron, Lavril, et al. 2023). The Transformer architecture has reached and remained state-of-the-art in many Natural Language Processing applications using those language models as a basis. One of the most recent examples is ChatGPT [1].

Study of language models and text representations Textual representations extracted from Transformer-based models contain both syntactic and semantic information. Syntax relates to the structure of a text and the grammatical information it conveys, while semantics relates to the meaning behind the words used in a text. Before Transformer models, Word2vec (Mikolov et al. 2013) has shown that word embeddings encode semantic relationships between words. The contextual nature of Transformers (and recurrent neural networks before that) helps them learn word embeddings that encode structural syntactic information. This has been shown through their performances on many tasks requiring an understanding of syntax or semantics (A. Wang, Singh, et al. 2019). For instance, CoLa (Warstadt et al. 2019), which evaluates whether a sentence is linguistically acceptable, or natural language inference (Williams et al. 2017), which evaluates whether a sentence can be inferred from another.

However, our understanding of the specific capabilities of language models is still limited. A new field of study, called Bertology (Rogers et al. 2021), consists in studying BERT and other similar models to have a more precise understanding of their capabilities. In this paper, they explain that the understanding of a language model can be split into three categories: syntactic understanding, semantic understanding

1. https://chat.openai.com/

and the understanding of world knowledge. The latter is knowledge relating to the understanding of the world. For instance, it can consist in common sense knowledge that humans learn during their life, but also more specific knowledge relating to events or people. They give an overview of results from several works studying the understanding of transformer-based language models. For instance, among the studies evaluating syntactic understanding, agreement tasks (Schijndel et al. 2019) have shown that transformer-based language models learn enough syntactic information to agree the verb with the subject, even in complex sentences. Regarding semantic understanding, Transformer models have some understanding of semantic roles (Ettinger 2020), i.e., they can associate a corresponding verb depending on the type of subject (e.g., person or animal). However, they struggle with other semantic concepts, such as negation (Ettinger 2020) and numbers (Wallace et al. 2019). BERT can extract world knowledge from its training corpus, but has difficulties using this knowledge to reason, for instance by associating correct physical attributes to objects (Forbes et al. 2019). Since then, the development of transformer-based models pre-trained on more data, with more parameters, has led to a better understanding of syntax, semantics and world knowledge (Min et al. 2021). However, these limitations remain for models of a comparable size and pre-training dataset to BERT.

In addition, studying language models and their representations is complex and can have several pitfalls. Indeed, results on an evaluation task are not always sufficient to conclude on the capabilities of the model for this task. Indeed, BERT has shown poor generalization abilities. This is especially the case on cloze-style tasks, where the model is asked to generate a masked word to complete a sentence. The use of plural instead of singular in a sentence can completely change its performance on such a task (Ravichander et al. 2020). In order to also evaluate the robustness of language models, some alternative evaluation methods have been proposed. For instance, Ribeiro et al. 2020 propose a checklist to test the capabilities and robustness of Natural Language Processing models. Chapter 3 will study those methods in more detail.

Beyond the study of the capabilities of language models, the question of their understanding of the language remains. Indeed, Emily M Bender et al. 2020 question the meaning of a language model's 'understanding', as human children perceive and interact with the world to learn their language. As a result, a machine learning model would need some kind of similar perceptual grounding, using data from other modalities such as images, to understand words based on multimodal concepts.

1.1.5. Transformers in Computer Vision

Before Transformers, self-attention was already used in Computer Vision combined with CNNs. Some models have even used a CNN architecture as a backbone of a Transformer architecture. For example, DETR (Carion et al. 2020), for object detection, uses a CNN to create image features which are then used as input of a transformer. However, the use of Transformers-only models was developed later than in Natural Language Processing. To this day, there is no clear superiority of the Transformer architecture or CNNs in Computer Vision tasks such as object classification (Arkin

et al. 2023).

Visual pre-training tasks Similarly to language models, visual transformers are pre-trained on large amounts of data. This pre-training can either be supervised, through the task of image classification, or self-supervised. Several self-supervised tasks have indeed been adapted or developed for visual transformers. Some methods based on pre-training tasks developed for CNN models or Language Transformers have reached good performances on the visual Transformer architecture.

— Contrastive pre-training: MoCo (K. He, Fan, et al. 2020) learns a visual encoder through contrastive learning by building a large dynamic dictionary. The model must discriminate between views of the same image. DINO (Caron et al. 2021) develops a self-distillation technique. The teacher is the same model as the student, with different weights. Different views of the same image are given in input, using different data augmentation techniques on the input image. The similarity between the output of the student and that of the teacher is computed.

— Generative pre-training: In M. Chen et al. 2020, the input is first resized to a low resolution and reshaped to a sequence of pixels from a custom color palette. The model is then trained to predict those 'pixels' in an autoregressive way, similar to GPT pre-training. Several works have also adapted BERT pre-training to Visual Transformers, training models to reconstruct image patches (K. He, X. Chen, et al. 2022) or predict the corresponding discrete token (Bao, L. Dong, et al. 2021). In Kakogeorgiou et al. 2022, the authors introduce a self-distillation technique to optimize the masking of image patches using the teacher's attention weights.

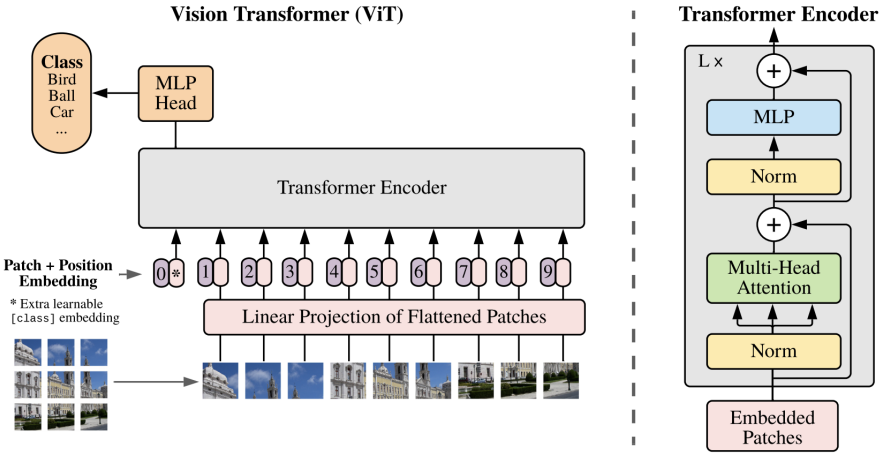


Figure 1.6. – ViT architecture (Dosovitskiy et al. 2021)

Visual Transformer (ViT) With the success of Transformer models in Natural Language Processing, Dosovitskiy et al. 2021 introduced a Visual Transformer (ViT) which

reached state-of-the-art results in several computer vision tasks. As transformers need larger amounts of training data than CNNs, ViT is pre-trained on a 300 million images dataset. The authors apply the BERT encoder transformer architecture with minimal changes. The input image is split into several $16 * 16$ pixel patches, which are then projected, as shown in Figure 1.6. The transformer receives a 1-dimensional sequence of embeddings. The positional encoding is adapted to take into account the 2-dimensional nature of the vision modality. Similarly to BERT, it also uses a classification embedding to represent the whole instance, i.e., the image. Contrary to its language counterparts, ViT is pre-trained using a supervised task of image classification. It can then be fine-tuned on specific computer vision tasks or datasets.

Other architectures Several other transformer-based architectures have been developed for computer vision, to enable more efficient pre-training. DeiT (Touvron, Cord, Douze, et al. 2021) uses distillation to improve the efficiency of Visual Transformers, and is pre-trained on a smaller dataset of 1.2 million images, with a CNN teacher model. To that end, a 'distillation' embedding in addition to the 'classification' embedding. As the transformer architecture lacks the ability of CNNs to learn local features, some models have sought to reproduce their hierarchical structure. For instance, in Z. Liu et al. 2021; X. Dong et al. 2022, authors use hierarchical blocs, which reduces the number of parameters in later layers to create higher-dimensional representations.

Study of visual representations Though the use of Transformers for computer vision is a more recent development than in Natural Language Processing, studies have started to compare visual representations created by Transformers to those created by CNNs. For example, Raghu et al. 2021 find that Transformers extract global information earlier than CNNs, and need a large quantity of data to learn to aggregate local information at lower layers, which CNNs do easily. A comparison of CNNs and transformer with the same training protocols has shown that while Transformers are no more robust than CNNs to adversarial attacks, they are more robust to out of distribution examples (Bai et al. 2021). Studying more specific capabilities of Transformers, Naseer et al. 2021 find that they are robust to severe occlusions and less biased towards local textures than CNNs. Additionally, they how good shape recognition abilities.

1.2. Architecture of Vision-Language Transformers

Following the success of the pre-trained models in Natural Language Processing tasks, researchers have built similar models for vision-language multimodal tasks. The goal is to pre-train a self-supervised vision-language model on a large amount of image-text data. The pre-trained model learns to extract visual, textual and multimodal concepts from texts and images. In this thesis, we are interested in pre-trained vision-language models able to learn general-purpose multimodal representations,

rather than models learning representations specific to one task. Indeed, such representations could be fine-tuned on supervised multimodal tasks with limited data. In addition, multimodal models able to extract a large amount of relevant information would need less resources to fine-tune.

First, we explain different architectures of vision-language transformers in section 1.2. Then, we introduce the datasets and tasks used to pre-train those models in section 1.3. Finally, we present the evaluation of vision-language transformers in section 1.5.

Most vision-language transformer models are pre-trained using instances composed of an image and its associated caption. However, they have different ways of encoding the text and image data, of merging their representations, and different pre-training tasks used to update the weights of the model. Figure 1.7 presents a general overview of vision-language transformer pre-training.

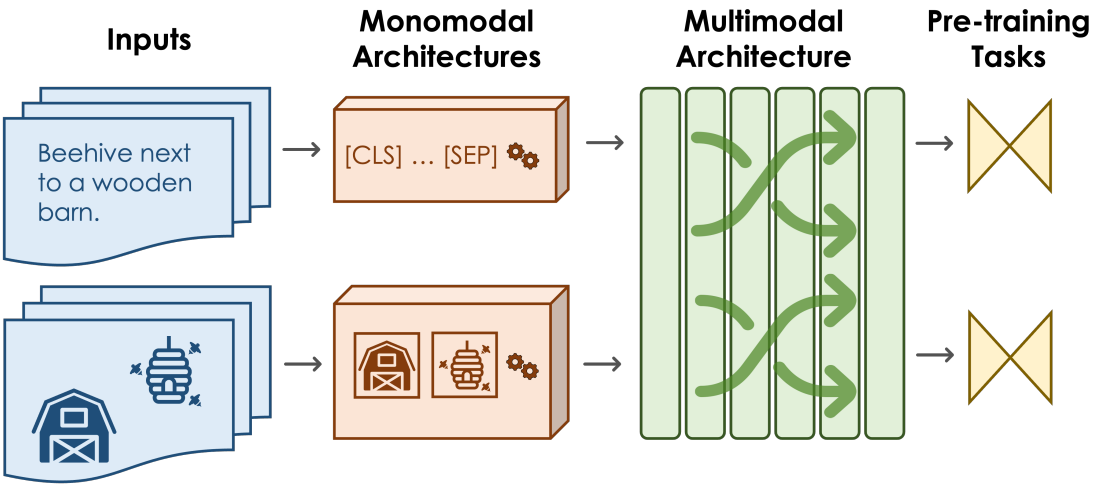


Figure 1.7. – Overview of a Vision-Language transformer model's pre-training

1.2.1. Monomodal Architecture

Text encoder In most cases (Y.-C. Chen et al. 2019; Tan et al. 2019), the text is encoded using a process similar to the BERT model. While most models use BERT cloze-style pre-training tasks, some also use autoregressive tasks similar to GPT pre-training, using visual representations as context in addition to previous words. The text, most of the time a caption, is tokenized into word pieces. The resulting token embeddings are added to position embeddings that encode the position of a token in the sequence. In addition, a modality embedding can be added to represent the textual modality. The position embedding and modality embedding can be learnable parameters.

Some models (J. Li, R. Selvaraju, et al. 2021) use a pre-trained language transformer to encode the textual modality. The output of the transformer is then used as input of the multimodal transformer. The weights of the language transformer can be updated

during the pre-training, J. Yang et al. 2022 use a momentum encoder to this end. LXMERT (Tan et al. 2019) also uses a BERT-style single modality transformer for text, without loading pre-trained weights.

Image encoder Several methods have been developed to encode the visual modality for vision-language Transformers:

— Object detector: The image is preprocessed using an object detector, which outputs a sequence of object representations. Most of the time, this object detector is frozen, which means that the weights are not updated during pre-training. The bounding box of each object is used to compute a positional encoding, which is added to the object representation. While earlier models (Y.-C. Chen et al. 2019) use the same frozen Region-based Convolutional Neural Network (R-CNN) object detector (Faster RCNN) (Ren et al. 2015) to process images, VinVL (P. Zhang et al. 2021) has shown that an object detector trained on more data and categories produces better results in downstream tasks.

— Patch features: Inspired by ViT Dosovitskiy et al. 2021, ViLT W. Kim et al. 2021 and other models use patch embeddings to encode the image modality. Square pixel patches are extracted from the image and projected linearly. Position embeddings, which are usually learnable parameters, are added to this projection.

— Convolutional features: The image can be processed using a CNN to get visual features, like Zirui Wang et al. 2022. Those features can be directly used as input of the multimodal Transformer, with added positional embeddings. In Flamingo (Alayrac et al. 2022), a resampler is trained on top of the frozen visual encoder, in order to reduce the dimension of the visual features.

— Visual transformers: Similar to the textual modality, some works (J. Yang et al. 2022; J. Li, R. Selvaraju, et al. 2021) use a pre-trained vision transformer to process the visual input of a multimodal transformer. LXMERT (Tan et al. 2019) also uses an object relationship transformer without loading the pre-trained weights. In this case, the processing of the visual modality, as well as the positional embeddings, depends on the transformer.

Integration of text and image data Most models consider the text and image data either in separate architectures or as two successive parts of the same sequence in the same architecture, some models use other approaches. In the latter case, modality-specific embeddings are added to the input to differentiate between text and image data, in addition to their respective position embeddings. Alayrac et al. 2022 integrate visual information in a language transformer using cross-attention layers. Zhengyuan Yang et al. 2021 integrate visual information to text tokens by using $< obj >$ tokens with bounding box information with relevant object words.

1.2.2. Multimodal Architecture

Types of multimodal attention Once mono-modal text and image representations are computed, there exist several approaches to merge the two modalities (Figure

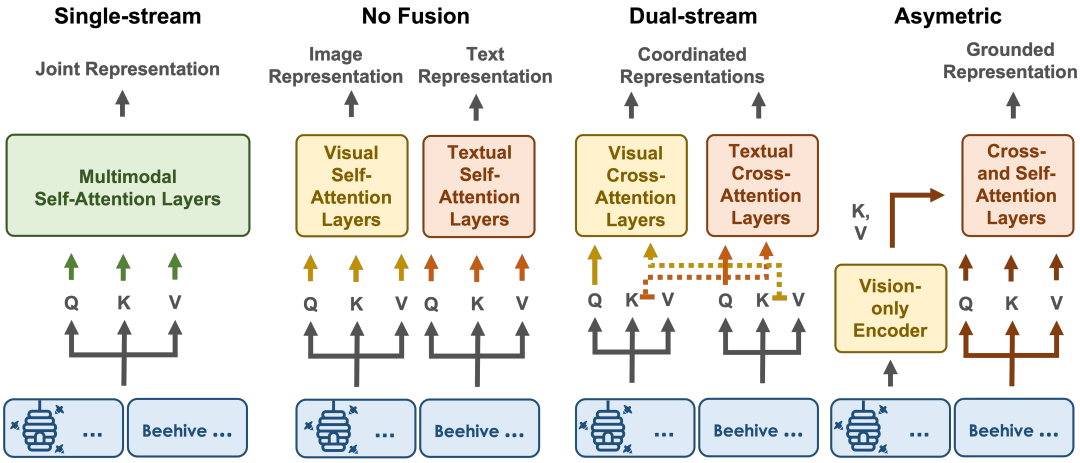


Figure 1.8. – Types of cross-modal architecture for vision-language models

1.8).

— Single-stream architecture or merged attention (Y.-C. Chen et al. 2019): This approach is the most favored when building vision-language transformers. It consists in treating the two modalities as two parts of the same sequential input. Self-attention is computed on the whole sequence, regardless of the modality. A modality embedding is summed to the mono-modal representations in order to differentiate between the two modalities.

— Double-stream architecture or co-attention (J. Lu, Batra, et al. 2019b; Tan et al. 2019): Contrary to the single-stream architecture, those models do not concatenate the modalities into one input sequence. They use them as input of two modality-specific transformer blocks that do not share parameters. They compute cross-attention by using a modality to attend to the other, and alternate cross-attention mechanisms with intramodality self-attention mechanisms.

— No multimodal attention: Some models, such as CLIP (Radford, J. W. Kim, et al. 2021), only use the Transformer architecture for each modality, and do not merge the textual and visual representations. The model uses contrastive pre-training to link image and text representations.

— Inserting cross-attention into monomodal transformers: Some models use asymmetric cross attention, using one modality as query and the other as key and values. This is the case of Alayrac et al. 2022, which use frozen visual and textual encoders, and adds asymmetric cross attention layers to the language transformer. FIBER (Dou, Kamath, et al. 2022) inserts cross attention into both the language transformer and the visual transformer. FIBER cross-attention layers can also be removed for tasks such as retrieval.

Use of transformer decoder Most models follow the same general transformer architecture as BERT (Devlin et al. 2019) or ViT (Dosovitskiy et al. 2021), and only use a Transformer encoder. However, some models such as SimVLM (Zirui Wang

et al. 2022) also use a decoder architecture in order to pre-train the models using generative pre-training tasks. While encoder models are more helpful for classification and retrieval tasks, they are limited for generative tasks. Similarly, encoder-decoder models have limitations for non-generative tasks.

In BLIP (J. Li, D. Li, Xiong, et al. 2022), authors propose an architecture that combines both cases and can be used in both generative and retrieval downstream tasks. They use a modular architecture, using monomodal encoders, a text encoder with image co-attention and a text decoder with image co-attention trained with different pre-training tasks.

Task-specific architecture Similar to monomodal models, Vision-Language models usually use fully connected layers on top of the final layers of the transformer decoder, such as classification heads or language modeling heads. These heads can project the hidden states of the final layer into logits for classification tasks. A task-specific pre-training loss such as the Cross Entropy loss is then computed using those logits.

1.2.3. Discussion

Evolution of vision-Language transformer architectures Table 1.1 compares the architecture of Vision-Language Transformer models. Earlier models used mostly used a single- or dual-stream encoder and extracted features using a pre-trained object detector model. The development of visual transformers has led the way towards the use of two mono-modal transformers as well as an additional single-stream multimodal transformer. Another developing trend is the use of very large language models supplemented with multimodal information (Z. Peng et al. 2023; S. Huang et al. 2023). This new type of architecture can take advantage of pre-trained large language models at minimal fine-tuning cost, and remain usable in language-only contexts. In addition, this method can be adapted to other modalities beyond images. However, there is no consensus yet on the best performing and most efficient architectural designs for multimodal transformers. Figure 1.8 shows an overview of several architectures.

Using monomodal transformers Dou, Xu, et al. 2022 compare different architectural choices. They use an architecture based on monomodal transformers and a double stream multimodal transformer. They find that the choice of different pre-trained visual transformers can significantly impact the results of a model. While the use of a pre-trained textual transformer also improves the results, two different language transformers will not have an important difference in results.

Input features: impact of image resolution In several works (Dou, Xu, et al. 2022; Gui et al. 2022), authors have found that using higher-resolution images during fine-tuning significantly improves results on downstream tasks. This echoes

Model	Text Enc.	Image Enc.	Transformer	Size
UNITER (Y.-C. Chen et al. 2019)	emb.	OD	single-stream	86M/303M [a]
VisualBERT (L. H. Li et al. 2019)	emb.	OD	single-stream	100M
VLBERT (Su et al. 2020)	emb.	OD	single-stream	—
Unicoder-VL (G. Li et al. 2020)	emb.	OD	single-stream	—
Unified VLP (L. Zhou et al. 2020)	emb.	OD	single-stream	—
PixelBERT (Zhicheng Huang, Z. Zeng, B. Liu, et al. 2020)	emb.	CNN	single-stream	144 M
OSCAR (X. Li et al. 2020)	emb.	OD	single-stream	86M/303M
VinVL (P. Zhang et al. 2021)	emb.	OD	single-stream	86M/452M
UNIMO (W. Li et al. 2020)	emb.	OD	single-stream [b]	—
VILLA (Gan et al. 2020)	emb.	OD	single stream	—
VL-T5 (Cho et al. 2021)	emb.	OD	single stream [c]	220M
ViLT (W. Kim et al. 2021)	emb.	patch emb.	single-stream	87 M
MDETR (Kamath et al. 2021)	RoBERTa	CNN	single-stream	—
SOHO (Zhicheng Huang, Z. Zeng, Y. Huang, et al. 2021)	emb.	ResNet	single-stream	—
VLP-IMF (Xue et al. 2021)	emb.	TF	single-stream	180M
CLIP-ViL (Shen et al. 2022)	emb.	CLIP	single-stream	—
FLAVA (Singh, R. Hu, et al. 2021)	TF	TF	single-stream	—
TCL (J. Yang et al. 2022)	TF	TF	single-stream	210 M
SimVLM (Zirui Wang et al. 2022)	emb.	patch emb.	single-stream [c]	—
X-VLM (Y. Zeng et al. 2021)	6l BERT	Swin TF	single-stream	216 M
VLC (Gui et al. 2022)	emb.	patch emb.	single-stream	86M/307M
VL-BEIT (Bao, W. Wang, L. Dong, and F. Wei 2022)	emb.	patch emb.	single-stream [d]	—
LEMON (X. Hu et al. 2021) [e]	emb.	OD	single-stream	13M/675M
Unitab (Zhengyuan Yang et al. 2021)	RoBERTa	ResNet	single-stream [c]	—
OFA (P. Wang et al. 2022)	emb.	CNN	single-stream [c]	33M/940M
KOSMOS (S. Huang et al. 2023)	emb.	frozen CLIP	single-stream	1.6B
CLIP (Radford, J. W. Kim, et al. 2021) [f]	TF	ViT	no fusion	151 M
Align (Jia et al. 2021)	BERT	EfficientNet	no fusion	—
Florence (Yuan et al. 2021)	12l TF	CoSwin TF	no fusion [g]	893M
LXMERT (Tan et al. 2019)	TF	OD + TF	dual-stream	181 M
ViLBERT (J. Lu, Batra, et al. 2019b)	TF	OD	dual-stream	221 M
12-in-1 (J. Lu, Goswami, et al. 2020)	TF	OD	dual-stream	270 M
Ernie-ViL (F. Yu et al. 2021)	emb.	OD	dual-stream	—
ALBEF (J. Li, R. Selvaraju, et al. 2021)	6l BERT	12l ViT	dual-stream	210 M
VLMO (Bao, W. Wang, L. Dong, Q. Liu, et al. 2021)	emb.	patch emb.	dual-stream	—
METER (Dou, Xu, et al. 2022)	RoBERTa	ViT	dual-stream	—
CoCa (J. Yu et al. 2022)	TF	TF	dual-stream [c]	2.1 B
BLIP (J. Li, D. Li, Xiong, et al. 2022)	TF	TF	dual-stream [c]	252 M
FIBER (Dou, Kamath, et al. 2022)	emb.	patch emb.	dual-stream	
Flamingo (Alayrac et al. 2022)	emb.	frozen NFNet	asymmetric	3B/9B/80B

Table 1.1. – Architecture of Vision-Language models: Types of image encoder, text encoder, multimodal transformer as well as number of parameters when available. emb. refers to embedding, TF refers to transformer, OD refers to object detector.

a. Like other models using a pre-trained and frozen R-CNN, the object detector weights are not counted.

b. Also uses monomodal transformers for monomodal representations

c. The model also uses a decoder for text generation

d. Model uses Beit tokenizer for patch classification

e. LEMON is a captioning model.

f. CLIP is used as an image-text retrieval model.

g. A dual-stream adapter is added to pre-train the model

other works in computer vision showing that higher resolution often leads to better performances for CNNs (Kannojia et al. 2018). For visual transformers, having high image resolution during pre-training is not necessary, as fine-tuning them with higher resolution can be sufficient (Touvron, Cord, El-Nouby, et al. 2022).

Impact of attention Several works have found no significant differences in the use of single- or dual-stream multimodal transformers. However, Dou, Xu, et al. 2022 find that dual-stream attention improves performances in their setting: two pre-trained monomodal transformers and a multimodal transformer. This improvement is more visible on image-text retrieval tasks. In Hendricks, Mellor, et al. 2021, authors also find marginally better results with co-attention instead of merged attention. With the evolving architectural designs of vision-language models, we still do not know what type of attention is more efficient.

1.3. Pre-training of Vision-Language Transformers

Vision-language Transformer models are pre-trained on large amounts of parallel image-text data, on a combination of textual, visual and multimodal pre-training tasks. While some models also use video-based datasets (Alayrac et al. 2022), those will not be the focus of this section.

1.3.1. Pre-training Dataset

Summary of image-text datasets A limiting aspect of vision-language multimodality is the amount of available data for training. Indeed, very large image-text datasets are necessary to pre-train vision-language transformers. Several datasets have been created to help vision-language multimodal tasks, and many of them have been used for the pre-training of vision-language transformer models. They are composed of images and associated text, which most of the time is a one-sentence caption.

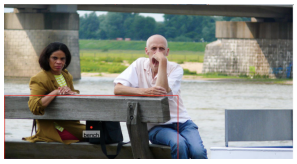
Dataset	Size (Images / Captions)	Annotation	Example
MSCOCO (T.-Y. Lin et al. 2014)	111 k / 558 k	Manual	A horse carrying a large load of hay and two people sitting on it. 
Visual Genome (Krishna, Yuke Zhu, et al. 2016)	103 k / 5 M	Manual	Park bench is made of gray weathered wood 
CC (3M) (Sharma et al. 2018)	3 M / 3 M	Crawled	a worker helps to clear the debris. 
CC (12M) (Changpinyo et al. 2021)	12 M / 12 M	Crawled	<PERSON> was the first US president to attend a tournament in sumo's hallowed Ryogoku Kokugikan arena. (AFP photo) 
SBU (Ordonez et al. 2011)	1 M / 1 M	Crawled	Man sits in a rusted car buried in the sand on Waitarere beach 
LAION (Schuhmann, Vencu, et al. 2021)	400 M / 400 M	Crawled	cat, white, and eyes image 

Table 1.2. – Details of commonly used and publicly available pre-training datasets at the time of writing, with one image-text example from each dataset.

There are two main types of datasets: automatically collected datasets and human annotated datasets (see Table 1.2). Before the use of very large datasets from Common Crawl, they were built up from manual annotations. The largest human-annotated datasets are MS COCO (T.-Y. Lin et al. 2014) and Visual Genome (Krishna, Yuke Zhu, et al. 2016). As those datasets use images from similar sources, they may have images in common.

— The MS COCO dataset is mainly composed of 'non-iconic' images, which are images with multiple objects instead of one centered object. The associated texts correspond to image descriptions provided by different annotators, with five captions per image in order to include diverse possible descriptions. Annotators are instructed to describe all important parts of the scene in at least eight words, which gives rich captions.

— Visual Genome is composed of images similar to MS COCO. However, the dataset is an object-oriented dataset. Each image has several annotations corresponding to region descriptions.

However, human annotated datasets are expensive to create and thus limited in size. The use of larger models in the last few years has led the need for bigger datasets. To increase the amount of available image-text data, other datasets have been created using web crawled data, such as SBU (Ordonez et al. 2011), LAION (Schuhmann, Vencu, et al. 2021) and the Conceptual Captions (Sharma et al. 2018; Changpinyo et al. 2021) datasets. They are noisier and less stable, as some data may not be available anymore.

— SBU images and captions are collected using Flickr, filtered so that the captions match visual content in the image.

— The two Conceptual Captions (CC) datasets are automatically collected from the web, using 'alt-text' fields as captions, with some filters on text and image quality. Classifiers are used to ensure that some text tokens can be associated with image components. For example, authors require the presence of specific parts of speech. The correspondence between image and text is ensured by using vision models to assign labels to images and compare them to text. The CC 3M dataset transforms the text to get captions similar to standard captioning data, for instance by removing named entities. CC 12M does not apply any transformation on the text.

— LAION-400M is an image-text dataset obtained using crawled web data, with more flexible filtering criteria. The dataset is filtered using CLIP embeddings and to ensure that text and image pairs are above a similarity threshold. A new version (Schuhmann, Beaumont, et al. 2022), made of 5.85 billion image-text pairs, is now available.

While most models are pre-trained on available datasets, some also use private data. This is the case of CLIP. Indeed, this model is pre-trained on a private dataset of 400 million image-text pairs. Some models use even more data. Align (Jia et al. 2021), for example, uses a noisy dataset with minimal filtering consisting of 1.8 billion image-text pairs.

Some models also use vision-only datasets to complement image-text data. Indeed,

VinVL (P. Zhang et al. 2021) uses the Open Images dataset (Kuznetsova et al. 2020). While we do not mention multimodal documents in this section, recent models Z. Peng et al. 2023 have taken to using datasets based on multimodal documents for vision-language pre-training. For this pre-training scheme, rather than using image/text pairs, image and text inputs are interleaved.

Ethical issues The selection of a dataset for vision-language pre-training can raise ethical issues, depending on how the data is collected and filtered.

— The use of web crawled data raises the issue of consent in acquiring data. For instance, web-crawled datasets such as CC (Sharma et al. 2018; Changpinyo et al. 2021) or LAION (Schuhmann, Vencu, et al. 2021) contain images that are not explicitly licensed. Some datasets also collect images and text that include personal information without requiring consent from the individuals. Datasets can also include images from problematic sources and illegal data (Birhane et al. 2021).

— In the case of human-annotated datasets or web-crawled datasets that require human evaluations, it is important to consider their compensation and working conditions, and especially psychological impact of harmful content (Díaz et al. 2022).

— Datasets can subject to harmful bias, especially in the case of minimally filtered data. One of the most common forms is the lack of representation. It usually targets some societal groups (D. Zhao et al. 2021), and is called representational bias. It can have a significant impact in downstream applications (Birhane et al. 2021). To mitigate this bias, it is important to consider the sources of visual data, and when there are human annotators, the diversity of those annotators in terms of culture and experiences (Díaz et al. 2022).

Data augmentation In order to better take advantage of the available data, data augmentation techniques have been used for both images and text. For the visual modality, data augmentation used for vision-language models follows the methods developed for vision-only models. Automated data augmentation methods can significantly improve the results of deep learning models and is used in many state-of-the-art models. In Cubuk et al. 2019, authors investigate data augmentation techniques and build RandAugment, a new technique with fewer parameters to tune. This method consists in choosing a number n of data augmentation techniques from a list (e.g., rotation, brightness, contrast) and setting the intensity of the augmentation methods. The n augmentation methods are then applied to the image.

Although less common, some data augmentation techniques have also been developed for the textual modality. Some models also augment textual data with back translation (Edunov et al. 2018), which consists in using an intermediary language to translate a text once and back. The goal is to change the syntax without changing the meaning. In BLIP (J. Li, D. Li, Xiong, et al. 2022), authors introduce a new data augmentation technique that uses a multimodal decoder to create captions

from web images and a multimodal encoder to filters out noisy text samples from the pre-training dataset. According to the authors, this technique is analogous to knowledge distillation, because a pre-trained captioning model distills its knowledge to the vision-language transformer.

Choice of pre-training dataset In Hendricks, Mellor, et al. 2021; Singh, Goswami, et al. 2020, authors compare how the use of different pre-training datasets affects the pre-training. Using more images will lead to better performances, but larger datasets are not always better, as similarity between pre-training and evaluation data is an important factor.

For example, Hendricks, Mellor, et al. 2021 find that using SBU captions (Ordonez et al. 2011) will lead to worse performances than a smaller dataset such as MSCOCO. This is likely due to the noise of the dataset. Indeed, they find that the SBU captions dataset has less overlap between detected objects and words than other datasets. They also find that the similarity in language between pre-training and fine-tuning datasets significantly impacts model performances, even with the same pre-training images. J. Li, D. Li, Xiong, et al. 2022 also find that noisy datasets affect negatively the pre-training, and that more diverse captions lead to better results.

1.3.2. Pre-training Tasks

In this section, we focus on pre-training tasks that use an input image V and a text W, $(W,V) \in \mathscr{D}$, with $\mathscr{D}$ the pre-training dataset. We denote the text tokens w and the image tokens v. In the case of tasks using masked tokens, we denote w_m and v_m those masked tokens. We note $w_{\tilde{m}}$ and $v_{\tilde{m}}$ for their non-masked counterparts.

Language-based pre-training We call language-based pre-training tasks those that use textual information as labels to train the model. However, as vision-language models are multimodal, input can be composed of both an image and a caption, and textual as well as visual information can be used for these tasks.

— MLM: As explained in section 1.1.4, MLM consists in predicting masked or otherwise corrupted textual tokens, which means minimizing the MLM loss (Equation 1.3). Vision-Language models have different ways of applying this task. While most follow the protocol of BERT (Devlin et al. 2019), some change the masking percentages or use whole-word masking instead of token-based masking (W. Kim et al. 2021).

$$\text{MLM} = -\mathbb{E}_{(W,V)\in\mathscr{D}} \log \mathbf{P}(w_m \mid w_{\tilde{m}}, v) \tag{1.3}$$

— Autoregressive pre-training or Prefix Language Modeling (PLM): It consists in generating the next word in a sentence. SimVLM (Zirui Wang et al. 2022) uses a generative pre-training task with the image and previous words as context, as in Equation 1.4. It differs from the traditional autoregressive pre-training because

it enables the use of bidirectional attention on the context sequence.

$$\text{PLM} = -\mathbb{E}_{(W,V)\in\mathscr{D}} \log \mathbf{P}(w_{\geq T} \mid w_{<T}, v) \tag{1.4}$$

This task is especially helpful for captioning and text generation, and some models use it as the sole learning objective. This is the case in recent large multimodal models, such as KOSMOS (S. Huang et al. 2023). This model focuses on conversations aspects with multimodal inputs.

— Contrastive language pre-training (Contrastive Learning (CL)): It is a form of intramodal contrastive learning, introduced in vision-language models by TCL (J. Yang et al. 2022). The goal is to learn the semantic differences between variation of a text input and other inputs. It is based on the InfoNCE loss. Equation 1.5 computes the InfoNCE loss with W, V, and $\widetilde{V}$ as set of k negative textual examples and τ a temperature hyperparameter and sim a similarity function between two tensors. The model uses dropout as a data augmentation technique to get several views of the same text. Another variation of this task introduced by TCL is Local Mutual Information Maximization. Instead of considering only global [CLS] features, this task will encourage high mutual information between the global [CLS] feature and local tokens.

$$\text{InfoNCE} = -\mathbb{E}_{(W,V)\in\mathscr{D}} \log \frac{\exp(sim(V,W)/\tau)}{\sum_{n=1}^{k} \exp(sim(V,\widetilde{W_k})/\tau)} \tag{1.5}$$

While most models use MLM, recent very large language models have focused on generation tasks, with sometimes no other pre-training task. Indeed, as the PLM task uses both textual and visual information, it can be used with no additional task to extract and generate multimodal concepts.

Vision-based pre-training Similarly to the language-based task, the vision-based tasks use visual information as labels. As a result, different types of visual input will lead to different possible visual tasks. For example, vision-Language transformers that use image features based on object detectors, tend to use tasks based on object categories. In some cases (Dou, Xu, et al. 2022; W. Kim et al. 2021) authors do not use any vision-based pre-training tasks. Indeed, they found no visual task that improves the performance.

— Masked region classification: This task is used by UNITER (Y.-C. Chen et al. 2019) and other models that use object detectors to provide object regions as visual features. Similar to MLM, it consists in predicting the object category of masked region features. A portion of object representations are masked, and the model predicts their label. There are two ways to obtain labels from the object detector. Either the most likely object category is used as ground truth, or the object distribution at the output of the object detector is used as a soft label. In the first case, the loss used is the Cross Entropy loss, like MLM. It is shown in equation 1.6 with CE the cross entropy loss, y the classification labels and $f(v)$

the output of the task-specific functions applied to the visual tokens. The goal of this function is to project into the relevant number of classes and normalize. In the second case, KL-divergence can be used to minimize the distance between the two probability distributions.

$$\text{MRC} = -\mathbb{E}_{(W,V)\in\mathscr{D}} \sum_m CE(y_m, f(v_m)) \tag{1.6}$$

— Masked feature regression: Similar to MLM, this task consists in predicting the input masked image features. The label used for this task is the masked input features, and the L2 norm between the label and an output of the same dimension is computed to obtain the loss. This task is mostly used by models that use object detectors to get visual features (Y.-C. Chen et al. 2019). Equation 1.7 shows the MFR loss with $f(v)$ the output of the task-specific visual head and $r(v)$ the input feature of a region.

$$\text{MFR} = -\mathbb{E}_{(W,V)\in\mathscr{D}} \sum_m \left\| f(v_m) - r(v_m) \right\|_2^2 \tag{1.7}$$

— Masked patch classification: Inspired by the visual transformer BEIT (Bao, L. Dong, et al. 2021), VL-BEIT (Bao, W. Wang, L. Dong, and F. Wei 2022) uses it to compute discrete labels. A percentage of image patches are masked, and the model must predict the discrete token corresponding to the masked patches. Another variation of this task is proposed by Dou, Xu, et al. 2022, doing patch classification with in-batch negatives.
— Masked patch prediction: Inspired by the visual transformer masked auto-encoder (K. He, X. Chen, et al. 2022), Gui et al. 2022 mask a percentage of the input image patches and reconstruct those missing patches using a transformer decoder. The loss is computed using the L2 norm between the input pixels and reconstructed pixels.
— Contrastive visual pre-training: Similarly to contrastive language pre-training, it is introduced in TCL (J. Yang et al. 2022) as an intramodal contrastive learning task, also based on the InfoNCE loss. The different visual views of the same image are obtained using standard visual data augmentation techniques. The Local Mutual Information Maximization variation was also used in TCL for the image modality.

The use of visual tasks mostly depends on the type of visual input given to the model. Many recent models, especially those using visual representations from pre-trained models, have found it unnecessary to use visual pre-training tasks.

Cross-modal pre-training Some models, such as OSCAR (X. Li et al. 2020) and Ernie-Vil (F. Yu et al. 2021) use additional inputs. These inputs can be tags or scene graphs that enable them to pre-train on additional pre-training tasks. Most cross-modal pre-training tasks use as weak supervision the fact that images and their captions are two views of the same instance in different modalities. Figure 1.9 shows

different multimodal pre-training tasks.

— Image Text Matching (ITM): This is the most commonly used multimodal task. Similar to the next sentence prediction task used by BERT (Devlin et al. 2019), models use a [CLS] token to represent the whole input. For this task, the input is constituted half the time of a matched image-text pair, and the other half of an unrelated image and text pair selected randomly in the dataset. This task is a binary classification task to predict whether the image and text match, the binary cross entropy loss is used. Equation 1.8 shows this loss, with $f(V,W)$ the output of the [CLS] token through the task-specific layers, and l the label: 1 for matching, 0 for non-matching pairs.

$$\text{ITM} = -\mathbb{E}_{(W,V)\in\mathscr{D}}[y\log(f(V,W)) + (1-y)\log(1-f(V,W))] \tag{1.8}$$

— Word Region Alignment (WRA): The goal of this task is to get a fine-grained alignment between image features and word tokens. It is introduced by UNITER (Y.-C. Chen et al. 2019), which uses object regions as image features. As a result, this task computes an alignment between the visual representation of objects and words. This task uses Optimal Transport to learn a transport plan to optimize the alignment between image and text. It consists in computing an approximation of the optimal transport distance between an image text pair, using the IPOT algorithm (Y. Xie et al. 2020). This distance is used as an alignment loss, represented in Equation 1.9, where T is the transport plan and c is the cost function, which can be the cosine distance. Similarly to ITM, the image and text input pair are half of the time unrelated. As a result, if the image and text pair match, the distance is positive and should be minimized, while if they do not match, the distance is negative and should be maximized.

$$\text{WRA} = \min_{T} \sum_{i} \sum_{j} T_{i,j}\, c(w_i, v_j) \tag{1.9}$$

— Multimodal contrastive pre-training: This task consists in computing a similarity score between a text and an image. The goal is to bring closer global [CLS] features of matching text and images. In TCL (J. Yang et al. 2022), the model maximizes the mutual information between matched image-text pairs by minimizing the InfoNCE loss (Oord et al. 2018).

— Supervised multimodal tasks: Contrary to the other tasks, these tasks need a specific dataset and can also be used as a fine-tuning task. This is the case of VQA (Y. Goyal et al. 2017), which LXMERT (Tan et al. 2019) uses during its pre-training. The model gets an image and question as input, and must answer the question. Most of the time, it is done as a classification task, with all answers in the training dataset matched to labels, but it can also be treated as a generative task.

— Multimodal Next Token Prediction (MNTP): This pre-training task consists in predicting in a left to right manner interleaved textual tokens and visual representations. It is for instance used by the KOSMOS (S. Huang et al. 2023) model. In the case of KOSMOS, only discrete textual tokens are used for the loss. However,

the KOSMOS-2 (Z. Peng et al. 2023) model also considers bounding box location information in the loss.

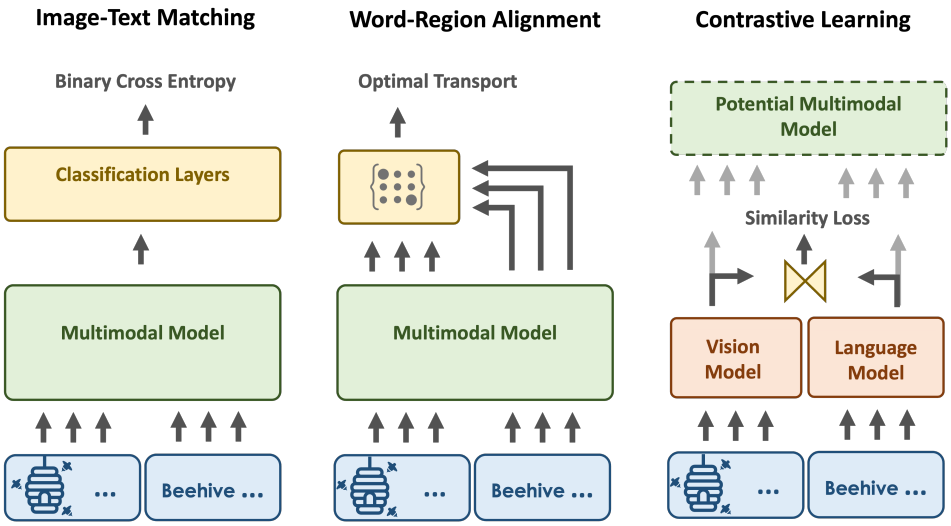


Figure 1.9. – Examples of multimodal pre-training tasks

While image-text matching remains a staple of vision-language models, it is sometimes insufficient to learn deeply multimodal representations. As a result, the use of multimodal tasks based on contrastive learning has increased in recent years. Figure 1.10 shows the different contrastive tasks used by the TCL model (J. Yang et al. 2022).

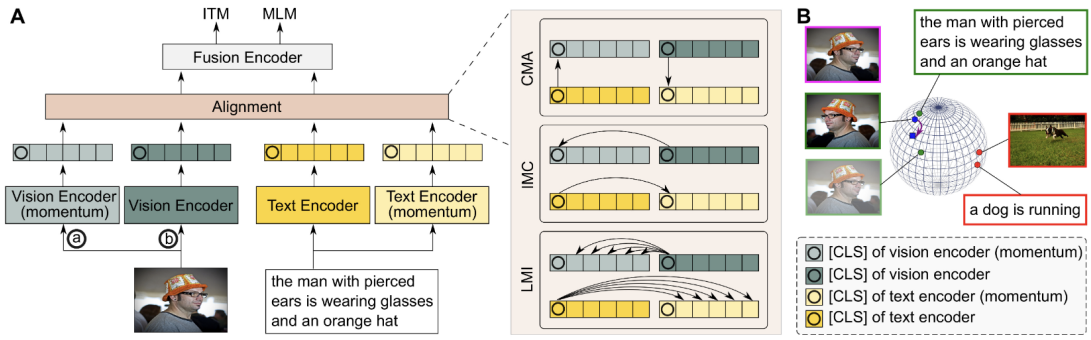


Figure 1. (A): An overview of our framework which consists of a vision encoder, a text encoder, and a fusion encoder. Each encoder has a paired momentum encoder updated by the momentum-based moving average. For the image input, we apply two separate data augmentation operators (a and b) which are sampled from the same family of augmentations. The alignment module contains three contrastive objectives (i.e., CMA, IMC, and LMI) for both cross-modal and intra-modal representation learning (make it easier for the fusion encoder to learn joint multi-modal embeddings). (B): The motivation of leveraging both cross-modal and intra-modal supervision. The original image (pink) is augmented to two different views (green). For CMA only, the middle image only has a positive text example (green) and treats other texts (red) as negatives. Its embedding (blue cirble) would be close to its positive text example. By incorporating IMC, it has two positive examples (one text and one image) and two sets of negative examples (one from text and one from image) and tends to learn more reasonable embeddings (blue square).

Figure 1.10. – Pre-training tasks of TCL (J. Yang et al. 2022)

Ablation on pre-training tasks Many authors have performed ablation on pre-training tasks, to understand the role of a task in the performance of a model. The use of MLM and ITM usually leads to an important increase in performance, while the use of visual tasks does not always increase performance(Dou, Xu, et al. 2022; W. Kim et al. 2021).

Pre-training protocol The pre-training protocol of vision-language transformers can vary a lot in addition to the different choices in pre-training tasks and datasets. Some models use load pre-trained mono-modal models, either by loading their weights at the beginning of the training, or by using them as 'frozen' models to compute mono-modal representations. Some models (Xue et al. 2021) implement more complex masking techniques for pre-training tasks that generally use random masking. VILLA (Gan et al. 2020) implements adversarial training strategies by introducing perturbations in the embedding space. Some models, such as TCL (J. Yang et al. 2022) and ALBEF (J. Li, R. Selvaraju, et al. 2021) use momentum encoders. Those aim at updating more robustly parts of the model when pre-training on large and noisy datasets. For example, they can help learn more robust mono-modal representations. Finally, model pre-training can also vary in size, following the number of epochs and the resources. Table 1.3 summarizes those pre-training tasks and protocol for vision-language transformers.

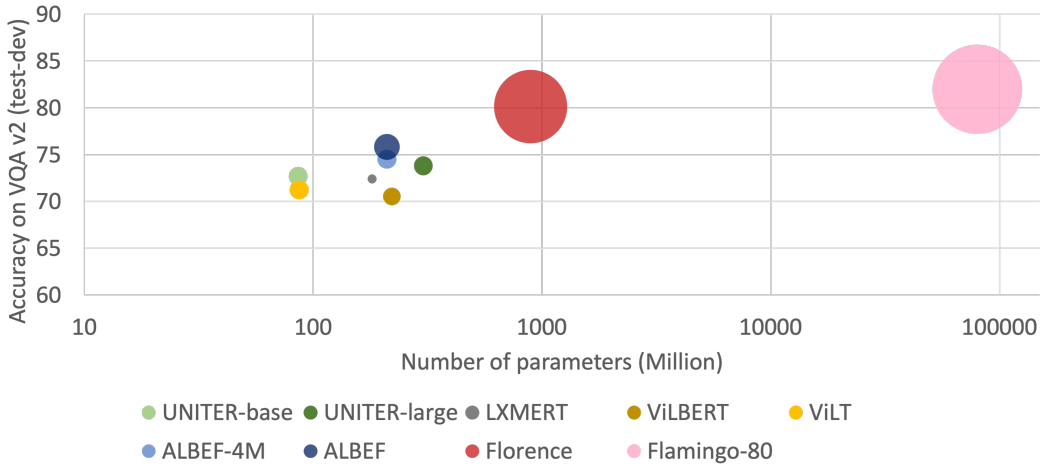


Figure 1.11. – Accuracy of selected vision-language transformers on VQA (VQAv2 test-dev) in relation to the number of parameters of the model (million) and the number of images of the dataset (size of the point)

1.3.3. Conclusion

The recent popularity of vision-language transformer models has led to the development of various models, with different architectural choices, and pre-training protocols. Those differences can make it difficult to evaluate the impact of individual

Model	Textual	Visual	Multimodal	Dataset	Duration
UNITER (Y.-C. Chen et al. 2019)	MLM	MRR+MRC	ITM+WRA	4M	L: 3.7k V100 h [a]
LXMERT (Tan et al. 2019)	MLM	MRC	ITM+VQA	200k	960 Titan Xp h
VisualBERT (L. H. Li et al. 2019)	MLM	—	ITM	100k	96 V100 h
ViLBERT (J. Lu, Batra, et al. 2019b)	MLM	MRC	ITM	3M	—
VLBERT (Su et al. 2020)	MLM	MRC	ITM	3M	—
Unicoder-VL (G. Li et al. 2020)	MLM	MRC	ITM	3.8M	—
Unified VLP (L. Zhou et al. 2020)	MLM+PLM	—	—	3M	—
12-in-1 (J. Lu, Goswami, et al. 2020)	MLM	MRC	ITM+others	3M	960 V100 h
PixelBERT (Zhicheng Huang, Z. Zeng, B. Liu, et al. 2020)	MLM	—	ITM	200k	—
OSCAR (X. Li et al. 2020)	MLM	MTC	ITM	4M	—
VinVL (P. Zhang et al. 2021)	MLM	MTC	ITM	6M	—
UNIMO (W. Li et al. 2020)	—	—	CL	4.1M	L: 15.4k V100 h
Ernie-ViL (F. Yu et al. 2021)	MLM	MRC	ITM+SGP	3.8M	—
VILLA (Gan et al. 2020)	MLM	MRC	ITM	4M	—
VL-T5 (Cho et al. 2021)	MLM+PLM	Gr	ITM+VQA	200k	384 2080 Ti h
ViLT (W. Kim et al. 2021)	MLM	—	ITM+WRA	4M	—
Align (Jia et al. 2021)	—	—	CL	1.8B	—
MDETR (Kamath et al. 2021)	—	BBox	CL	200k	5.4k V100 h
SOHO (Zhicheng Huang, Z. Zeng, Y. Huang, et al. 2021)	MLM	MPC	ITM	200k	—
VLP-IMF (Xue et al. 2021)	MLM	MFR	ITM	200k	—
CLIP (Radford, J. W. Kim, et al. 2021)	—	—	CL	400 M	74k V100 h
CLIP-ViL (Shen et al. 2022)	MLM	—	ITM/VQA	200k	960 A100 h
ALBEF (J. Li, R. Selvaraju, et al. 2021)	MLM	—	ITM/CL	4M / 14M	—
FLAVA (Singh, R. Hu, et al. 2021)	MLM	MPC	CL/ITM	68M	
TCL (J. Yang et al. 2022)	MLM+CL	CL	ITM/CL	4M / 15M	—
SimVLM (Zirui Wang et al. 2022)	PLM	—	—	1.8B	—
VLMO (Bao, W. Wang, L. Dong, Q. Liu, et al. 2021)	MLM	—	ITM+CL	4.1M	L: 9.2k V100 h
METER (Dou, Xu, et al. 2022)	MLM	MPC	ITM	4.1M	1.5k A100 h
X-VLM (Y. Zeng et al. 2021)	MLM	BBox	ITM+CL	3.8M / 15M	768 A100 h
VLC (Gui et al. 2022)	MLM	MAE	ITM	4M / 5.6M	—
VL-BEIT (Bao, W. Wang, L. Dong, and F. Wei 2022)	MLM	MPC	—	4M	—
Florence (Yuan et al. 2021)	MLM	Object RL	ITM/CL	900M	123k A100 h
CoCa (J. Yu et al. 2022)	PLM	—	CL	1.5B	246k TPU h
Flamingo (Alayrac et al. 2022)	PLM	—	—	2B+	553k TPU h
BLIP (J. Li, D. Li, Xiong, et al. 2022)	PLM	—	ITM+CL	14M / 129M	—
LEMON (X. Hu et al. 2021)	PLM	—	—	200M	—
UniTab (Zhengyuan Yang et al. 2021)	PLM	BBox	Multitask	200k	—
OFA (P. Wang et al. 2022)	MLM	MFR+OD	ITM+others	15M	—
FIBER (Dou, Kamath, et al. 2022)	MLM	—	ITM+CL	4M	540 V100 h
KOSMOS (S. Huang et al. 2023)	—	—	MNTP	3 B+	—

Table 1.3. – Details of the pre-training protocol of Vision-Language models, relating to pre-training tasks (textual, visual and multimodal), dataset size and training time. Dataset size corresponds to the number of distinct images in the pre-training dataset (or a close approximation). When two figures are given, they correspond to two pre-training configurations (small/large). Pre-training duration is reported in GPU hours. [b]

a. L refers to the large model, similar to BERT-Large

b. MRR refers to Masked Region Regression, MRC refers to Masked Region Classification, MTC refers to Masked Tag Classification, BBox refers to Bounding Box prediction, Gr refers to Grounding, RL refers to Representation Learning, MAE refers to Masked Auto Encoder, MPC refers to Masked Patch Classification, OD refers to Object Detection, SGP refers to Scene Graph Prediction, MFR to Masked Feature Regression, and MNTP to Multimodal Next Token Prediction.

pre-training choices on the performances of a model, as several factors can impact performance in various ways. One point that seems to stand out among those factors

is the importance dataset size, and to a lesser extent, the number of parameters. We illustrate this through a scatter plot of the accuracy of vision-language models on VQAv2 (test-dev) in relation to the number of parameters of the model and the number of images of the dataset (Figure 1.11). To assess the individual impact of those factors, it is important to carefully chose the evaluation methodologies and tasks.

1.4. Evaluating Vision-Language Transformers

Evaluating the performance of a deep learning model is a complex issue to solve, especially for models with various possible applications requiring different capabilities. Faced with this problem, Natural Language Processing researchers have developed multiple ways to evaluate language models. First, they evaluate language models using metrics such as perplexity, which evaluates the ability of a model to predict a text sequence. However, such a metric does not exist in vision-language multimodality. Moreover, perplexity has flaws, and does not always reflect the performance of a model on a downstream task. To that end, linguistic multitask benchmarks such as GLUE (A. Wang, Singh, et al. 2019) have been developed, to test a language models on varied aspects of Natural Language Processing. The goal is to test the 'general-purpose' language understanding of a model. This helps researchers estimate if a pre-trained model is appropriate for a specific task or domain.

However, as this method usually requires fine-tuning after pre-training, it does not directly reflect the performance of pre-training. In addition, results on a downstream task can be difficult to interpret, as they are greatly influenced by factors such as bias in the evaluation dataset. As a result, other ways to evaluate language models have been implemented. We introduce in this section the most commonly used tasks. In chapter 3, we study different evaluation methods, and introduce the concept of BERTology and its application to vision-language multimodality.

Downstream tasks are also used to evaluate vision-Language models. They reflect multiple possible applications related to vision-language multimodality, such as reasoning and retrieval, in order to test general vision-language understanding. Figure 1.12 shows an example of evaluation on selected fine-tuning tasks, from the CoCa paper (J. Yu et al. 2022).

To evaluate a pre-trained vision-language model on a specific task, it must first be fine-tuned on this task. This means that the weights of the model are updated to reflect a possible domain change or to prepare for the new task. However, this method of evaluation also has limits. Indeed, language models have reached better results than human performance on some evaluation tasks, while their true capabilities to understand language remain limited compared to humans. This can be due to bias in the dataset. This is why the evaluation benchmark GLUE has been improved to reflect the gap between human-level and computer-level comprehension, leading to SuperGLUE (A. Wang, Pruksachatkun, et al. 2019).

While there is no benchmark for vision-language pre-trained models similar to GLUE, several downstream tasks are commonly used to evaluate different aspects of

vision-language models. However, not all models are evaluated on the same tasks, as some of them are oriented towards specific aspects of vision-language multimodality, such as retrieval or object detection. We give here a brief introduction to various vision-language downstream tasks, which are used to compare the reasoning, retrieval and generative capabilities of vision-language models.

	VQA		SNLI-VE		NLVR2	
Model	test-dev	test-std	dev	test	dev	test-p
UNITER [26]	73.8	74.0	79.4	79.4	79.1	80.0
VinVL [27]	76.6	76.6	-	-	82.7	84.0
CLIP-ViL [73]	76.5	76.7	80.6	80.2	-	-
ALBEF [36]	75.8	76.0	80.8	80.9	82.6	83.1
BLIP [37]	78.3	78.3	-	-	82.2	82.2
OFA [17]	79.9	80.0	90.3†	90.2†	-	-
VLMo [30]	79.9	80.0	-	-	85.6	86.9
SimVLM [16]	80.0	80.3	86.2	86.3	84.5	85.2
Florence [14]	80.2	80.4	-	-	-	-
METER [74]	80.3	80.5	-	-	-	-
CoCa	**82.3**	**82.3**	**87.0**	**87.1**	**86.1**	**87.0**

Table 6: Multimodal understanding results comparing vision-language pretraining methods. †OFA uses both image and text premises as inputs while other models utilize the image only.

Figure 1.12. – Results of state-of the art models on selected tasks, from (J. Yu et al. 2022).

Visual Reasoning There exist several types of visual reasoning tasks.

— Natural Language Visual Reasoning (NLVR) (Suhr et al. 2019) is a very commonly used reasoning task. This task consists in giving as input two photographs and an affirmation to a vision-language model. The model must classify this affirmation as correct or incorrect using both images. Contrary to other evaluation tasks, the use of two images can require some adaptation, which can differ depending on the fine-tuning protocol. Figure 1.13 shows a positive instance of the NLVR2 dataset.

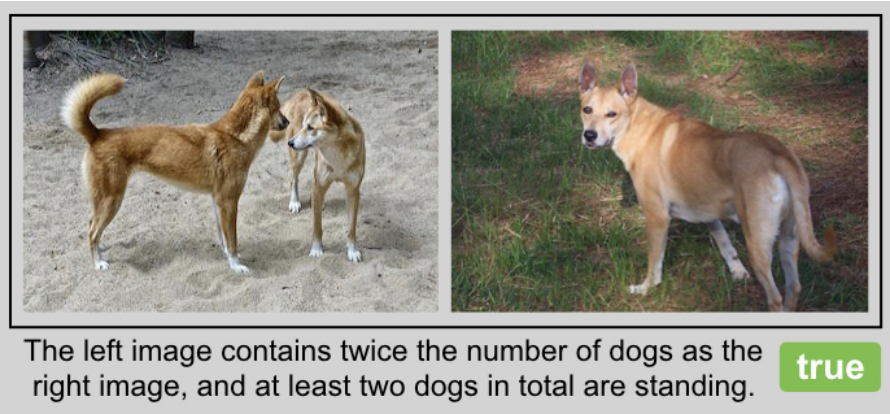


Figure 1.13. – Example from the NLVR2 dataset (Suhr et al. 2019)

— Visual Entailment (SNLI-VE) (N. Xie et al. 2019) is a fine-grained multimodal reasoning task. The model is given a text hypothesis and an image, and must classify the pair into one of three classes: Entailment, Neutral, Contradiction. However, as the first dataset has noisy labels, another one has been proposed with some corrections (Do et al. 2020).
— Visual Spatial Reasoning (F. Liu, Emerson, et al. 2022): A recent task has been introduced to evaluate a model's ability to apprehend relational information between objects. This is a classification task where the model must assess whether a sentence correctly describes the spatial relationships present in the image.
— Binary Image Selection (Hexiang Hu et al. 2019) is a task that consists in asking a model to select the correct image with respect to a sentence among a pair of semantically similar images.

Visual Question Answering VQA consists in giving a model an image and an associated question. The model predicts an answer for this question. It can be treated as a classification task, with a dictionary of possible answers, or as a generative task, where a model generates an answer. There exist several Visual Question Answering datasets. The most commonly used datasets for the Visual Question answering task are VQA (Antol et al. 2015) and its later iteration VQA v2 (Y. Goyal et al. 2017), which aims to increase the reliance of the models on visual clues. However, studies have shown that the dataset suffers from bias as the distribution between training and test data is the same. As a result, a study has created a new split of the dataset (A. Agrawal et al. 2018). The use of an out of distribution test sets prevents the models from relying on bias. A new dataset (Hudson et al. 2019) has also been proposed to avoid the over-reliance on bias, and with further metrics to understand the failures of the models. Another dataset (Bogin et al. 2021) focuses on the compositionality of the image, to assess whether a model has more complex reasoning abilities, for example aggregation and quantification.

It can be difficult to compare different models accurately on these tasks, as some models (Y.-C. Chen et al. 2019; Tan et al. 2019) use data from other visual question answering datasets such as Visual Genome (Krishna, Yuke Zhu, et al. 2016), which makes the model comparison not always fair.

Visual Commonsense Reasoning The goal of this task (Zellers et al. 2019) is to evaluate the ability of a model to understand common sense knowledge and use it to reason over images and language. This task is a multiple-choice question where the model needs to choose the justification of an answer. However, models sometimes use the word co-occurrences between question and answer more than visual clues (Ye et al. 2021). Indeed, syntax changes that do not impact semantics can cause severe drops in performance.

Vision Language Navigation Vision-Language Navigation is a main subfield of vision-language multimodal tasks. The goal of this task is a direct application of

vision-language models to robotics. A recent survey (Jing Gu et al. 2022) introduces the key methods and datasets of this field.

Visual Grounding Similar to the Natural Language Processing task *referring expressions*, visual grounding determines the part of the image that relates to a specific text portion. This can be done using an object detector model to get the possible bounding boxes, or by generating bounding box coordinates. Several datasets are available for this task (L. Yu et al. 2016; Kazemzadeh et al. 2014). However, Akula et al. 2020 show models do not always need a deep understanding of linguistic structures to perform well on this task. They develop a new dataset for this task. Its goal is to make it necessary for the model to understand linguistic structure to achieve good results. Following on this idea that the most commonly used referring task does not correctly evaluate visual reasoning, Z. Chen et al. 2020 create another dataset for this task, using visual distractors and taking into account the compositionality of the image. A Flickr phrase grounding dataset (Plummer, Liwei Wang, et al. 2015) is also available, to ground the mentions of an entity in an image. Several other variations of this task exist, such as grounded action recognition (Pratt et al. 2020).

Image-Text Retrieval The two types of vision-language retrieval tasks are image-to-text retrieval and text-to-image retrieval. They both are classification tasks. The models need to match one text (respectively image) among several, to an image (respectively text). There is not always one ground truth for the image, as multiple texts can correspond to a single image. Models can have different ways of achieving this task, either by treating it as multiple image-text matching tasks, by computing similarity metrics, or using a combination of both.

There are multiple datasets available for this task, such as COCO (T.-Y. Lin et al. 2014) and Flickr30k (Young et al. 2014a). The models can be fine-tuned, or the task can be treated as zero-shot or few-shot retrieval. The scores usually show whether a model manages to retrieve the correct instance among its 1st, 5th and 10th most likely instances.

Image Captioning This task is a generative task, contrary to the previous classification tasks. As a result, models which were pre-trained without generative pre-training tasks must be fine-tuned on a generative task. There are two main datasets used for this task: COCO (T.-Y. Lin et al. 2014), which is widely used, and NoCaps (H. Agrawal et al. 2019), which specifically tests a model's ability to describe unseen objects. A common evaluation metric for image captioning is CIDEr Vedantam et al. 2015. The metric compares the generated sentences to human-made references. In particular, it compares n-grams between the different sentences.

Object Detection Some models are also evaluated on object detection tasks, such as COCO (T.-Y. Lin et al. 2014), to prove that they generalize to many applications. Object detection is evaluated by comparing the detected bounding box and the gold

bounding box. Indeed, Intersection over Union (IoU) compares the area of overlap of the bounding boxes with the area of union. A threshold is established to differentiate between correct and incorrect predictions. The average precision (AP) corresponds to the area under the curve of the precision-recall curve. From then, the mAP, or mean average precision, is computed by averaging the average precision over all object classes.

Discussion Fine-tuning tasks are helpful to compare the performances of vision-language models. Indeed, their difficulty is usually appropriate to compare vision-language models on difficult tasks that remain easily understandable for a human. They can also be used as few-shot evaluation tasks to evaluate large vision-language models. This is the case of KOSMOS-1 (S. Huang et al. 2023) and KOSMOS-2 (Z. Peng et al. 2023), which are evaluated in a few-shot manner. However, building such a task presents its difficulties, as the reliance on textual bias can distort the result of the models. In addition, it can be hard to interpret them to point out potential strengths or weaknesses of the models, due to the broad and complex nature of those tasks. Moreover, the evaluation protocols can vary, in terms of dataset and training time used for fine-tuning. This can make it hard to compare the pre-training of models.

1.5. Currently Investigated Challenges

Several directions continue to be explored in the growing field of vision-language transformers. Below are a few questions that have emerged, which are not exhaustive. These research directions also apply to other multimodal models, not only vision-language transformers. For instance, models based on speech and video may face similar challenges.

Multilingual vision-language models While multilingual vision-language datasets are lacking, the use of video transcript written in various languages (P.-Y. Huang et al. 2021) has been explored. Indeed, they can be used to create multilingual multimodal datasets. The use of a translation-based language-only pre-training task has also been shown to improve multilingual multimodal performances (Jain et al. 2021). The use of larger models and datasets could lead to the development of very large multilingual models, in a similar trend to that observed for language-only models. However, the noisy image-text datasets available for multilingual models can also hinder their pre-training.

Compressing vision language models In an opposite trend to that of very large models, some researchers have also sought to compress models while maintaining their performances. Indeed, vision-language transformer models are large and resource consuming, which can be detrimental to many downstream applications. Fang et al. 2021 introduce distillation techniques in both pre-training and fine-tuning to compress a large vision-language model. As vision-language models continue to grow

in scale, it is important to conduct research into various methods for compressing vision-language models.

Modularity One problem of multimodal transformers is that the applications are very diverse, and a pre-trained model optimized for a specific application can perform poorly on another. For example, models are rarely optimized for both generation and classification tasks. Some researchers have introduced modularity into their models, reflecting the use of adapters in natural language processing. A way to accomplish this is to use multiple modules for each possible modality during pre-training and select relevant modules for a specific task. J. Li, D. Li, Xiong, et al. 2022 tackle the problem of optimizing for both generation and classification. They use an architecture composed of encoder-only and encoder-decoder modules to pre-train a model. The for the model to be adaptable to a more varied range of applications.

Knowledge injection While vision-language models continue to improve, a remaining difficulty is how to compensate for knowledge that is not present in the pre-training dataset. Many vision-language multimodal tasks rely on common sense knowledge or knowledge about the world that is difficult to obtain using only still images and captions. To overcome this difficulty, models have tried to inject knowledge from outside the pre-training dataset into models using different methods. In Shevchenko et al. 2021, authors use an additional task to inject this knowledge. This task consists in using outside knowledge bases to align visual-linguistic embeddings with embeddings of entities extracted from a knowledge base. Further works could include outside sources of information could expand the possible applications of vision-language models.

Prompts The use of prompts can also help models improve performances on specific tasks without requiring additional fine-tuning. First introduced in Natural Language Processing (T. Shin et al. 2020), the idea is to use minimal data after pre-training to improve results for zero-shot downstream tasks. In K. Zhou et al. 2022, the authors introduce an efficient way to adapt prompt learning for vision-language models. Another use of prompt learning could be domain adaptation. Indeed, vision-language transformers have difficulty generalizing to other domains outside their pre-training. This is especially the case of CLIP (Radford, J. W. Kim, et al. 2021), which has been pre-trained on a very large dataset but shows poor performance on unseen classes. C. Zhao et al. 2022 propose a method to generate domain invariant prompt. Their goal is to improve prompts for a better generalization of CLIP to other domains.

Other approaches to multimodality Indeed, vision and language are not the only modalities to have benefited from the Transformer architecture. Models have also been developed to process sound (Verma et al. 2021) and video (Lei et al. 2021; Linjie Li et al. 2020), among other modalities. This has led to the development of other transformer-based multimodal models. Some of those models have been built in a

modular design, so that modalities could be added or removed depending on the application. Their goal is to pre-train different modalities and be able to use them in a diverse range of downstream tasks. One can imagine that multimodal models using more modalities than only vision and language would help learn better multimodal representations. However, this requires resources in additional modalities. Video data, which can combine text, audio, and vision, may help lead the way towards such models.

While our study focuses on vision-language transformers that produce multimodal vision-language representations, there exist other sorts of state-of-the-art vision-language models. Mainly, text-to-image generative models have reached great results in recent years with the introduction of DALL-E (Ramesh et al. 2021), Imagen (Saharia et al. 2022) and Stable Diffusion (Rombach et al. 2022). While those models are intrinsically different, progress in vision-language research is linked, and better vision-language transformer models could help evaluate and produce data for text-to-image models.

1.6. Conclusion

Vision-language transformer models have been developed to learn general text-image multimodal representations. Those can be applied to a wide range of tasks, as shown in Figure 1.14.

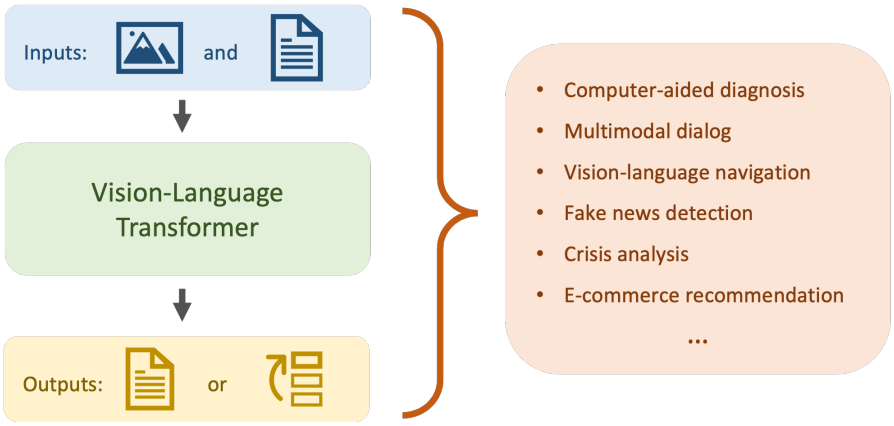


Figure 1.14. – Potential use-cases of a vision-language transformer model

Some of those tasks are detailed in Chapter 2. Through this chapter, we have undertaken a survey of state-of-the-art vision-language transformers, as they are at the time of writing. Since this field is quickly evolving, this survey may not be complete. Figure 1.15 summarizes important aspects of vision-language transformers.

Architecture of vision-language models Following the progress in the application of the transformer architecture to the visual modality, vision-language models

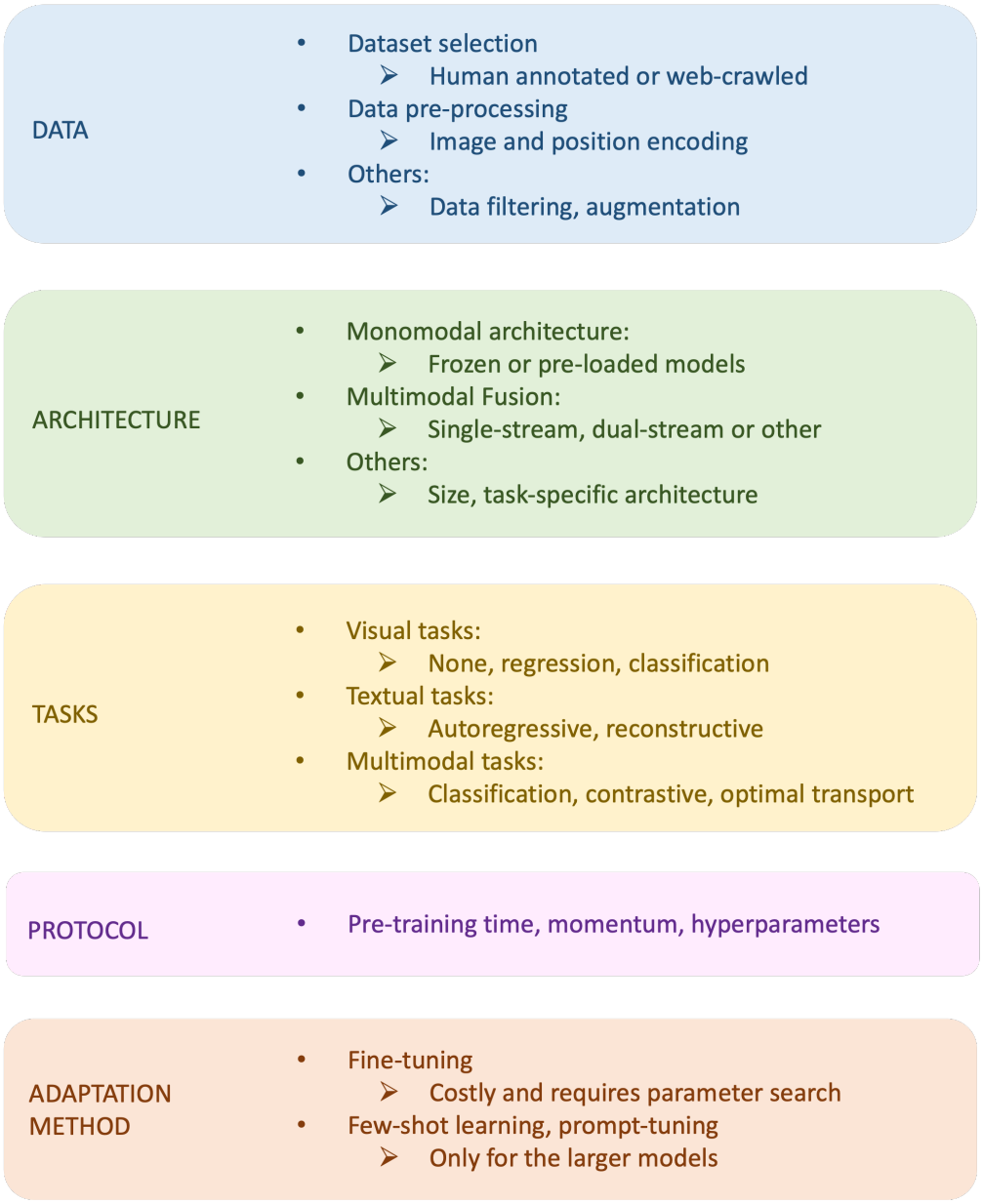


Figure 1.15. – Overview of vision-language pre-training choices

have used a transformer-only architecture instead of relying on object detectors. Additionally, a new trend has developed to use visual information as a complement of Large Language Models, by freezing some layers of language models or incorporating visual information with the textual input. However, there is no consensus yet on which architecture is more efficient for vision-language pre-training. Some studies done on models with an equal number of parameters and with the same pre-training protocol have shown no significant difference between the architectures. It is all the more difficult to draw any conclusion regarding the architecture that most state-of-the-art models do not share the same pre-training protocol, datasets, or tasks.

Pre-training datasets of vision-language models The datasets on which vision-language models are pre-trained can be considered a limiting factor. Indeed, gathering a text-image dataset of good quality is expensive and time-consuming. Due to the cost of human-annotated datasets, the current trend is to use automatically annotated datasets crawled from the web. However, this can raise several ethical problems, in particular related to privacy. In addition, it has been shown that the quality of pre-training datasets significantly impacts the performance of a pre-trained model. As a result, using minimally filtered automatically annotated datasets for pre-training may be inefficient.

Pre-training tasks of vision-language models Several tasks have been developed to pre-train vision-language models, inspired from existing tasks of monomodal models. While the commonly chosen language and multimodal tasks, MLM, PLM, ITM and Contrastive Learning, significantly improve the results of vision-language models, it is not the case of visual tasks. Indeed, no consensus has emerged for visual pre-training yet.

Evaluation of vision-language transformers Multiple vision-language tasks have been created to evaluate the multimodal performance of vision-language models, from vision-language reasoning to retrieval. Studies have shown that some of those datasets exhibit bias. This prevents them from accurately evaluating the performance of vision-language models. In addition, those tasks evaluate broad understanding, but rarely focus on the performance of vision-language models on precise capabilities, making it hard to analyze and interpret results in terms of the abilities of a model.

These observations regarding vision-language transformers raise several questions that we attempt to answer in the rest of this work.

First, vision-language transformer models have greatly increased in performance during the last few years. However, while fine-tuning tasks enable us to compare different models on a general level, they are rarely granular enough to help get a deeper analysis of those models. In particular, they do not highlight the potential weaknesses of vision-language models regarding their capabilities. For instance, a

model might be able to associate words describing object categories to corresponding visual information, but have difficulty doing the same for words related to size. Yet, to use a model in real-world applications, it is essential to avoid blind spots in a model's capabilities, or at least to be aware of them.

In this work, we hope to start answering the following question: What methodology should we use to evaluate vision-language models aimed at a wide range of real-world applications? We argue that the evaluation of general-purpose vision-language models should rely first on their evaluations on granular capabilities. This would help identify potential weaknesses and apprehend the scope of those models. To our knowledge, there has been no attempt to list the vision-language capabilities that can be used to characterize the performance of a deep learning model. Thus, in Chapter 2, we propose a first version of a taxonomy of vision-language capabilities. The goal of this taxonomy is to help build an evaluation of vision-language models as exhaustive as possible.

Then, we are interested in building appropriate tasks to evaluate such capabilities. At the start of this thesis, few tasks had been proposed to evaluate specific capabilities of vision-language models. As a result, there was not much hindsight on the inner workings of those models, in particular in terms of their potential weaknesses. In addition, the increasing development of new vision-language transformers, with different architectures and pre-training protocols, raised the question of how those methods compared to each other on granular capabilities. For instance, a model might have a better than another on some capabilities. Indeed, results of state-of-the-art models on fine-tuning tasks are not sufficient to draw conclusions regarding how architecture or protocol choices affect a model's performance on a specific capability.

In Chapter 3, we aim to answer the following questions: How do state-of-the-art models compare on several specific visual, textual and multimodal capabilities? What are their strengths or potential limiting factors? To that end, we get inspiration from Bertology methods and build several probing and evaluation tasks to diagnose state-of-the-art vision-language models.

Finally, our experiments enable us to make several hypotheses regarding the strengths and potential limiting factors of vision-language models. However, as described through this chapter, state-of-the-art models are pre-trained on different protocols, and comparing them is difficult. Indeed, two models often vary in pre-training datasets, architectures, pre-training tasks and other parameters, such as pre-training length. Thus, it is difficult to interpret the results, and to pinpoint which parameter has the most impact.

In Chapter 4, we aim to answer the following question: How do the various implementation choices impact the capability of this model ? In order to explore more in depth our hypotheses, we pre-train in a vision-language model on several evaluation protocols. We study more precisely several aspects of the pre-training, such as pre-training datasets and tasks.

2. Methodology for a Taxonomy of Vision-Language Capabilities

Table of Contents

2.1. Introduction

As explained in Chapter 1, vision-language models pre-trained to be adapted to many applications are a recent development. They are a result of the application of the Transformer architecture to multimodal machine learning. As in monomodal machine learning, the use of such models raises the question of how to evaluate them. This question is all the more prevalent with the emergence of large-scale pre-trained models, i.e., foundation models.

Indeed, the development of foundation models in the last few years has enabled new state-of-the-art performances across many tasks in computer vision and Natural Language Processing (NLP) (Radford, J. Wu, et al. 2019; Yuan et al. 2021). Yet, monomodal models have shown to be limited in their ability to perform real-world tasks (Emily M. Bender et al. 2021), as they are not sufficiently grounded in real-world experiences to be able to grasp multimodal concepts. Multimodality can be considered as an effective approach to ground models and reach a better understanding of human semantics. This has resulted in a growing focus on multimodal foundation models. In this chapter, we specifically consider vision-language models, which use visual and textual inputs (Tan et al. 2019; Y.-C. Chen et al. 2019; W. Kim et al. 2021; S. Huang et al. 2023; W. Wang et al. 2023; Alayrac et al. 2022; J. Li, D. Li, Savarese, et al.

2023). These models have been tested on a wide range of tasks, from image-to-text generation to cross-modal retrieval or classification. Yet, recent work has brought to light weaknesses in their understanding of multimodal concepts, i.e., concepts that cannot be captured by a single modality. For instance, vision-language models have a limited multimodal understanding of position (Rösch et al. 2022a; Salin et al. 2022), vision-language dependencies (Nikolaus, Salin, et al. 2022) and word order (Thrush et al. 2022). This has prompted the creation of dedicated evaluation tasks to assess those capabilities (Yuksekgonul et al. 2022; Z. Ma et al. 2023). Although benchmarks have also attempted to consider a wider spectrum of vision-language capabilities (Parcalabescu, Cafagna, et al. 2021; Z. Ma et al. 2023), no attempt has been made to provide an exhaustive evaluation of those models.

Drawing inspiration from the work that has been accomplished in the monomodal model evaluation, we aim at starting a discussion on the evaluation of vision-language foundation models. Our goal is primarily to reach a better explainability of models aimed at real-world applications. While other important aspects that should be considered when evaluating a model, such as environmental and societal impact, they are not the focus of this work. General-purpose models are notoriously more difficult to evaluate than task-specific models. Indeed, the latter can be reliably evaluated on one specific task. Foundation models, on the other hand, are applicable to many tasks and domains. Thus, they must be evaluated on their whole scope of application. While researchers have developed benchmarks committed to a comprehensive evaluation of monomodal foundation models (A. Wang, Singh, et al. 2019; Liang et al. 2022; Zhai et al. 2019), to our knowledge, there has been no such proposal in the case of vision-language models. In the Section 2.2, we describe we compare evaluation methods of foundation models, and describe current evaluation methods of multimodal models. In this work, we are specifically interested in the evaluation of *General Multimodal Understanding*, i.e., the capability of a model to extract textual and visual information to understand multimodal concepts, with good generalization ability. To that end, we argue that it is essential to assess the performance of those models on a wide range of specific capabilities. Those capabilities should cover various multimodal concepts, and especially those that are necessary for real-world applications. As such, Section 2.3 proposes a methodology to determine vision-language capabilities using real-world application as a basis. In Section 2.4, we make a first attempt at a taxonomy of vision-language capabilities for the evaluation of multimodal models. Figure 2.1 shows a summary of the proposed taxonomy. Finally, Section 2.6 presents the limits of this taxonomy. In this chapter, we report work from the forthcoming paper: Salin, E., Ayache, S. & Favre, B. Towards an Exhaustive Evaluation of Vision-Language Foundation Models, In ICCV Workshops 2023. This chapter aims at being a first step towards an exhaustive evaluation of such models.

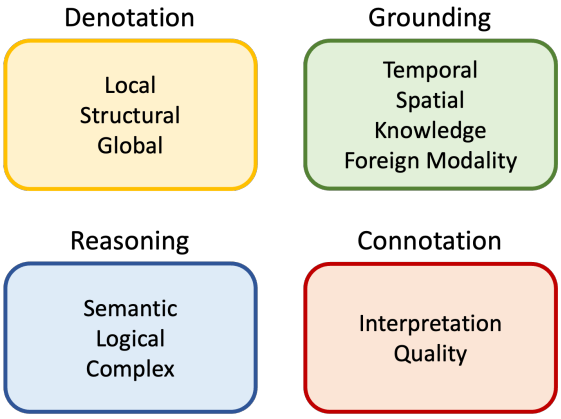


Figure 2.1. – Summary of the suggested taxonomy

2.2. Evaluating Foundation Models

In this work, we consider vision-language foundation models. The goal of vision-language foundation models is to serve as the basis of multiple tasks by learning general representations of texts or images on a large amount of data. The question of foundation model evaluation has still no answer. In particular, researchers can have different goals when evaluating a foundation model. One of those goals can be the comparison to human intelligence. In that respect, it is important to focus on its generalization ability and its capacity to solve previously unseen tasks (Chollet 2019). Yet, the evaluation of a foundation model also aims to reach a better understanding of its precise capabilities possible scope. Indeed, foundation models are being used in real-world environments, where failures can have considerable consequences. Those are more likely to happen if users are unaware of their potential weaknesses, or the extent of their reliability.

2.2.1. Monomodal Foundation Model Evaluation

There have been standardization efforts in the evaluation of general-purpose models in Natural Language Processing and Computer Vision, following the development of multitask models. In this section, we present a non-exhaustive description of those methods. The fast development of language models has led to benchmarks designed to test the multitask abilities of those models. For instance, GLUE (A. Wang, Singh, et al. 2019) and SuperGLUE (A. Wang, Pruksachatkun, et al. 2019) have gathered complex tasks to compare models to human performance. Similar benchmarks have been developed in Computer Vision. For instance, VTAB (Zhai et al. 2019) aims to evaluate representation learning algorithms on a diverse range of 19 tasks (e.g., object counting, location recognition, fine-grained classification, disease classification) in diverse domains.

However, these benchmarks offer limited insight on the explanation of a model's performance. To reach a better understanding of those black box models, new methods

have been developed (Rogers et al. 2021). Among those methods, there has been an emergence of studies evaluating specific skills using probing tasks or other evaluation methods (Conneau et al. 2018; Raghu et al. 2021). These have been established to understand how a model extracts information. From such works, Bertology has emerged to understand BERT-related models, i.e., pre-trained language models based on the transformer architecture. Pre-training transformers is resource-consuming, and the way they operate is opaque. Thus, ablation studies are insufficient to understand their inner workings, and many questions remain unresolved. Various evaluation and probing techniques were developed to reach a finer understanding of what transformer really learn. In Rogers et al. 2021, authors gather numerous findings relating to the understanding of transformer language models through a review of this field. They coin the term 'Bertology' to describe the evaluation of model capabilities and the study of their inner workings. Thus, they start the movement towards a more systematic study of transformer models. We have described some of those findings in section 1.1.4. In particular, Bertology can be used to explain how the pre-training of a model impacts its predictions and the representations it learns, through the use of evaluations such as probing tasks.

Yet, probing tasks have also shown that they can lack in robustness, being highly dependent on syntactic variations (Ravichander et al. 2020). This has led to the development of methods to stress test NLP models such as Checklist (Ribeiro et al. 2020) or HELM (Liang et al. 2022) with regard to robustness, but also bias and fairness. Similar studies have also tested the robustness and bias of models learning visual representations (Hendrycks et al. 2019; Zeyu Wang et al. 2020).

With the emergence of foundation models, the question of evaluation methods shifted from fine-tuning to few-shot evaluations on many tasks, which is less resource consuming. For instance, Y. Wang et al. 2022 develop 1600 few-shot evaluation tasks for generative language models. While some studies focus on gathering numerous evaluation tasks (L. Gao et al. 2022), others have chosen to evaluate those models on human examinations rather than machine learning benchmarks (Zhong et al. 2023). For the visual modality, Florence (Yuan et al. 2021) and CLIP (Radford, J. W. Kim, et al. 2021) authors also use a wide range of visual and vision-language tasks and datasets to assess their models. Some methods tackle the evaluation problem from a capability-centric perspective (Srivastava et al. 2022), or attempt to build a taxonomy for the evaluation of language foundation models (Liang et al. 2022). This enables a more precise explanation of their performances. However, building a comprehensive evaluation benchmark is complicated, due to the variety of possible applications. As a solution, authors rely on existing work in the field (Liang et al. 2022). Thus, it is not aimed to be frozen but to evolve with the inclusion of new applications (Srivastava et al. 2022).

Other difficulties impact the evaluation of foundation models. First, the metrics used to evaluate those models are not always appropriate, especially in the case of generative models, either for texts (T. He et al. 2022) or for images (Borji 2019). The use of human evaluation enables researchers to avoid the flaws of metrics, but lack in standardization ability. In addition, the evaluation of foundation models relies on data

dependent on bias and subjectivity (Lamiroy 2014). The use of appropriate datasets and metrics to evaluate on a task and the development of exhaustive evaluation methods are decisive to diagnose and analyze foundation models.

2.2.2. Vision-Language Foundation Model Evaluation

It can be difficult to assess the understanding of a multimodal model. Indeed, those models rely on spurious correlations, and may use information taken from one modality without relying on the other. This has been shown in vision-language models, where visual information can be ignored in favor of textual bias (Y. Goyal et al. 2017). Therefore, to trust a vision-language model's performance in a real-world application, it is important to be aware of what concept this model is able to understand at a multimodal level. In recent years, several benchmarks have been developed (W. Zhou et al. 2022; Bugliarello, F. Liu, et al. 2022) to evaluate vision-language models. Beyond fine-tuning, introduced in Chapter 1, several other methods can be used to adapt and evaluate vision-language transformer models. Figure 2.2 illustrates some of those methods.

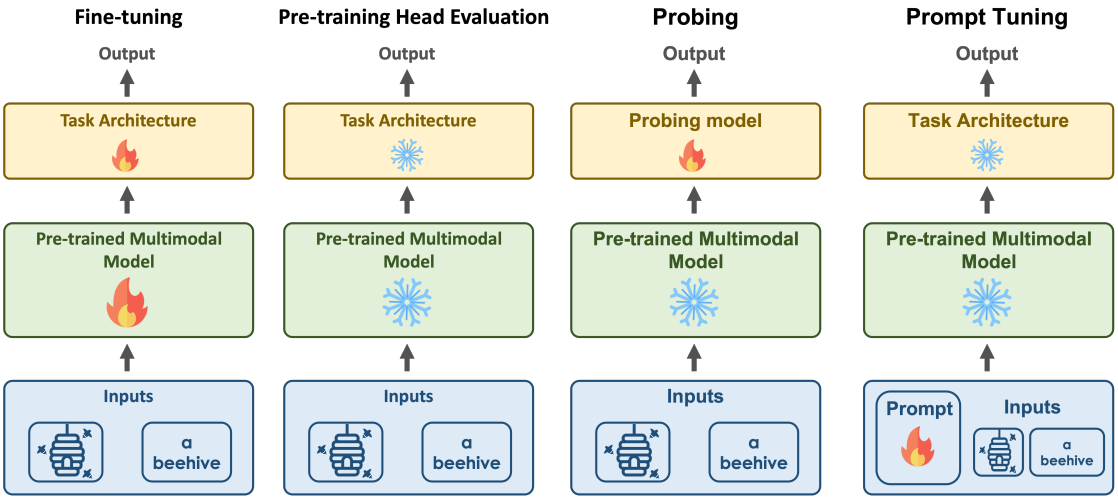


Figure 2.2. – Different methods of evaluation for Vision-Language Transformers

Those methods can be used to evaluate different aspects of a model. Fine-tuning is usually used to compare different models on complex tasks, and can show if a model has significantly improved upon previous state-of-the-art. On the other hand, vision-language evaluation has also been inspired by Bertology methods introduced in the previous section. For instance, some methods specifically evaluate whether a pre-trained model understands a specific multimodal concept. We describe in this section different evaluation methods applied to vision-language model, beyond fine-tuning. In particular, probing and pre-training head evaluation are appropriate methods to evaluate a specific capability. We also give an overview of studies of vision-language transformers, such as the study of attention weights, which can give us insight on those models.

Pre-training head evaluation This method consists in using the task-specific heads used during pre-training to evaluate a pre-trained model. This type of evaluation aims at evaluating a precise capability of a model. To that end, a dataset is specifically created by selecting relevant examples. For instance, language models can be directly evaluated using 'fill in the blanks' tasks. In this manner, Ettinger 2020 study hypernym and negation understanding by creating appropriate datasets. Several such vision-language tasks have been proposed, most of them, concurrently or after our own work.

— Foil It! (Shekhar et al. 2017) is a dataset designed to test the semantic understanding of object categories, by using highly similar 'foil' captions as negatives, requiring a fine-grained understanding of instances. However, it has been shown that the task can be solved by language-only models using textual bias (Parcalabescu, Cafagna, et al. 2021).

— In Parcalabescu, Gatt, et al. 2021, the authors test the counting abilities of vision-language models, in particular by using a Masked Language Modeling direct evaluation. They find that models do not have a good ability to count and mostly rely on bias in the training data.

— Hendricks and Nematzadeh 2021 rely on probes to study verb understanding in pre-trained transformer-based models. They show that vision-language models have difficulty with fine-grained multimodal understanding. Among subject, object and verb understanding, the latter is the most difficult for models. They also notice that large but noisy datasets are counterproductive compared to smaller manually curated datasets.

— To evaluate vision-language compositionality, Winoground (Thrush et al. 2022) is created using manually selected examples. The method consists in selecting two image-caption pairs, where each caption has the same bag of words in a different order. This ensures that the task evaluates the ability of the model to take into account word order. Results show that none of the models do better than chance. Diwan et al. 2022 further discusses the performances of models on this task, and conclude that the main difficulty of this task is not the ability to understand word order. Indeed, the models also need to understand small details of images and perform complex reasoning, such as spatial reasoning.

— Valse (Parcalabescu, Cafagna, et al. 2021) is a new benchmark designed to test the ability of vision-language models to understand various linguistic and semantic phenomena. This study shows that state-of-the-art vision-language models manage to identify objects in images, but have difficulty to understand other semantic aspects, such as counting and spatial relationships.

— In T. Zhao et al. 2022, the authors develop a set of tasks to evaluate the understanding of objects, attributes, and relations. They alter captions and images to create the dataset.

Probing evaluation The goal of probing is also to provide a better understanding of the capabilities of a model. More specifically, probing consists in studying what information is encoded in the representations learned by a model. Probing tasks

have been first developed to analyze language models through benchmarks such as SentEval (Conneau et al. 2018). For example, Hewitt and Manning 2019 show that syntactic parse trees can be inferred from ELMO and BERT representations. Although explainability has been largely explored in vision models, the use of probing tasks is more limited. Recently, Basaj et al. 2021 have developed a visual probing framework by constructing visual equivalents to words based on superpixels. It then translated language probing tasks such as sentence length and semantic odd man out to the vision modality.

Few studies have been conducted using probing models on vision-language transformers. Those were introduced concurrently or after our work:

— Lindström et al. 2021 probe the vision-language non-transformer models on their abilities to understand object categories and to count objects in an image. They specifically study Visual Semantic embeddings, and notice the importance of linguistic information in multimodal tasks. However, they do not analyze transformer-based models in their study.

— In Rösch et al. 2022b, vision-language models are probed to evaluate how much positional information is extracted through those models. They find that positional information is extracted, but the model does not learn to rely on it in downstream tasks.

Studying a Model's Internal Weights Beyond studying specific capabilities, other methods can be used to gain more understanding of vision-language models, such as the study of its internal weights. For transformer models, this can provide clues as to the relationships between different tokens of an input sequence. In J. Cao et al. 2020, the authors analyze the internal weights of some vision-language transformers, to study how the transformer architecture impacts the learning process of single-stream and dual stream models. They observe the role of each layer, and the fusion of the vision and language modalities across layers. They notice the prevalence of the language modality over the visual modality, which we study further in this chapter, by evaluating the multimodal nature of representations. L. H. Li et al. 2020 analyze attention heads and conclude that some of them can ground language tokens to their corresponding image regions. On a model fine-tuned for VQA, Sikarwar et al. 2022 find that words in the questions, and specifically nouns, are what drives the visual attention of cross-modal attention layers. In Ilinykh et al. 2021, the authors study attention weights of captioning models. They find that models based on an object detector form local dependencies in lower layers. Indeed, object features from a same 'thematic cluster', which are usually closer together, attend to each other. Furthermore, objects that are referenced in text are more attended to in later layers. Xue et al. 2021 build a metric to quantify the interaction between modalities in a transformer-based model. Yet, previous work has shown that the analysis of transformer weights may not be a reliable way to explain the predictions of transformer models. However, the study of internal weights is not limited to transformer models. In computer vision, explainability methods use CNN gradients in order to explain the predictions of a model (R. R. Selvaraju et al. 2017). While this field is still relatively recent, it is especially

relevant for models aimed at real-world use. Indeed, it could help to understand the predictions of a model on a case-by-case basis.

Studying the Balance between Modalities A vision-language transformer model is pre-trained using inputs from two modalities: images and texts. The relationship between inputs of different modalities can be complex. In a recently proposed taxonomy of image-text relationships (Otto et al. 2019), authors argue that these relationships can be classified into eight categories. They propose three metrics to evaluate text-image relationships: Cross-modal mutual information, semantic correlation, and status, which indicates whether a modality is subordinate to another. In the case of equal status between modalities, image and text can be *uncorrelated*, *complementary*, or *contrastive*. If the image is subordinate to the text, it can count as an *illustration*, while the opposite can be an *anchorage*. In the case of generic image-text datasets, the relationship between text and image is *anchorage*. Negative examples can be built to form *uncorrelated* or *contrastive* relationships. A study on vision-language transformers (S. Frank et al. 2021) investigates the relationship between the two modalities, and whether one is subordinate to another. They build a set of diagnostic tasks that they test on early transformer models, and find that they use more visual clues for text prediction than textual clues for visual predictions, seeming to show an unbalanced relationship between modalities. A question that is recurrent is whether one modality should be more reliable than another. Indeed, K. Liu et al. 2018 explore different ways of combining modalities, in particular to avoid noise from a 'weaker' modality.

An important point when studying the role of each modality is the impact of monomodal bias on the predictions made by the model. Indeed, vision-language models can show an overreliance on the textual modality, at the expense of the visual modality (Y. Goyal et al. 2017). As a result, it is important to pay attention to the distribution of each modality when building a dataset. Specifically, the textual modality can be significantly subject to spurious correlations.

Impact of bias Several studies have analyzed harmful bias in computer vision (Steed et al. 2021) and NLP (Blodgett et al. 2020) models. Deep learning models are known to reproduce gender or racial bias present in their training data. Hendricks, K. Burns, et al. 2018 find that vision-language models are also subject to gender bias. F. Liu, Bugliarello, et al. 2021point out that most datasets are heavily biased towards western culture, showing that the good performance of vision-language models may be restricted to a narrow domain of North American or Western European datasets.

Discussion The study of vision-language transformer models is still at its early stages, and various methods are being developed to provide insight into their inner workings. Among them, several methods are aimed at evaluating a specific capability of models. As the field of vision-language multimodality is less mature than those of language only or vision-only machine learning, there is also a lack of hindsight on what issues vision-language foundation models will be facing. To our knowledge, there

has been no attempt at evaluating a broad coverage of vision-language capabilities.

2.3. Methodology

In this chapter, we discuss an exhaustive evaluation of vision-language foundation models, to help point out precise failures in the multimodal understanding of foundation models. With access to such information, users would be able to make an informed decision on the use of a model. To that end, we must first consider the intended use of vision-language foundation models. We argue that the goal of foundation models is to reach a general understanding of the domain over which they operate. In the context of vision-language models, this understanding is visual, textual, and multimodal. However, in this work, we mainly focus on the multimodal understanding of such models.

We aim to get a precise overview of the general multimodal understanding of a vision-language foundation model. To that end, it would be interesting to study its performances on a diverse set of multimodal capabilities, with an extensive coverage of the necessary capabilities for real-world applications. Such methods have indeed proven beneficial in NLP and computer vision to understand the inner workings of large black-box models. Indeed, a more granular evaluation will help to point out limiting factors of vision-language models. Contrary to current works in NLP, we do not focus on specific complex tasks (e.g., retrieval, inference, generation, question answering) but on the capabilities required for multimodal understanding. Indeed, we argue that the goal vision-language evaluation aimed at real-world applications should be to diagnose a model on the range of capabilities necessary for those applications, rather than proving a higher accuracy on a complex task. For instance, this would mean to have more precise insight on their multimodal understanding of some concepts (e.g., position, object interaction). To that end, we propose a taxonomy of vision-language capabilities, to establish the possible range of evaluation for a vision-language model. Indeed, the goal of this taxonomy is to cover a broad range of vision-language capabilities necessary for real-world applications. This would help users and researchers identify potential blind spots of vision-language models. In this section, we explain our methodology for the categorization of vision-language capabilities into the taxonomy, and how we determine granular vision-language capabilities relevant in real-world applications. We encourage the evaluation of foundation models to go from a task-centric perspective to a capability-centric perspective, by creating a list of vision-language capabilities needed for real-world applications.

2.3.1. Categorization

Indeed, multiple types of broad abilities are required when a foundation model performs a vision-language task. The categorization of granular vision-language capabilities into those broad abilities can help identify potential blind spots. To organize those abilities, we draw a parallel with the human understanding. Indeed,

we refer to visual literacy, which studies the human understanding of images, to help us establish different stages of visual literacy for machine learning systems. There is no clear definition of what it means to be visually literate, due to the complex nature of the concept Kedra 2018. Visual literacy is defined by aggregating sets of skills in two main categories: 'denotation' and 'connotation' Bardin 1975. *Denotation* refers to the perception of visual elements in an image, while *Connotation* associates the image with an ideological or affective meaning. However, those abilities are not sufficient to evaluate the capabilities of a model. Indeed, a may fail where humans see no difficulty. As a result, we propose four broad categories of vision-language capabilities, with the following definitions.

Definition 2 (grounding) *Capabilities requiring the use of information that is not directly accessible using the inputs (2D image and text), or the understanding of concepts that cannot be described using those modalities (e.g., time, space, knowledge, sound, mathematical documents).*

Definition 3 (reasoning) *Capabilities requiring the application of abstract thinking or logic to the analysis of an image-text instance.*

Definition 4 (connotation) *Capabilities related to the subjective analysis of a text-image instance, from symbolic interpretation to qualitative evaluation.*

Definition 5 (denotation) *Text explicitly depicts or refers to image elements and does not require grounding, reasoning or evoke connotation.*

2.3.2. Determining Vision-Language Capabilities

In order to build this taxonomy, we must consider the context in which it operates, meaning the current state of the vision-language field. Indeed, the evaluation of vision-language foundation models should be to be appropriate, considering the use cases and challenges of vision-language models. We are inspired by HELM (Liang et al. 2022), which uses conference tracks to assess the coverage of their evaluation. However, vision-language machine learning is less mature than Natural Language Processing, and not all challenges have been identified. By precisely analyzing the context, we can identify relevant vision-language capabilities at a granular level. As a result, before establishing a taxonomy, we must first consider what use such models can have, i.e., their potential real-world applications. These can help us identify needs of those systems, or even dangers related to harmful use of those models.

Vision-Language Applications Since foundation models are aimed at real-world applications, we identify some works from current research that have pinpointed possible application of vision-language machine learning. These could be a use case for vision-language transformer models. There are a growing number of complex applications, with challenges that have not yet been resolved. For instance, there is

an increasing number workshops related to vision-language applications in machine-learning conferences. We list in this section some of those real-world applications based on vision-language data. For each application, we identify one capability necessary for this application that could pose a challenge for vision-language models.

— **Multimodal Dialog** Das, Kottur, et al. 2017: Use textual and visual context for dialog with a user.
Example challenge: Understand the subjective meaning of some instances, such as jokes, memes ({Connotation}).

— **Fake News Detection** Jing et al. 2021: Identify fake news in social media.
Example challenge: Understand the intent behind a specific text-image combination ({Connotation}).

— **Vision-Language Navigation** A. Burns et al. 2022: Understand natural language instructions in a visual environment.
Example challenge: Understand if there is a mismatch between a text command and the available visual information ({Reasoning}).

— **Tools for Visually Impaired People** Zongming Yang et al. 2022: Help a visually impaired person navigate or answer questions on an image.
Example challenge: Precisely describe the structure of a scene ({Denotation}).

— **Crisis/Event Analysis** M. Li et al. 2022: Understand a crisis, the relevant actors and its context based on text-image data.
Example challenge: Understand spatial and temporal context of a text-image instance ({Grounding}).

— **Video Summarization** Plummer, M. Brown, et al. 2017:
Vision-language models can be used in some cases to complement applications based on video. *Example challenge*: Describe visual elements relevant to temporal data, in still images ({Grounding}).

— **Computer-assisted Food Analysis** Shukor et al. 2022: For instance, it can consist in image-text retrieval applied to food, and can have applications in health and nutrition.
Example challenge: Understand the temporal and spatial structure of text-image food or recipe data ({Grounding}).

— **Biomedical Vision-Language Processing** Boecking et al. 2022: Interpreting visual and textual biomedical data for clinical care.
Example challenge: Understand and reason on complex biomedical semantics ({Grounding}& {Reasoning}).

— **Agriculture** Y. Cao et al. 2023: Identify plant disease for agricultural purposes and differentiating between healthy and diseased plants.
Example challenge: fine-grained classification from limited examples ({Denotation}).

— **Autonomous Driving** Marathe et al. 2023: For instance, vision-language models can help design datasets geared towards autonomous driving. Specifically, they can design instances that are not present in sufficient quantity in real datasets.
Example challenge: Semantic understanding of events such as weather, accidents or other incidents ({Denotation}& {Grounding}).

— **E-commerce Recommendation** W. Shin et al. 2022: Product recommendations

based on textual and visual information. There are several possible subtasks such as product matching, classification, clustering.
Example challenge: Associate text to the corresponding semantic information using visual data despite limited grammatical structure ({Denotation}).

— **Multimodal Hate Speech Detection** Y. Chen et al. 2022: Detecting hate speech that is present in multimodal data.
Example challenge: Understanding subjective and ambiguous meaning of text-image data ({Connotation}).

— **Remote Sensing Understanding** Wen et al. 2023: Study of satellite mages in correlation with text data.
Example challenge: Differentiate semantically between atmospheric visual data and relevant ground visual data ({Grounding}).

— **Market Prediction** Wimmer et al. 2023:
Predict the evolution of the stock market using text and image data. *Example challenge*: Identify patterns in time series data represented using text or images ({Grounding}).

A vision-language model aimed at being applied to many tasks and domains would have to be evaluated on challenges linked to those various applications. However, we argue that instead of designing a task for each application, the model should first be evaluated on capabilities related to those challenges. Indeed, some of those challenges are related to the understanding of similar vision-language capabilities. This is for example the case of the challenges identified for Hate Speech Detection, or Fake News Detection, that can both be correlated with detecting the intent of an instance. Thus, they could be tackled together, by evaluating them on this specific capability. Some of those challenges should be tackled as a common goal, and that it should reflect in the evaluation of those models. However, the complex nature of those applications makes it difficult to compile a list of capabilities for model evaluation.

In this section, we propose several methods to determine vision-language capabilities related to a real-world application. We study more precisely several of the identified real-world applications. Our goal is to get as complete a picture as possible of the capabilities involved in those tasks: news captioning, medical VQA (Abacha et al. 2019) and vision-language navigation (Shridhar et al. 2020) to determine associated vision-language capabilities. As observed previously, those applications do not cover the whole range of vision-language multimodality, but they offer insight into different capabilities relevant to multimodality. For each of those applications, we proceed with a method to identify related vision-language capabilities. These methods could then be applied to other vision-language applications to identify capabilities.

Manually studying relevant data: the case of news-related data Vision-language foundation models can be used with news-related data for fake-news detection algorithms. We study the capabilities necessary for such applications from a data-centric perspective: we collect examples and manually identify relevant capabilities. News-related data varies across cultures, periods, and topics of interest. We choose to study examples from selected newspapers to extract different types of

multimodal interaction, as well as capabilities needed for a vision-language system to understand those examples.

In order to get a comprehensive perspective of news data, we select 5 online news sources from several countries and varying demographics. We restrict ourselves to English language newspapers.

— The New York Times, a daily American newspaper [1]
— Daily Mail, a daily British tabloid [2]
— Wall Street Journal, a daily American business newspaper [3]
— France 24, a French international news network [4]
— Al Jazeera, a Qatari international news network [5]
— Global Times, a daily Chinese English-language newspaper. [6]

We select three dates and study a captioned image from those newspapers for each of those dates, selecting a topic at random for each example. These examples vary across topics: ranging from business to culture. The details of those examples are available in Appendix A.

We notice that news images and their captions follow two main different types. Either the image is described by the caption, with possibly a bit of context added by the text, or the image is used as an illustration of the text, and the link between text and image is less direct. Following the vocabulary introduced by (Otto et al. 2019), we call the first text-image relationship *anchorage* and the second situation *illustration*. The examples are evenly split along those two categories. From those examples, we extract several capabilities necessary for a good understanding of the instances:

— Object Recognition {Denotation}: Understand the content of an instance. For instance, in the case of war reporting, it is important to differentiate between systems belonging to two armies.
— Text Understanding {Grounding}-{Reasoning}: Understand written text in an image, and its role with respect to the object it is written on. For instance, texts written on a protest board or a shop window have widely different intents.
— Named Entity Recognition {Grounding}: Link famous people or monuments in an image to the corresponding entity.
— Semantic Role Understanding {Grounding}: Understand the role of both objects and people. For instance, understanding the job of someone with respect to the context.
— Sentiment Understanding {Denotation}-{Grounding}: Understand the stance, gaze, expressions and interaction of a person (or animal) with their environment.
— Structural Understanding {Denotation}: This can relate to the understanding of an image structure (e.g., counting, understanding position). For instance, it can help understand how each element relates to each other (e.g., interaction

1. https://www.nytimes.com/
2. https://www.dailymail.co.uk/
3. https://www.wsj.com/
4. https://www.france24.com/en/
5. https://www.aljazeera.com/en
6. https://www.globaltimes.cn

between people).
— Context Grounding {Grounding}: Identify when the picture was taken, where it was taken, or the event it depicts. This is particularly important in the case of news reports.
— Image Interpretation {Connotation}: Some instances show a discrepancy between text and image, which can help understand the intent of the journalists. For instance, the use of the words 'is investigated' in a caption gives a new meaning to a picture.
— Style understanding {Connotation}: This can relate to the understanding of art or style, and the understanding of iconography.
— Multimodal document understanding {Denotation}: Beyond the image-caption relationship, news-related data is often multimodal at a document level. Thus, the understanding of multimodal documents is essential for a better apprehension of this data. For instance, title, texts, and image descriptions are often complementary.

Relying on existing datasets: the case of medical data Vision-language foundation models can be used as part of multiple real-world applications, as detailed in Section 2.3.2. Those applications often require specific technical knowledge to understand the underlying challenges. To compensate for our lack of technical knowledge, we can rely on existing tasks and datasets to identify relevant capabilities. In this section, we specifically study Computer-Aided Diagnosis systems as an example. The main goal of those systems is to help communication between users. These systems can provide doctors with another tool to reach a medical diagnosis or help communication. Some datasets have already identified relevant problems of vision-language multimodality applied to medical data. The domain of medical data is particularly specific, and the marked difference between instances in the natural domain and medical instances shows that using pre-trained general models for medical applications may not be appropriate. Instead, models are pre-trained specifically for medical purposes (Rasmy et al. 2021). However, we think that understanding the capabilities necessary for downstream medical applications can help design more appropriate architectures and tasks for general-purpose models, which can then be specifically pre-trained for medical applications. To that end, we refer to the question types identified in medical VQA tasks (Abacha et al. 2019).
— Data Collection Context {Denotation}: In medical imaging, data can vary following what is being observed, which machine is used for the observation, the options and angles.
— Object Recognition {Denotation}: Recognize different organs or body parts, and to be able to segment them.
— Semantic Object Understanding {Grounding}-{Reasoning}: Differentiate between 'normal' or 'abnormal' organs. It requires a good understanding of their typical texture, color, size and position.
— Focus Understanding {Denotation}: Understand the main 'abnormality' in an image, which requires the system to understand the focus of a medical instance.

— Knowledge Grounding {Grounding}: Medical technical knowledge is necessary to describe images and differentiate technical terms, as well as a good understanding of medical taxonomy.
— Logical Reasoning {Reasoning}: The system may need to perform logical reasoning to aggregate multiple factors.
— Multi-source understanding {Denotation}-{Reasoning}: Summarize and compare several sources of data.

The use of deep learning in the medical domain is very sensitive, and better explaining the performances of foundation models can help practicians be assured of their trustworthiness. While we focus here on capabilities of models, other aspects should be considered, such as their robustness to noise and meaningless alterations of data. They should also be controlled for bias linked to age, gender, machine used for imaging, and other characteristics.

Relying on extensive research in a field: the case of vision-language navigation Vision-language foundation models can be used to build agents that can interact with their environment using human language and visual information. This is an important field of robotics and embodied artificial intelligence. This field is referred to as vision-language navigation (VLN). To identify relevant vision-language capabilities, we rely on extensive research (Jing Gu et al. 2022) that studied the challenges and problems related to this field. To be able to perform VLN, a system must have a good understanding of:

— Spatial Understanding {Denotation}-{Grounding}: Understand the position of an agent relative to other objects in the scene, as well as the depth and size of other objects. This skill depends on the point of view of the system (ego or 3rd person).
— Space-based Reasoning {Reasoning}: The ability to design a path based on available information.
— Object Recognition {Denotation}: Recognize objects in the scene.
— Object Role Understanding {Grounding}: A model should be able to recognize the role objects, and understand their associated physics. Specifically, some objects can be obstacles, and others can be interacted with.
— Object State Understanding {Grounding}: Recognize the state objects, and the semantic change in those states. For instance, a cup can be empty or full and will not have the same role depending on its state.
— Action Understanding {Grounding}-{Reasoning}: Understand the sequence of actions necessary for a task, and their effect on the environment. For instance, washing something implies changing the state of an object from 'dirty' to 'clean'.
— Structure Understanding {Denotation}: Recognize the structure of a scene and the dependency between objects.
— Intent Understanding {Connotation}: Understand the intent, even in the case of a misalignment between modalities. The model must be able to understand the intent despite this discrepancy.

In addition, the data provided to a VLN system may not always have good quality,

framing or lighting. As a result, it is important for such a system to be robust to different types of data quality, detail, and complexity of images.

Discussion In this section, we study a few diverse applications of vision language systems to determine a set of skills necessary for vision-language systems. In addition to downstream applications, we also rely on previous works in computer vision and NLP (MacCartney 2009; X. V. Lin et al. 2018; Bowman et al. 2015; Zhilin Yang, Qi, et al. 2018; Camburu et al. 2018) to identify relevant capabilities to complete the taxonomy. Due to the breadth of the vision-language field, it is difficult to enumerate all possible vision-language capabilities. Thus, it is difficult to create an exhaustive taxonomy in terms of the coverage of vision-language capabilities. To further this study, a more in-depth look at several other applications detailed in Section 2.3.2 could help provide a more complete understanding of vision-language skills. Before using a vision-language foundation model on a real-world application, we encourage studying this application to uncover relevant vision-language capabilities.

Beyond this methodology, it is important to evaluate the coverage of the taxonomy, and especially of identified vision-language capabilities. We propose in this chapter a method to evaluate the coverage of this taxonomy. Indeed, we project the vision-language capabilities on to existing evaluation tasks. This can help us identify or specify novel vision-language capabilities related to existing tasks, and recognize blind spots that current tasks do not evaluate. We detail the taxonomy in Section 2.4, and provide a projection of the taxonomy into existing evaluation tasks in Section 2.5.

2.4. Taxonomy

In this section, we propose a preliminary attempt at a taxonomy of vision-language capabilities. We supplement the previously determined capabilities (Section 2.3.2) using previous work in NLP, computer vision and cognitive sciences to build a taxonomy of vision-language capabilities.

Denotation The capabilities of a vision-language model to explicitly associate a text and an image are conditioned on its ability to take into account information at different levels.

At a **local** level, denotation capabilities evaluate the understanding of a single element of a text-image instance, independently of the rest. Among the previously determined capabilities, object identification is such an ability. A parallel can be made with the Communicative Development Inventories (CDIs) (Fenson et al. 2007), where recognizing objects such as animals or vehicles is among the first skills evaluated for children. Several datasets have focused on the evaluation of object categories (Shekhar et al. 2017; Parcalabescu, Cafagna, et al. 2021). A related category that appears in CDIs is the understanding of descriptive words (e.g. 'dark', 'blue'). We infer from it the capability to detect basic descriptive attributes. This capability is often included in more complex tasks, alongside other capabilities Johnson et al. 2017; Kafle et al. 2017.

Local denotation skills:

— Basic Property Detection: *Def.* The ability to detect the presence of a basic property (e.g., color, texture) and associate it to a corresponding word.
Ex. Associate the color red with the word 'red'.

— Object Perception: *Def.* The ability to differentiate between objects, both at coarse and fine-grained level. Includes the understanding of the continuity of an object (e.g., segmentation).
Ex. Identify a flower from its picture.

At a **structural** level, denotation capabilities evaluate the understanding of the dependency between an element and the rest, or between several elements of an instance, i.e., the compositionality of an instance. As a whole, those skills also require local understanding, because the model needs to understand each element individually. An instance depends, in addition to the individual elements, on the structure of those elements. Although we have identified, in the previous section, the need for structural understanding of an instance, we specify here more granular capabilities using as basis previous work in vision-language multimodality. As the structure of text and that of an image are radically different, we first consider the understanding of the two structures individually: scene understanding and syntactic understanding. Scene understanding, which also groups positional understanding and counting, is an active field of research in vision-language multimodality (Johnson et al. 2017; Parcalabescu, Cafagna, et al. 2021; Salin et al. 2022). Similarly, the multimodal understanding of syntax remains part of ongoing research, as works have shown the difficulty of vision-language models to understand word order at a multimodal level (Thrush et al. 2022). In addition, understanding the multimodal alignment between individual textual and visual elements is also important. Multiple capabilities are related this, such as the understanding of multimodal dependencies (Nikolaus, Salin, et al. 2022) and coreferences (Z. Chen et al. 2020).

Structural denotation skills:

— Syntactic Understanding: *Def.* The ability to grasp the syntactic structure of a sentence and deduce the relation between different words using visual information. Includes the resolution of polysemy.
Ex. Differentiate 'bear' as a verb or a noun.

— Scene Understanding: *Def.* The ability to grasp the structure of an image using textual information. Includes counting and positional understanding (i.e., the ability to understand depth, distance, and position between objects in the referential of the image).
Ex. Count people in a crowd.

— Multimodal Alignment Understanding: *Def.* The ability to correctly associate textual elements using visual information. The textual elements can be non-explicit (i.e., co-reference resolution). Includes understanding the static interaction between people and objects in an instance.
Ex. Associate a predicate to the correct noun.

At a **global** level, denotation capabilities evaluate the understanding of the whole instance. Two main capabilities determined in the previous section correspond to this

category: the ability to understand the document type (e.g., the context behind the data collection) or the focus. However, to our knowledge, besides domain-specific datasets, no multimodal dataset evaluates these precise capabilities.

Global denotation skills:

— Document Type Understanding: *Def.* The ability to detect the topic of an instance, its source (e.g., author, machine used to capture it), its date or its style.
Ex. Specify how a medical image was captured.

— Focus Identification: Understanding what elements are or are not the focus of an instance using its textual and visual information.
Ex: Identify which person is the focus of a newspaper image/caption pair.

Denotation skills characterize factual understanding of a vision-language instance and its components. We listed in this section several skills that, to our knowledge, are necessary to establish this understanding of a vision-language instance. This list omits the ability to ground the instance in the world or use knowledge that is specific to a domain.

Grounding We identified several subtypes of the grounding category: **temporal**, **spatial**, **knowledge** and **foreign modalities** grounding.

First, **temporal** grounding capabilities evaluate a model's ability to understand the situation of an instance in time. The ability of action understanding, context understanding and object state understanding described in the previous section are related capabilities. Several datasets already evaluate the grounding in time of a model, through tasks such as event captioning or procedural understanding (Krishna, Hata, et al. 2017; Yagcioglu et al. 2018), but not all capabilities are covered.

Temporal grounding skills:

— Temporality Perception: *Def.* The ability to detect if time affects the instance. For the image modality, it includes whether an object/structure changes state and position in the immediate past or future. For the textual modality, it means using text information (e.g., verb tense) to detect temporality.
Ex. Detect which element of an instance are moving.

— Object State Understanding: *Def.* The ability to associate the state of an object with corresponding words and differentiate the role of an object depending on its state.
Ex. Differentiate between an empty or full glass.

— Temporal Extrapolation: *Def.* The ability to extrapolate the past or future structure of a scene using multimodal information.
Ex. Understand that a glass will break if pushed.

— Time Period Identification: *Def.* The ability to identify a specific period in a multimodal instance.
Ex. Recognize that an instance depicts medieval times.

Then, **spatial** grounding capabilities evaluate a model's ability to understand a scene as part of a wider spatial context. Among the applications studied in the previous section, it is especially useful in Vision-Language Navigation, but also in context understanding. Several datasets and tasks focus on spatial grounding capabilities,

mainly relating to 3D understanding (Gordon et al. 2018; Chou et al. 2020; Chang et al. 2017; Kolve et al. 2017).

Spatial grounding skills:

— Spatial Understanding: *Def.* The ability to ground an instance in the world using textual and visual information. Includes the understanding of perspective, depth, size and spatial referential.
Ex. Recognize that a plane in the sky is the same size as at the airport.

— Physical Spatial Understanding: *Def.* The ability to understand how physics affect the position of objects in an image. Includes occlusion, obstacles, contact.
Ex. A partially hidden object is still the same.

— Spatial Extrapolation: *Def.* The ability to extrapolate the spatial context not seen in the instance using multimodal information.
Ex. Extrapolate what is behind the photograph taking a picture.

— Location Identification: *Def.* The ability to recognize known places using multimodal information.
Ex. Recognize a specific country using street furniture.

In addition, technical or cultural **knowledge** can be necessary to understand a vision-language instance. This can be relevant to context understanding in news data, or to the understanding of medical data. In the case of technical grounding, evaluations specific to the domain are necessary (X. He et al. 2020; Xiaosong Wang et al. 2017; Biten, Gomez, et al. 2019; Ramisa et al. 2017).

Knowledge grounding skills:

— Semantic Grounding: *Def.* The ability to exploit knowledge from semantic relations (e.g., roles, synonyms, antonyms, and hypernyms).
Ex. Understand that 'robin' and 'bird' can refer to the same element.

— Technical Grounding: *Def.* The ability to exploit knowledge from a specific domain (e.g., medical). Includes the understanding of specialized objects, technical terms, events, or specific named entities. *Ex.* Associate visual information to the term 'pneumothorax'.

— Cultural Grounding: *Def.* The ability of a model to understand the cultural context of an instance, with respect to textual or visual elements, and differentiate across cultures.
Ex. A mask can mean a medical mask or a mold that represents someone else. The latter, following cultures, can be traditional, religious, used for theater or for carnivals.

— Symbolic System Grounding: *Def.* The ability to recognize symbols and characters in an image. Ranges from Optical Character Recognition to the ability to recognize the meaning of a symbol.
Ex. Describe signs held at a demonstration.

Finally, vision-language models can also be evaluated on their understanding of other **foreign modalities** not present in the instance. For instance, they can be used in applications which refer to time series, such as financial data understanding. In this case, evaluation tasks for those capabilities are very specific and depend on the domain.

Other multimodal grounding skills:

— Human Senses Grounding: *Def.* Detecting and associating words or objects that can refer to human senses not linked to vision, such as hearing, touch or taste.
Ex. Associate a waterfall with the word 'loud'.

The use of grounding can be necessary for some vision-language applications. The understanding of temporality and other forms of grounding is complex, and requires precise data to be appropriately evaluated. If a vision-language model is destined at being used in this context, evaluating it on granular capabilities can be necessary to understand weaknesses. In addition, if the training data permits it, other multimodal foundation models designed with those specific modalities in mind may lead to better performances on those tasks.

Reasoning We identify a few reasoning tasks necessary for vision-language models, using as inspiration existing monomodal tasks (MacCartney 2009; X. V. Lin et al. 2018; Bowman et al. 2015; Zhilin Yang, Qi, et al. 2018; Camburu et al. 2018).

First, some reasoning capabilities can require a good understanding of **semantic** knowledge, which can be useful in applications requiring some kind of technical knowledge such as medical assisted diagnosis. We can for instance list the detection of abnormality. However, there is to our knowledge no dataset evaluating multimodal knowledge-based reasoning.

Semantic reasoning skills:

— Abnormality Detection: *Def.* The ability to detect an abnormal instance. Includes making the distinction between something rare and something unrealistic. Can be local, structural, or global.
Ex. Detect that an object is at an unrealistic position.

— Mismatch Detection: *Def.* The ability to spot if information is missing from one of the two modalities.
Ex. Detect that a sentence asks a question about an object which isn't present in the image.

Then, reasoning skills can be based on **logic**, or the understanding of mathematical concepts. Several evaluation tasks have focused on logical and mathematical reasoning (Cherian et al. 2023), as such tests are used as a metric to measure human intelligence. Other skills linked to logical reasoning are those based on comparison between instances. Those are well known in natural language processing, being evaluated through tasks such as natural language inference (Do et al. 2020).

Logic reasoning skills:

— Logical Operations: *Def.* The ability to understand logic operations (e.g., negation, *or*, *and*).
Ex. Understand 'no' in 'There is no cat'.

— Comparison: *Def.* The ability to compare two parts of an instance. Can also be applied between multiple instances.
Ex. Compare the size of two objects in an image.

— Multimodal Inference: *Def.* The ability to detect whether one instance can be entailed from another.

Ex. Use context and a medical image to assist in a diagnosis.

— Mathematical Reasoning: *Def.* The ability to use topological, geometrical, arithmetical or algebraic skills.
Ex. Answer a math-related IQ question.

Finally, some reasoning capabilities are more **complex**, due to the use of abstraction or several steps of reasoning. For instance, this is the case of multi-hop reasoning that can be encountered in vision-language navigation. As such tasks are complex and specific, they are mostly evaluated with respect to the relevant application domain. We also group in this subcategory the ability to perform introspection, i.e., to explain the reasoning of a prediction, which is an active field of research (Kayser et al. 2021; Zellers et al. 2019; Das, H. Agrawal, et al. 2017).

Complex reasoning skills:

— Extrapolation: *Def.* The ability to complete an instance from incomplete visual or textual information. Includes the ability to distinguish between extrapolation and hallucinations.
Ex. Deduce part of an obstructed text in an image without hallucinating.

— Multi-hop Reasoning: *Def.* The ability to perform reasoning using multiple steps.
Ex. Path computing in vision-language navigation.

— Introspection: *Def.* The ability to explain the prediction of a task.
Ex. Explain the reasoning when answering a question.

These reasoning capabilities can be complemented by other monomodal capabilities transferred to multimodality. Some of those tasks can require task-specific data or fine-tuning, and be difficult to achieve using only a foundation model.

Connotation The skills listed in this section may not be useful to all applications of vision-language models, as they rely on individual interpretation of multimodal instances. In addition, their evaluation is subjective and can vary depending on the annotations.

The connotation capabilities can evaluate a model's ability to **interpret** the meaning or intent of an instance. In particular, this relates to the previously identified capability of intent understanding. Some related evaluation tasks interpret the emotion (Mathews et al. 2016) or the style techniques (Perronnin 2012).

Interpretation connotation skills:

— Symbolism Understanding: *Def.* The ability to understand the intent behind the symbolism in multimodal elements (e.g., metaphors).
Ex. Associate a person holding a scale with 'justice'.

— Ambiguity Understanding: *Def.* The ability to understand voluntary ambiguity (e.g., optical illusions, word plays).
Ex. Understand that an image shows a duck or a rabbit.

— Sentiment Understanding: *Def.* The ability to understand the emotions evoked by an instance. Includes the detection of humor and irony.
Ex. Understand that the gap between an image and its associated text conveys humor.

In addition to interpretation, connotation capabilities can also relate to the **qualitative** evaluation of an instance. These are mostly evaluated using user judgment, and evaluate stylistic appreciation (Radev et al. 2015; T. Levinboim et al. 2019).

Critical connotation skills:

— Stylistic Appreciation: *Def.* The ability to evaluate whether stylistic elements are appropriately and consistently used.
Ex. Criticize the symmetry in an image.

— Effectiveness Evaluation: *Def.* The ability to evaluate whether an instance is effective at expressing its intended meaning.
Ex. Evaluate whether a cartoon transmits the intended message.

In the connotation category, we list several capabilities for which we have found no related evaluation tasks. Those are inspired from human evaluation methods of visual literacy. This is why they often rely on interpretation and assessment of instances. These skills can be used in real-world applications where the interpretation of an instance is important, such as applications related to art.

2.5. Evaluating Foundation Models Using the Taxonomy

The taxonomy presented in the previous section aims at providing a guideline for an extensive evaluation of vision-language foundation models, taking into account their real-world applications. We argue that foundation models should be evaluated when possible on granular capabilities, more easily interpretable than complex tasks. The various capabilities should have the broadest possible coverage, with respect to real-world applications. Indeed, it is important to be aware of the main weaknesses of a foundation model, as well as the scope of tasks and datasets it can be applied to. By proposing this taxonomy, we hope that models are evaluated on a wide range of tasks that covers their intended use cases, to highlight their strengths and possible weaknesses.

However, a model should not necessarily be evaluated on all possible capabilities. Indeed, depending on the application and domain of a vision-language foundation model, it can be unnecessary to evaluate it on every possible capability, and all capabilities may not have the same usefulness. For instance, a foundation model geared towards medical assisted diagnosis would have no use for connotation capabilities. Some skills are important for most real-world applications, such as capabilities belonging to the Denotation category. However, other capabilities mostly depend on the intended use of the model. In the case of some applications such as Vision-Language Navigation, the models should also be evaluated on *Spatial Grounding*, *Temporal Grounding* and *Complex Reasoning* capabilities. Models used for applications such as chit-chat would also need to be evaluated on *Connotation* capabilities.

In Table 2.1 and 2.2, we give a projection of vision-language evaluation tasks in our proposed taxonomy. However, the goal of this taxonomy is not to help compute a ranking score from an aggregation of tasks, but to bring back the focus on multimodal

Category	Subtype	Datasets	Description
{Denotation}	Local	GQA (Hudson et al. 2019), Foil it! (Shekhar et al. 2017), TDIUC (Kafle et al. 2017), VQA (Y. Goyal et al. 2017), VALSE (Parcalabescu, Cafagna, et al. 2021), Toolbox (T. Zhao et al. 2022)	Object and attribute recognition
	Structural	GQA (Hudson et al. 2019), Daquar (Malinowski et al. 2014), CLEVR (Johnson et al. 2017), TDIUC (Kafle et al. 2017), Probing (Salin et al. 2022), VALSE (Parcalabescu, Cafagna, et al. 2021), Toolbox (T. Zhao et al. 2022)	Position understanding and counting
		Winoground Thrush et al. 2022	Understanding word order
		Noun-Predicate Dep (Nikolaus, Salin, et al. 2022), Abstract Semantics (Zitnick et al. 2013), CREPE (Z. Ma et al. 2023), ARO (Yuksekgonul et al. 2022)	Understanding compositionality
		Cops-ref (Z. Chen et al. 2020), RefCOCO (Kazemzadeh et al. 2014), CLEVRRef (R. Liu et al. 2019), VALSE (Parcalabescu, Cafagna, et al. 2021)	Multimodal referring expressions
{Grounding}	Temporal	Dense Event Captioning (Krishna, Hata, et al. 2017), RecipeQA (Yagcioglu et al. 2018)	Event and procedure understanding
	Spatial	IQUAD (Gordon et al. 2018), VQA360 (Chou et al. 2020), Matterport3D (Chang et al. 2017), AI2-THOR (Kolve et al. 2017), RemoteSensing (X. Lu et al. n.d.)	Spatial understanding (3D & aerial)
	Knowledge	OK-VQA (Marino et al. 2019), TDIUC	Object role understanding
		TextVQA (Singh, Natarjan, et al. 2019), SceneText VQA (Biten, Tito, et al. 2019), TextCaps (Sidorov et al. 2020)	Optical character recognition
		OK-VQA (Marino et al. 2019)	VQA with cultural knowledge
		GoodNews (Biten, Gomez, et al. 2019), BreakingNews (Ramisa et al. 2017)	News-related tasks with NER
		PathVQA (X. He et al. 2020), Chest Xrays (Xiaosong Wang et al. 2017)	Medical tasks

Table 2.1. – Projection of a range of existing vision-language evaluation tasks in the suggested taxonomy — Denotation and Grounding

understanding capabilities relevant to real-world vision-language applications. Using several pre-defined tasks for the evaluation of vision-language foundation models may encourage a focus on raising the performance on those tasks. Yet, we argue that they should be used for an introspective evaluation, to establish a diagnosis of a foundation model.

Selecting an evaluation dataset The datasets presented in Tables 2.1 and 2.2 may not always be appropriate for the multimodal evaluation of models. Indeed, among the existing evaluation tasks for vision-language models, some of them evaluate an aggregate of complex skills more or less directly linked to a specific capability. They may not be granular enough to identify potential blind spots. Another aspect is that

Category	Subtype	Datasets	Description
{Reasoning}	Logical	E-SNLI-VE (Do et al. 2020), NLVR2 (Suhr et al. 2019)	Multimodal inference and comparison
		SMART (Cherian et al. 2023)	Logical and mathematical reasoning
	Complex	E-vil (Kayser et al. 2021), VCR (Zellers et al. 2019), VQA-HAT (Das, H. Agrawal, et al. 2017)	Explanations for VQA
		Visual Dialog (Das, Kottur, et al. 2017), FashionIQ (H. Wu et al. 2021), GuessWhat?! (De Vries et al. 2017)	Dialog with multimodal context
{Connotation}	Interpretation	AVA (Perronnin 2012)	Image style understanding
		SentiCaps (Mathews et al. 2016)	Caption generation with sentiments
	Quality	New Yorker Caption Contest (Radev et al. 2015), ICQD (T. Levinboim et al. 2019)	Rating Caption quality
		DPC (Jin et al. 2019), VizWizQuality (Chiu et al. 2020), AVA (Perronnin 2012), Aesthetic Cap (Ghosal et al. 2019), VILA (Ke et al. 2023)	Image Quality Evaluation

Table 2.2. – Projection of a range of existing vision-language evaluation tasks in the suggested taxonomy — Reasoning and Connotation

they may not truly evaluate multimodal understanding of a concept. Indeed, some of those tasks present considerable textual bias. This can hamper the multimodal evaluation of those models. For instance, Parcalabescu, Cafagna, et al. 2021 show that a language model can achieve good performances on 'Foil it!' (Shekhar et al. 2017). In other cases, the task itself may not be built with multimodality in mind. This is the case for datasets of the connotation category, where the evaluation of instance quality can often be associated to a vision-only task. Indeed, the difference between monomodal and multimodal capabilities can be blurry, as shown by the use of vision-language models to perform vision-only tasks (Radford, J. W. Kim, et al. 2021). It is difficult to characterize what capabilities belong to *multimodal understanding*. What is most important is how they are evaluated. Indeed, the presence or absence of an object could be said to be solely monomodal in the case of an image classification task. However, in the context of establishing a link between an image and a text, this capability can be evaluated at a multimodal level. This is why some capabilities that we present in this taxonomy may belong to both multimodal and monomodal understanding.

2.6. Limits of the Current Taxonomy

This taxonomy is aimed at guiding the evaluation of foundation models for real-world applications. However, the use of such a taxonomy also presents its limitations. First, it may not reflect the possible applications of vision-language foundation models, and may be more specifically biased towards existing tasks. Indeed, capabilities were selected from a range of English language vision-language applications. Those may hide challenges or needs more present in other languages or cultures. In Tables 2.1 and 2.2, we give an overview of vision-language evaluation tasks related to the categories listed in the taxonomy. These evaluation tasks are not evenly distributed through the categories, and this taxonomy can help us identify potential gaps in the evaluation of vision-language models. These gaps can be due to the lack of interest, available data or known research challenges, but still hide potential blind spots of those models. This taxonomy is not final, but the gaps can also be used to guide the way towards other evaluation tasks relevant for vision-language applications. The taxonomy we presented in this section establishes a set of skills relating to vision-language multimodal understanding. However, evaluation tasks for foundation models may not necessarily fit into this taxonomy. Indeed, there can be overlap in the skills that different tasks evaluated. In addition, more complex skills are built on simpler skills. For instance, most reasoning skills require first an understanding of denotation skills. As a result, this taxonomy is not intended to be complete, but a first step towards building a more comprehensive evaluation of multimodal foundation models.

What is not evaluated through this taxonomy Although we focus this paper on *capabilities* of vision-language models, other factors are to take into account to provide a comprehensive evaluation of a foundation model. For instance, a foundation model should have a good ability to generalize to unseen examples from different domains. This diversity could be ensured by selecting instances from a broad range of semantic categories. For instance, vocabulary from Communicative Development Inventories for various cultures (M. C. Frank et al. 2017) can be used to ensure diversity, as well as images from diversified sources. In addition, we do not mention limiting bias and ensuring fairness and robustness, which are major aspects of foundation models evaluation, and should be considered when building evaluation tasks and datasets. In this taxonomy, we also do not take into account the type of task (e.g., generation, classification). Yet, the evaluation of a multimodal capability can vary depending on the type of task used. Finally, as this taxonomy is based on a sample of tasks that is not necessarily representative of all possible vision-language applications, it is also incomplete. It is intended to evolve, and to be more specified, for instance regarding the various uses of a foundation model.

Evaluating the taxonomy An important question is how to evaluate such a taxonomy, in terms of its coverage. Indeed, it is difficult to be both granular and exhaustive. One could study a range of tasks from Section 2.3.2 in the same way as Section 2.3.2 to ensure a coverage of all necessary capabilities. This is particularly difficult to assess,

as it depends on how models are used in downstream applications. This taxonomy is incomplete, and is aimed at evolving with the improvement of vision-language foundation models and the creation of new applications. Indeed, this will lead to new challenges. In addition, evaluating a model on the whole taxonomy is time- and resource-consuming, this is why our goal in presenting this taxonomy is above all to serve as a guideline.

2.7. Conclusion

Foundation models are notoriously difficult to evaluate. To our knowledge, no exhaustive evaluation method of vision-language foundation models has been developed yet. Most evaluation methods at the time of this work are based on complex fine-tuning tasks. Though they enable comparison of vision-language models due to their complexity, they offer limited insight into the inner workings of vision-language models. In this chapter, we focus on the evaluation of the *General Multimodal Understanding* of vision-language models. We argue that such a method should aim at evaluating a wide range of precise multimodal capabilities, to apprehend the possible weaknesses of such models. Indeed, more complex tasks may not highlight the blind spots in understanding of specific multimodal concepts, because they often combine multiple different capabilities. Our goal is to shift the evaluation of vision-language model towards diagnosis and analysis, in addition to standard state-of-the-art model comparison. As such, it would be easier to identify blind spots of vision-language models, in order to possibly remedy to them. In addition, being aware a model's limitations can prevent harmful use of this model.

To that end, we propose a methodology to build a taxonomy of vision-language capabilities. Figure 2.3 summarizes this methodology. We rely on vision-language tasks to establish vision-language capabilities useful for vision-language applications. We also relate this taxonomy to existing evaluation tasks. However, the use of such a taxonomy also presents its limitations, due to potential bias in determining useful capabilities. First, it may not reflect the possible applications of vision-language foundation models. The taxonomy is also influenced by existing vision-language tasks. In particular, it may also be biased towards a specific view of vision-language relations that is not necessarily consistent across cultures. In addition, it is difficult to evaluate the coverage and relevance of such a taxonomy. In the future, it would be interesting to strengthen this taxonomy with additional perspectives, and to complete its coverage of vision-language real-world applications.

Since the beginning of pre-trained vision-language models, few studies related to vision-language capabilities have focused on the evaluation of specific capabilities. Evaluating a model on all identified vision-language capabilities is not feasible, both in terms of resources and available data. However, it is essential to investigate potential blind spots of vision-language models. This is all the more important that the field of vision-language multimodality is not yet mature, and identifying weaknesses is essential to guide future research in the domain. Thus, we investigate in Chapter 3

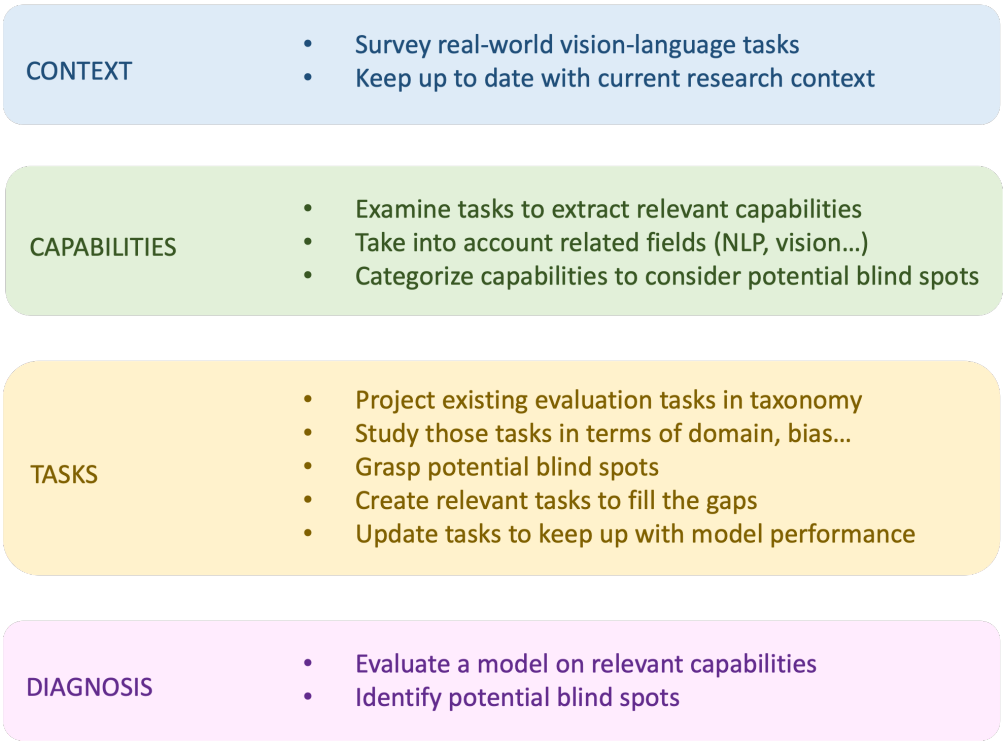


Figure 2.3. – Methodology for the evaluation of vision-language models using the taxonomy

several monomodal and multimodal capabilities, and compare the performances of state-of-the-art vision-language models. In particular, we study several denotation capabilities, which are especially relevant for most vision-language applications.

3. Investigating Capabilities of Vision-Language Models

Table of Contents

3.1. Introduction

At the start of this thesis, vision-language transformer models were a new development, and had not yet reached the size of current billions-of-parameters *foundation* models. Since then, their state-of-the-art results on various vision-language fine-tuning tasks have cemented the architecture in the field of vision-language multimodality. However, while fine-tuning tasks are useful to compare the ability of vision-language models on complex tasks, they are not sufficient to pinpoint precise weaknesses. Indeed, these weaknesses may not be easily perceptible through the overall performance of the model on these tasks. In addition, as presented in chapter 1, there is a wide variety of pre-trained vision-language models, with different architectures, pre-training protocols, and datasets. These choices can be optimized for some vision-language tasks while impacting negatively another aspect of the multimodal understanding of vision-language models. As a result, fine-tuning tasks are insufficient to ascertain what information pre-trained models extract in the representations they learn, and what textual, visual and multimodal concepts they understand.

Vision-language models are expected to be used in many real-world applications. Thus, we should aim towards a better explainability of their performances. As discussed in the previous chapter 2, evaluating a model on specific capabilities help us reach a better understanding of its scope and potential weaknesses. Through this chapter, we aim to better understand the inner workings of vision-language transformer models, identify and study potential weaknesses. To that end, we compare different vision-language models to deduce potential limiting factors affecting their understanding of various capabilities. Through this chapter, we focus on capabilities introduced in the denotation category of the proposed taxonomy (Chapter 2). Indeed, those are required for almost all vision-language downstream tasks, and do not need expert annotations. Consequently, weaknesses identified through the studies presented in this chapter can have a significant impact on a wide range of uses for vision-language models.

First, we detail two evaluation methodologies presented in Chapter 2: probing and pre-training head evaluation. Indeed, in this chapter, we use those two methods to compare the performance of state-of-the-art vision-language models.

Then, we use a range of probing tasks to study the understanding of state-of-the-art vision-language models on a range of basic monomodal and multimodal tasks to compare their strengths and weaknesses. We probe the representations of pre-trained and fine-tuned models to highlight eventual weaknesses, and compare how different choices in training and architecture may affect these representations. We specifically study capabilities applicable to most vision-language tasks:

— Syntax understanding (Bigram Shift, Part of Speech Tagging)
— Object counting
— Fine-grained object classification
— Color differentiation
— Position understanding
— Size understanding

— Minimal difference understanding (objects, actions)

We aim to answer the following questions: Do vision-language models encode information relevant to those textual, visual and multimodal capabilities in their representations? If so, how do pre-trained models compare to fine-tuned models? How do training design choices impact the information encoded by those models? Do those models rely both on textual and visual information, or is there an overreliance on one modality at the expense of the other? To answer those questions, we build datasets for each of these tasks and evaluate whether the necessary information is extracted in vision-language representations of three state-of-the-art models. We analyze the results of monomodal and multimodal probing tasks to identify how vision-language models compare to their monomodal counterparts. We also design a methodology to assess the impact of textual bias on vision-language model performances, that we call mismatched tasks. We also compare pre-trained models to models fine-tuned on two common tasks to give us insight on the role of fine-tuning. In this chapter, we report results recently published in the following paper: "Salin, E., Farah, B., Ayache, S., & Favre, B. (2022, June). Are vision-language transformers learning multimodal representations? a probing perspective. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 10, pp. 11248-11257).". [1]

In a complementary study, we focus on the understanding of position. Indeed, we find that vision-language models have difficulty understanding the concept of object position at a multimodal level. We aim to answer the following questions: Do the models understand the concept of position at a monomodal level? If so, what are the reasons behind their failure at understanding position at a multimodal level? What measures should be taken during pre-training to improve their multimodal understanding of position? To answer those questions, we build tasks using synthetic data based on the CLEVR dataset (Johnson et al. 2017). We conduct several experiments to apprehend how information related to the concept of position is encoded in the representations of those models. In this chapter, we report results published in 'Salin, E. (2022, November). 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) (pp. 181-187). CNRS.'

Finally, we evaluate how vision-language models understand multimodal dependencies. Indeed, at the time of our work, the evaluation of vision-language models had been specifically focused on their understanding of individual entities. The study of more fine-grained multimodal concepts is still not very represented in this field. We aim to answer the following questions: Do vision-language models understand fine-grained multimodal dependencies? Is their understanding of this concept correlated to their understanding of more coarse-grained multimodal concepts? How do different pre-training choices affect their understanding of this concept? To answer those questions, we create a dataset to evaluate vision-language models on the task

1. This work was carried out in collaboration with Badreddine Farah, a master student whom I supervised during his internship.

of image-text matching (explained in chapter 1). This dataset is specifically focused on evaluating a model's understanding of noun-predicate dependencies. We compare several state-of-the-art models to see how they are affected by various possible pre-training choices. The goal of this task is to gain a finer understanding of how a model learns multimodal dependencies, which remains a major unresolved capability for vision-language transformers. Indeed, while vision-language models have been shown to identify objects, their understanding of the compositionality of an instance remains limited. Our goal is to identify possible factors that promote the understanding of multimodal dependencies. In this chapter we report results recently published in the following paper: "Nikolaus, M., Salin, E., Ayache, S., Fourtassi, A., & Favre, B. (2022). Do Vision-and-Language Transformers Learn Grounded Predicate-Noun Dependencies?"

Finally, with hindsight on the creation of evaluation tasks and the evaluation of vision-language models, we discuss the limitations of those methods. We propose guidelines for the creation of future vision-language tasks.

3.2. Evaluation Methodologies

In this chapter, we aim to investigate pre-trained vision-language models. We are especially interested in how training choices affect monomodal and multimodal capabilities. Indeed, when studying pre-trained vision-language transformers, it is interesting to compare their architectures, pre-training tasks and various protocols on multiple capabilities. Some choices may positively impact the performance of a model on specific tasks, while negatively impacting it on others.

As explained in the previous chapter, a vision-language model can be used in many tasks, and diverse capabilities can be required. Thus, we aim to evaluate a selection of capabilities of a vision-language model, and compare models on those capabilities. To that end, we use appropriate methods, described in the previous chapter (Section 2.2.2): probing and pre-training head evaluation. These methods, inspired by Bertology, aim to evaluate models on specific tasks or capabilities.

3.2.1. Pre-training Head Evaluation

As explained in the previous chapter, pre-training head evaluation consists in creating a dataset and evaluating a model on this dataset using one of the pre-trained task-specific heads. Most vision-language models can be evaluated using an image-text matching task using the ITM task-specific head (Chapter 1). We can illustrate this method through the task of *object category understanding*. To test the ability of a vision-language model to understand object categories, one would create a dataset of positive and negative examples. The positive examples would have a caption matching an image, and negative examples would have one of the objects in the image not matching the caption. Then, the text-image inputs are given to the model for inference only. The task-specific head of the model computes the image-text match-

ing probability for each image-text pair in the dataset. One can then aggregate the accuracy of the model on the object understanding task, which can be treated as a binary classification task.

Figure 3.1. – Caption: a man kneeling down to play with a small dog. Instance from MSCOCO T.-Y. Lin et al. 2014

In the case of *object category understanding*, one could use as a positive instance the image/caption pair shown in Figure 3.1. An appropriate negative instance could be the same image, with the altered caption: 'A man kneeling down to play with a small *cat*.'

However, several studies have shown that pre-training head evaluation can be biased, in particular due to an overreliance of vision-language models on textual bias. Indeed, it is not robust to a minor change in the dataset. For instance, using plural instead of singular can considerably alter the results (Ravichander et al. 2020). In the case of the object categories example, depending on the way the negative example is constructed, some negative instances may seem obvious for a model. This leads the model to rely mainly on textual information, rather than multimodal information. Thus, while the task is aimed at evaluating a multimodal capability, it is impacted by monomodal bias. For instance, a negative example could be built consisting of the same image (Figure 3.1) and the caption : 'A man kneeling down to play with a small *horse*.' It is grammatically correct, and the *dog* and *horse* objects belong to the same broad category of mammals, but the first is more likely to be found in the dataset than the latter. Thus, the task can be biased, due to unbalanced object categories in the pre-training dataset.

Thus, in Section 3.5, we alter the methodology to remove the impact of monomodal bias, and create a balanced dataset.

3.2.2. Probing

Probing tasks are used to study what information is encoded in representations learned by a deep learning model. For transformer models, it can be any hidden state layer, or the final layer representations (before they are fed to task-specific heads).

In the case of vision-language models, the goal is to evaluate whether the model has learned to encode relevant textual, visual and multimodal information. To that end, we study the representations produced by a vision-language model *VL* using a classifier using a probing task. A probing task is created to study a specific aspect of the representations. For instance, the task can evaluate whether the representations

produced by the vision-language model extract information related to the color of objects. A dataset (X, y) is designed, to evaluate that aspect of the model. Then, the representations of each instance X of this probing dataset are computed, using the vision-language model with its parameters frozen. A probing model P is then trained on this task using as input the representations produced by the vision-language model and the corresponding labels y of the dataset. The probing model P can be trained for a classification or a regression task. The model is then evaluated by comparing its predictions $\hat{y}$ (Equation 3.1) to the gold labels.

$$(P \circ VL)(X) = \hat{y} \tag{3.1}$$

For a (vision-language model, probing model) pair (P_1, VL_1) to obtain significantly better result than another pair (P_2, VL_2) on a probing task, it means that the model VL_1 has managed to extract more information relevant to the task from the input than VL_2. The probing model P is often a linear classifier, to evaluate information linearly extracted from VL's representations. Indeed, P learns a linear transformation of the model representations R (Equation 3.2).

$$\hat{y} = RA^T + b, \text{ with } A \text{ the weight and } b \text{ the bias} \tag{3.2}$$

If a more complex probe reaches a significantly better result than a linear probe, it should mean that the model VL does not directly use the information, even if it is present in the features. In addition, information that is linearly accessible at a layer of the model VL may not be accessible by probing the representations from another layer of the same model VL.

However, interpreting probing results can be complex. Several factors can impact probing results, such as potential bias in the probing task dataset. Indeed, the probing model P could use potential bias present in the probing dataset to achieve good results, without taking into account the quality of VL representations. In Hewitt and Liang 2019, authors advocate for the use of control tasks, to make sure that the performance tested through probing is that of VL and not that of P.

Additionally, the representations studied can belong to other layers than the final layer of VL, in order to have more insight on the role of the different layers of a model.

In Section 3.3, we use probing tasks to analyze the performance of vision-language models on monomodal and multimodal capabilities. As we also aim to better apprehend the reliance on monomodal bias of those models, we create a new methodology, called mismatched tasks.

3.2.3. Discussion

Several important factors are to consider when evaluating vision-language models. It is important to create balanced tasks, so that the good performance of a model does not hide an overreliance on one modality. As vision-language models are heavily reliant on textual bias, it is important to carefully build datasets to limit the bias. While newer vision-language models show remarkable performances on some evaluation

tasks, this may hide some difficulties on select aspects of vision-language multimodality.

Probing and pre-training head evaluation are two methods that are useful to understand and compare different vision-language models. On one hand, probing enables us to evaluate whether information or concepts are present in a model's layer representation. It can help us reach a better explainability of the representations produced by those models, especially in terms of their ability to extract information. It can be used to compare models with significant architectural differences. On the other hand, pre-training head evaluation enables us to compare the performance of models on one of their pre-training tasks, such as Image-Text Matching. It shows a direct reflection of the performance reached with pre-training, on specific instances. The success or failure of a model on those instances can help us better understand the pre-training of those models. While both methods have their limitations, which we analyze later, they can be used to provide clues by comparing the performances of various models. As such, they can be used to provide a better understanding of the inner workings of vision-language models. As, at the start of this thesis, our understanding of vision-language transformer models was limited, we decide to use such methods to provide insights on the capabilities of vision-language models. First, we focus on monomodal and some basic multimodal capabilities belonging to the Denotation category introduced in Chapter 2. Indeed, such capabilities are often required in many vision-language applications, and a potential weakness could have a major impact. Then, due to our results showing a limited multimodal understanding of position, we study the positional understanding of vision-language models. Finally, we study their understanding of fine-grained multimodal dependencies.

3.3. Probing Monomodal and Multimodal Capacities

In this section, we study the monomodal and multimodal capacity of vision-language models by investigating what is encoded in their representations. While previous studies have shed light on some particular aspects of transformer-based vision-language models, they lack a more systematic analysis of monomodal biases that impede the nature of the learned representations, which we study through these experiments. Inspired by probing tasks developed in the Natural Language Processing field, we probe three state-of-the-art vision-language models: UNITER (Y.-C. Chen et al. 2019), LXMERT (Tan et al. 2019) and ViLT (W. Kim et al. 2021). UNITER is a single-stream with Faster R-CNN visual features, LXMERT is a dual-stream with Faster R-CNN visual features, and ViLT is a single-stream which does not use Faster R-CNN visual features. Although there are now other more recent alternatives, at the start of this thesis, these were representative of the different types of vision-language transformer models. These models have been introduced in chapter 1. All three models are pre-trained on the MLM language pre-training task. ViLT does not have visual pre-training, while LXMERT and UNITER have two different variations of visual pre-training based on faster R-CNN object regions. All three models are pre-trained on the ITM multimodal

pre-training task. ViLT and UNITER are pre-trained using the same aggregation of manually annotated and web-crawled datasets (MSCOCO, Conceptual Captions, Visual Genome). LXMERT mainly uses manually annotated datasets (MSCOCO, Visual Genome). UNITER and ViLT are in addition pre-trained on WRA (which does not use object regions in the case of ViLT, but pixel patches), while LXMERT is also pre-trained on VQA using the relevant available datasets.

In this section, we explore what information is learned and forgotten between pre-training and fine-tuning. We also compare vision-language models to their monomodal counterparts and evaluate the impact of textual bias on models performances. Our goal is to identify possible limiting factors in current statue-of-the-art models. To answer those questions, we probe both pre-trained and fine-tuned models.

We propose probing tasks and collect associated datasets to evaluate those models on monomodal and multimodal capabilities. We have made the set of monomodal and multimodal probing tasks, as well as all software developed for this study, available for further research [2].

Similar to our study, Shekhar et al. 2017 build a dataset to evaluate if the text and image information in vision-language models are both deeply integrated. Contrary to this dataset, we do not study the ability of a model to differentiate between objects from the same super-categories, but focus on multimodal concepts such as color, size, and position.

3.3.1. Methodology

In this section, we aim at evaluating the representations of vision-language models at a textual, visual and multimodal level. We write VL_{pre} a pre-trained transformer-based vision-language model, such as UNITER, LXMERT or ViLT. This model can be fine-tuned on a task T, for example VQA or NLVR2, as explained in Chapter 1. We note $VL_{fine(T)}$ the vision-language model fine-tuned on task T. As fine-tuning tasks are used to embed new knowledge in the model, it is interesting to evaluate more dedicated semantics or abilities related to a specific task.

Evaluation We use the *probing* methodology detailed in section 3.2.2 to study the representations of VL_{pre} and $VL_{fine(T)}$ models. To evaluate the representations of vision-language models on a probing task p, we build a training dataset $S_p = \{(X_j, Y_j)\}_{j=1}^{n_p}$, drawn i.i.d. from $\mathscr{D}^{\mathrm{p}}_{\mathscr{X}\times\mathscr{Y}}$ where $\mathscr{X}$ and $\mathscr{Y}$ relate to the datasets needed to probe the models. The first step of our method is to compute the final layer representations $VL_{pre}(X)$ or $VL_{fine(T)}(X)$ of an instance $X_j = (x_j^{image}, x_j^{caption})$ of S_p through the vision-language transformer-based model. If the probing task p studies the instance at a global level, we use the representation of the classification token [CLS] as input for p. If p studies the representations of each word, the representation of word tokens are used as input for p. The second step of our method is to use the representations R_j of the [CLS] or word tokens as input of a linear probing model P_p trained using

2. https://github.com/ejsalin/VLm-probing

the $\{(R_j, Y_j)\}_{j=1}^{n_p}$ dataset. As VL_{pre} or $VL_{fine(T)}$ are not trained on the probing task, the probing model P_p can only rely on linearly separable information the model has already learned to extract during pre-training or fine-tuning. As a result, the performance of P_p will reflect the capability of VL_{pre} and $VL_{fine(T)}$ models to extract the information needed for the probing task p.

The models VL and the set of probing tasks P are described in the following sections. Figure 3.2 illustrates the methodology for the object counting task V-ObjectCount of P.

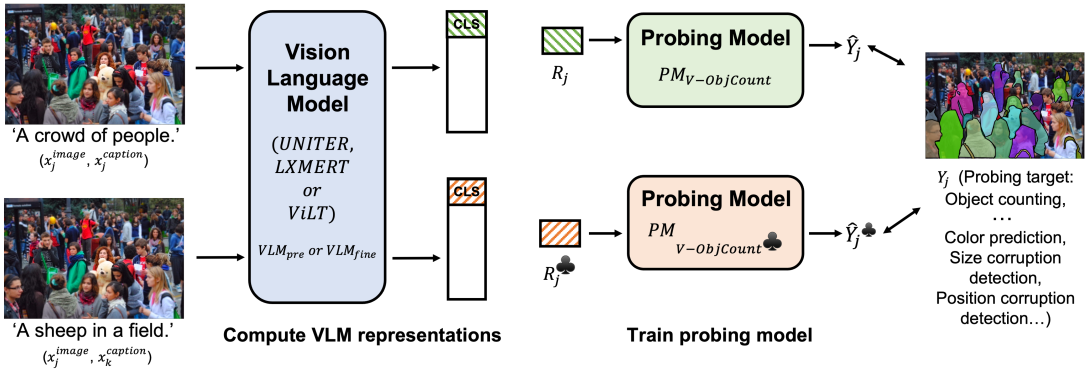


Figure 3.2. – Probing methodology: The first step is to compute the final layer representations of the image/caption input using the chosen Vision Language Model. Then, we use the final layer [CLS] or word token representations R_j to train a linear probing model on the probing task. This example illustrates the methodology using the visual probing task V-ObjectCount and an image from MS-COCO. This task consists in counting the number of objects in an image using $(x_j^{image}, x_j^{caption})$ as input and Y_j as target. The notation $\clubsuit$ indicates the corresponding task (V-ObjectCount) with mismatched instances (i.e., a caption that does not match the image and label), which uses $(x_j^{image}, x_k^{caption})$ as input and serves as a baseline.

A baseline: the mismatched task Previous works studying vision-language capabilities have shown to be subject to monomodal bias. This is the case of the previously mentioned Foil it! dataset. Due to this, models may reach good results by relying only on one modality, most often the textual modality. Indeed, BERT or other language only models can reach results significantly above chance on the Foil it! dataset. In order to better understand what part of a model performance is due to multimodal understanding or monomodal bias, we decide to design a new methodology, based on what we call a *mismatched* task. The goal of this task is to evaluate whether a model uses information from both modalities. We call this task the mismatched task.

We want to study how much vision-language models rely on the language and vision

modalities when building text-image representations. To that end, for each task p build on S_p, we create another corresponding task $p^{\clubsuit}$ with *mismatched* image and caption pairs. The dataset $S_{p^{\clubsuit}}$ is built using S_p by associating the label with the image (respectively caption) and selecting randomly a mismatched caption (respectively image). For better comparison, all models use the same mismatched datasets.

If p is a language-oriented task, each instance of $p^{\clubsuit}$ will be $(x_k^{image}, x_j^{caption}, Y_j)$, with the caption corresponding to the label and a wrong image. If the performance of P_p is similar to the performance of $P_{p^{\clubsuit}}$, we can deduce that the representations $R_j^{\clubsuit}$ given by *VL* are not affected by visual 'bias'. Similarly, when p is a vision-oriented task, each instance of $p^{\clubsuit}$ will be $(x_j^{image}, x_k^{caption}, Y_j)$, with the image corresponding to the label and a wrong caption. If the performance of P_p is similar to the performance of $P_{p^{\clubsuit}}$, we can deduce that the representations $R_j^{\clubsuit}$ given by *VL* are not affected by linguistic 'bias'.

For a multimodal probing task p, as we want to study the presumed prevalence of language over vision in model decisions, each instance of $p^{\clubsuit}$ will be $(x_k^{image}, x_j^{caption}, Y_j)$, with the caption corresponding to the label and a mismatched image. If *VL* extracts multimodal information, rather than only linguistic information, P_p should reach better performance than $P_{p^{\clubsuit}}$. This is a way to control whether $P_{p^{\clubsuit}}$ mainly uses textual information or if it also uses multimodal information.

3.3.2. Datasets

The set of probing tasks P, summarized in Table 3.1, is composed of language-oriented tasks L, vision-oriented tasks V, and multimodal tasks M. We build each probing task and dataset (except from the already existing flower-102 task) to evaluate the mono-modal or multi-modal performance of a vision-language model on a specific capability. The tasks consist of regression, binary or multiclass classification problems. For each of them, a linear layer is trained as in Hewitt and Manning 2019.

Ideally, one would probe vision-language models on all capabilities that have an impact on down-stream tasks, as described previously in the taxonomy (chapter 2). However, in this section, we restrain ourselves to a few textual, visual and multimodal capabilities belonging to the **Denotation** category. In particular, we choose and build tasks that are easy to implement on new datasets. For the language and vision capabilities, we rely on tasks well understood in past work. For multimodal capabilities, as there is at the time of our work a current lack of vision-language tasks evaluating specific capabilities, we create new tasks assessing multimodal capabilities that we think are especially relevant for vision-language models. We explain the tasks in the following section.

Language probing tasks: L For language-oriented probing tasks, we choose already existing language probing tasks and adapt them to a subset of 3,000 instances from Flickr30k (Young et al. 2014a). We mainly use the Flickr30k dataset as it a well-known image-text dataset that does not overlap with the pre-training datasets of the

Task	Description	Input Repr.	Type (Metric)	Dataset	Size	Maj. (%)
L-Tagging	Part-of-speech tagging	Word	Multiclass (acc.)	Flickr	3,000	24.06
L-BigramShift	Bigram shift detection	[CLS]	Binary (acc.)	Flickr	3,000	50.20
V-Flower	Fine-grained classification	[CLS]	Multiclass (acc.)	Flower-102	7,169	0.98
V-ObjectCount	Object counting	[CLS]	Regression (MSE)	MS-COCO	2,424	—
M-Color	Color prediction	[MASK]	Multiclass (acc.)	Flickr	3,000	25.30
M-Size	Size identification	[CLS]	Binary (acc.)	Flickr	2,552	50.67
M-Position	Position identification	[CLS]	Binary (acc.)	Flickr	2,626	53.75
M-Differences	Adversarial captions	[CLS]	Binary (acc.)	MS-COCO	700	50.00

Table 3.1. – List of probing tasks. *Repr.* is the representation vector used as input of the probing task. *Instances* indicates the number of image/caption instances used for the probing task. *Maj.* is the majority baseline. As LXMERT uses a different pre-training subset of MSCOCO than UNITER, there is a risk of overlap for this task. To avoid overlap between pre-training dataset and evaluation task, another subset is selected for LXMERT. However, LXMERT may not be directly comparable with ViLT or UNITER for those tasks.

three studied models. It is a varied dataset manually annotated with descriptive captions, which ensures that the caption matches the image. We choose tasks appropriate for the relatively simple structure of captioning datasets, and easy to transfer to a new dataset.

— **Part of Speech Tagging (L-Tagging):** This task evaluates a *Structural* capability relating to *Syntactic Understanding*. Part of Speech Tagging consists in associating a word with its corresponding part of speech label, such as *verb*. To create a gold standard for this task, we annotate the Flickr30k dataset using the *en_core_web_sm* SpaCy tagger (Honnibal et al. 2015), which performs at 97% accuracy on Ontonotes. This English-language tagger is multi-task CNN with GloVe vectors trained on common crawl. As a result, though we do not use the gold labels for this task, these labels are an appropriate approximation to estimate the ability of vision-language models to understand part-of-speech tags. We use the 34 categories made available by the Spacy tagger, which are fine-grained part-of-speech tags specific to the English language. This task evaluates the syntactic knowledge present in the representation of individual word tokens. In particular, it is interesting to evaluate whether the syntactic structure of a sentence is encoded in the representations of a vision-language model. As a result, we train a linear classifier $P_{\text{L-Tagging}}$ using contextual word token representations given by *VL*.

— **Bigram Shift (L-BigramShift):** This task evaluates a *Structural* capability relating to *Syntactic Understanding*. Bigram Shift (Conneau et al. 2018) consists in determining whether two consecutive words have been swapped. For example, in the phrase 'People *at relaxing* the park.', tokens from the bigram ('relaxing', 'at') have been swapped to create a negative example caption. As this evaluates the global correctness of a sentence, we use the [CLS] token representation given

by the last layer of the transformer. A random bigram is modified in half of the captions, as on Figure 3.3.

Figure 3.3. – Bigram shift task illustration: Shifted caption: 'Two men in green shirts are **in standing** a yard.'

Vision probing tasks: *V* In order to probe the vision capability of the models, we selected two tasks. The first is an object counting task to assess if information on the general structure of an image is present in the representation. The second is a fine-grained object classification task to evaluate whether the representations also retain information on fine details of objects.

— **Flower Identification (V-Flower):** This task evaluates a *Local* capability relating to *Object Perception*. This is a fine-grained object classification task which consists in classifying flower pictures into 102 categories. We use the 102-Flower dataset Nilsback et al. 2008, an example is given in Figure 3.4. This visual task is based on the perception of differences among two objects belonging to the same category. As UNITER and LXMERT use faster R-CNN object representations, this task aims at testing the limits of such design and comparing it with other designs (ViLT). As there is no caption available for this task, we use an empty caption. The linear classifier $P_{\text{V-Flower}}$ uses the representation of the [CLS] token.

— **Object Counting (V-ObjectCount):** This task evaluates a *Structural* capability relating to *Scene Understanding*. We build this object counting task on a subset of 3,000 instances of the MS-COCO dataset T.-Y. Lin et al. 2014. Since models are pre-trained using this dataset, we use the appropriate data splits in order to avoid the risk of overlap between pre-training and evaluation. The labels are created by counting the number of objects in its manual annotations. The notion of what can be considered an object is subjective. Indeed, objects can be annotated in some images and not in others depending on different factors such as size and visibility. For instance, not all people in a crowd are annotated. Some

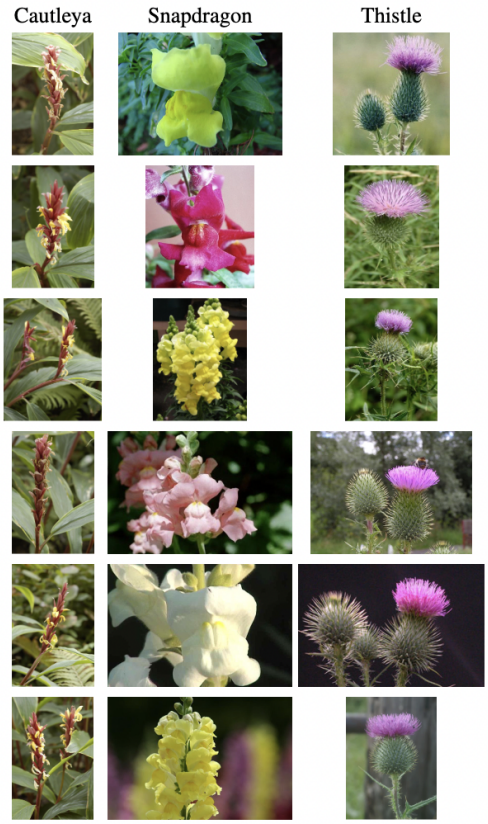
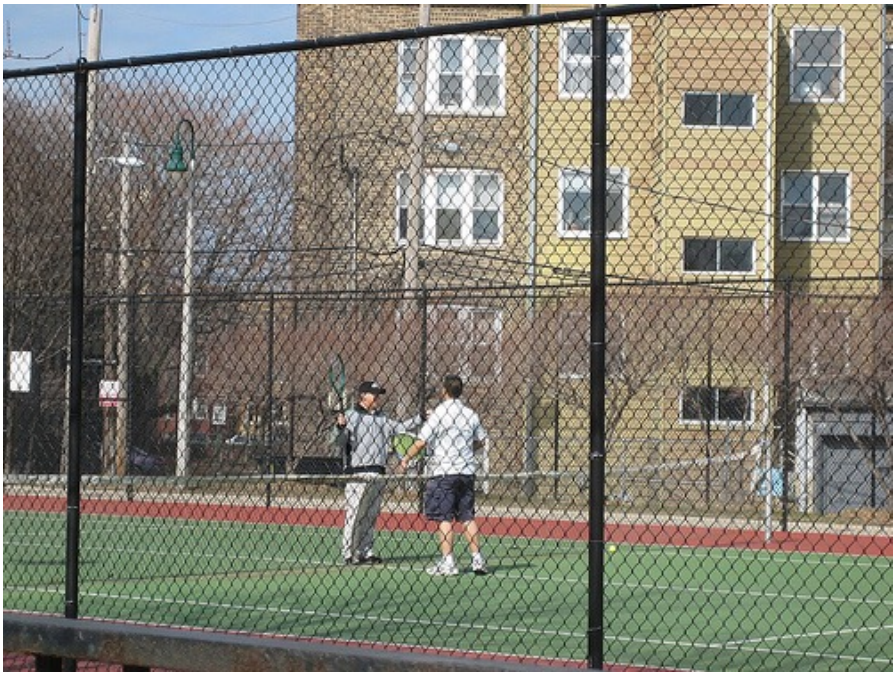


Figure 3.4. – Examples of different classes of the Flower-102 dataset, from Nilsback et al. 2008

object categories are not at all considered, depending on the design choices of the dataset authors. For instance, clouds are rarely considered as objects. In the case of the MSCOCO dataset, objects are defined by a list of 80 classes ranging from vehicles to animals. The linear regression model $P_{\text{V-ObjectCount}}$ also uses the [CLS] token representation. As there can be clues in the caption indicating how many objects are in the image, some multimodal information present in the representation can be used for this task, which makes the use of a baseline important. Indeed, another variation of this task could have been considered: counting objects that are both in the caption and the image. This would have made the task more language-based. UNITER and LXMERT use visual representations based on object regions. Through this task, we aim to verify whether it helps them apprehend the structure of a scene better than other types of models such as ViLT.

Multimodal probing tasks: M To evaluate the multimodal information present in vision-language representations, we evaluate multimodal *Denotation* capabilities. However, as evaluating all possible capabilities can be time-consuming, we restrain



Task	Text Input
M-Color	Two men standing behind a tall [MASK] fence.
M-Size	Two men standing behind a short black fence.
M-Position	Two men standing in front of a tall black fence.
M-Differences	Two men running behind a tall black fence.

Figure 3.5. – Example of modified captions for the multimodal probing tasks, using the caption 'Two men standing behind a tall black fence' as the original (image from Flickr30k).

ourselves to a few significant ones that can be more easily adapted from existing datasets. To that end, we study the understanding of color, size, and position. We create datasets to evaluate the understanding of those attributes.

To compensate for the fact that we cannot evaluate all *Denotation* capabilities, we also create a more general task. This task is not linked to a specific capability. To that aim, in addition to the three attribute-specific tasks, we create a task that assesses how well the model captures minimal, linguistically likely differences in two multimodal instances.

The creation of the four tasks consists in altering the caption of half of the instances to create negative examples. Then, we evaluate the performance of a model in distinguishing positive and negative examples. The probing datasets are carefully designed to avoid textual bias. Figure 3.5 lists changed captions for the multimodal tasks with an example picture.

— **Color Identification (M-Color):** This task mainly evaluates two *Denotation* capabilities: the *Local* capability relating to *Basic Property Detection* in order to

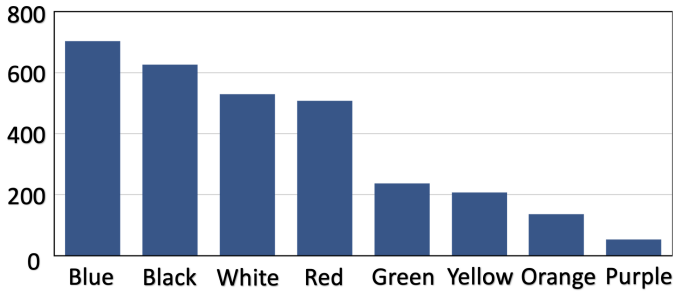


Figure 3.6. – Colors distribution for task M-Color

understand color, and the *Structural* capability relating to *Multimodal Alignment* in order to align a color with the correct caption element. To this end, we select 8 common colors that are unlikely to be ambiguous: blue, red, black, white, yellow, orange, green, purple. A subset of 3,000 instances from Flickr30k that contain those colors is used for evaluation. We do not control for text bias, and use the text-only and mismatched baselines to analyze the results. For each instance, a color word is masked with [MASK] in the caption. The last layer representation of this token by *VL* is used as input of the linear classifier $P_{\text{M-Color}}$ in order to predict the missing color as in the MLM pre-training task. The goal is to check whether vision-language representations associate text and visual features to determine the masked color. Figure 3.6 represents the color distribution for the M-Color task.

— **Size Identification (M-Size):** This task evaluates a *Denotation* capability: the *structural* capability *scene understanding*, and specifically the understanding of the size of objects. Indeed, this task aims at assessing if object size is a multimodal concept included in *VL* representations. We want to force the probing model to use multimodal cues instead of textual bias for this task, so we build the dataset to minimize the possibility of using only linguistic cues. Instances are selected if their caption contains size adjectives (large, big, long, tall, or small, little, short, narrow). Then, we select among those captions 56 concrete object categories (e.g.: car, tree, bird). Those are present in the test set with opposite size adjectives (i.e., large vs. small) subject to a relatively balanced prior . To ensure this balance, the least frequent variant represents at least 10% of the occurrences in the subset. For example, there are no examples in the dataset describing a rock as 'small' while more than 100 describe one as 'large', leading this category to be left out. By comparison, if we compare 'small' and 'large' dogs, 37% of dogs are 'large'. This ensures that the model has at least limited ability to exploit the object category bias to determine its size. We then manually create negative instances by switching the adjective with its opposite. The resulting dataset is a subset of 2,552 instances of Flickr30k. A linear binary classifier $P_{\text{M-Size}}$ is trained to determine if the caption has been modified, using the [CLS] token representation.

— **Position Identification (M-Position):** This task evaluates a *Denotation* capabil-

ity: the *Structural* capability *Scene Understanding*, and specifically the understanding of the position of objects. Indeed, this task aims at assessing if object position is a multimodal concept present in vision-language representations. For this task, we also minimize the possibility of exploiting linguistic bias. Captions are selected based on their use of positional expressions (bottom, top, inside, outside, left, right, up, down, towards, away from, over, under, behind, in front of). Then, we select among those captions 16 different contexts where an expression and its opposite are both present in the dataset in similar proportions. For example, the top/bottom pair is unbalanced since there are 2,362 occurrences of top and 161 occurrences bottom in the dataset. There are 177 occurrences of 'at the top' and 78 occurrences of 'at the bottom', which is more balanced. To ensure relative balance, we select expressions where the least frequent variant represents at least 30% of the occurrences in the subset. This ensures that the model will need to rely on the image to determine the position. The negative instances are created by switching an expression with its opposite. The resulting dataset is a subset of 2,626 instances from Flickr30k. We train a linear binary classifier $P_{\text{M-Size}}$ to determine if the caption has been modified using the [CLS] token representation.

— **Minimal Differences (M-Differences):** This task does not evaluate a specific capability: It is based on the task of Image-Text Matching, in the way that it consists in determining if a caption matches an image. However, the examples of this dataset are crafted in order to be challenging. To that end, we use a subset of the MS-COCO dataset. For each caption from an MS-COCO subset, we automatically select words corresponding to visually relevant grammatical categories (nouns, verbs, adjectives, numbers). For each of those words, a new caption is created with minimal differences, as only that specific word changes from the original caption. Then, a likely replacement is selected from the top of the distribution output by the text-only BERT model. For instance, the word can be a noun: 'A close up of a *car* on a dirt ground with a tree in the background .' replaces 'A close up of a *horse* on a dirt ground with a tree in the background .' In other cases, it is a number: 'A black and white photo of *five* zebras near one another' replaces 'A black and white photo of *two* zebras near one another'. This means that the captions, although wrong, are believable for a language model. This minimizes the possibility for multimodal models to rely on text bias. The adversarial instances are then manually screened for semantic and syntactic correctness prior to inclusion in the dataset. As a result, the words replaced in the test set are mainly object related. They are related to people (15%), or to the 79 other MS-COCO categories (35%) or referring to other objects (26%). They can also be noun or adjectives qualifying objects (10%), verbs (6%), words expressing quantity (6%) and others (2%). Contrary to the other tasks, as BERT is used to generate the captions with minimal differences, the multimodal concepts that are evaluated are diverse. We train a linear binary classifier $P_{\text{M-Differences}}$ to determine if the caption has been altered, using the [CLS] token representation from the last layer.

3.3.3. Experimental Setup

We study three state-of-the-art vision-language models: UNITER (single-stream with Faster R-CNN visual features), LXMERT (dual-stream with Faster R-CNN visual features), and ViLT (single-stream which does not use Faster R-CNN visual features). These models differ in transformer architecture and pre-training tasks. The training protocol of those also models vary, and they are not pre-trained on the same datasets. Our goal is to analyze the results of those models on our probing tasks, to understand how their pre-training process affects the representations learned by those models.

Each VL is studied as a pre-trained model and as a fine-tuned model on fine-tuning tasks $T = VQA$ and $T = NLVR$. We choose these tasks as they differ from the tasks used for pre-training and therefore require non-trivial model fine-tuning. They are also well known when evaluating vision-language models and necessitate fine-grained multimodal understanding. VQA is a visual question answering task, while NLVR2 consists in determining whether a sentence is correct using a pair of images as input. Our goal is to explore the effect of the tasks T on probing task performances. For instance, we aim to know if those fine-tuning tasks help the vision-language models learn some concepts better than the pre-training process.

In addition, we compare the performance of vision-language transformers to the monomodal baselines BERT, ResNet (K. He, X. Zhang, et al. 2016) and ViT (Dosovitskiy et al. 2021). This gives us an understanding of the performance that can be reached using a single modality.

For all three models, we use the available checkpoints for VL_{pre}. For UNITER and LXMERT, we fine-tune the models using authors' instructions, to obtain $VL_{fine(VQA)}$ and $VL_{fine(NLVR)}$. In the case of ViLT, we use the available pre-trained and fine-tuned checkpoints. For the BERT and ViT baselines, we follow the same protocol as the vision-language VL models. Their representations are of dimension 768. For the ResNet baseline, we use the final layer representation, which is of dimension 2048. We use the pre-trained models from Pytorch and Hugging Face (Wolf et al. 2020) for the experiments.

The probing model P is a linear model trained over 30 epochs for M, V and L-BigramShift tasks and 50 epochs for L-Tagging, with a learning rate of 0.001. We use MSE loss to train $P_{\text{V-ObjectCount}}$ and report RMSE as a metric to evaluate V-ObjectCount, and the cross entropy loss for all other probing tasks, with accuracy as a metric. The results of each probing task are averaged over 5 runs. We trained the models on a cuda75-capable GPU.

3.3.4. Results

This section is organized according to the model analyzed and to the modality of the probing task.

3.3.4.1. Pre-trained Models

Table 3.2 shows the results of VL_{pre} representations for Language, Vision and Multimodal probing tasks.

Modality	Task Name	UNITER	LXMERT	ViLT	BERT	ResNet	ViT
Language	L-Tagging	94.66	95.13	***96.27***	95.57	—	—
	L-Tagging♣	90.86	95.36	***96.16***	—	—	—
	L-BigramShift	*80.89*	70.65	72.08	**86.33**	—	—
	L-BigramShift♣	***76.25***	72.22	71.05	—	—	—
Vision	V-Flower	71.82	75.56	*91.34*	—	86.83	**99.66**
	V-ObjectCount	5.49	5.49	***4.90***	6.27	4.96	5.67
	V-ObjectCount♣	7.20	7.31	*5.44*	—	**4.96**	5.67
Multimodal	M-Color	***86.27***	71.21	85.97	37.02	—	41.19
	M-Color♣	34.80	*39.33*	35.69	—	—	**41.19**
	M-Size	57.15	***58.43***	55.45	55.66	—	51.76
	M-Size♣	***56.06***	55.05	52.10	—	—	51.76
	M-Position	***55.92***	54.62	48.95	56.52	—	52.78
	M-Position♣	54.06	***54.68***	52.37	—	—	52.78
	M-Differences	***79.71***	72.60	73.4	53.46	—	—
	M-Differences♣	51.92	***61.25***	56.4	—	—	—

Table 3.2. – Pre-trained models probing: Results of the pre-trained vision-language models on the language, vision and multimodal probing tasks. Reported metric is mean square error for V-ObjectCount (lower is better) and accuracy for all other tasks. ♣ indicates mismatched instances. In bold are the best results, in italics are the best results for vision-language models.

Language (pre-trained) The language probing tasks are part-of-speech tagging (L-Tagging) and bigram shift (L-BigramShift). Results for the L-Tagging task are close for all models. For L-BigramShift, BERT reaches the best results with an accuracy of 86.33, and UNITER has a higher performance than the others vision-language models. We notice that using wrong images as input for these tasks impacts negatively UNITER.

Vision (pre-trained) The visual probing tasks are fine-grained classification (V-Flower) and object counting (V-ObjectCount). We notice that the ViLT reaches significantly better results than both UNITER and LXMERT models. V-Flower is an image-only task, so we use an empty caption. On the V-Flower task, ViLT is better than the ResNet baseline.

On the V-ObjectCount task, the metric symbolizes the average object count error for different models. The results show that using the associated caption significantly improves the object counting results. It shows that VL_{pre} models use linguistic cues for V-ObjectCount. The performance of ViLT drops when using mismatched captions,

but it remains better than the vision-only baseline ViT. UNITER and LXMERT, however, barely reach this performance using the right caption.

Multimodality (pre-trained) Table 3.2 shows the results for the four multimodal probing tasks.

On the color prediction (M-Color) and adversarial examples (M-Differences) tasks, vision-language models reach much higher results than the monomodal baselines. UNITER and ViLT have better performance than LXMERT for M-Color, with an accuracy of 86.27 and 85.97, while LXMERT reaches 71.21. UNITER also has significantly better results than the other two models for the M-Differences task. We notice that LXMERT has better results when using the wrong image, on the M-Differences♣ and M-Color♣ tasks. It seems that the LXMERT performances are both lower and more dependent on linguistic cues than UNITER and ViLT, which extract more visual information.

For the M-Size and M-Position tasks, UNITER and LXMERT yield similar results, while ViLT shows the worst results on those tasks. However, all results are close to the monomodal baselines. It seems to show that vision-language models have a hard time extracting visual information related to size and position. On these tasks, it seems that bias in text data is linked to the performances of the models. Thus, the concepts of size and position seem not to be very well understood at a multimodal level by pre-trained vision-language models.

3.3.4.2. Fine-tuned Models

Table 3.3 shows the results of the $VL_{fine(VQA)}$ and $VL_{fine(NLVR)}$ models for the visual, textual and multimodal probing tasks.

Language (fine-tuned) We notice that fine-tuning negatively impacts model performance on L-BigramShift, and especially for LXMERT. For L-Tagging, all fine-tuned models except LXMERT$_{fine(NLVR)}$ have similar performances to pre-trained models.

The performances of UNITER with wrong images are the only one which show an improvement, reaching the level of their respective 'normal' tasks. It seems to show that the gap in performance of UNITER$_{pre}$ for mismatched instances is due to a specificity of its pre-training protocol.

The lower performances for the NLVR fine-tuned models could be because the NLVR task is used to having two images as input, contrary to pre-training and probing tasks. The lower performance of fine-tuned LXMERT models could show that LXMERT forgets, more easily than the others, the linguistic knowledge it has learned through pre-training.

Vision (fine-tuned) For the V-Flower task, we notice an improvement of the fine-tuned UNITER models compared to the pre-trained models. ViLT performances were already high, and decreased slightly. However, LXMERT only improves with VQA fine-tuning.

Modality	Fine-tuning	Task Name	UNITER	LXMERT	ViLT	Baseline
Language	VQA	L-Tagging	93.84	94.14	94.79	**96.27**
		L-Tagging♣	93.80	94.73	94.89	**96.16**
		L-BigramShift	79.48	65.40	69.43	**80.89**
		L-BigramShift♣	**76.92**	62.74	68.32	76.25
	NLVR	L-Tagging	94.37	88.44	95.60	**96.27**
		L-Tagging♣	94.38	88.49	95.50	**96.16**
		L-BigramShift	72.74	57.10	67.18	**80.89**
		L-BigramShift♣	72.34	57.82	67.10	**76.25**
Vision	VQA	V-Flower	82.91	78.80	**93.11**	91.34
		V-ObjectCount	4.98	5.13	5.20	**4.90**
		V-ObjectCount♣	6.49	6.85	5.87	**5.44**
	NLVR	V-Flower	82.78	74.23	91.23	**91.34**
		V-ObjectCount	4.95	5.65	4.92	**4.90**
		V-ObjectCount♣	6.22	6.94	5.50	**5.44**
Multimodal	VQA	M-Color	83.39	82.60	81.23	**86.27**
		M-Color♣	35.75	37.18	33.49	**39.33**
		M-Size	**64.23**	60.85	58.62	58.43
		M-Size♣	55.96	**56.60**	55.24	56.06
		M-Position	**57.55**	56.25	53.43	55.92
		M-Position♣	**55.27**	54.51	52.84	54.68
		M-Differences	78.37	74.90	70.07	**79.71**
		M-Differences♣	49.42	61.06	52.27	**61.25**
	NLVR	M-Color	82.52	78.00	83.18	**86.27**
		M-Color♣	36.17	37.02	33.83	**39.33**
		M-Size	**63.19**	59.28	54.20	58.43
		M-Size♣	**59.28**	54.57	53.17	56.06
		M-Position	**56.99**	56.23	52.80	55.92
		M-Position♣	54.42	**55.09**	53.31	54.68
		M-Differences	77.50	68.46	68.12	**79.71**
		M-Differences♣	53.17	53.46	57.42	**61.25**

Table 3.3. – Results of the fine-tuned vision-language models on textual, visual and multimodal probing task. Reported metric is mean square error for V-ObjectCount (lower is better) and accuracy for all other tasks. ♣ indicates mismatched instances. Baseline column reports the best result for a pre-trained vision-language model for the same task.

On the V-ObjectCount task, UNITER and LXMERT also show improvements. UNITER fine-tuned models reach the performance of the ResNet baseline with 4.98 for UNITER$_{fine(VQA)}$. However, LXMERT$_{fine(NLVR)}$ is also worse than the pre-trained model for this task. Additionally, the results using the wrong caption also improve, showing that the increase in performance relies partly on a better extraction of visual information.

Fine-tuning improves the vision performance of UNITER and, to a lesser extent, LXMERT. This seems to show that VQA and NLVR rely on visual information that is not linearly accessible within the pre-trained models. On the other hand, it seems that fine-tuning does not improve the visual capacity of ViLT, which was already similar

in terms of performance to the visual baselines for the pre-trained model. It shows that the vision performances of UNITER and LXMERT pre-trained models seem to be lacking, which could point out that the visual pre-training of those models is a limiting factor. Our hypothesis is that it is easier to extract information from the textual input than the Faster R-CNN features, making UNITER and LXMERT rely more on text than image.

Multimodality (fine-tuned) On the color (M-Color) and adversarial (M-Differences) tasks, we notice that fine-tuned UNITER and ViLT models have slightly lower performances than their pre-trained counterpart, while LXMERT shows generally an increase in performance, except LXMERT$_{fine(NLVR)}$ for the M-Differences task. For LXMERT especially, VQA fine-tuning leads to better performances than NLVR fine-tuning. UNITER remains the overall best model for those tasks.

For the size (M-Size) and position (M-Position) tasks, we notice a slight increase in performance for all models. This is more noticeable for the M-Size task, while M-Position results remain close to the mismatched image baseline. ViLT has the worst results on those tasks. The improvement on these tasks could be because fine-tuning datasets are more focused on the concepts of size and position than pre-training datasets. These results seem to show that the models, and UNITER in particular, manage to extract additional visual information relevant to size, but also keep using textual bias. This seems to show that fine-tuning models using datasets related to position and size can increase their reliance on textual bias linked to these concepts, while also increasing their multimodal performance.

3.3.5. Discussion

Language capabilities Language-oriented probing seems to show that vision-language models have slightly worse syntactic understanding than language-only models such as BERT. This could be due to the less varied syntactic structure of the captioning datasets used for pre-training. UNITER shows overall better performances. Our hypothesis is that UNITER is less robust to non-parallel image/text data because all pre-training tasks use parallel data, while LXMERT also pre-trains with VQA.

Vision capabilities Vision-oriented probing seems to show that visual pre-training is a limiting factor for vision-language models based on Faster R-CNN features, as UNITER and LXMERT show significantly worse performance than ViLT. We think that the models rely on textual information because they cannot extract accurate visual information from the representation. This limiting factor is consistent with what has been found in other studies. Indeed, VinVL P. Zhang et al. 2021 shows that a better object detection model leads to better downstream tasks results.

Multimodal capabilities Multimodal probing shows that pre-trained vision-language models are able to capture some multimodal information, with UNITER reaching

the best performances. While ViLT has shown better results on visual probing than UNITER, this has not translated to the multimodal probing tasks. In particular, its weaker performance in the M-Differences task could be due to the absence of an object prediction task during the pre-training of ViLT. This could limit its semantic understanding of objects.

However, concepts related to object size and position are still not well understood by those models. These are harder to grasp because they are relatively subjective and depend on the context and annotator. For those concepts, the models still almost exclusively rely on linguistic cues, resulting in a performance drop when they cannot rely on textual bias. In additional ablation studies, we use non-curated size and position datasets to see how the models perform when there are more linguistic clues. We notice that on this dataset, UNITER pre-trained representations reach an accuracy of 71.66 on the M-Size probing task, and of 65.69 when using wrong images. For the M-Position probing task, the model reaches 73.18 using the right images and 72.68 using the mismatched images. This shows that using linguistic cues is helpful on these tasks on less controlled datasets. The performance of the position task seems to show that visual information regarding this concept is even less accessible in representations than size-related information. It could show that the current visual pre-training is not enough to understand the positional relationship between objects at a multimodal level. This is especially true for ViLT, which shows the worst performances on those tasks.

Impact of fine-tuning Contrary to our expectations, fine-tuning does not necessarily lead to better cross-modal probing performances. The improvements in performance on probing tasks are specific and not consistent from one model to another. This seems to point out that architecture and model pre-training are particularly important to understand multimodal concepts. Also, concepts that are not well understood by a pre-trained model will not have much improvement with fine-tuned models. This may also be due to the relatively small size of fine-tuning datasets compared to pre-training datasets. One other aspect is that fine-tuning datasets also generate textual bias that can add to the already existing textual bias of pre-trained models.

Impact of linguistic bias Finally, our results seem to show that for some concepts, models prioritize text other vision and multi-modal information. Indeed, multimodal performance is similar to monomodal performance in tasks such as position or size understanding. As it is, vision-language models are susceptible to textual biases present in the data, and do not seem to extract visual information. This seems to show that overreliance of a model on linguistic clues is detrimental to the multimodal understanding of this model. Future work related to vision-language evaluation should be cautious to take into account the presence of textual bias, by carefully balancing the dataset and/or including monomodal baselines.

Limitations The tasks we created for the evaluation of monomodal and multimodal capabilities have several flaws. For instance, the multimodal probing tasks are created by changing a caption to provide a negative image-text pair. This could induce bias in the creation of the data. In addition, the position and size datasets are too difficult. Indeed, it is difficult to compare the performances of the three vision-language models. Finally, the use of models pre-trained using different protocols limits our ability to pinpoint the cause of potential weaknesses.

The visual pre-training of vision-language transformers seems to be weaker than their monomodal counterparts. In particular, the results in this section indicate that future work could focus on the adaptation of vision-language pre-training to fine-grained multimodal concepts such as position and size. In the next section 3.4, we investigate more precisely the reasons behind their poor multimodal understanding of position. We question if this is due to the fact that models do not extract the necessary visual information, or whether it is due to the multimodal aspect of positional understanding.

3.4. Probing Positional Understanding

In the previous section, we have shown that vision-language models do not reach a better understanding of position and size than monomodal models on these probing tasks. Several hypotheses could explain that result.

— The complexity of the probing dataset: the use of real-world images could be too complex for current vision-language models. In particular, some position and size-related are subjective or require an understanding of 3D space, such as 'small' or 'behind'.
— A difficulty to extract visual information related to spatial understanding. The pre-training of vision-language transformers may not help them extract relevant visual information related to size and position in their last layer representations. This could mean that the encoding of position used by the models, or the visual representations used as input of the models, are not appropriate.
— A difficulty to combine visual and textual information related to size and position. This could mean that the datasets and tasks used during vision-language pre-training are not sufficient to learn a multimodal understanding of size and position.

In this section, we aim at providing insight on the lack of spatial understanding of vision-language models. To that end, we build new probing tasks to evaluating the *Scene Understanding* capability, and more specifically the spatial understanding of vision-language.

3.4.1. Methodology

The use of real-world data restricts our control of this data, in particular concerning the position and size of those objects. In addition, human annotations used in the previous study can be biased and lack in precision. To study more precisely the multimodal understanding of position and size, we build a new dataset of synthetic images. This dataset is based on CLEVR (Johnson et al. 2017). We also automatically generate captions to create several multimodal probing tasks. We evaluate UNITER (Y.-C. Chen et al. 2019), ViLT (W. Kim et al. 2021) and LXMERT (Tan et al. 2019) on these tasks to grasp their multimodal understanding of spatial relationships. In particular, our goal is to study the role of visual information in their multimodal understanding of position. We use a similar experimental setup as the previous section.

Understanding of size The pre-training data used for vision-language models is particularly biased regarding the description of the size of objects. Indeed, the adjectives used to qualify size are particularly unbalanced in different contexts. For an annotator, the main reasons to refer to object size are:

— If the size of an object in the image is especially unusual with respect to the annotator's frame of reference for this specific object. Annotators rarely describe the size of an object when it is considered ordinary.
Ex: 'A small plane'

— If the noun and adjective pair is part of a commonly used expression.
Ex: 'A little girl'

— To emphasize the size of an object, i.e., when size is an important aspect of this object.
Ex: 'A huge skyscraper.'

In the first case, the most commonly used adjective is the one that is unusual compared to the common size distribution of a given object. In the two last cases, the opposite effect is visible, and the objects are usually only used with an adjective corresponding to the usual size of this object. For all cases, the size adjectives are not evenly distributed, and make it difficult for a model to understand the overall concept of size.

As a result, the multimodal understanding of size is hindered by the subjectivity and bias of the dataset, which reflect our own use of size-related words. This leads vision-language models to rely too much on textual context, and not enough on visual information in the case of size.

Understanding of position The understanding of position is less affected by the bias real-world data. In particular, annotators will have no preference in the use of 'left' and 'right' when describing the position of objects. As a result, we aim in this section to understand why vision-language models fail at multimodal position understanding. We focus on the position of objects in an image and relative to other objects. Indeed, other variations, such as the understanding of 3D can also be evaluated. In this section, we use synthetic data to have better control over the positioning of objects.

Model	UNITER	LXMERT	ViLT
Color	**88.4**	**80.0**	**91.1**
Material	62.9	54.9	58.5
Shape	68.0	55.6	63.0

Table 3.4. – Accuracy of UNITER, LXMERT and ViLT on a task evaluating the multimodal understanding of attributes. Task performed on a synthetic image dataset inspired by CLEVR.

3.4.2. Datasets

We study the multimodal spatial understanding of LXMERT, ViLT and UNITER, using a dataset of 5714 synthetic images inspired by CLEVR. Each image is made of one or two simple objects, as shown in Figure 3.7. Objects are composed of several attributes (color, material, shape) that can be used to describe and distinguish them from other objects in the image. As our goal is to evaluate the understanding of position of objects, we first have to decide what attribute to use to refer to an object. To that end, we evaluate the multimodal understanding of the color, size, and shape attributes on synthetic images for UNITER, LXMERT and ViLT. The concept of color is the one that the models manage to grasp best. We show in Table 3.4 that the model understands the color attribute better than material or shape. Thus, we choose to use the concept of color to refer to an object in the textual description. Consequently, in the probing datasets used for our experiments, when two objects are present in an image, their color is necessarily different.

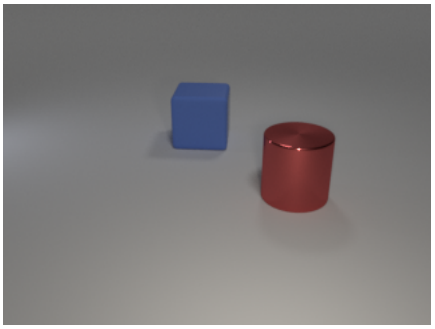


Figure 3.7. – Example of a synthetic image with two objects

Spatial relationships We create several probing tasks to test the spatial understanding of these models at different levels of text anchoring in the image. Several spatial relationships are considered:

— The position of an object in the reference frame of the image ('left/right' and 'front/back');

— The position of an object relative to another object in the reference frame of the image ('left/right' and 'front/back');

— The distance between two objects in the reference frame of the image('far/near').

The position of objects is annotated precisely using pixel coordinates, so that images can be selected and descriptions can be automatically generated according to a template. When we consider the position of an object in the reference frame of the image (first case), thresholds are defined to mark whether an object is to the left, right, front, or back. A margin is used to eliminate ambiguous situations, i.e., objects too close to the center of the image. The same type of margin is used to eliminate ambiguous situations in the two other cases.

Probing tasks Each image in the dataset is associated with a text description in English, which is automatically created according to a template. These descriptions vary according to the tasks under consideration. Thus, the following probing tasks are carried out:

— *Monomodal understanding of the position of an object in the reference frame of the image*: These are two binary classification tasks evaluating the position of a single object in the reference frame of the image. The classes considered are (left, right) for the first task and (front, back) for the second. The use of visual information is sufficient for this task, and textual information is not necessary.
Example caption: 'The object is blue.'

— *Monomodal understanding of distance between two objects*: This is a binary classification task evaluating the distance between two objects in the image. The classes considered are (far, near). All captions use the same adjective to describe distance between objects ('far'). The use of visual information is sufficient for this task, and textual information is not necessary.
Example caption: 'The yellow object far from the blue object.'

— *Multimodal understanding of the position of an object in the reference frame of the image*: These are two binary image-text-matching tasks that evaluate the position of a single object in the reference frame of the image, for two-object images. The model evaluates the position of the object referenced in the caption. The textual description refers to the object by its color. The model uses textual information to identify the object, and visual information to determine its position.
Example caption: 'The object is green.'

— *Multimodal understanding of the relative position of two objects*: These are two binary image-text matching tasks to assess whether a caption correctly describes the relative position of two objects. Visual anchoring of the text is necessary. The model must use visual information to determine position and associate it to the relevant objects.
Example caption: 'The yellow object is to the right.'

— *Multimodal understanding of the distance between two objects*: This is a binary image-text matching task to assess whether a caption correctly describes the distance between two objects. Image anchoring of the text is necessary, as the model must use visual information to check whether the textual description of the distance is true.

Example caption: 'The red object is next to the black object.'
Experiments are carried out using a linear probing model. The results are an average of 5 trials with different seeds.

Description	Nb Objects	Type	Binary Classification Labels	Size
Position in the reference frame	1	Monomodal	Left/Right, Front/Back	(1287, 1254)
Position in the reference frame	2	Multimodal	Matching, Non-matching	(905, 805)
Relative position of two objects	2	Multimodal	Matching, Non-matching	(1057, 923)
Distance between two objects	2	Monomodal	Close/Far	585
Distance between two objects	2	Multimodal	Matching, Non-matching	585

Table 3.5. – List of positional probing tasks. Size is counted in number of image sentence pairs, each image is unique. For datasets relating to position, the size is reported for the (left/right, front/back) couple of datasets.

Table 3.5 summarizes the datasets created for the experiments.

3.4.3. Results

Monomodal understanding of the position of an object in the reference frame of the image We observe that the models have a good monomodal understanding of an object's position in the image reference frame. Their performance is slightly inferior to that of monomodal model ResNet (K. He, X. Zhang, et al. 2016), as shown by the results in table 3.6. We can see from these results that all models succeed in extracting position information from visual data. LXMERT outperforms the other models.

Task	ViLT	UNITER	LXMERT	ViT	ResNet
Left/Right	83,94	93,21	96,76	90,66	**99,4**
Front/Back	97,04	92,64	98,61	98,47	**99,31**

Table 3.6. – Evaluation of monomodal understanding of object position in the image reference frame (Accuracy).

Multimodal understanding of the position of an object in the reference frame of the image Vision-Language models have worse performances in multimodal absolute position understanding tasks, but outperform monomodal references (Table 3.7). UNITER has significantly worse performances than the other models. [3].

Multimodal understanding of the relative position of two objects The task of multimodal understanding of the relative position of an object shows the weaknesses of the multimodal anchoring of Vision-Language models. Indeed, the table 3.8 shows

3. The data for the this classification is unbalanced, which explains why ViT and ResNet perform better than 50%.

Task	ViLT	UNITER	LXMERT	ViT	ResNet
Left/Right	65,92	56,09	**73,12**	52,8	50,44
Front/Back	**92,23**	82,28	89,31	80,22	75,86

Table 3.7. – Evaluation of the multimodal understanding of object position in the image reference frame (Accuracy).

that the results of these models are no better than monomodal baselines. LXMERT is the only vision-language model to outperform monomodal baselines significantly, in the case of the (left/right) image text matching task.

Task	ViLT	UNITER	LXMERT	ViT	ResNet
Left/Right	46,48	49,83	**71,14**	49,83	50,74
Front/Back	**55,75**	50,04	50,67	52,58	48,67

Table 3.8. – Evaluation of multimodal understanding of the relative position of one object to another (Accuracy).

Monomodal and multimodal understanding of the distance between two objects We compare the results of monomodal and multimodal evaluation of the distance between two objects in table 3.9. These results show that the Vision-Language models succeed in extracting the visual information needed to evaluate the distance between two objects, even if the results obtained are inferior to the monomodal visual references ViT and ResNet. However, multimodal evaluation shows that there is no anchoring of distance-related words in the image, leading to poor multimodal performances.

Model	ViLT	UNITER	LXMERT	ResNet	ViT
Monomodal evaluation	66.44	70.10	74.23	**92.79**	66.92
Multimodal evaluation	**52.50**	50.77	51.44	**52.60**	44.42

Table 3.9. – Monomodal and multimodal evaluation of the distance between two objects. For the first task, the text does not use distance-related words, only the image is relevant to provide positional information. The second task involves assessing whether a caption (e.g., 'the red object is next to the blue object') is true (compared with 'the red object is far from the blue object') (Accuracy).

3.4.4. Discussion

Monomodal and multimodal positional understanding These results show that the models succeed in extracting visual information relating to the position of an

object in the reference frame of the image. However, they have more difficulty in combining it with textual information relating to position. As a result, their performances in monomodal position evaluations are significantly better than those in multimodal position evaluations. In particular, tasks assessing multimodal understanding of relative positions between two objects are more complex for a model than classifying position of an object within the image frame, as they require an understanding of the compositionality of the instance. These results point towards a possible weakness of vision-language models in their understanding of multimodality because of the compositionality of multimodal instances. Such weaknesses could extend beyond the understanding of position. As a result, future work could study the *Multimodal Alignment* capability of vision-language models. We investigate this capability in the next section 3.5.

The case of distance seems to point that models do not associate the visual information related to position to corresponding words. This lack of understanding could be due to the fact that the concept of 'distance' is more contextual than that of position. Thus, the low amount of data and the variability of contexts involving those words (i.e., 'close', 'far') makes it difficult to associate it to the corresponding visual information.

Impact of pre-training protocol on the understanding of position In addition, LXMERT has a better multimodal understanding of spatial concepts, such as left and right. Unlike UNITER and ViLT, LXMERT uses VQA as a pre-training task by combining several datasets Antol et al. 2015; Hudson et al. 2019; Yuke Zhu et al. 2016. An analysis of this data shows that they contain a significant number of position-related questions, in particular relating to the concepts of 'left' and 'right', which LXMERT manages to anchor relatively well in the image. Thus, these models can extract visual information relating to the position of an object in the image relatively easily, and use it in a multimodal context. However, they have more difficulty extracting textual information relating to the position of objects and anchoring position-related words in the image. These results seem to show that visual information relating to position is correctly extracted by the models, meaning that the poor understanding of the concept of position does not stem from the positional encoding of the models. The differences in performance of UNITER, ViLT and LXMERT seem to show that this is due to the tasks and datasets used during pre-training. A multimodal task with precise annotations requiring position-related reasoning, like VQA used by LXMERT, seems necessary to anchor position-related words in the visual information.

Limitations In this section, we use synthetic data to evaluate monomodal and multimodal position understanding. While it offers us increased control on the position of objects in an image, our data uses limited categories of objects. It would be interesting to investigate whether these findings are robust to other types of image composition (e.g., more complex images) or varied objects categories. One could imagine that more complex images would make the understanding of object position more difficult, in particular if it involves more fine-grained differentiation between objects than

differences in color.

We have investigated why Vision-Language models have poor multimodal understanding of position. By probing the models, we observe that these weaknesses are not primarily due to model architecture, visual representations or poor representation of object position, but to pre-training tasks and data. Future work may therefore investigate the design of such tasks in order to improve multimodal understanding of these concepts.

3.5. Evaluating Noun-Predicate Dependencies

In the previous sections, we have studied how vision-language models grasp, or fail to grasp, monomodal and multimodal capacities. In particular, they have shown difficulty in understanding some multimodal concepts, such as position, and seem to have difficulty understanding the compositionality of instances in that regard. Indeed, we have seen that models perform worse on the position understanding tasks when they also require an understanding of compositionality, as is the case in relative position understanding. In this section, we focus on how well vision-language models grasp compositionality in a more general manner. We build a task to evaluate how well vision-language models understand multimodal dependencies. The results of the first study have shown an overreliance of vision-language models on textual bias when they have trouble combining visual and textual information. As a result, we carefully design the task to ensure that models do not 'cheat' by relying on textual bias.

In this section, we explore how well vision-language models learn predicate-noun dependencies across modalities (see example in Figure 3.8) using the method of pre-training head evaluation. To this end, we create an evaluation set that contains carefully selected images and pairs of sentences with minimal differences. Given an image and two predicate-noun sentences, the models need to find the correct sentence corresponding to the image. Crucially, they can only succeed by taking into account the dependencies between the visual concepts in the image corresponding to the noun and predicate in the sentence. As it has been shown that visual reasoning performance in several tasks can be spuriously augmented by capitalizing on textual biases in the training data Y. Goyal et al. 2017; A. Agrawal et al. 2018; Hendricks, K. Burns, et al. 2018; J. Cao et al. 2020, we counterbalance our evaluation dataset to control for linguistic biases. Code to reproduce the analyses and run the evaluation on new models is publicly available at https://github.com/mitjanikolaus/multimodal-predicate-noun-dependencies.

Evaluation of grounded syntax Akula et al. 2020 test for sensitivity to word order in referring expressions. Similarly, Thrush et al. 2022 study the ability of vision-language models to consider word order by designing adversarial examples that require differentiating between similar image and text pairs, while the text pairs only differ in their word order. Their results suggest that state-of-the art models still lack

Target sentence:
A man is wearing a hat.

Distractor Sentence:
A man is wearing glasses.

Figure 3.8. – We evaluate vision-language models on their ability to track predicate-noun dependencies that require a joint understanding of the linguistic and visual modalities. The task is to find the correct sentence (choosing between the target and distractor) that corresponds to the scene in the image. In this example, the models should connect the predicate 'is wearing a hat' to 'man'. A model that does not track dependencies would judge the distractor sentence 'A man is wearing glasses' as equally likely, as there is a man is the image, as well as a person that is wearing glasses.

precise compositional reasoning abilities, related to the capability of *Multimodal Alignment*.

L. H. Li et al. 2020 study so-called *syntactic grounding* of VisualBERT. They show that certain attention heads of the transformer architecture attend to entities that are connected via syntactic dependency relationships. However, such probing experiments do not necessarily indicate how much a model is actually *using* the encoded information when making predictions.

In our work, we test a range of state-of-the-art models specifically on their ability to track predicate-noun dependencies. Crucially, we test the models in a much more controlled setting compared to previous work. Our setup involves visual distractors as well as a control task, disentangling the challenge of understanding syntactic dependencies from simpler object and predicate recognition. Additionally, we strictly control for any possible linguistic bias by counterbalancing all evaluation examples.

3.5.1. Methodology

We construct an evaluation dataset that is suited for evaluating the sensitivity of visually grounded predicate-noun dependencies in a zero-shot setup.

The data consists of pairs of triplets, and each triplet consists of an Image I, a target sentence S_1, and a distractor sentence S_2. Target and distractor sentences are minimal pairs, i.e., one sentence differs from the other only with regard to either the noun (e.g., 'A girl is sitting.' vs. 'A man is sitting.', Figure 3.9) or the predicate (e.g., 'A man is wearing a hat.' vs. 'A man is wearing glasses.', Figure 3.8).

The images always contain visual distractors, meaning that both the noun and the predicate of the distractor sentence are present in the image. However, distractors do not have an appropriate noun-predicate relationship. For instance, in the sentence 'A man is wearing glasses', there is a man in the image, who is not wearing glasses, and a person wearing glasses, who is not a man. Thus, it is necessary to take into account the *dependency* between noun and predicate to distinguish the target and distractor sentences (Figure 3.8).

Controlling for linguistic biases As mentioned earlier, vision-language models have shown to rely on textual bias instead of using visual information (Y. Goyal et al. 2017; A. Agrawal et al. 2018; Hendricks, K. Burns, et al. 2018; J. Cao et al. 2020). For example, if a training dataset contains more often the phrase 'a girl is sitting' than 'a man is sitting', a model might prefer the caption 'a girl is sitting' during evaluation. This would only based on linguistic co-occurrence heuristics, irrespective of the visual content. In our evaluation methodology, we control for potential linguistic biases by pairing every triplet with a corresponding counterbalanced example where target and distractor sentence are flipped. More specifically, for every triplet (I_1, S_1, S_2), there exists a corresponding triplet (I_2, S_2, S_1), as depicted in Figure 3.9. In that way, a model that omits the use of visual information cannot succeed in the task (Nikolaus and Fourtassi 2021).

Evaluation We evaluate pre-trained models on their image-text matching performance in a zero-shot setting, i.e., without any further training. For each triplet, we test whether the models give a higher similarity score for the correct sentence than for the distractor sentence. We calculate accuracy for each pair, i.e., the model needs to succeed for both the example and the counterbalanced example triplet.

For each pair of triplets $(t_1, t_2) = ([I_1, S_1, S_2], [I_2, S_2, S_1])$, we calculate the following score:

$$f(t_1, t_2) = \begin{cases} 1, & \text{if } s(I_1, S_1) > s(I_1, S_2) \\ & \text{and } s(I_2, S_2) > s(I_2, S_1) \\ 0, & \text{otherwise} \end{cases}$$

where $s(I, S)$ denotes the similarity between an image I and a sentence S. To obtain the similarity score, we use the softmaxed output of the image-text matching pre-training heads of the models. [4]

The final accuracy is the average score over all pairs in the evaluation set. Chance performance is at 25%. Figure 3.10 describes this methodology.

4. For the model CLIP, we feed the image and both sentences at the same time, and obtain a similarity score for both sentences, where $s(I_1, S_1) = 1 - - - s(I_1, S_2)$.

Figure 3.9. – Counter-balanced evaluation: Each triplet has a corresponding counter-example, where target and distractor sentence are flipped.

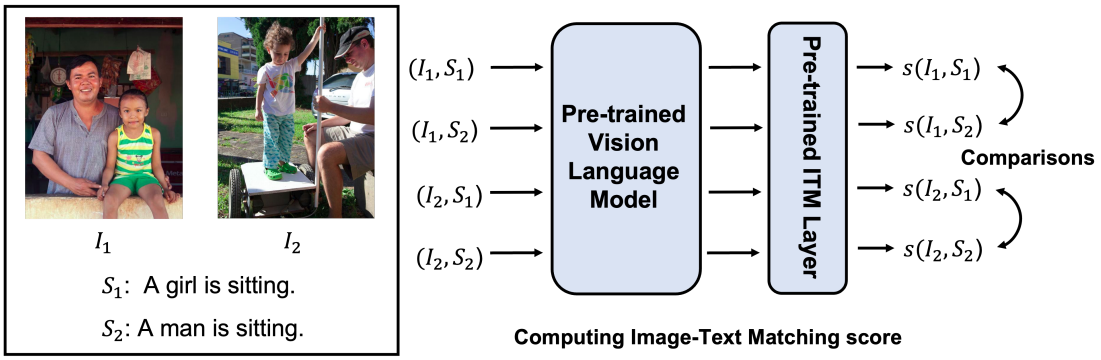


Figure 3.10. – Methodology for the counterbalanced evaluation with vision-language models. The model succeeds if the similarity scores fall into one of four possible configurations. Success: $s(I_1, S_1) > s(I_1, S_2) \wedge s(I_2, S_1) > s(I_2, S_2)$; Failures: $s(I_1, S_1) < s(I_1, S_2) \wedge s(I_2, S_1) < s(I_2, S_2); s(I_1, S_1) > s(I_1, S_2) \wedge s(I_2, S_1) < s(I_2, S_2); s(I_1, S_1) < s(I_1, S_2) \wedge s(I_2, S_1) > s(I_2, S_2)$.

A topline: the cropped task In order to explore the effect of the visual distractors on this noun-predicate dependency task, we additionally evaluate all models in a *cropped* task. We reduce the image to the bounding box of the target object. Thus, the cropped image usually [5] only contains the target object, and no more visual distractors (i.e., the referent of the noun or the predicate in the distractor sentence is no longer present in the cropped image). To succeed at this (simpler) task, the model no longer

5. If the bounding boxes of the target and visual distractor object overlap to a high degree, the cropped image might still contain (parts of) the distractor object.

needs to capture the predicate-noun dependency, it just needs to ground the single words correctly. We use this task to estimate how much the performance of the models is affected by the ability to ground nouns and predicates in our evaluation dataset, in comparison to the (more sophisticated) ability of understanding predicate-noun *dependencies*.

3.5.2. Dataset

Through this task, we evaluate a part of the *Multimodal Alignment* capability, introduced in the structural part of the *Denotation* category presented in the previous chapter. In particular, we focus on the understanding of noun-predicate dependencies. Ideally, the task would be focused on a diverse range of syntactic dependencies. However, due to the available annotations present in current image datasets, we focus on a specific range of noun-predicate dependencies, most concerning human actions. Our hope is that our work is transferable to a more diverse range of noun-predicate and syntactic dependencies.

Automatic pre-filtering Our evaluation dataset is based on Open Images (Kuznetsova et al. 2020). We filter the images based on existing human-annotated object and relationship labels and bounding boxes. The objects refer to people, animals, as well as inanimate objects. The relationships can either describe an action that an object is engaged in (e.g., WOMAN — SIT), or an action linking two objects (e.g., MAN — WEAR — GLASSES). All nouns in the selected relationships for our dataset refer to people, due to lack of sufficient annotations for other kinds of agents.

We look for images that contain a target object-relationship pair as well as a distractor object-relationship pair. Either the target and distractor object are the same, but the relationships differ, or vice versa (as in the example in Figure 3.8). We consider labels that occur at least 100 times in the dataset. As some labels are similar and sometimes used interchangeably by the annotators, we create groups of synonyms for some labels and treat labels within a group as identical in the following. The groups of synonyms can be found in Table 3.10. Further, we verify that the bounding boxes of the target and distractor objects are big enough (at least 20% width and 20% height of the image). In addition, the bounding box sizes of target and distractor objects don't differ by more than a factor of 2. Finally, we ensure that there is at least one counter-example for each triplet before starting the manual image selection phase.

Manual selection We manually select suitable images after the automated prefiltering, in order to ensure high quality of each example and in particular to verify that the distractor sentences are indeed incorrect given the images. This step is crucial, because many of the annotations in Open Images are incomplete. An image may contain, for example, a woman that is sitting but not annotated as such (in this case, we disregard the image for our evaluation set). We select pairs of examples and counter-examples and ensure that there are no duplicate images within the set of images for each object-relationship pair.

'Table', 'Desk', 'Coffee table'
'Mug', 'Coffee cup'
'Glasses', 'Sunglasses', 'Goggles'
'Sun hat', 'Fedora', 'Cowboy hat', 'Sombrero'
'Bicycle helmet', 'Football helmet'
'High heels', 'Sandal', 'Boot'
'Racket', 'Tennis racket', 'Table tennis racket'
'Crown', 'Tiara'
'Handbag', 'Briefcase'
'Cart', 'Golf cart'
'Tree', 'Palm tree'
'Football', 'Volleyball (Ball)', 'Rugby ball', 'Cricket ball', Tennis ball'

Table 3.10. – Groups of label synonyms for dataset creation. Each line corresponds to one group.

Sentence generation We generate target and distractor sentences based on the object and relationship annotations from Open Images.

We construct English sentences using a template-based approach. Given an object and a relationship, we add the indefinite article (a/an) in front of each noun and use all verbs in present progressive tense, as this is most frequent in image-text datasets. In cases where multiple connecting predicates between a verb and a noun are plausible (e.g., 'a man wearing glasses' vs. 'a man with glasses'), we choose the construction that occurs most frequently in the Conceptual Captions training data (Sharma et al. 2018). This dataset is most commonly used for training vision-language transformers. For example, from WOMAN — IS — SIT we generate 'a woman is sitting.'; and from MAN — HOLD — CAMERA 'a man is holding a camera.'.

Linguistic robustness This template-based approach is necessary for our controlled evaluation. However, the choice of the exact template for the construction of the sentences may influence the results. For example, Ravichander et al. 2020 found that results of some probing experiments can vary substantially with slight changes in wording. As such, we evaluate the models, additionally, using a slightly different template. We vary the original templates by using the definite article ('the') at the beginning of sentences, and using verbs in simple present instead of present progressive tense (e.g., 'the woman sits.' or 'the man holds a camera.').

Final evaluation set The final evaluation set contains 2584 triplets. For 1486 of these triplets, the distractor sentence contains an incorrect predicate and for the other 1098 triplets, the distractor contains an incorrect noun. Figure 3.11 shows the number of triplets for each noun and predicate. For a given noun or predicate, we count all pairs that contain this concept in at least one of the two sentences, i.e., cases in which correct understanding of a concept is useful for making the correct decisions.

As the dataset was manually filtered and requires only rather simple understanding of the images, we assume human performance to be close to 100%. To verify this claim, we had a one person annotate a randomly sampled subset of 500 triplets. For each triplet, the annotator was asked to judge which of the two sentences describes the image better. The resulting performance was at 100%.

A note on perceived gender annotations Our evaluation dataset uses annotations from the Open Images dataset, which rely on the appearance of people to annotate their perceived gender. We use the provided annotations, and the resulting biases are unfortunately reproduced in our evaluation set.

In Salminen et al. 2018, gender classification from face pictures by human annotators shows an inter-annotator agreement greater than 95%. True gender cannot be classified, and high inter-annotator agreement does not imply a correct gender choice, but we expect the gender annotations of Open Images to be reliable enough to be used as a basis for our analyses.

3.5.3. Experimental Setup

We consider a range of state-of-the-art vision-language models that are pre-trained using text, image, and multimodal pre-training objectives on corpora of parallel image and text data. All models use the transformer architecture (Vaswani et al. 2017), but vary in terms of pre-training data and objectives, image encoders, and multimodal fusion approaches.

In addition to their image and text pre-training objectives, the models commonly make use of an image-text matching objective, where the models are asked to predict whether a given sentence describes an image or not. We leverage the output of the corresponding pre-training head for calculating image-text similarities for our task. [6]

We evaluate LXMERT (Tan et al. 2019), UNITER (Y.-C. Chen et al. 2019), ViLBERT (J. Lu, Batra, et al. 2019a), Oscar (X. Li et al. 2020), VinVL (P. Zhang et al. 2021) [7], ViLT (W. Kim et al. 2021), and CLIP (Radford, J. W. Kim, et al. 2021). We could not evaluate the original VL-BERT (Su et al. 2020), because of the lack of image-text matching loss. Also, we did not evaluate the original VisualBERT (L. H. Li et al. 2019), as their implementation of the image-text matching loss requires one correct caption (in addition to a possibly faulty caption). This is not available for our dataset. Both models were evaluated in the controlled conditions using VOLTA (see Section 3.5.4.2).

6. The multimodal pre-training objective of Oscar does not involve matching and mismatching images and descriptions, but only matching and mismatching sequences of object tags. Therefore, we evaluate the checkpoint that has been fine-tuned for image-text retrieval.

7. VinVL is pre-trained on parts of the Open Images dataset, and we found that most images used in our evaluation set are indeed part of the VinVL pre-training dataset. Because of this, results of VinVL are not directly comparable to the other models (even though we test the model here with novel sentences with respect to the images). Our results show, however, that VinVL compares very closely to Oscar, which has the same architecture, suggesting that the access to the images during training does not substantially affect the model's performance. This is also possibly the case for CLIP, for which the training data is not public

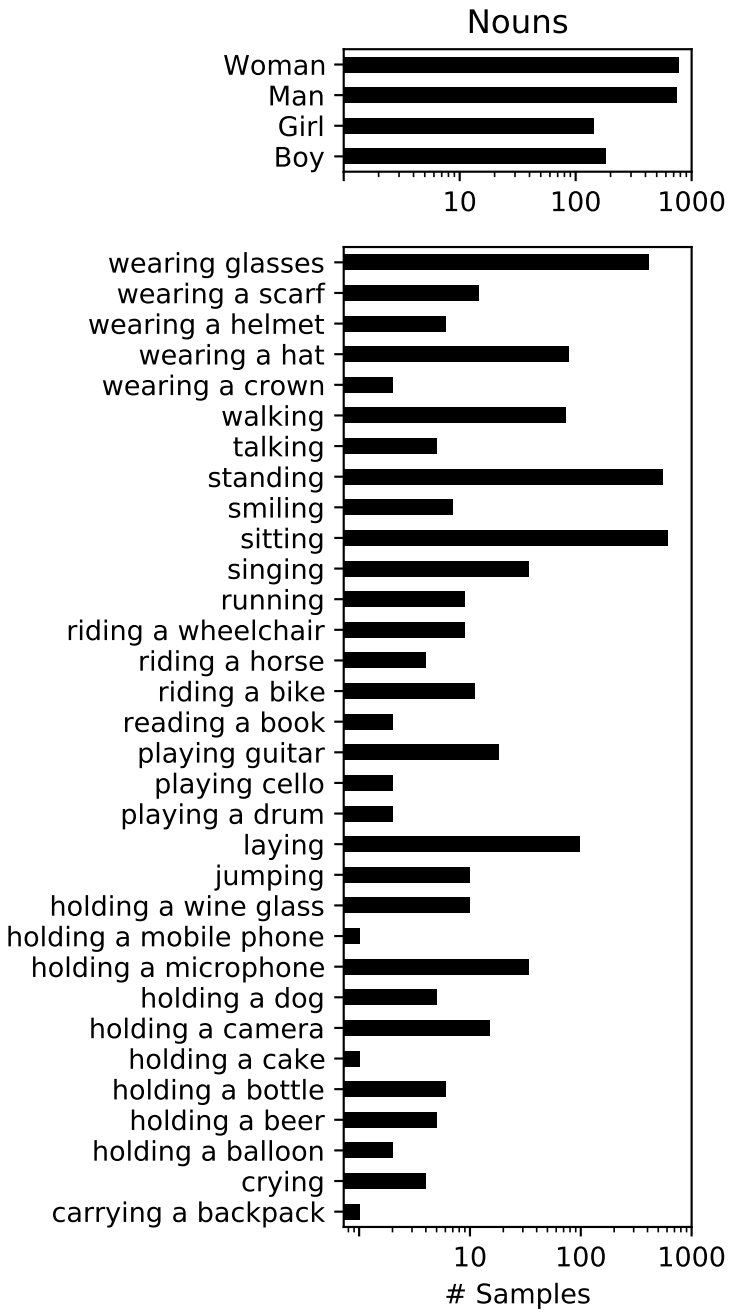


Figure 3.11. – Number of triplets per concept. Note that the x-axis is on logarithmic scale.

Table 3.11 details the multimodal datasets used for vision-language models. Dataset sizes as reported in the corresponding papers. Note that these sizes are also affected by the fact that some models leverage validation sets for pre-training, while others constrain the data to the training sets. Also, different approaches for dataset overlap

detection have been applied. The pre-training data size for CLIP is reportedly 400M image-text pairs, an order of magnitude more data than for most of the other models, which do not surpass 10M image-text pairs in size. The architectures and pre-training tasks of these models are described in more details in Chapter 1.

								Total size	
Model	CC	COCO	SBU	VG	QA	F30K	OI	# images	# image-text pairs
LXMERT		✓		✓	✓			0.18M	9.18M
UNITER	✓	✓	✓	✓				4.16M	9.59M
ViLBERT	✓							3.10M	3.10M
ViLT	✓	✓	✓	✓				4.05M	9.85M
Oscar	✓	✓	✓		✓	✓		4.10M	6.50M
VinVL	✓	✓	✓		✓	✓	✓	5.65M	8.85M

Table 3.11. – pre-training datasets of the tested models: CC (Sharma et al. 2018), COCO (T.-Y. Lin et al. 2014), SBU captions (Ordonez et al. 2011), VG (Krishna, Yuke Zhu, et al. 2016), F30K (Young et al. 2014b), OI (Open Images), and QA (including VQA2.0 (Y. Goyal et al. 2017), GQA (Hudson et al. 2019), and VG-QA (Yuke Zhu et al. 2016)).

Vision-language model implementations We test all models using the evaluation methods and data described above. To test the models, we use two protocols: we test original pre-trained checkpoints made publicly available by the authors: this is the *Original implementation*. In addition, weaker evaluate models that are trained in controlled (and therefore more directly comparable) conditions, as proposed in the VOLTA framework Bugliarello, Cotterell, et al. 2021. In this setup, all models are trained on Conceptual Captions using the same pre-training objectives (masked language modeling, masked object classification, and image-text matching) and use the same image features, extracted from a Faster R-CNN. Not all models are available with this setup. Thus, we evaluate all models for which pre-trained weights are available.

3.5.4. Results

3.5.4.1. Original implementations

The results for the original implementations of the models are shown in Table 3.12. We find that only some models perform substantially above chance, notably ViLT, UNITER and LXMERT. In the *cropped* task, performance is much higher for all models, with VinVL and ViLT reaching the highest performance. This gap in performance between the *full* and *cropped* tasks indicates that while those models can match nouns and predicates in the image with the corresponding words rather well, they struggle to take into account the dependencies between them. This also indicates that performance in the *cropped* task may not be correlated with performance for more fine-grained multimodal tasks.

Model	Full	Cropped
LXMERT	**0.57**	0.69
UNITER	0.54	0.64
ViLBERT	0.28	0.66
ViLT	0.40	**0.75**
Oscar	0.32	0.67
VinVL	0.30	0.76
CLIP	0.20	0.59
Chance	0.25	0.25

Table 3.12. – Accuracy of models trained in original conditions when provided the full images and when only exposed to the target object in the *cropped* task.

3.5.4.2. Controlled training conditions

Model	Full	Cropped
CTRL_UNITER	0.24	0.63
CTRL_LXMERT	0.20	0.56
CTRL_ViLBERT	**0.27**	**0.66**
CTRL_VL-BERT	0.24	**0.66**
CTRL_VisualBERT	0.20	0.64
Chance	0.25	0.25

Table 3.13. – Accuracy of models trained in controlled conditions when provided the full images and when only exposed to the target object in the *cropped* task.

The results for controlled pre-training conditions are presented in Table 3.13. We find that under these controlled conditions, all models perform comparably and generally around chance level. It is therefore not straightforward to draw any conclusions regarding the effect of model architecture from these results. In the *cropped* task, performance is much higher, with ViLBERT and VL-BERT reaching the highest performance. The performance gap between the two tasks (i.e., *full* vs. *cropped*) is substantially larger than for the original implementations, suggesting that the models are even less sensitive to predicate-noun dependencies under these controlled training conditions. ViLBERT performs similarly compared to the controlled conditions. This could be explained by the fact that this implementation is only trained on Conceptual Captions, like in the controlled setup.

3.5.4.3. Control experiments

Model	Full	Cropped
LXMERT	**0.55**	0.70
UNITER	0.54	0.66
ViLBERT	0.26	0.67
ViLT	0.34	0.72
Oscar	0.32	0.65
VinVL	0.30	**0.74**
CLIP	0.21	0.58

Table 3.14. – Accuracy of models trained in original conditions when provided the full images and when only exposed to the target object in the cropped task. Results with alternative sentences.

Linguistic robustness control task Results with alternative sentences are shown in Table 3.14. We find that overall result patterns are highly similar to those with the original sentences in Table 3.12. This shows that the performance of models on this task seems robust when syntactic variations are applied.

CLIP pre-training dataset control task The pre-training data of CLIP is not publicly available. As it was automatically scraped from the internet, we believe the quality (i.e., descriptiveness) of its captions to be comparable to that of Conceptual Captions. In additional, we also study the performance of CLIP models trained on different datasets using a range of publicly available model checkpoints. The performance of CLIP remains below chance level for all tested checkpoints. This might be because all available checkpoints are all trained on rather noisy data, or because the architecture and pre-training objectives of CLIP do not allow it to learn grounded predicate-noun dependencies. Table 3.15 presents the accuracy scores of multiple publicly available checkpoints for CLIP trained on different training data.

Performance for nouns vs. predicates We compare performance for pairs in which the sentences differ with respect to the noun, to sentences with a different predicate in Table 3.16. Overall patterns show a slightly better performance for cases in which the noun was switched, especially in the *cropped* task. This is in line with findings that vision-language models are better at grounding nouns than verbs (Hendricks and Nematzadeh 2021).

Analysis of individual nouns and predicates For a given concept (noun or predicate), we consider all pairs that contain this concept in at least one of the two sentences, i.e., cases in which a model's understanding of a concept is instrumental for making the correct decision.

Figure 3.12 shows the per-concept accuracies of the best performing model, LXMERT.

Visual Encoder	Dataset	Accuracy
RN101	YFCC-15M	0.18
RN101	400M	**0.21**
RN50	cc12m	0.18
RN50	400M	0.20
RN50	YFCC-15M	0.17
ViT-B-32	laion2b_e16	0.21
ViT-B-32	laion400m_e31	0.20
ViT-B-32	laion400m_e32	0.19
ViT-B-32	400M	0.20
ViT-L-14	400M (336px)	0.20

Table 3.15. – Accuracy (Full) of CLIP models with varying visual encoders and pre-training data on the noun-predicate dependency task with full images.

	Full		Cropped	
Model	Noun	Predicate	Noun	Predicate
LXMERT	**0.60**	**0.55**	0.78	0.62
UNITER	**0.60**	0.50	0.76	0.56
ViLBERT	0.27	0.28	0.74	0.59
ViLT	0.44	0.37	0.80	**0.72**
Oscar	0.36	0.30	0.75	0.62
VinVL	0.33	0.28	**0.83**	0.71
CLIP	0.21	0.19	0.69	0.52

Table 3.16. – Accuracy of models for cases in which the distractor sentence contains a different noun or a different predicate. We report scores for all models in their original implementations.

We observe large variation in accuracy scores of predicates, and less variation for nouns. We could not find any simple reasons that explain the predicates' variability. For example, verbs can have good or bad performances (e.g., 'running' vs. 'talking'), and the same can be said for predicates that are composed of both verb and noun (e.g., 'holding a bottle' vs. 'wearing a helmet'). That said, factors that may influence model performance on specific nouns or predicates are further discussed in Section 3.5.4.3.

Additionally, we observe that for some concepts, the models perform better if the concept is the target, and for others, performance is better if it is the distractor. This is, e.g., the case for the pair 'sing' vs. 'stand', where the models consistently perform better if 'sing' is the target predicate.

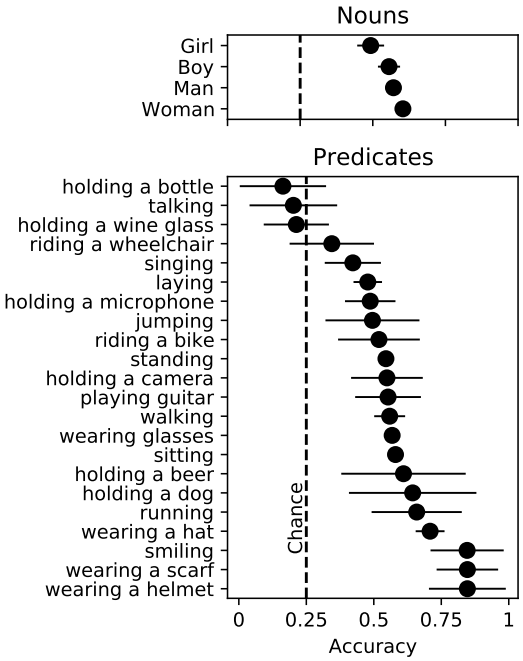


Figure 3.12. – Per-concept accuracies for LXMERT. We display nouns and predicates for which we have at least 10 evaluation triplets. Standard deviation calculated using bootstrapping (100 re-samples).

Confounding Factors We also discuss possible factors influencing the models' performances beyond the pre-training and architecture of the models. In Table 3.17 we show the correlation scores for several confounding factors as described in Section 3.5.4.3.

3.5.5. Discussion

The role of pre-training data Within the set of evaluated models, we do not find evidence for a correlation between the size of the pre-training dataset and the model's ability to capture predicate-noun dependencies. Despite being trained on comparable or even larger amounts of data, ViLT, Oscar and VinVL perform substantially worse than LXMERT and UNITER. CLIP performs below chance level, despite having, by far, the largest pre-training dataset.

Datasets that are composed of highly descriptive captions seem to be advantageous for the learning of noun-predicate dependencies. Indeed, for datasets such as COCO (T.-Y. Lin et al. 2014) or VQA (Antol et al. 2015), the images are not only strongly associated with the captions or question–answer pairs (as they were crowd-sourced specifically for the tasks), but also precise and detailed in nature. In contrast, Conceptual Captions (Sharma et al. 2018) is composed of images with captions that were automatically collected from web pages, and therefore generally rather broad descriptions of the image content.

Model	Bounding box size	Distance to center	Perplexity	OD confidence
LXMERT	-0.03 (p=0.12)	0.03 (p=0.08)	-0.01 (p=0.48)	0.30 (p=0.00)
UNITER	-0.09 (p=0.00)	0.09 (p=0.00)	0.05 (p=0.01)	0.26 (p=0.00)
ViLBERT	0.11 (p=0.00)	-0.16 (p=0.00)	0.05 (p=0.02)	0.22 (p=0.00)
ViLT	-0.01 (p=0.73)	0.03 (p=0.13)	0.05 (p=0.02)	0.26 (p=0.00)
Oscar	0.12 (p=0.00)	-0.17 (p=0.00)	0.06 (p=0.00)	0.15 (p=0.00)
VinVL	0.12 (p=0.00)	-0.12 (p=0.00)	0.05 (p=0.01)	0.04 (p=0.05)
CLIP	0.14 (p=0.00)	-0.24 (p=0.00)	0.08 (p=0.00)	0.17 (p=0.00)

Table 3.17. – Correlations between difference in similarity and various factors related to targets and distractors: difference in bounding box size, distance from the image center of the bounding boxes, perplexity, and confidence scores of the bounding box as calculated using a Faster R-CNN object detector model. OD refers to object detector.

ViLBERT and models trained in the controlled conditions are only trained using Conceptual Captions, and the resulting performances are around chance level. UNITER and LXMERT perform much worse compared to their original training setups. One main difference for these two models in their original implementation compared to the controlled condition is that they are trained on richer datasets with respect to the language modality, leveraging more descriptive captions. The original pre-training datasets for UNITER and LXMERT are also larger in terms of the number of image-text pairs. However, LXMERT is actually trained on much fewer unique images than in the controlled conditions (180K vs. 3.1M), and the datasets used are not web-crawled. Therefore, we assume that the sheer size is not the driving factor of performance. [8]

This observation is coherent with what Hendricks and Nematzadeh 2021 found when studying verb understanding of vision-language models. They compare performance of the same model when trained on Conceptual Captions or COCO. They find that the model trained on COCO performs better, despite Conceptual Captions being bigger and closer to the task in terms of image and language distribution.

These results suggest that, when considering multimodal dependencies, having a high-quality pre-training dataset with less noise and more descriptive textual data could be more important than having a larger dataset. Highly descriptive textual data is essential to learn precise predicate-noun dependencies.

The role of pre-training objectives While models such as ViLT, Oscar, and VinVL are trained on datasets that are comparable in size and quality to those of LXMERT and UNITER, they still perform substantially worse on the task. One explanation could be that contrary to the other models, UNITER and LXMERT both have multimodal

8. Another difference that could cause the performance drop is the choice of pre-training objectives, as discussed in the upcoming paragraph.

pre-training objectives *in addition* to image-text matching: Visual question answering for LXMERT and word-region alignment for UNITER. [9] This could help the models to establish finer multimodal dependencies. Indeed, ViLT and VinVL show better results than UNITER and LXMERT in the *cropped* task (indicating that their object/predicate recognition performance even surpasses that of the other models), but worse results in the *full* task. Our hypothesis is that the pre-training objectives of UNITER and LXMERT enable them to learn more fine-grained multimodal dependencies than ViLT and VinVL, even though their performance on the *cropped* task is worse. Most directly comparable are probably the cases of ViLT and UNITER, which are both trained on the same datasets, but with different pre-training objectives.

This gap in performance should not only be due to the training data associated with the additional pre-training objectives, as VinVL also uses data from Visual Question Answering task, but without training on the objective.

The impact of the multimodal pre-training objectives can be an additional explanation for the drop in performance of CTRL_UNITER and CTRL_LXMERT. Indeed, those were only trained using image-text matching as a multimodal pre-training objective. The gap in performance between those controlled models and the original models indicate that using more precise multimodal pre-training objectives and better annotated datasets can greatly improve the learning of multimodal dependencies.

The lack of suitable multimodal pre-training objectives could also offer an explanation for the poor performance of CLIP in our task.

The role of image encoders The authors of ViLT and VinVL motivate their work by suggesting that improved image features are mandatory for improved multimodal reasoning of vision-language transformers. Here, we observe that these improved features only translate to better results in the *cropped* task (where ViLT and VinVL perform best). We speculate that the improved image encoders allow for a better understanding of visual entities, but not necessarily of the dependencies between them. In order to obtain more conclusive interpretations regarding the role of image features, we require more targeted experiments which control for other confounding factors present here (such as different pre-training objectives).

The role of model architecture In addition to the pre-training objectives, the worse performance of CLIP compared to the other models could also be because it does not support inter-modal fusion of features within the model. Indeed, image and text are processed in separate submodules that prevent inter-modal interaction). This shortcoming of CLIP is also discussed in W. Kim et al. 2021, where the authors find representations from CLIP not to be useful for the more advanced multimodal reasoning task NLVR2 (Suhr et al. 2019).

9. ViLT also uses a word-patch alignment objective similar to word-region alignment. However, the patches are not based on regions detected by an object detector. Therefore, the loss cannot leverage any semantic labels for the patches during training, making this multimodal objective probably less useful.

However, there seems to be no major effect of architecture with respect to multimodal fusion in the case of single and dual stream transformers. LXMERT and UNITER have comparable performances, even though one is a dual-stream transformer and the other a single-stream transformer.

Object salience In most of the images in our evaluation set, the target and distractor people in the image are not of equal size, nor equally salient (sometimes one is more in the foreground than the other). We explore whether there is an effect on the models' decisions by correlating the models' predictions with target and distractor bounding box size and location.

More specifically, we measure the difference in similarity for target and distractor sentence $s(I_1, S_1) - - - s(I_1, S_2)$ and correlate it with the difference in bounding box size of the target and distractor object. Further, we also correlate it with the difference of distances from the center of the image.

For LXMERT, we find no significant correlation (Bounding box size: Pearson $r = -0.03$, $p = 0.16$, bounding box distance to center: Pearson $r = 0.03$, $p = 0.14$). Correlation scores for other models can be found in Table 3.17. While there are statistically significant correlations for some models, these are small and of varying direction. The largest correlations are found for CLIP (Bounding box size: Pearson $r = 0.14$, $p < 0.01$, bounding box distance to center: Pearson $r = -0.24$, $p < 0.01$), indicating that the performance of CLIP could be affected, to some extent, by object salience.

Concept recognizability We also correlate the models' similarity judgments differences to differences in concept recognizability. For concept recognizability, we use the object or attribute confidence score for a given concept in an image from a Faster R-CNN Ren et al. 2015 trained on Visual Genome. If there are multiple objects/attributes with the corresponding label in the image, we take the maximum confidence.

For most models, we find a small positive correlation (see Table 3.17), indicating that the models' similarity judgments are affected by the varying degree to which the concepts are recognizable in the image.

Linguistic biases Another aspect, already mentioned earlier, is that models' performance could be affected by linguistic biases in the training data, such as the frequency and co-occurrence of words and phrases. Indeed, the study presented in Section 3.3 has shown that models tend to have an overreliance on textual bias.

To explore this possible effect, we correlate the difference in similarity for target and distractor sentence with the difference of target and distractor sentence perplexity. We calculate the perplexity for each sentence using a single-modality BERT model (*bert-base-uncased*), that was fine-tuned for 3 epochs on the textual data of Conceptual Captions.

For LXMERT, we find no significant correlation (Pearson $r = -0.01$, $p = 0.48$). For the other models, we find very small positive correlations (see Table 3.17). We conclude

that the models do not rely only on shallow heuristics of the training data in the textual modality.

Limitations The range of concepts evaluated in this section is rather small and therefore not representative for the understanding of grounded predicate-noun dependencies in general. More targeted data collection will be necessary in order to obtain more large-scale evaluation datasets. Additionally, our zero-shot evaluation paradigm introduces a possible mismatch between training and evaluation. Indeed, models are trained using pairs of images and descriptions where the descriptions often describe *all* salient parts of the image, whereas in our evaluation set the descriptions focus on only *one* aspect/person in the image. In the *cropped* condition, the images are not representative of the typical photographic framing of image-text corpora, which could deteriorate our results. That said, random cropping is a frequent data augmentation technique in computer vision research, where it has been successfully applied to improve generalization performance (Krizhevsky et al. 2012).

Our controlled experiments and analyses on a range of recent models reveal that their capability to track such dependencies is variable. Some models (e.g., LXMERT and UNITER) show performance above chance level and others (e.g., CLIP) performing even below chance. In contrast to the recent trends focused on increasing pre-training data and using simple general-purpose pre-training objectives T. Brown et al. 2020; Devlin et al. 2019, we observe that best performance is achieved with high-quality pre-training data and more fine-grained pre-training objectives. More specifically, our results suggest that multimodal pre-training objectives have a major impact on the model's learning of grounded predicate-noun dependencies. Models that include more targeted objectives such as visual question answering and word region alignment in addition to the general image-text matching objective show better performance. In addition, having highly descriptive pre-training datasets seems to help with learning fine-grained multimodal dependencies. In comparison, models trained on larger, web-scraped datasets do not perform well. In future work, the proposed highly controlled evaluation protocol can be used to conduct more targeted studies regarding the role of model architecture, pre-training objectives, as well as training data quality and quantity.

3.6. Discussion on Vision-Language Evaluation

In this chapter, we have investigated the performance of state-of-the-art vision-language models on various monomodal and multimodal tasks, evaluating in particular some *Denotation* capabilities listed in the taxonomy of Chapter 2. These tasks aim to analyze the pre-training of vision-language transformers. The results have shown vision-language models have difficulty to reach a multimodal understanding of position, size, and compositionality, which relate to the structural capabilities of *Scene Understanding* and *Multimodal Alignment*. In addition, results have shown that

some model architectures based on object detectors are detrimental to the extraction of visual information. However, the main factors that impact the understanding of capabilities seem to be the differences in pre-training tasks and datasets. For instance, image-text matching on noisy datasets appears to be insufficient to learn capabilities linked to the understanding of compositionality. In particular, the use of very noisy datasets during pre-training seems to be detrimental to the learning of fine-grained multimodal dependencies. Figure 3.13 summarizes our analyses drawn through this chapter.

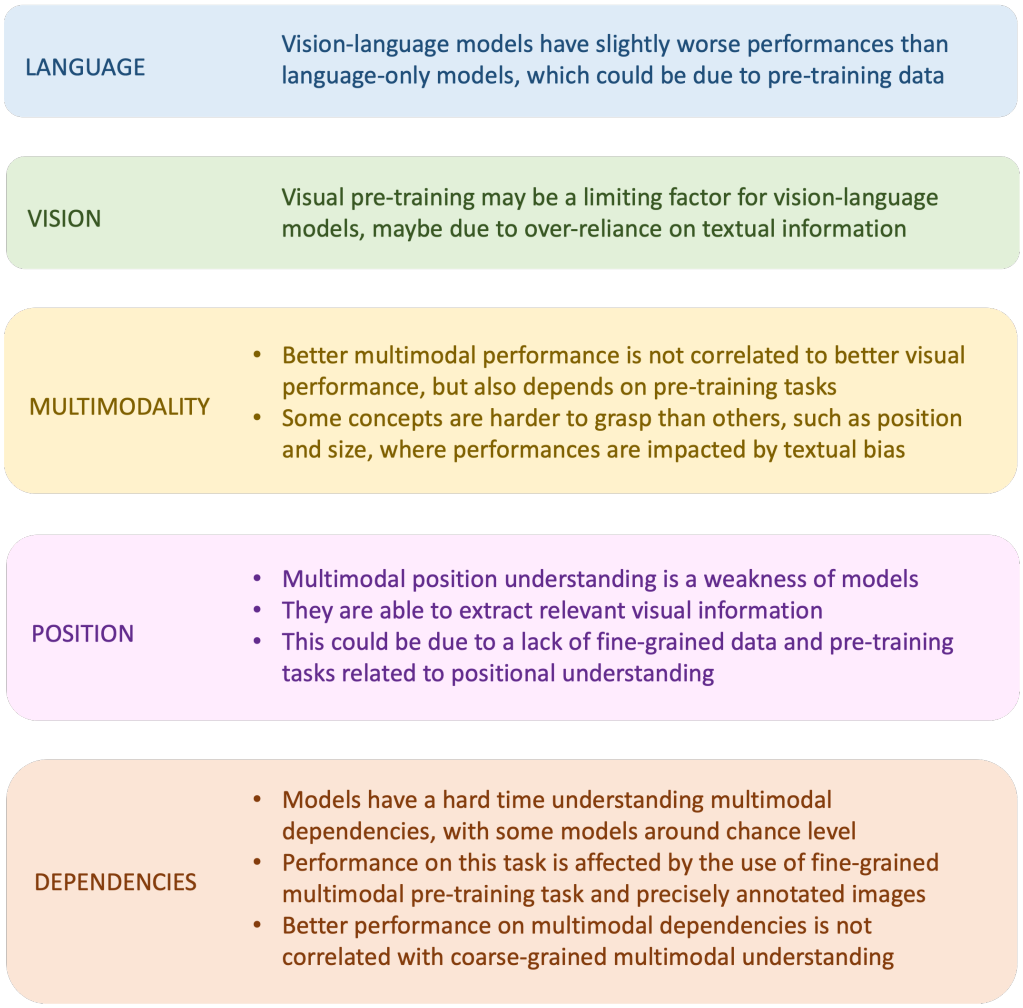


Figure 3.13. – Summary of the evaluation of state-of-the-art vision-language models on multimodal and monomodal capabilities

In this chapter, we have created probing and pre-training head evaluation tasks and datasets based on previous research, in particular in the field of Bertology. Although the methods used in this chapter have enabled us to have more insight on the workings of vision-language models, they have substantial limits. In terms of datasets, we use already available datasets from the web. Some scenes, actions, and cultures are disproportionately represented in our evaluation dataset. As proposed in

F. Liu, Bugliarello, et al. 2021, it is important to pursue further work on more diverse datasets. This point is especially important as we have seen that textual bias can have a significant impact on the predictions of vision-language models.

In terms of methods, we use probing and pre-training head evaluation tasks. First, we have used probing tasks to study vision-language models in sections 3.3 and 3.5. They are useful to explore what information models extract from the data, and can enable comparison of various models trained in different ways, as they only focus on the representations themselves. However, the probing methodology requires training a linear model, which puts constraints on the size of the dataset. In addition, probing models may take advantage of artifacts in the dataset during the training, leading to incorrect interpretation of results. This requires the classes to be carefully balanced in the case of classification tasks, and makes the use of baseline models necessary. Moreover, the results of a probing task are dependent on the type of probing model used. As such, results using a linear probing model can significantly differ from non-linear probing. As a result, the results obtained through probing must be carefully interpreted.

Then, we have also evaluated models on their pre-training tasks, in this case image-text matching, in section 3.5. This method requires fewer data than probing, as no training set is needed, only a test set. In those experiments, models can also rely on spurious correlations due to pre-training dataset biases. Thus, this can lead to a bad interpretation due to the instability of a model. This has been observed with fill-in the blank models relying on the MLM pre-training task (Ravichander et al. 2020). This requires additional robustness evaluations, for example with variations of syntax. This method enables the evaluation of the model on a task, but does not precisely pinpoint possible cases of failure. The cause could be that a model does not extract the necessary information with the deeper layers. It could also mean that its pre-training head does not correctly use all available information. We could reproduce the first two studies using the pre-training head evaluation method. However, the baseline of the $mismatched^{\clubsuit}$ task would not be applicable. Indeed, the pre-training head has not been trained on monomodal instances, which would considerably impact the results. Due to this, the impact of textual bias would solely be evaluated by comparing to monomodal baselines.

Though these methods have flaws, they let us draw potential conclusions on the weaknesses of vision-language transformer models. In particular, we can draw hypotheses regarding what part of the pre-training negatively impacts the models concerning specific capabilities. For instance, these weaknesses may be due to the pre-training tasks used by those models, as well the datasets used during pre-training. As a result, we encourage the use of similar methods, with carefully balanced datasets, to test specific monomodal or multimodal capabilities of vision-language models. This would enable researchers in the fields to have a better understanding of their scope and weaknesses.

For future work, it would be interesting to use those methods to evaluate more recent large vision-language models, such as Flamingo Alayrac et al. 2022 and GPT-4, provided the necessary computing power. For this purpose, access to different model

outputs would be needed, as well as a precise idea of the pre-training protocol used for each model. For vision-language models based on large monomodal models, it would be especially interesting to study their resilience to monomodal bias and their ability to understand fine-grained multimodal concepts. Indeed, by extrapolating results from Section 3.5, we can make the hypothesis that models would improve on coarse-grained multimodal understanding as evaluated by the *cropped* setting. However, their improvement on fine-grained understanding such as the *full* setting, or on concepts such as positional understanding could require more descriptive image-text datasets.

In this chapter, our experiments enable us to develop credible hypotheses. However, they are limited in terms of the comparison of the different vision-language models. Indeed, most models use different architectures and pre-training protocols, rendering their comparison less evident. For instance, this makes it particularly difficult to isolate the effect of pre-training tasks from the effect of the pre-training datasets. A better understanding of all factors influencing model performances could be achieved by training models on comparable conditions, and changing only one factor at a time. In the next chapter, we build on these hypotheses to devise new pre-training protocols and perform ablation studies to understand the impact of each pre-training choice on the performance of a model.

4. Studying Vision-Language Transformer Pre-training

Table of Contents

4.1. Introduction

Through this thesis, we aim to reach a better understanding of the inner workings of vision-language models, through the lens of transformer-based models. In Chapter 2, we argued that a comprehensive evaluation of vision-language models would need to include the evaluation of a wide range of specific capabilities, in order to better apprehend potential weaknesses. In Chapter 3, we studied various monomodal and multimodal capabilities of state-of-the-art vision-language models. Those capabilities belong to what we identified as the *Denotation* category, and are at the basis of many vision-language applications. We have identified how different state-of-the-art models perform on those various capabilities. From those experiments, we made hypotheses regarding limiting factors in their pre-training. In particular, pre-training tasks and datasets seem to have a significant impact on vision-language model performances.

The results we obtained in Chapter 3 are not always comparable. For instance, one of the goals of the experiments is to compare the different pre-training tasks and architectures of pre-trained state-of-the art models. However, the pre-trained state-of-the-art models we evaluate in the previous chapter use different kinds of pre-training protocols. Those protocols vary in preprocessing, architectures, tasks, and datasets, as well as various hyperparameters. This makes it difficult to clearly compare the performance of vision-language transformer models on various tasks. Indeed, it is difficult to conclude on which pre-training tasks are the most useful for a multimodal capability if the models we compare do not have the same hyperparameters or pre-training datasets. Likewise, it is difficult to conclude on the impact of pre-training data for models which also have different architectures. As such, the conclusions we reached regarding the impact of different pre-training choices on performances could be validated through further studies.

In this chapter, we attempt to conduct a more systematic study of pre-training factors. More precisely, our goal is to investigate the importance of different pre-training factors for specific monomodal and multimodal capabilities. In order to have better control over the different pre-training factors, we realize ablation studies by pre-training a vision-language model using specific pre-training protocols. However, pre-training vision-language transformer models is resource- and time-consuming. As a result, we pre-train smaller scale models in terms of training time and datasets. Although those models are not comparable to the state-of-the-art models, we hope to establish findings that can hold up on bigger models and datasets. In the past few years, there has been a growing debate on how the abilities of transformer models vary with scale. Indeed, with factors of scale, models seem to exhibit whole new skills that are unknown to smaller-scale models (J. Wei et al. 2022), that have been called *emergent abilities*. In that case, the study of smaller models would be counter-productive, as conclusions would maybe not hold for larger-scale models. Yet, some argue that rather than exhibiting new skills, the choice of the evaluation protocol, and in particular the metrics, are what makes the increase in performance seemingly abrupt (Schaeffer et al. 2023). In this chapter, we focus on specific capabilities, and hope that the difference in performances are more easily interpretable than with more complex tasks. Then, with these results, we make hypotheses regarding larger vision-language models.

4.1.1. Hypotheses

Several aspects of vision-language pre-training are especially relevant to model performance. First, we are interested in how pre-training length, i.e., the number of pre-training steps used to pre-train a model. Indeed, we are limited in resources for these ablation studies and do not pre-train ablation models as long as their state-of-the-art counterpart. As a result, we study how model performance improves or degrades at several pre-training epochs on different tasks.

Then, we study the impact of pre-training tasks, i.e., the various losses optimized during the pre-training of a model. As shown in chapter 1, vision-language models can

be pre-trained on a variety of textual, visual and multimodal pre-training tasks. The goal of those tasks is to help models extract textual and visual information, as well as combine that information to get multimodal representations. There is no consensus yet what best pre-training are the most useful, especially for models that are not aimed at a specific vision-language task. Thus, we study how several pre-training tasks impact model performances.

Finally, the previous chapter has shown that data is a major aspect of vision-language pre-training. There are two major types of pre-training datasets: manually annotated datasets and web-crawled datasets. These datasets vary in size as well as quality. In this chapter, we question how the different aspects of a text-image dataset impact the pre-training of a vision-language model.

To compare the different pre-training strategies, we will in particular evaluate the models on the denotation tasks presented in Chapter 3, to answer the following questions:

— How does pre-training length impact the capacities of a model? Does the model overfit or does it continually improve?
— What is the impact of textual and multimodal pre-training tasks on the performance of a model?
— What kind of visual pre-training would be more appropriate to learn for a multimodal vision-language model?
— Which aspects of a dataset affect the performances of a model?

4.1.2. Experimental Setup

In this chapter, we realize our experiments based on ViLT (W. Kim et al. 2021 [1]). At the time of our experiments, several vision-language models that now exist were not yet developed or publicly available. Contrary to other vision-language transformer models such as UNITER (Y.-C. Chen et al. 2019) or LXMERT (Tan et al. 2019), the performances of ViLT are not dependent on the extraction of visual features through pre-trained object detectors. Indeed, VinVL (P. Zhang et al. 2021) has shown that the choice of object detector can significantly impact the performance of a vision-language transformer based on object representations. As such, we use ViLT to study pre-training, and especially visual pre-training, which has less potential limiting factors due to the visual representations used as input. In addition, ViLT, like UNITER or LXMERT, is of a similar size to BERT-base, which makes it is more easily adaptable to ablation experiments than larger models. This is especially true compared to larger models, which require more pre-training steps.

We realize pre-train variations of the ViLT model on a node of four V-100 GPUs from the Jean Zay cluster. The original ViLT checkpoint given by the authors is pre-trained using 64 GPUs.

1. https://github.com/dandelin/ViLT

Data preprocessing The data preprocessing of ViLT is based on BERT (Devlin et al. 2019) for the textual modality and ViT (Dosovitskiy et al. 2021) for the visual modality. For the visual modality, an image is split into N patches of $(n * n)$ pixels, which are flattened into a sequence of N embeddings. The patch size is usually set to $n = 16$ or $n = 32$. In our experiments, we use a patch size of $n = 16$ and an image size of 224 pixels based on the ViT model with the same parameters, compared to the original ViLT model of $n = 32$ and image size of 384. We use the image augmentation method RandAugment presented in Cubuk et al. 2019. This method is dependent on two hyperparameters: M the magnitude of image transformations and N the number of image transformations. Following literature, we use N=2, M=9.

Architecture ViLT is equivalent to BERT-base in terms of the number of parameters. In order to avoid expensive hyperparameter tuning, we keep the same transformer architecture, which is composed of 12 attention layers of multi-head self-attention, 12 attention heads, and a hidden size of 768. In this chapter, our model is a single-stream transformer model.

Pre-training tasks As explained in Chapter 1, the tasks used to pre-train ViLT are:
- A language-specific task based on Masked Language Modeling (MLM),
- An Image-Text Matching multimodal task (ITM),
- An optimal transport multimodal loss based on the UNITER loss Word Region Alignment, adapted to pixel patches instead of object regions (WRA). The goal of this task is to align pixel patches to word tokens as a fie-grained multimodal task.

Datasets Due to limitations in storage and computing power, we try to pre-train vision-language models in a less resource-consuming way. To that end, we use two different dataset aggregations.

For the first dataset D_{COCO}, we only use the COCO dataset (T.-Y. Lin et al. 2014). This dataset is human annotated and smaller than large-scale web-crawled datasets. This enables pre-training on many epochs, even with less resources. For that protocol, we set a batch size of 176. Each epoch is constituted of 3363 steps, which results in around 600000 instances per epoch.

For the second dataset $D_{\text{COCO+VG+CC}}$, we use three datasets: COCO, Conceptual Captions (Sharma et al. 2018) and Visual Genome (Krishna, Yuke Zhu, et al. 2016). This pre-training dataset enables us to study pre-training that takes advantage of a larger amount of data, including web-crawled data. For that protocol, we set a batch size of 4096. Each epoch is constituted of 2158 steps, which results in around 9 million instances per epoch. As a result, models pre-trained using this protocol are pre-trained on a considerably smaller number of epochs.

As a comparison, the original ViLT model is pre-trained on 300000 steps with a batch size set to 4096, which results in 800 million instances in total during the pre-training.

Pre-training protocol We set hyperparameters following literature, and use the Adam optimizer (J. Li, D. Li, Xiong, et al. 2022). Changes in hyperparameters may not affect all pre-training strategies the same way. To avoid expensive hyperparameter tuning, we do not realize an extensive search of hyperparameters. However, they may not be optimal for our experiments.

Evaluation tasks In Chapter 2, we argue for an exhaustive evaluation of vision-language models based on granular capabilities. To that end, we introduced a taxonomy of vision-language capabilities. In Chapter 3, we create tasks to study the performances of vision-language models on a few capabilities belonging to the *Denotation* category, as well as tasks studying monomodal capabilities. In this chapter, we examine how variations in pre-training affect the performance of a model on some of those tasks. We do not evaluate the models on tasks where ViLT results are at chance level, as no conclusion can be drawn if the task is too difficult for a smaller-scale model. We select a textual and a visual evaluation task, as well as tasks evaluating multimodal capabilities. The following probing and pre-training head evaluation tasks, described in detail in chapter 3 are used for the evaluation:

— the bigram shift task (L-BigramShift),
— the flower-102 task (V-Flower),
— the color identification task (M-Color),
— the minimal caption differentiation task (M-Differences),
— the multimodal dependencies task (M-Dependencies).

In the case of the probing tasks (L-BigramShift, V-Flower, M-Color, and M-Differences), we do not only evaluate the last layer representations of the model in this chapter. Indeed, we also evaluate the representations extracted from the 4th, the 8th and the 12th (i.e., last) layer for each model. We train linear probing models with gradient descent using the Adam optimizer. We report the mean of five different seeds.

4.2. Pre-training Length

In this section, we study the impact of pre-training length on the performances of a model. As pre-training length and dataset size are intrinsically linked, we also want to take into account the size of the pre-training dataset. Indeed, studies have shown that pre-training length and dataset size enable better performances for transformer-based models. This has encouraged the development of vision-language models pre-trained on larger web-crawled datasets.

Another aspect we study in this section is the use of preloaded monomodal models. Indeed, a growing trend for vision-language transformers is the use of pre-trained monomodal modals as basis of multimodal models. In some cases, they are used as frozen models to extract monomodal features. In other cases (e.g., ViLT), they initialize transformer weights, as explained in Chapter 1. Indeed, the original ViLT model is initialized with the weights of a pre-trained ViT model.

We aim to establish in what capacity the use of a preloaded model improves the pre-training of ViLT. To that end, we pre-train the vision-language model with and without preloading a ViT model. We note *with ViT* the models whose weights were initialized with ViT, and *without ViT* the other models. We study the pre-training performance of the model pre-trained on two different datasets:

— D_{COCO}, at epochs 21, 37, 54, 80 and 106;

— $D_{\text{COCO+VG+CC}}$, at epochs 1, 3, 5, 7 and 9.

4.2.1. Results

Multimodal Dependencies: M-Dependencies As explained in Chapter 3, we build the multimodal dependencies task to evaluate two different types of capabilities. First, the *full* task evaluates the ability of a model to take account of noun-predicate dependencies with a distractor in the image, i.e., fine-grained multimodal dependencies. Then, the *cropped* task removes the distractors from the image, which leads to an easier image-text matching task, which evaluates more coarse-grained multimodal understanding.

	Full		Cropped	
Dataset D_{COCO}	without ViT	with ViT	without ViT	with ViT
Epoch 21	0.18	0.24	0.23	0.27
Epoch 37	0.24	0.24	0.23	0.33
Epoch 54	0.22	0.22	0.30	0.30
Epoch 80	0.23	0.23	0.35	**0.40**
Epoch 106	**0.25**	**0.25**	**0.37**	**0.40**
Dataset $D_{\text{COCO+VG+CC}}$	without ViT	with ViT	without ViT	with ViT
Epoch 1	0.15	0.18	0.11	0.13
Epoch 3	0.18	0.12	0.14	0.10
Epoch 5	0.15	0.19	0.14	0.17
Epoch 7	**0.23**	**0.21**	**0.20**	**0.21**
Epoch 9	0.19	0.20	0.19	0.20
Original ViLT	—	0.40	—	0.75

Table 4.1. – Accuracy of ablation models on the multimodal dependency tasks. Models are pre-trained on D_{COCO} and $D_{\text{COCO+VG+CC}}$ with and without ViT preloading, at different epochs. *Full* refers to the task with full images, and *Cropped* refers to the task with cropped images, i.e. without distractor objects.

Table 4.1 shows that the ablation models do not reach the performance of the original ViLT model on the M-Dependencies task. Indeed, the results in the *full* setting do not reach random performance, and are considerably worse than the original ViLT

model. However, we can see a slight increase in performance for models trained at later epochs. Compared to the base model pre-trained on D_{COCO} at epoch 106, the original model has a pre-training length accounting for 13 times more text/image instances.

In the *cropped* setting, initializing weights with a ViT model enables the model to reach significantly better performances for the model pre-trained on D_{COCO}. In addition, the accuracy of this model continually increases with the epochs, and seems to stabilize at epoch 106. This seems to indicate that pre-training models on a considerably long time, reaching the 100 of epochs, does not lead to overfitting even with smaller datasets such as the COCO dataset. However, the increase in performance seems to drop with later epochs, which could indicate a reduced efficiency in pre-training.

Compared to the model pre-trained on $D_{\text{COCO+VG+CC}}$ at epoch 9, the original ViLT model has a pre-training length accounting about 10 times more pre-training steps. At this point, our model has not yet managed to reach results above chance level on this task. The model pre-trained on $D_{\text{COCO+VG+CC}}$ shows worse result than all studied epochs of the model pre-trained on D_{COCO}. This seems to show that a high number of pre-training steps is necessary to reach a good multimodal understanding on this task.

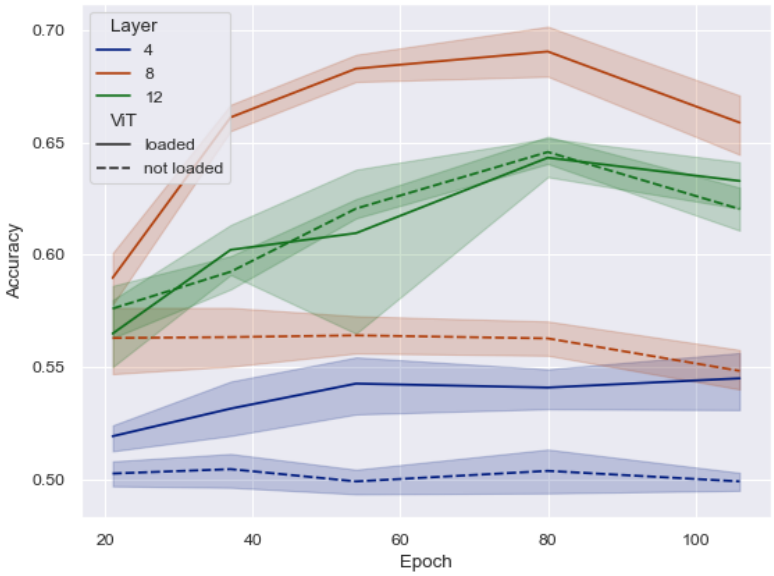


Figure 4.1. – Results of models pre-trained on the D_{COCO} dataset for the bigram shift probing task. Results include models with and without ViT preloading, at attention layers 4, 8 and 12.

Bigram Shift: L-BigramShift The bigram shift task evaluates a model's ability to differentiate between grammatically correct and incorrect word order in a caption. Figure 4.1 shows the results of the model trained on D_{COCO}. Upper layers are better than the fourth layer are taking into account this kind of syntactic information. While the model with ViT shows better performance for the 8th layer, the results for the 12th layer of both models are comparable. Results seem to slightly decrease for the last epoch, perhaps due to overfitting. Models do not reach the original ViLT accuracy of 72, but the results are significantly higher than chance for upper layers. This shows that the limited size of the COCO dataset is sufficient to learn to extract syntactic information related to word order.

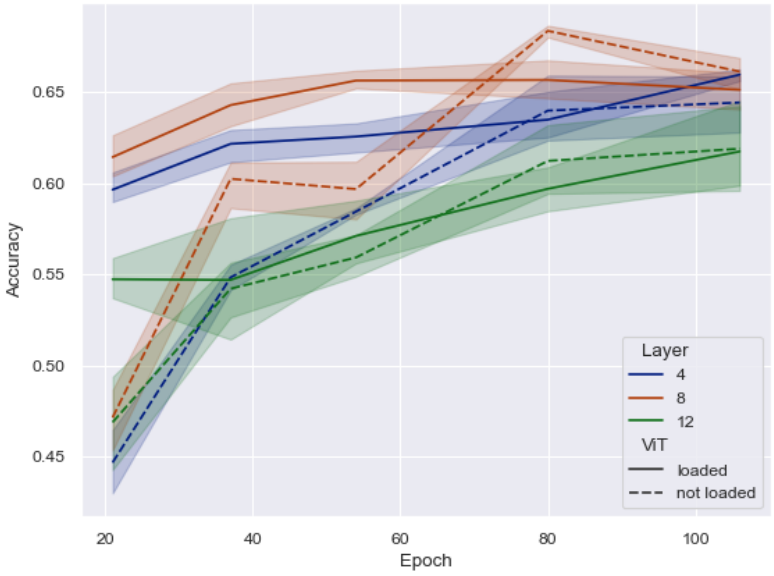


Figure 4.2. – Results of models pre-trained on D_{COCO} for the Flower-102 probing task. Results include models with and without ViT preloading, at attention layers 4, 8 and 12.

Flowers-102: V-Flower The Flower-102 task evaluates a model's ability to perform fine-grained object classification, by taking into account the shape and color of an object. This task is solely visual. Figure 4.2 shows the results of the model trained on D_{COCO}. While the models with ViT reach significantly better performances than those without ViT in early epochs, this gap disappears by epoch 80, with the stabilization of their performances. In addition, all layers have results above chance level (i.e., 0.01), showing that the ability to distinguish between the details of an object is acquired from the early attention layers, and this information remains in later layers, though slightly reduced. However, all models perform significantly worse than the original ViLT model, which reaches an accuracy of 91 on this task.

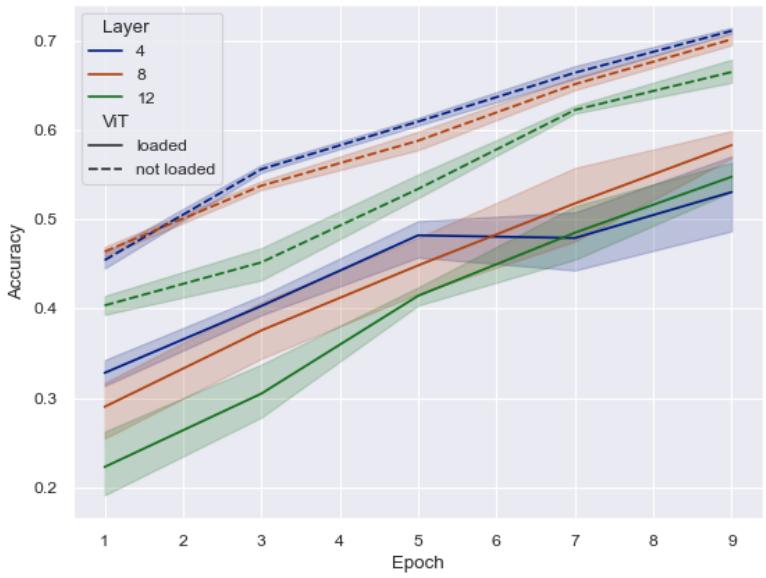


Figure 4.3. – Results of models pre-trained on $D_{\text{COCO+VG+CC}}$ for the Flower-102 probing task. Results include models with and without ViT preloading, at attention layers 4, 8 and 12.

Figure 4.3 shows the results for the dataset $D_{\text{COCO+VG+CC}}$. With this pre-training protocol, the models without ViT reach better results at epoch 9 than the models pre-trained only the COCO dataset at epoch 106. While this is reduced for the models with ViT, the performances at epoch 9 show no sign of stabilization, but continue increasing, contrary to the D_{COCO} models. This seems to indicate that the reduced size of D_{COCO} in terms of images (i.e., 100k) compared to the $D_{\text{COCO+VG+CC}}$ dataset or the dataset used for the original pre-training is detrimental to visual performances. In particular, it significantly impacts the model's performance on fine-grained object classification tasks.

Color: M-Color The color probing task evaluates a model's ability to associate the correct color to an object, by masking a color word in a sentence. As a result, this task is multimodal. Figure 4.4 shows that apart from the early epochs, the model with ViT and the one without ViT have similar results. The representations from the last layer encode slightly more relevant information in the case of the model with ViT initialization. The models perform significantly worse than the original ViLT, which reaches an accuracy of 86 on this task. However, they reach results well above chance level, or monomodal levels (37 for BERT and 41 for ViT), for the 12th layer representations. Indeed, the last layer reaches significantly better results than the lower layers, and continues to improve with later epochs. This seems to suggest that continuing pre-training could lead to further increase in performance for the 12th

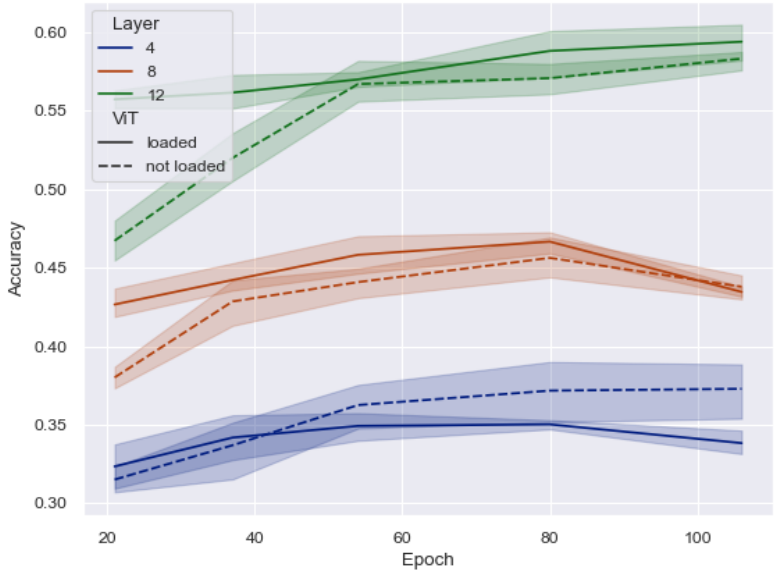


Figure 4.4. – Results of models pre-trained on the COCO dataset for the color probing task. Results include models with and without ViT preloading, at attention layers 4, 8 and 12.

layer.

Minimal captions differentiation: M-Differences This probing task evaluates the ability of a model to distinguish small differences between texts and images. These can consist in a difference in noun, verb, or adjective. Figure 4.5 shows that the models with ViT reach better performances on this task than their counterparts without ViT, for almost all epochs and layers evaluated. Results do not reach the performances of the original model, which reaches 73.4 accuracy on this task. Only the 12th layer show results significantly above chance. The performance on this task increases significantly from the 21st epoch to the 106th epoch, showing that the information necessary for this task is better extracted at later epochs. In this aspect, this task is different from the color probing task, where the increase in performance is less noticeable. This seems to indicate that the capabilities necessary for this task are acquired at a later stage during pre-training. However, the results of the different epochs on this task are more volatile, which makes them less interpretable.

4.2.2. Discussion

Impact of dataset size The models pre-trained on COCO, a relatively small dataset, are able to learn coarse-grained multimodal dependencies, as well as textual and visual capabilities. However, they do not learn more fine-grained multimodal capabilities

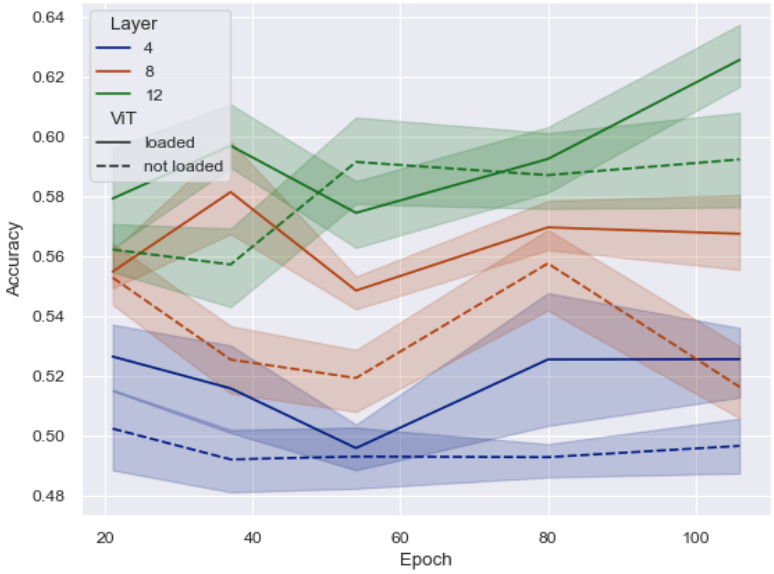


Figure 4.5. – Results of models pre-trained on the COCO dataset for the minimal differences probing task. Results include models with and without ViT preloading, at attention layers 4, 8 and 12.

such as multimodal noun-predicate dependencies. These results also show that the size of pre-training dataset has a major impact on the performance of a vision-language transformer. Indeed, for the multimodal capabilities, as well as the visual capability, the gap with the original ViLT model is significant. On several tasks, the slower increase in performance in later epochs seems to show that a model pre-trained on the COCO dataset only would not reach the results of the original ViLT model. While this worse performance could be impacted by other factors such as the size of input images, the reduced size of the pre-training dataset D_{COCO} seems to have major impact.

Possibility of overfitting While no overfitting has been noticed for the visual and multimodal capabilities, the results of the bigram shift task suggest a possibility of overfitting of the model for the language capability. Indeed, there is a slight decrease in accuracy for the last epoch. This could be due to the relatively small size of the pre-training dataset compared to the large number of epochs it has been pre-trained on. This may be mitigated with data augmentation techniques for the textual input during pre-training, but would benefit from the use of larger datasets with more diverse annotations. For the visual capability, the results of the Flower-102 task suggest a stabilization of the performances, but no overfitting yet, despite the relatively small size of the D_{COCO} dataset.

Impact of dataset variability The variability of a dataset significantly impacts a model's performance on the visual task. Indeed, this is shown by the better performance of models pre-trained on the $D_{\text{COCO+VG+CC}}$ dataset compared to those pre-trained on D_{COCO}. One important point is that those models are not pre-trained using visual pre-training tasks. The results on the Flower-102 probing task could suggest that the lack of visual pre-training of the model is a limiting factor for vision-language transformer models. The use of visual pre-training could considerably affect the performance of a model on this task.

Impact of ViT initialization The multimodal dependency task also suggests a significant gap in performance between the models initialized with or without ViT. Almost all tasks show better results for the models with ViT initialization for models pre-trained on D_{COCO}. This could corroborate the trend exhibited in Chapter 1 of using pre-trained multimodal models to initialize multimodal models. However, the structure of a multimodal transformer model could be significantly different from that of a monomodal model, especially in terms of what information is extracted at which layer. Indeed, probing the different layers of a model seems to indicate that upper layers encode more multimodal information than lower layers, while lower layers retain more local monomodal information, such as details needed for fine-grained object classification tasks. This could explain the fact that using a ViT model initialization does not improve performances for visual tasks. Thus, initializing a multimodal model with a pre-trained monomodal model might be beneficial for some capabilities but hindering other capabilities.

In later experiments, we pre-train the models on the first D_{COCO} dataset for 107 epochs, with the weights initialized using ViT.

4.3. Pre-training Tasks

After pre-training length, we are interested in studying how pre-training tasks affect the performances of vision-language transformers. The original ViLT model is pre-trained using three pre-training tasks: a textual pre-training task MLM, as well as two multimodal pre-training tasks, ITM and WRA. It is important to note that while we differentiate between textual and multimodal pre-training, MLM is also designed to encode multimodal information. Indeed, the combined use of visual input and masked textual input could lead the model to rely on visual information to predict masked words. In this section, our goal is to better understand the role of each pre-training task in the monomodal and multimodal capabilities acquired by the models. To that end, we pre-train five models using several combinations of those pre-training tasks.

— ITM
— MLM
— ITM and WRA

— MLM and ITM
— MLM, ITM and WRA

We compare how those models perform relatively to each other, to identify if the pre-training tasks improve or deteriorate the performance of the model.

4.3.1. Results

Multimodal Dependencies: M-Dependencies Due to the fact that the multimodal dependencies task is based on the pre-trained ITM task-specific head, we only compare models which have been pre-trained on ITM.

Pre-training tasks	Full	Cropped
ITM	0.23	0.31
ITM & WRA	0.22	0.27
MLM & ITM	0.23	0.39
MLM & ITM & WRA	**0.25**	**0.40**
Original ViLT	0.40	0.75
Chance level	0.25	0.25

Table 4.2. – Accuracy of the models with different pre-training task combinations on the M-Dependencies task. *Full* refers to the task with full images, and *Cropped* refers to the task with cropped images, i.e., without distractor objects. Chance level reaches 25% due to the use of counterbalanced pairs in the computing of the results (see Chapter 3).

The results of Table 4.2 show that while ITM alone is sufficient to learn some coarse-grained multimodal capabilities, MLM helps the model better extract multimodal information. Indeed, the MLM+ITM pre-training shows significantly better results than the ITM only pre-training. While WRA does not improve the results as much as MLM, the complete pre-training MLM + ITM + WRA shows the best results. However, the use of WRA without MLM seems to hinder the performances of the model on this task. This could be due to the fact that the WRA pre-training task helps the model learn to associate textual and visual tokens. Thus, it should be important to correctly extract textual and visual information for this pre-training task to be relevant. This could indicate that better visual pre-training could help further improve the effect of the WRA pre-training task.

Bigram Shift: L-BigramShift Table 4.3 shows the results of models pre-trained on the different pre-training strategies on the bigram shift task. Though the MLM task is the most relevant pre-training task, its use in combination with the ITM task enables the model to reach better performances in later layers. Not using the MLM task leads the model to have results around chance level. This suggests that multimodal tasks such as Image Text Matching may not be efficient at learning fine-grained monomodal

Task	L-BigramShift			V-Flower			Agg.
Model	Layer 4	Layer 8	Layer 12	Layer 4	Layer 8	Layer 12	Layer 12
ITM	0.51 ± 0.01	0.51 ± 0.01	0.51 ± 0.01	0.55 ± 0.02	0.30 ± 0.06	0.13 ± 0.05	0.32
ITM & WRA	0.50 ± 0.01	0.49 ± 0.01	0.50 ± 0.00	0.16 ± 0.06	0.05 ± 0.03	0.01 ± 0.00	0.26
MLM	**0.62** ± 0.04	0.59 ± 0.08	0.58 ± 0.06	0.34 ± 0.04	0.09 ± 0.06	0.19 ± 0.09	0.39
MLM & ITM	0.54 ± 0.02	**0.67** ± 0.02	**0.68** ± 0.02	0.63 ± 0.01	**0.69** ± 0.02	**0.64** ± 0.02	**0.66**
MLM & ITM & WRA	0.54 ± 0.02	0.66 ± 0.02	0.63 ± 0.01	**0.66**± 0.00	0.65 ± 0.01	0.62 ± 0.03	0.63
Original ViLT	—	—	0.72	—	—	0.91	0.82
Majority class	—	—	0.50	—	—	0.01	0.26
BERT	—	—	0.86	—	—	—	—
ViT	—	—	—	—	—	0.997	—

Table 4.3. – Accuracy of models pre-trained on the different pre-training task combinations for the monomodal probing tasks, at attention layers 4, 8 and 12. Reported results are a mean over 5 runs of the accuracy. Agg. refers to the mean of the Layer 12 accuracy for the two monomodal tasks.

capacities such as word order by themselves. However, they can help complement textual pre-training tasks. As such, a multimodal model pre-trained without language pre-training would have a hard time extracting information related to word order, which is important in many multimodal tasks.

Flowers-102: V-Flower Table 4.3 shows the results for the Flower-102 task. ViLT is not pre-trained on a visual pre-training task, which could improve the results on this task. The worst pre-training strategy for this task seems to be the combination of ITM and WRA, especially for the last layer. The best results are obtained for the upper layers with ITM+MLM pre-training, followed by the complete pre-training. While ITM is more efficient than MLM on this task, MLM pre-training offers results significantly above the 0.01 chance level for all layers. The results also suggest that the efficiency of the WRA pre-training task is greatly influenced by the other tasks it is combined with, and in particular the language task.

In the last section, we concluded that early layers extract more visual information related to fine-grained classification than later layers. Those results indicate that this may be related to the use of the WRA pre-training task. Indeed, the results obtained with the upper layer representations using ITM + MLM + WRA pre-training are significantly lower than that of the MLM + ITM pre-training. This could be due to the fact that WRA uses last layer representations of the visual (and textual) embeddings, and optimizes them for purposes that are not directly related to visual understanding. Indeed, the results for the bigram shift task also show that the WRA pre-training degrades syntactic performances, especially for the last layer.

Minimal captions differentiation: M-Differences Table 4.4 shows the results of the different pre-training strategies on the minimal captions task. The ITM pre-training is more useful than MLM pre-training, and there is not a jump in performance between the ITM and ITM + MLM pre-training. The performances of the MLM only model seem to show that the model does not overly rely on textual bias on this

Task	M-Differences			M-Color			Agg.
Model	Layer 4	Layer 8	Layer 12	Layer 4	Layer 8	Layer 12	Layer 12
ITM	0.54 ± 0.01	0.57 ± 0.07	0.59 ± 0.03	0.36 ± 0.03	0.34 ± 0.06	0.28 ± 0.03	0.44
ITM & WRA	**0.58** ± 0.03	**0.59** ± 0.04	0.55 ± 0.06	0.26 ± 0.07	0.21 ± 0.03	0.14 ± 0.06	0.35
MLM	0.50 ± 0.03	0.47 ± 0.05	0.54 ± 0.03	**0.38** ± 0.01	0.39 ± 0.01	0.53 ± 0.01	0.54
MLM & ITM	0.51 ± 0.02	0.56 ± 0.02	0.61 ± 0.02	0.34 ± 0.01	0.35 ± 0.01	**0.60** ± 0.01	**0.61**
MLM & ITM & WRA	0.53 ± 0.02	0.57 ± 0.01	**0.63** ± 0.01	0.34 ± 0.01	**0.43** ± 0.00	0.59 ± 0.01	**0.61**
Original ViLT	—	—	0.73	—	—	0.86	0.80
Majority class	—	—	0.50	—	—	0.25	0.38
BERT	—	—	0.53	—	—	0.37	0.45

Table 4.4. – Accuracy of models pre-trained on the different pre-training task combinations for the multimodal probing tasks, at attention layers 4, 8 and 12. Reported results are a mean over 5 runs of the accuracy. Agg. refers to the mean of the Layer 12 accuracy for the two multimodal tasks.

task. These results seem to confirm the hypothesis that upper layers encode more multimodal information than lower layers. The only exception is the pre-training strategy WRA + ITM, which seems to reach worse performances in the last layer. This could mean that using WRA without a corresponding monomodal pre-training task is counter-productive and does not help the model learn multimodal capabilities.

The pre-training strategies that use MLM in addition to multimodal pre-training tasks show a significant jump in performance between the late layer and the other layers. This may show that the semantic textual information needed to differentiate between minimally wrong and right captions is more easily accessible at later layers.

Color: M-Color Table 4.4 shows the results of the different pre-training strategies on the color identification task. All strategies except the combination of ITM and WRA are above majority-class level (25%). This is unexpected, because the color task uses text token representations, which should mean that the MLM pre-training task should be needed for results above chance. This suggests that relevant information is encoded in representations corresponding to text tokens, even when there is no language pre-training.

The language-only baseline (i.e., last layer representations of BERT) reaches an accuracy of 0.37 on this task (see Chapter 3). In the case of models without language pre-training, the lower layer performs better than upper layers, which are significantly below the language-only baseline. The models trained using only language pre-training reach above the language-only baseline in the upper layers, showing that MLM also encodes visual information. The combination of language and multimodal pre-training tasks, reach the best results. Similarly to the language-only pre-training, the last layer reaches results significantly above lower layers. This suggests that later layers encode more multimodal information.

This task does not highlight a significant role of the WRA pre-training task. As discussed in paragraph 4.3.1, this could be due to WRA pre-training hindering textual representations, or a lack of visual pre-training hindering the role of the WRA task.

In additional experiments 4.5, we study the role of textual bias for model perfor-

Model	Layer 4	Layer 8	Layer 12
ITM	0.22 ± 0.04	0.25 ± 0.06	0.19 ± 0.04
ITM & WRA	0.17 ± 0.06	0.18 ± 0.04	0.19 ± 0.02
MLM	**0.29** ± 0.01	**0.31** ± 0.01	**0.31** ± 0.01
MLM & ITM	0.28 ± 0.01	**0.31** ± 0.01	**0.31** ± 0.01

Table 4.5. – Accuracy of models pre-trained on the different pre-training task combinations for text-only color identification probing task, at attention layers 4, 8 and 12. The only inputs given are captions, images. Reported results are a mean over 5 runs of the accuracy.

mances on the color task. To that end, we only give captions to the model, and withhold visual information. As a result, models predict the color that is the most appropriate given the textual context only. The results show that without MLM, the models do not reach better performances than the majority class. On the contrary, the use of MLM leads the models to rely more on textual bias. The textual information linked to the masked color word seems to be stable across layers. The results do not change with the addition of ITM to the MLM task. This means that the increase in performance observed for the text-image color tasks (Table 4.4) between the MLM and ITM + MLM settings is not due to textual bias but to visual information. As such, the textual bias for the color task seems mostly due to the MLM pre-training task. In addition, textual information seems more accessible, from lower layers, than relevant visual information.

4.3.2. Discussion

In this section, we have studied the effect of different pre-training task combination of textual, visual and multimodal capabilities. The MLM language task and ITM multimodal task are the most relevant, useful for vision, language and multimodal capabilities.

Extracting visual and textual information The MLM task seems able to extract visual information in addition to textual information, at least regarding colors. However, the combination of ITM and MLM seems required to extract more fine-grained visual information, as shown by the V-Flower task. This could also be due to a lack of visual pre-training task to help extract relevant visual information. In the next section 4.4, we study potential visual pre-training tasks.

The multimodal pre-training tasks ITM and WRA are not precise enough to distinguish texts based on word order, which seems to highlight the need for a textual pre-training task.

Impact of the WRA pre-training task The case of the WRA pre-training task is a bit more complex, as it does not always improve performances. In particular, the

combination of ITM+WRA seems to worsen the ability of a model to extract visual information. For the M-Differences task, its use improves multimodal performances. Indeed, this task evaluates minimal differences between textual and visual information, which directly relates to the role of WRA to align corresponding visual and textual tokens. This could be due to the fact that it causes visual representations to have degraded monomodal capabilities in favor of multimodal capabilities. However, the use of the MLM pre-training task in addition to the ITM+WRA pre-training seems to soften the impact of WRA on visual performances. This could suggest that visual pre-training could help improve the effect of WRA pre-training.

Role of attention layers While the combination of textual and multimodal pre-training tasks leads to better results in multimodal probing tasks, the jump in performance is mainly visible at later layers. Indeed, the M-Color task suggests that the combination of visual and textual information in textual tokens improves for upper layers, at least for models using MLM pre-training. It would be interesting to see whether the use of visual pre-training of the same type as MLM would lead to similar observations. This confirms the fact that more multimodal information is encoded in later layers.

Consequence for larger vision-language models The ITM and MLM pre-training tasks complement each other for the learning of multimodal information, while ITM does not help encode word order. This suggests that larger vision-language models pre-trained using only a multimodal task similar to ITM would have difficulty to encode information linked to word order. This information is especially relevant for vision-language applications, especially for the learning of fine-grained multimodal dependencies. In addition, while MLM helps the model encode multimodal information, the combination of MLM and ITM offers the best results. There could be several reasons behind those results. One is that both tasks focus on different aspects of multimodal interactions, and the combination of both helps the model extract more relevant visual and textual information. Another reason could be that the use of both the ITM and MLM pre-training tasks leads the model to observe more textual and visual data combinations than only one task or the other. Indeed, by masking the text, MLM affects captions at a local level, while ITM affects image-text data pairs. This result could suggest that models pre-trained using only a textual task, such as the Prefix Language Model task introduced in 1, could be hindered by their lack of an additional multimodal pre-training task.

It would be interesting to conceive a theoretical model of vision-language pre-training tasks to better assess the efficiency of such tasks, and how to combine them optimally.

4.4. Visual Pre-training

The previous sections suggest that model performances are limited by the lack of visual pre-training task. Indeed, this could hinder the ability of a model to extract visual information, which is relevant for both visual and multimodal capabilities. In this section, we design a visual pre-training task and evaluate how it affects model performances on those capabilities. Inspired by previous work in the computer vision field, we decide to rely on teacher-student methods. Teacher-student methods, or distillation methods, transfer knowledge from a teacher model to a student model. In many cases, they compress help compress model size, by training a student model from a larger teacher model. Yet, it can also be used to make pre-training more efficient. Indeed, Touvron, Cord, Douze, et al. 2021 use a distillation mechanism for to pre-train a data efficient visual transformer.

In this section, our goal is to efficiently learn visual representations of image patches. To that end, we use frozen teacher networks to learn representations corresponding to masked patches. The goal of our pre-training task is for the vision-language model to learn similar last layer representations of pixel patches as the teacher network. Indeed, this could help the model extract visual features relevant for visual tasks such as semantic classification.

This task, which we call Masked Patch Regression (MPR), is similar to the UNITER task of Masked Feature Regression. With this task, we randomly mask a portion m (20%) of visual patches v. We project the last layer representations of these patches vl_s to the dimension of the representations of the teacher network with a linear layer. Finally, we compute the mean square error between the projected representations and the teacher network representations v_t, which we call the MPR loss 4.1.

$$\text{MPR} = -\mathbb{E}_{(W,V)\in\mathcal{D}} \sum_{m} \left\| f(v_m) - r(v_m) \right\|_2^2 \tag{4.1}$$

This loss is added to the other pre-training tasks, MLM, ITM and WRA. In this section, we compare the performances of different teacher networks: ViT, ViTMAE (K. He, X. Chen, et al. 2022), ResNet (K. He, X. Zhang, et al. 2016) and VGG (Simonyan et al. 2014). We use pre-trained models made available by Hugging Face. We train four vision-language transformer model with each of those networks as teacher, and compare their performances. To evaluate the impact of masking percentage of visual patches, we also test the ViT teacher model with a masking proportion of 60%.

4.4.1. Results

For comparison purposes, we add report the results without the use of the MPR loss. In the following tables, we denote the *MPR Teacher* of these results as *None*. It corresponds to the MLM+ITM+WRA results of the previous section (Section 4.3).

Multimodal Dependencies: M-Dependencies Table 4.6 shows that the use a visual pre-training task can slightly improve multimodal performances on the mul-

MPR Teacher	Full	Cropped
ResNet	0.25	0.39
VGG	0.24	0.37
ViTMAE	**0.25**	**0.42**
ViT	0.23	**0.42**
ViT (60%)	0.24	0.41
None	0.25	0.40

Table 4.6. – Accuracy of the models with different teacher model for the visual pre-training task, evaluated on the multimodal dependencies task. *Full* refers to the task with full images, and *Cropped* refers to the task with cropped images, i.e. without distractor objects.

timodal dependencies task. However, this is not the case for all variations of the visual task. Indeed, for the *cropped* setting, the transformer teacher models improve results while the CNN teacher models deteriorate the results. The use of the visual pre-training task does not help the model reach results above chance level on the *full* setting.

MPR Teacher	Layer 4	Layer 8	Layer 12
ResNet	0.64 ± 0.01	0.64 ± 0.02	0.60 ± 0.02
VGG	0.63 ± 0.01	0.70 ± 0.01	0.61 ± 0.03
ViT	**0.66** ± 0.01	**0.71** ± 0.01	**0.68** ± 0.03
ViT (60%)	0.65 ± 0.01	0.70 ± 0.02	0.67 ± 0.03
ViTMAE	0.65 ± 0.02	0.63 ± 0.02	0.60± 0.03
None	**0.66** ± 0.00	0.65 ± 0.01	0.62 ± 0.03

Table 4.7. – Accuracy of models pre-trained using different visual pre-training tasks for the Flower-102 probing task, at attention layers 4, 8 and 12. Reported results are a mean over 5 runs of the accuracy.

Flowers-102: V-Flower Table 4.7 shows the results of the visual pre-training task on the flower-102 probing task. The only teacher model that significantly improves results over the original pre-training combination is ViT. The lack of improvement for CNN teacher models could be due to the fact that ViLT does not extract visual information the same way as CNN models. Indeed, the internal structure of ViLT is based on attention layers, while VGG and ResNet are based on convolutional layers. Thus, learning to encode visual information similarly to CNN models may not be the most appropriate for transformer models. The difference between ViT and ViTMAE could be explained by the fact that ViT is pre-trained as a classification model, while ViTMAE is pre-trained in an auto-encoding manner. Thus, ViT features could encode more semantics related to fine-grained visual classification than ViTMAE features. The

masking percentage used for the visual pre-training task does not seem to significantly impact the way the models extract visual information.

Task	M-Differences			M-Color			Agg.
MPR Teacher	Layer 4	Layer 8	Layer 12	Layer 4	Layer 8	Layer 12	Layer 12
ResNet	0.35 ± 0.01	0.37 ± 0.01	0.60 ± 0.01	0.52 ± 0.01	0.56 ± 0.01	0.63 ± 0.01	0.62
VGG	0.35 ± 0.02	0.38 ± 0.02	0.62 ± 0.02	0.52 ± 0.00	0.53 ± 0.01	0.61 ± 0.01	0.62
ViT	**0.41** ± 0.02	0.37 ± 0.01	**0.63** ± 0.01	0.51 ± 0.01	0.56 ± 0.01	0.63 ± 0.01	0.63
ViT (60%)	0.38 ± 0.02	0.38 ± 0.01	**0.63** ± 0.02	**0.54** ± 0.01	**0.59** ± 0.01	**0.65** ± 0.01	**0.64**
ViTMAE	0.35 ± 0.01	0.33 ± 0.00	0.60 ± 0.01	**0.54** ± 0.01	0.55 ± 0.02	0.63 ± 0.01	0.62
None	0.34 ± 0.02	**0.43** ± 0.01	0.59 ± 0.01	0.53 ± 0.01	0.57 ± 0.00	0.63 ± 0.01	0.61

Table 4.8. – Accuracy of models pre-trained using different visual pre-training protocols for the multimodal probing tasks, at attention layers 4, 8 and 12. Reported results are a mean over 5 runs of the accuracy. Agg. refers to the mean of the Layer 12 accuracy for the two multimodal tasks.

Color: M-Color Table 4.8 shows the results of the visual pre-training task on the color identification probing task. The use of a visual pre-training task slightly improves the performances of the models for the last layer representations. Similarly to the task V-Flower, the ablation models using a ViT teacher shows the best performances. However, the improvement is limited, and the difference between the MPR + MLM + ITM + WRA pre-training and the MLM + ITM + WRA is not as marked as the difference between MLM + ITM + WRA and ITM + WRA pre-training.

Minimal captions differentiation: M-Differences Table 4.8 shows the results of the visual pre-training task on the minimal caption differentiation probing task. The ViT teacher model with 60% masking is the only one that slightly improves the performances, while other models do not affect the performances or even see them deteriorate, such as VGG. This shows that the designed pre-training task is not beneficial for this task.

4.4.2. Discussion

Impact of the visual pre-training task MPR The results show a small improvement by using the MPR task with the ViT teacher model. This is in particular apparent for the visual probing task, where the layer representations show significant improvement. However, such improvement is less noticeable, or even absent, for multimodal probing tasks. This could have several implications: maybe the additional visual information extracted using the ViT teacher model is not relevant for the multimodal probing task we test. This seems unlikely, because V-Flower is a fine-grained object classification task that relies a lot on color among other visual features. Thus, an improvement on this task could mean that information related to color is better extracted. Another hypothesis is that while ViT teacher models help ViLT extract relevant visual information, the multimodal pre-training does not deepen the interaction between

textual tokens and visual embeddings. Thus, rather than a visual pre-training task, what could be needed is another multimodal task that relies more on visual inputs.

Comparison of MPR variations Results seem to suggest that the use of ViT as a teacher model for the visual pre-training task leads to better results than the use of CNN-based models. It is maybe easier for a vision-language transformer to approximate features from a transformer-based model than a CNN-based model. The difference between ViT and ViTMAE for V-Flower could mean that the use of features pre-trained on classification tasks are more efficient in extracting relevant visual information than those pre-trained using auto-encoding. While our experiments do not exhibit a major impact of the masking percentage on the capabilities of the models, other hyperparameters could be relevant for this task: the choice of the teacher model, the layer used for the representations, as well as the weight of the task in the total loss compared to the other tasks. It would be interesting to test several configurations to evaluate how they impact the results of the models.

Limitations of the MPR task This method is very resource consuming, and though it provides a small improvement in several tasks, may not be efficient. Indeed, the use of a teacher model requires significantly more space. In addition, pre-training using the visual tasks requires, depending on the model, up to 1.6 times more hours of computing power in our configuration. An alternative would be to pre-compute patch representations of the teacher models prior to pre-training. However, that would also require increased space. As such, though these experiments confirm that the lack of visual pre-training task may be a limiting factor for visual but also multimodal performances, it does not provide yet a viable pre-training task.

Use of visual patches Another aspect is that the use of patches for visual pre-training is not necessarily the most intuitive for vision-language applications. Indeed, the fine-grained vision-language pre-training task WRA has been designed by UNITER authors, that use visual inputs based on Faster R-CNN object representations. ViLT, on the other hand, uses visual inputs based on pixel patches. While that offers more potential for the extraction of visual information, it seems to significantly hinder the WRA pre-training task, that aims at relating textual embeddings to visual embeddings. Indeed, visual embeddings corresponding to pixel patches may not be associated with semantic information, or, on the other hand, may be associated with too many words. Visual representations linked to objects may be more relevant for fine-grained multimodal understanding. This leads to a dilemma, as the use of a frozen faster R-CNN to preprocess images is a limiting factor of vision-language models, as explained by VinVL authors. More recently, models (Z. Peng et al. 2023) have used visual representations based visual embeddings, such as CLIP embeddings, for vision-language models based on large language models. While this could help make visual information more accessible, it is subject to the potential bias of the models used to extract the features. To mitigate that bias, it could be interesting to combine visual embeddings extracted

through CLIP-like models and other visual representations. For this purpose, one could rely on the extensive work in the domain of feature engineering in computer vision.

Consequence for larger vision-language models The question of the visual pre-training of vision-language transformer models is difficult, and no consensus has been established yet on the best way to pre-train vision-language transformers. This may be due to the fact visual pre-training is time-consuming and does not lead to jump in performances. Vision-language transformers such as TCL have favored the use of Contrastive Learning for monomodal tasks, instead of feature regression or object classification tasks. Used on very-large batches, it permits fine-grained differentiation between images. Other recent large multimodal models, such as KOSMOS (Z. Peng et al. 2023), opt for the use of frozen visual encoders to preprocess visual inputs. This way, they include visual information in the model without any visual pre-training task. However, one can imagine that the lack of visual pre-training or visually oriented multimodal pre-training task is a limiting factor for those models. The new variant of KOSMOS proposes a countermeasure by introducing bounding box location in the next token prediction task used by KOSMOS. It would be interesting for future work to design more fine-grained visual pre-training tasks adapted to such models.

4.5. Pre-training Dataset

Although the design of a pre-training task significantly impact model performances, the efficiency of a task is greatly influenced by the training data used for this task. Indeed, the impact of the pre-training dataset has already been highlighted in section 4.2. In this section, we specifically focus on the pre-training datasets used by vision-language transformers, and try to assess how they impact the representations computed by those models. In section 1.3.1, we have introduced the different types of pre-training datasets used by vision-language models. They mainly consist of human annotated datasets, that are smaller in scale, and web-crawled datasets, that are automatically filtered using various methods. In this section, we aim to assess how datasets affect vision-language transformers. However, we are limited in terms of resources and computing power. Consequently, we study the literature of vision-language models to try to establish how text-image pre-training datasets impact the performances of vision-language models. The next section is based on the following article: 'Salin, E. (2023). État des lieux des Transformers Vision-Langage: Un éclairage sur les données de pré-entraînement. In 30e Conférence sur le Traitement Automatique des Langues Naturelles (pp. 14-29). ATALA.'

4.5.1. Impact of Pre-training Data: a Survey

The question of what kind of data should be used to pre-train vision-language models is essential for more efficient pre-training and better performing models.

Indeed, it could help establish protocols for collecting, filtering and processing text-image data. It could also enable a pre-training of vision-language models with fewer resources. Many studies agree that using larger datasets improves the performance of vision-language transformers. This is particularly visible in ablation studies carried out for different models, showing a significant improvement in model performance on downstream tasks with the increase in scale of pre-training data (J. Li, R. Selvaraju, et al. 2021; J. Yang et al. 2022). However, it is interesting to explore other features of a corpus that may influence model performance. We have identified some of these features, and then grouped them into five categories, described below.

Variability In this thesis, we call the *variability* of a vision-language dataset the presence of a diverse range of images and texts in the dataset. Several metrics can refer to the variability of a dataset, such as the number and distribution of object categories present in the whole dataset or the size of the vocabulary.

High variability in the data makes it easier to use the pre-trained model on a wide variety of downstream tasks. Indeed, many of these tasks may use data similar to the pre-training data. For this reason, CLIP ensures a variety of semantic objects in its pre-training data (Radford, J. W. Kim, et al. 2021). To that end, images are collected to cover Wikipedia semantic objects. The authors evaluate CLIP on numerous tasks and observe, in a zero-shot setting, competitive results with specialized models on these tasks. However, the authors also observe that CLIP shows poor generalization on out-of-distribution data, such as images out of the pre-training domains. Thus, the greater the variety of visual elements covered by the pre-training dataset, the better the model's performance. Moreover, the BLIP model (J. Li, D. Li, Xiong, et al. 2022) uses automatically generated captions to augment the pre-training data. The authors find that generating captions with greater variability increases model performance, rather than generating more likely captions.

Accuracy In this thesis, we call the *accuracy* of a vision-language dataset the degree to which the visual part and the textual part of an instance correspond to each other. Though it can be affected by monomodal characteristics of the data, such as grammar, we primarily refer to the text-image relationship. Depending on the type of instance, this characteristic can be subjective. It is best evaluated, if possible, using human evaluation.

BLIP authors (J. Li, D. Li, Xiong, et al. 2022) find that using inaccurate data during pre-training has a negative effect on performance, and develop a filtering technique to eliminate it. In addition, by studying model performance on different datasets, Hendricks, Mellor, et al. 2021 show that a model pre-trained on SBU Ordonez et al. 2011 performs less well on downstream tasks than those trained on smaller datasets, such as MS COCO. They correlate this with the fact that SBU data shows less overlap between text objects and words than other datasets. This seems consistent, as the filtering method used to generate SBU relies little on text-image similarity. Furthermore, Hendricks and Nematzadeh 2021 shows that models trained on manually

annotated data such as MS COCO (T.-Y. Lin et al. 2014), which are less noisy, are more sensitive to slight semantic differences between two instances than models trained on automatically collected data, such as Conceptual Captions (Sharma et al. 2018).

Combinatoriality In this thesis, we call *combinatoriality* the number of elements described by an annotation and the complexity of their interactions. Indeed, some images can show only a single object, while others show multiple objects with various relationships. In a dataset, annotations may mostly focus on the central point of an image, while in other datasets may emphasize the relationships between the various objects. As such, combinatoriality can widely vary depending on the type of dataset. The combinatoriality of the instances of an image-text dataset can be estimated using: part of speech tags and dependencies, the number of objects in an image, the correspondence between objects in the image and in the caption.

In chapter 3, we have found that the presence during pre-training of more descriptive, manually annotated captions can help models better understand multimodal dependencies. Similarly, we have found that the use of datasets favoring spatial reasoning seems necessary for multimodal understanding of position concepts, as LXMERT does Tan et al. 2019 by using visual reasoning datasets during pre-training (VQA (Antol et al. 2015), GQA (Hudson et al. 2019), VG-QA (Yuke Zhu et al. 2016)).

Bias Machine learning models amplify biases present in their datasets (J. Zhao et al. 2017). Indeed, data and annotations are two of the five main sources of bias (Hovy et al. 2021). In particular, vision-language transformers are prone to gender bias (Hendricks, K. Burns, et al. 2018). These models also sometimes rely on textual bias rather than visual information (Y. Goyal et al. 2017). We have also noticed this fact in chapter 3. In addition, the pre-training datasets are generally heavily biased in favor of western culture, and the performance of these models drops on examples outside this domain (F. Liu, Bugliarello, et al. 2021).

Similarity between pre-training and fine-tuning Singh, Goswami, et al. 2020 show that similarity between pre-training and fine-tuning data can strongly impact task performance. Similarly, Hendricks, Mellor, et al. 2021 find that when taking two datasets with the same images, the one with a higher language similarity (calculated using perplexity) with the downstream task data will lead to better performance in that task. Thus, if the method of annotating the evaluation images varies greatly from the pre-training images, the models may observe a drop in performance.

In the case of vision-language models aimed at many tasks, it is difficult to predict what the pre-training dataset should be. In particular, vision-language multimodality has shown a bias towards certain types of evaluation datasets during the last few years, such as COCO or Flickr. These datasets are based on manual annotations, with a focus on object descriptions. However, this data could be significantly different from real-world data in many applications.

4.5.2. Results

In this section, we compare pre-training with a manually annotated dataset COCO, and a web-crawled dataset, Conceptual Captions. The models pre-trained on the COCO dataset use the same configuration as previous experiments. It is pre-trained for 107 epochs, reaching 360k steps. For the Conceptual caption dataset, we use the same batch size but pre-train the models for a higher number of steps, reaching a total of 645k steps. Due to the size of the pre-training dataset, it amounts to 20 epochs. For comparison purposes, we also report the results of the COCO model for epoch 21, which amounts to 71k steps.

Dataset	Full	Cropped
COCO (epoch 106)	**0.25**	**0.40**
COCO (epoch 21)	0.24	0.27
Conceptual Captions	0.16	0.21

Table 4.9. – Accuracy of the models with different pre-training datasets for the multimodal dependencies task. *Full* refers to the task with full images, and *Cropped* refers to the task with cropped images, i.e., without distractor objects.

Tables 4.9 and 4.10 show the results of the models pre-trained on the two different datasets, COCO and Conceptual Captions. These results show that the model pre-trained on Conceptual Captions reaches significantly worse results than the one pre-trained on COCO. Indeed, the model does not outperform chance level for the M-Differences task. This does not seem to be due to the smaller number of epochs, as the ViLT model pre-trained for 22 epochs also reaches better result on most tasks. The results could be impacted by a bias in the evaluation datasets. Indeed, the L-BigramShift, M-Differences, and M-Color tasks are based on datasets annotated by humans that follow similar instructions, which is a completely different type of image-text pair than the Conceptual Captions dataset. The images themselves may be biased towards scenes with multiple objects and interactions, contrary to the Conceptual Captions dataset. However, the results on the V-Flower task confirm the fact that the model pre-trained on Conceptual Captions encodes much less relevant visual information than the other models pre-trained on COCO. Thus, the use of the Conceptual Captions dataset, though larger than the COCO dataset, seems less efficient for those tasks. This suggests that dataset size is far from the only factor to take into account when choosing a pre-training dataset. Indeed, a dataset with manual annotations such as COCO will be more efficient and reach better performances. This could be due to several dataset characteristics mentioned in the previous section, such as accuracy and combinatoriality.

Task	Dataset	Layer 4	Layer 8	Layer 12
L-BigramShift	COCO (epoch 106)	**0.54**	**0.66**	**0.63**
	COCO (epoch 21)	0.52	0.59	0.56
	Conceptual Captions	0.51	0.57	0.58
V-Flower	COCO (epoch 106)	**0.66**	**0.65**	**0.62**
	COCO (epoch 21)	0.60	0.61	0.55
	Conceptual Captions	0.53	0.29	0.25
M-Color	COCO (epoch 106)	**0.34**	**0.43**	**0.59**
	COCO (epoch 21)	0.32	**0.43**	0.56
	Conceptual Captions	0.30	0.32	0.51
M-Differences	COCO (epoch 106)	**0.53**	**0.57**	**0.63**
	COCO (epoch 21)	**0.53**	0.55	0.58
	Conceptual Captions	0.49	0.51	0.51
Agg.	COCO (epoch 106)	—	—	**0.62**
	COCO (epoch 21)	—	—	0.56
	Conceptual Captions	—	—	0.46

Table 4.10. – Accuracy of models pre-trained using different visual pre-training tasks for visual, textual and multimodal probing tasks, at attention layers 4, 8 and 12. Reported results are a mean over 5 runs of the accuracy. Agg. refers to the mean over the four tasks.

4.5.3. Discussion

A need for a more efficient pre-training As shown in this chapter, the use of very large datasets, such as the one used by the original model, greatly improves the performance of pre-trained models. However, using large quantities of data comes at a considerable environmental and economic cost, which makes them difficult to democratize. This raises the question of how to improve the efficiency of pre-training. A significant leverage is to filter and transform pre-training datasets for more efficient pre-training. One such case is the BLIP model (J. Li, D. Li, Xiong, et al. 2022), which proposes methods for textual data augmentation. However, little is known of what aspects of pre-training datasets have the most impact on vision-language models.

Characteristics of pre-training datasets In addition to dataset size, other aspects of pre-training datasets can have a major impact on model performance. By analyzing various studies in the field, we group these characteristics into variability, accuracy, combinatoriality and bias. Our results show that manually annotated datasets are significantly more efficient than web-crawled datasets, and that the size of the dataset does not always compensate for its quality. We advocate more precise filtering of large text-image datasets, to better meet these criteria we have identified.

Consequence for larger vision-language models These results seem to confirm that dataset quality has a major impact on performance. With the increase in size of models and datasets, it becomes more difficult to evaluate the impact of different filtering methods. Thus, there is a growing need for standardization in filtering methods. The use of further data augmentation methods, such as used in BLIP, could be explored to improve dataset quality and pre-training efficiency.

Large vision-language models such as Flamingo (Alayrac et al. 2022) and KOSMOS use interleaved vision-language data, meaning text with visual representations included at appropriate positions. This has led to the development of datasets based on natural multimodal documents, which allow for interleaved text-image data. This type of dataset could be an answer to the issue of dataset combinatoriality, as natural documents usually allow for more furnished descriptions. However, this could also lead to an increase reliance on monomodal textual bias. A recently proposed dataset, OBELICS Laurençon et al. 2023, proposes a new dataset of this type. They propose steps to filter the quality of an image-text dataset, for instance to filter out non human-written text[2]. Indeed, an increasing quantity of non-human written text is present in web-crawled datasets. An in depth evaluation of this dataset to evaluate bias and accuracy of image-text relations could help devise further filtering rules.

Future work It would be interesting to validate those hypotheses by creating subsets of web-crawled datasets and comparing performances. For instance, such subsets could be created by selecting instances using part-of-speech tags, specific vocabulary or object count, in order to ensure combinatoriality. Variability could be evaluated by selecting instances based on vocabulary distribution or object categories. Bias could be evaluated with n-grams, or restrained with the use of pseudonymization methods. Meta information surrounding the instances could also be used. Accuracy could be tested by selecting instances with sentences the most correlated with objects detected by object detectors. While variability, combinatoriality, and bias can reasonably be measured using metrics, it is difficult to measure text-image accuracy without introducing some sort of bias due to the type of model used. Further studies would enable us to refine the given factors and propose filtering methods.

We hope that by emphasizing the importance of data quality rather than data quantity, the pre-training of vision-language models becomes more efficient, and less costly. Various methods have been developed to assess the quality of a text or image dataset, but few works have been done on multimodal datasets. It would be interesting to adapt some of those methods, to filter or process multimodal datasets.

2. However, the list of domain names used for the creation of the dataset shows a significant bias towards specific domain names, for instance: www.dailymail.co.uk

4.6. Conclusion

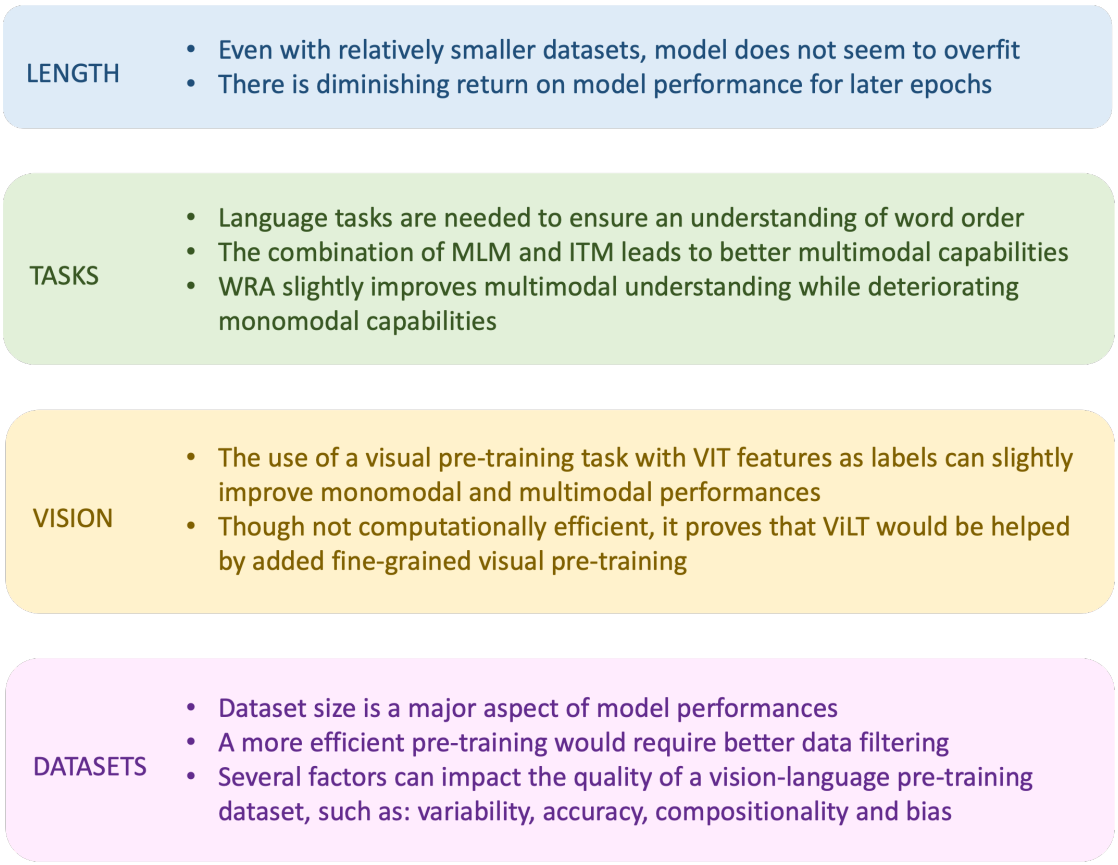


Figure 4.6. – Summary of the analysis of vision-language pre-training strategies described in this chapter.

In this chapter, we studied how different pre-training design choices can affect the performance of vision-language transformers on various capabilities. To that end, we realized ablation studies on the ViLT model, with the goal of reaching conclusions compatible with larger models. Figure 4.6 summarizes the findings.

Our results seem to show that the combined use of ITM and MLM significantly improve the ability of a model to encode multimodal information. The use of ITM only, would lead to a bad understanding of syntactic information such as word order.

In addition, while visual pre-training seems to be a limiting factor for ViLT, among other vision-language transformers, it is difficult to devise a visual pre-training task relevant for multimodal models. Indeed, the use of patches is not necessarily conductive to object-level representation. Thus, further work is required for the building of more efficient visual pre-training tasks.

Finally, we have also shown that while the size of a pre-training dataset and the length of pre-training are major factors that impact performances, data quality is also important. Thus, it would be interesting of directives for the filtering of web-crawled

pre-training data, in order to improve pre-training efficiency. To that end, additional experiments using subsets of web-crawled datasets could help devise filtering and data-processing methods.

In this chapter, we used the probing and evaluation tasks designed in Chapter 3 to study the impact of pre-training choices on model capabilities. It would be interesting to study in future works the understanding of other multimodal concepts not considered in this chapter. This would require the building of appropriate evaluation tasks, in particular to limit potential monomodal bias.

5. Conclusion

Through this thesis, we look into the pre-training of vision-language transformer models. Our goal is to better understand their impact and limitations. To that end, we study state-of-the-art models as they currently are at the time of writing this thesis, and develop methods applicable to future state-of-the-art models. We are especially interested in evaluating the multimodal understanding of vision-language models. We first remind you of the questions raised in the introduction that we aim to answer with the elements provided in this thesis.

1. What are vision-language transformers? What sets them apart from previous multimodal machine learning models? How have they evolved since their development?

Vision-language transformers stem from an application of the Transformer architecture to the field of vision-language multimodality. During the last few years, increasingly numerous variations of these models have been developed, using different multimodal architectures, pre-training tasks, and datasets. We describe various pre-training methods for vision-language transformers in Chapter 1. In a shift from previous multimodal models, they can be applied to many downstream tasks, by pre-training them on a large amount of text/image data. Thus, vision-language transformers have recently been scaled up, leading to the emergence of large vision-language models. These models can be said to belong to the category of foundation models.

Several research trends have emerged for the pre-training of vision-language models. In particular, recent models study the use of large language models as the basis of multimodal models, which require fewer resources than pre-training a large vision-language models from scratch. However, the field of vision-language multimodality is not yet stabilized. Indeed, there is no consensus yet on the most efficient pre-training for vision-language transformer models. Our understanding of those models is still at in its early stages.

2. What methodology should be used to evaluate such models, or in a broader sense, general-purpose multimodal models? What are the difficulties?

Our understanding of vision-language transformer models is still limited. Indeed, at the start of this thesis, evaluation methods did not offer much insight on the weaknesses of those methods. With the democratization of machine learning methods to real-world applications, it has become increasingly important to be able to explain the performances of a model. However, evaluation methods used for vision-language transformers lack in interpretability. Indeed, while fine-tuning tasks are useful to compare state-of-the-art models, they are insufficient to clearly pinpoint the weaknesses

of a model, as they evaluate complex multimodal reasoning. We argue that providing more granular evaluation methods, more easily interpretable, would be a major asset for a better understanding of those models. In particular, we propose a methodology for an evaluation of vision-language transformer models as exhaustive as possible. This method considers, in priority, the context of those models. Indeed, it is important to be aware of the context of use of a model, in terms of domains and applications. Then, a model should be evaluated on a range of granular capabilities with a coverage corresponding to those applications. To that end, we need to characterize the multimodal understanding of a vision-language model, and identify the possible granular capabilities. Therefore, we make a first attempt at a taxonomy of vision-language capabilities. We also propose a projection of existing evaluation tasks in this taxonomy. We think that it is important to aim for a more exhaustive evaluation methods for vision-language models, especially for foundation models aimed at a wide array of tasks. As explained in Chapter 2, this taxonomy is not yet exhaustive. We hope that this evaluation method could offer more insight regarding possible weaknesses of a model in real-world settings. This methodology should be transferable to other types of foundation models, and in particular to the use of other modalities.

3. How should vision-language tasks and datasets be created? What are the risks encountered when building such tasks?

In chapter 3, we design evaluation tasks and build corresponding datasets to evaluate state-of-the-art vision-language models on several textual, visual and multimodal capabilities. We use two evaluation methods for our experiments: probing evaluation and pre-training head evaluation. We are inspired by similar methods in Natural Language Processing, which aim at understanding the inner workings of transformer models. We specifically design methods to avoid the overreliance on monomodal bias, which can significantly impact the performance of vision-language models.

In Section 5.1, we discuss the design of evaluation tasks. Indeed, evaluating a vision-language model is a difficult endeavor, due to the complex nature of the models. Drawing from our experience in task design for vision-language evaluation, and from other works in the domain, we propose guidelines for the creation of vision-language evaluation tasks.

4. What are the strengths and weaknesses of vision-language transformer models? Are some aspects of multimodal understanding easier to understand for them than others?

Through the experiments in Chapter 3, we notice that vision-language model representations encode multimodal information useful for tasks such as color identification or the understanding of object categories. However, they have noticeable weaknesses in terms of visual capabilities and multimodal understanding. Indeed, several state-of-the-art models do not reach monomodal baselines for visual probing tasks, showing that visual pre-training may be a limiting factor, especially in terms of their ability to understand the details of an object. In addition, their pre-training does not help them encode multimodal information related to object size and position, where they rely considerably on textual bias.

Further experiments on synthetic data show that the poor understanding of position is not due to a lack of understanding positional information at a visual level. This weakness seems to be caused both by a difficulty to associate words to the corresponding visual information, and the compositionality of the instances (in the case of the relative position of two objects). Thus, an appropriate multimodal pre-training task and precisely annotated data could help multimodal models reach a better understanding of position.

Finally, we study more in depth their understanding of multimodal dependencies, through a noun-predicate dependency task. We find that though some models have learned to differentiate instances using multimodal dependencies, this is not the case for all state-of-the art models. Indeed, choices in pre-training can significantly impact the multimodal understanding of those models. Furthermore, the ability to understand multimodal dependencies seems to be uncorrelated to the understanding of more general multimodal concepts. Indeed, the use of noisy datasets seems to impact the ability of a model to learn fine-grained multimodal dependencies, where large amounts of data can be helpful to learn general multimodal concepts.

Through these experiments, we have evaluated vision-language transformers on several monomodal and multimodal capabilities. However, it would be interesting to propose a more exhaustive evaluation of multimodal capabilities, in order to pinpoint other potential weaknesses. In particular, the taxonomy proposed in Chapter 2 has several blind spots in terms of existing evaluation tasks, and building relevant datasets could help further improve the coverage of existing evaluation tasks.

5. How does the pre-training of a model affect its performances? What is the role of architecture, pre-training protocol, dataset and fine-tuning?

The experiments of chapter 3 enable us to draw some hypotheses regarding the pre-training of vision-language models. For instance, we find that fine-tuning does not seem to significantly impact the performance of vision-language models on the tested capabilities. However, these experiments have limitations, especially in terms of the comparison of different models. Indeed, state-of-the-art models are pre-trained using significantly different protocols, from datasets, to tasks and architecture. Due to this, we choose to focus chapter 4 on a single vision-language transformer, ViLT, and study how different pre-training protocols impact its performances. Results show that preloading a pre-trained monomodal model significantly helps vision-language model performances, even for multimodal capabilities. This result could explain the trend of using monomodal models as a basis for multimodal models, either to compute monomodal representations or as starting points of the multimodal transformer. Furthermore, experiments suggest that the combination of the linguistic pre-training task MLM and the multimodal pre-training task ITM significantly help multimodal capabilities. This could mean that using only one type of pre-training task, such as ITM or PLM, could be a limiting factor in the understanding of multimodal concepts. Though visual pre-training seems to be a limiting factor of many vision-language models, building an appropriate visual task for ViLT is complex. Indeed, our experiments show that while improvement in visual understanding can be noticed

with the MPR task using a ViT teacher model, this does not necessarily equate to a better understanding of multimodal concepts. Our hypothesis is that it is due to the visual representations used by ViLT. Indeed, pixel patches are used as visual input representations, which is not necessarily conductive to multimodal semantics. However, other options, such as the use of object detector features, have also shown weaknesses, especially in terms of fine-grained visual understanding in Chapter 3. Thus, we think that further work is needed to design, either to design visual pre-training tasks with more multimodal impact, or to elaborate a visual preprocessing architecture conductive to multimodal learning. Finally, we study the impact of pre-training datasets. We find that while dataset size has a major impact on vision-language model performance, the quality of a dataset is as important as its size for the tested capabilities. We identify several dataset characteristics that could influence vision language models, such as variability, accuracy, combinatoriality, and bias.

Most of these results can help us make inferences about newer models. However, several questions remain unanswered. Indeed, those experiments may not all be directly transferable to new models. Indeed, models based on large language models are significantly different from ViLT in terms of pre-training protocol. As such, it would be interesting to evaluate results from similar experiments with more recent, state-of-the-art, models.

6. What are the potential impacts of vision-language transformer models?

While performance is a major aspect of model evaluation, it is far from the only one. This is especially the case for models aimed at being used in real-world applications. Indeed, vision-language model evaluation should include a study of their societal and environmental impact. Though we do no focus on this in this work, we detail key points in Section 5.2.

In the next sections, we detail first guidelines for vision-language model evaluation, then describe ethical concerns involved in pre-trained vision-language transformers. Finally, we discuss several research challenges that could be investigated in future work.

5.1. Lessons Learned from Evaluating Vision-Language models: Task Design

In this section, we discuss the evaluation of vision-language models, with a focus on the building of evaluation tasks. More specifically, we study tasks aimed at evaluating a specific multimodal capability. The taxonomy presented in chapter 2 aims at providing a more precise and exhaustive evaluation of vision-language models. Indeed, they should be evaluated on a broad range of capabilities that covers their possible requirements. This would enable a better diagnosis of the performance of vision-language models.

We presented existing tasks that could be useful for the evaluation of the capabilities, linking them to the taxonomy. However, those tasks do not cover the whole range of the taxonomy. Thus, some capabilities mentioned in the taxonomy have no way of being evaluated yet. The identification of those capabilities and subsequent creation of relevant evaluation tasks could enable us to remedy the gaps in the evaluation of vision-language models. In addition, with the progress of vision-language machine learning, the difficulty of tasks may not always be appropriate for their intended use. This would also require the creation of new evaluation tasks. To answer this problem, we study in this section the creation of evaluation tasks targeting a specific vision-language capability.

The first stage of task creation is to define the task: its goal, potential applications, and basic design. The taxonomy created in chapter 2 can help in the selection of the capability to evaluate.

5.1.1. Task Difficulty

Evaluating multimodal understanding In the case of vision-language models, we are especially interested in the evaluation of multimodal skills. Those skills can be evaluated through different levels of multimodal understanding, depending on the design of the task. Indeed, in order to validate a vision-language model's understanding of a multimodal concept c, several types of tasks can be designed.

The experiments realized in chapter 3 regarding the concept of position can illustrate this. Indeed, a task can evaluate the visual understanding of position, through a classification task based on the position of an object in the image. If a model can correctly classify the instances, it means it successfully extracted the necessary visual information. Then, a task can evaluate the multimodal understanding of position, by determining whether a model can correctly associate the visual information with the words. A model that can manage to extract the relevant visual information may not be able to perform this task. Indeed, it may not be able to link this visual information to the corresponding words in the caption. In addition, the concept can be associated to other vision-language capabilities, like the ability to understand multimodal dependencies. For instance, using the same task with multiple objects requires the model to have several capabilities. The model needs to be able to identify the relevant object in a crowd of similar objects, to extract the visual positional information specifically related to this object, and to link it to the corresponding words. In that case, the inability of a model to answer correctly may not be due to its inability to understand the concept of position, and could be due to its inability to identify an object in a crowd.

As a result, it is important to be aware of what a task precisely evaluates: the extraction of visual or textual information, the use of multimodal concepts, the use of multiple capabilities simultaneously.

Other factors impacting difficulty A model's ability to understand vision and language concepts also depends on other factors that can increase the difficulty of

a task. The level of task difficulty must be carefully studied. If the task is too easy, it will not clearly reflect the weaknesses of the models. One should try to ensure that there are easy and difficult instances, for both humans and machines. Indeed, our experiments in Chapter 3 have shown that building too difficult tasks (in this case, the position, and size identification tasks in Section 3.3), can make comparison between models difficult. In Diwan et al. 2022, authors list several criteria of that lead to a vision-language model failing at a task. We use them as inspiration to compile a list of possible factors impacting difficulty:

— Details: When a task uses minute details to differentiate between two concepts. For instance, very small variations of position to differentiate between left and right;
— Domain: When an evaluation task uses images or text from a domain that is not well represented in the pre-training dataset. For example, paintings, when a model mainly trained on natural images.
— Quality: The use of pixelated images or text with mistakes can lead the model to make more errors.
— Complexity: The structural complexity of a scene and caption can greatly impact the difficulty of an instance. For instance, identification in a crowd is easier if there are fewer people.
— Ambiguity: Some tasks rely on subjective concepts, which can vary between different human annotators. For instance, the understanding of size ('small'/'big').
— The number of capabilities needed for the task. Some instances may require multiple capabilities, such as a fine-grained understanding of objects, position, and dependencies.

Depending on the intended use of a model, some of those criteria may be more relevant than others. For instance, robustness to low data quality may be a sought quality in one model, but not in others.

5.1.2. Task Design

The task should be generic enough in its implementation that it will be able to evaluate a wide range of vision-language models. Not all models will be able to access individual representations of words, or parts of the image. Indeed, all models are not pre-trained on the same tasks, nor do they have the same ways of encoding text or images, as seen in chapter 1. For instance, the study of object-based representations is not appropriate for models using pixel patches.

Most models have been pre-trained on a task similar to *image-text matching* or image-text *contrastive learning*. As such, they have been trained to match images and text using negative and positive pairs, with positive pairs indicating a matching relationship between image and text and negative pairs indicating a discrepancy. Thus, they could be evaluated on a similar type of task. Some other possible methods include generation methods, which can be used for example on models pre-trained on *prefix language modeling*. The new trend of using prefix language modeling as primary multimodal task may require other kinds of evaluation task, more adapted to

generative vision-language models. However, generative tasks are more difficult to evaluate, due to the lack of true gold labels.

Image-text matching with minimal differences In the case of image text matching, building image-text pairs with matching relationships or discrepancies makes it possible to evaluate whether a vision-language model correctly links or differentiates specific concepts. In Chapter 3, we created multiple tasks based on ITM. We aimed to ensure that negative image-text pairs only differ from a positive image-text pair with respect to a concept related to the capability we want to evaluate. Thus, the task is precise enough that the performance of the model on this task is a direct reflection of the ability of the model on that skill. Indeed, it evaluates whether the model can discriminate between two instances when only the concept which is specifically evaluated changes. However, using image-text matching with minimal differences does not prevent models from relying on monomodal bias. Thus, it is important to ensure that models cannot rely on bias to reach good performances.

5.1.3. Bias

Models often rely on spurious correlations from the training data, which can misrepresent their performance on evaluation tasks. Thus, it is important to limit the reliance on bias when evaluating models. Different factors will impact the quality of the dataset with regard to bias: its source, filtering and annotating process. Indeed, in Hovy et al. 2021, authors identify five main sources of bias in Natural Language Processing. Those that concern vision-language evaluation tasks the most are:

— Data: The selection of a data source can lead to representational bias and have harmful impacts on some groups of people. To counter that, it is important to pay more attention to the sources of the data, and whether it is balanced. Data selection can also lead to cultural bias, for example by increasing the visibility of western society at the detriment of other cultures.
— Annotations: Annotators rely on their experience when annotating a dataset. Those annotations can be skewed. A more representative selection of annotators can avoid some of that bias.
— Research design: Authors may make biased choices in their research. For instance, the use of the English language evaluation tasks is greatly favored compared to other languages.

Countermeasures can be used to limit the impact of bias. First, it is important to use a representative dataset for the desired application. In the case of general multimodal understanding, the dataset must be diverse, both semantically and syntactically. If the model is aimed at a real-world application, it is important that the dataset reflect the domain of this application, from the categories of semantic elements present in the inputs, to data quality.

Monomodal bias In the case of vision-language models, a major source of bias is 'monomodal bias'. Indeed, vision-language models often rely on monomodal bias if

they cannot extract relevant information from the input. We call 'monomodal bias' the reliance of a multimodal model on spurious training bias from a modality instead of factual information from the other modality. Vision-language models are pre-trained on unbalanced data, which may lead them to learn spurious correlations. One famous case is the color of a banana, which is disproportionately yellow in datasets (Cadene et al. 2019). When asked what is the color of a banana, a model might predict the answer yellow without using any visual information.

They can be several ways to ensure that a model may not reach significantly better results than another by only relying on only one modality:

— Balanced classes and out of distribution test sets: Some tasks and datasets lend themselves to the computation of specific statistics on the categories. Those can be used to determine whether categories are balanced. These methods do not entirely control for the use of monomodal biases, but they can also help avoid other categorical bias, or bring awareness to them during dataset creation.
— Baseline Experiments: It can also be helpful to use baseline experiments to make sure that the models use both visual information and textual information (e.g., Section 3.3).
— Counter-balanced Instances: This method, used in Nikolaus and Fourtassi 2021; Thrush et al. 2022, as well as Section 3.5, is useful to make sure that models do not rely on monomodal bias to achieve good performances in evaluation tasks. However, this method can add major requirements for the collection of the dataset for the task, which can limit its size and diversity.

To ensure that those models do not 'cheat' by relying on monomodal bias or spurious correlations from their training datasets, the evaluation dataset should be carefully balanced if possible using a 'counterbalanced' methodology. Additionally, monomodal models can be used to provide monomodal baselines of the task.

Subjectivity Another source of bias that specifically impacts vision-language models is the subjectivity of annotations. Indeed, most datasets are based on images and human- or computer- generated captions. Those captions reflect an interpretation of the images, but different annotators may lead to a completely different caption, both in terms of objects of interest, and the words used to describe them. There are three main factors to consider:

— The identity of the annotator: Our understanding of images is based on our knowledge and experiences, and can be deeply subjective.
— Guidelines: People will not necessarily focus on the same aspects of instances such as images, and annotation guidelines can impact the annotations.
— Context: The type of dataset from which an instance is extracted will usually impact its annotation, as people usually take context into account (e.g., social media vs. news articles).

As a result, there is often no 'gold' label in vision-language understanding. Our understanding of the world is biased, and this is reproduced in text-image data. A machine learning model hoping to reach a 'general' multimodal understanding would have to take into account the diversity of possible experiences. Some measures, such

as inter-annotator agreement, can help reach this goal. However, even if precautions are taken during the creation of the task and dataset, it is important to keep potential biases in mind during the analysis of the results.

5.1.4. Model Behavior

Beyond evaluating a vision-language model on a specific skill, it is important to check its behavior, and to study how different factors might impact the performance of a model on this skill. We adapt the checklist developed in Ribeiro et al. 2020 for Natural Language Processing to Vision-Language Multimodality.

Robustness A vision-language model uses textual and visual inputs to reach a prediction. However, to be able to trust the predictions of this model, it should be robust to several types of variations of the input data. An example of robustness benchmark is Adversarial Glue (B. Wang et al. 2021), which evaluates the robustness of language models to adversarial changes to the input data. In the case of an image-text matching task, a model should be robust to changes that do not affect the relationship between text and image. We list below some possible changes.

First, a vision-language model should be robust to textual variations that do not impact the text-image relationship, such as:

— Presence or absence of a phrase not directly related to the image
— Complexity of the syntax

Then, the model should also be robust to variations of the visual input that does not affect the text-image relationship:

— Presence or absence of a visual element not directly related to the text
— Complexity of the image structure

Finally, it should also be robust to changes in data quality.

Coherence of the model It is also important to evaluate a model's change of behavior between two semantically different instances, corresponding to the directional test in the Checklist methodology (Ribeiro et al. 2020). Indeed, it can be helpful to know whether a model is consistent in its predictions and shows coherent behavior, particularly for real-world applications on sensitive data.

In the case of an image-text matching task, the goal of those tests would be to evaluate whether a model's change of prediction is consistent with the difference in semantics. To that end, directional tests can be implemented to check the *consistency* of a vision-language model. One could evaluate model behavior on a spectrum of positive and negative instances, rather than on two positive and negative examples. This can for example be possible for the evaluation of the understanding of *position* and *size*. The goal is to check whether there is a coherent directional behavior of the model on a spectrum of instances, and whether there is a defined threshold between two opposite concepts.

In this section, we present several concerns related to task design for vision-language multimodality. Creating tasks with minimal differences in data is helpful to evaluate

the understanding of specific concepts. It is important to pay attention to bias, an especially monomodal bias, when building a dataset. Furthermore, tasks could consider robustness and consistency checks to make results more interpretable and provide insights on the models.

5.2. Ethical Concerns

In this thesis, we have emphasized the evaluation of models' capabilities. However, beyond the performance of vision-language models, several issues are to take into account. Namely, the use of vision-language transformer models pre-trained on large datasets raises several ethical concerns. We introduce in this section issues related to bias, data collection, environmental and societal impact.

Bias We have discussed the issue of bias in several sections (Section 1.3.1, Section 4.5.1, Section 5.1.3 paper), and bias in both Natural Language Processing and Computer Vision models has been largely documented during the past few years. In particular, bias can have societal impacts, due to a lack of representation in the data, or blind spots in the evaluation of models (e.g., related to the perceived gender, skin color, age, or social category of a person). This is called representational bias. The combined use of the vision and language modalities can make it more difficult to assess the impact of representational bias on a model. Ruggeri et al. 2023 find that state-of-the-art vision-language models show noticeable and potentially harmful bias, both for pre-trained and fine-tuned models.

Data collection We have also discussed in Section 1.3.1 issues related to data collection. Most vision-language models are nowadays, at least in part, trained on web-crawled image-text data. However, such data is usually collected not only without explicit consent from the interested parties, but also sometimes despite explicit non-consent. Initiatives have recently been implemented to avoid copyright issues when collecting training datasets. For instance, the Spawning API [1] mentioned in Chapter 4 aims at ensuring that models do not use images from URLs that have an opt-out status. This is in line with recent efforts of different governments, such as the European Union, to ensure a better protection of copyright and privacy on the internet. Yet, regulatory laws can evolve, and it would be interesting to devise other potential solutions to the issue of data collection.

The well-being of human annotators can also be severely impacted during data collection or model fine-tuning. This concern has already been raised for language-only models, such as ChatGPT [2]. The additional use of the visual modality raises its own issues, as when people have to work on potentially harmful content, such as NSFW or illegal images.

1. https://api.spawning.ai/spawning-api
2. https://time.com/6247678/openai-chatgpt-kenya-workers/

Environmental impact The pre-training and use of large-scale models such as transformers require a significant amount of energy (Strubell et al. 2019) and water (George et al. 2023). One of the stated goals of foundation models is to avoid the proliferation of large numbers of small-scale models. Indeed, one model could be applied to a wide range of tasks from different domains. However, in practice, foundation models continue to be developed, requiring each time a significant quantity of resources. In addition, there is a lack of awareness in the public as for the environmental impact of using such models. This leads to environmental impact in the form of energy consumption, which, depending on the country, can lead to significant green house gases emissions, as well as water consumption.

Societal impact Since the development of new generative deep learning methods such as transformers and diffusion models, there has been an increasingly widespread use of such models by industries and the public. However, the use of deep learning models in real-world applications can have a significant impact on society. Beyond potential flaws of those model such as harmful bias, such models can have an effect on people's livelihood. For instance, while ChatGPT is still recent, it has considerably affected the sector of education. In addition, the creation of deep fakes using such models is a very realistic threat for our societies. This should call for more interdisciplinary research on the impact of foundation models, as well as their potential real-world applications.

The past year has seen the development of vision-language foundation models aimed at the public. While the recent improvements of vision-language transformers are an opportunity for their applications to many real-world problems, it is important to consider their potential societal and environmental impact.

5.3. Future Work

As illustrated in this thesis, there are still many unanswered questions surrounding vision-language transformer models, and multimodal machine learning in general. We present in this section possible directions for future research. First, following the development of state-of-the-art large vision-language models, it would be interesting to evaluate those newer models. To that end, we could build on the methods developed in this thesis. We present in Section 5.3.1 interesting challenges concerning the evaluation of multimodal models. Then, we have shown through this thesis that there is no consensus yet on the architecture of multimodal models. From our results, we have determined several avenues of research that could be interesting to explore to build on vision-language transformers (Section 5.3.2).

5.3.1. Evaluation

Identifying real-world applications The recent improvement in performances of vision-language transformers has paved the way for new potential multimodal applications. However, the development of vision-language models is still somewhat disconnected from their use in real-world applications. It would be interesting to study in depth those applications, to identify use cases, domains, relevant data, risks, and challenges. Indeed, use cases could require data from distinct domains, and concern different aspects of text-image relationships. For instance, text-image relationships in multimodal documents such as news articles or education books are different from text-image relationships in vision-language navigation.

An in-depth study of potential vision-language applications for foundation models would answer important questions. In particular, it would be useful for more targeted evaluation of foundation models. In that respect, it could help complete the taxonomy introduced in 2. To that end, it would be necessary to obtain information from relevant specialists in different fields. Indeed, they could help anticipate possible challenges and risks regarding the use of multimodal foundation models in their field. This could help reduce risks during the development of such models, and before their introduction to the public. Those risks could consist in model flaws such as bias, but also in potential harmful use of those models.

Adapting the taxonomy for an exhaustive evaluation of multimodal foundation models The taxonomy presented in Chapter 2 is a first attempt of a more exhaustive evaluation of foundation models. In addition to the in-depth study of vision-language applications, one also could seek collaborations with researchers in linguistics and visual literacy to help better characterize the multimodal understanding of a model. Indeed, they could bring additional perspectives relevant to the evaluation of vision-language capabilities. Furthermore, the study of other modalities beyond the vision and language could help improve the existing taxonomy.

To build upon this taxonomy, it would be interesting to create a systematic evaluation protocol for multimodal foundation models. Indeed, such models have started being used by the public. A systematic evaluation protocol could help raise awareness on the weaknesses of those models, and help users choose between different models depending on their application. This protocol could recommend a selection of evaluation tasks depending on the use case of the model, in order to establish a good coverage of possible challenges, while restricting tasks unrelated to the application. Indeed, having a limited number of evaluation tasks to interpret would make it easier for end users to interpret the results. Thus, the most important characteristics of such a protocol would be, in our opinion, the coverage of relevant challenges and risks, and the interpretability of the results.

Task design for multimodal model evaluation There are still a number of blind spots that are not evaluated through vision-language tasks, or tasks whose difficulty make them irrelevant for state-of-the-art models. In addition, the concerns raised

in Section 5.1 regarding bias, and especially monomodal bias, make it difficult to interpret the results of some of those tasks. For instance, to our knowledge, few tasks evaluate semantic reasoning or the understanding of multimodal documents. We could also expand on a more detailed evaluation of multimodal dependencies or position, as these have been shown to be weaknesses of vision-language models.

Thus, future work could focus on the building of new evaluation tasks. It would be interesting for such tasks to also take into account the robustness of models, as explained in Section 5.1.4. Indeed, the robustness of a model is a major point of concern for use in real-world applications. This could help build trust in those models.

Another aspect of model evaluation that should be taken into account in task design is societal bias. Indeed, it would be interesting to investigate potential sources of bias, beyond those already identified (e.g., age, ethnicity, gender). The evaluation of model robustness could help identify some of those sources of bias.

Furthermore, there are still significant unanswered questions regarding the evaluation of multimodal generative models. Indeed, in this work, we have focused on evaluation based on classification. Evaluation of generation models is complex, both in the case of text generation and image generation. The standard remains human evaluation.

Building synthetic datasets Designing tasks evaluating the robustness of models could require building datasets using image-text pairs with minimal differences and counterbalanced examples, as explained in Section 5.1. However, such datasets are especially difficult to implement in the case of image data. Indeed, image data is more difficult and time-consuming to alter than textual data. In addition, existing manually annotated datasets carry a considerable risk of overlap between pre-training and evaluation datasets. This is in particular true for models pre-trained on very large-scale datasets using web-crawled data.

It would be interesting to explore the use of synthetic data for image-text evaluation datasets. Indeed, they have the added benefit of being able to control potential bias in the data. While some tasks (Johnson et al. 2017) have used synthetic data, they are usually not the focus of model evaluation, in particular due to the lack of object and scene variability. As such, the feasibility of such a methodology is not yet proven. There are several possible strategies for the creation of those datasets. On one hand, we could test the use of 3D rendered images, as used in the CLEVR dataset (Johnson et al. 2017), for more realistic datasets [3]. On the other hand, due to significant technological progress, image generation models could also be a viable solution for the creation of such datasets. It would be interesting to compare the two methods, in terms of building time, data control, and quality.

3. Some work was carried on this topic during my thesis, in collaboration with Nohayla Tanajyt, a master student whom I supervised during her internship.

5.3.2. Model Development

Vision-language models The fact that models are pre-trained on increasing amounts of data makes it difficult to design other pre-training methods with limited resources. Yet, there are still many possible research directions to build on state-of-the art vision-language models. Model compression can help train smaller models with similar performances. In addition to compressing large vision-language models, one could also try to compress a large language model. Indeed, several vision-language models now rely on large language models, which adapt by including visual embeddings as input. By compressing a large language model, one could attempt to reach similar linguistic performances in a smaller language model, and build upon this model to train a vision-language model.

It would also be interesting to design fine-tuning methods to help improve specific capabilities of vision-language models, for instance using adaptation methods similar to LoRa (E. J. Hu et al. 2021). It could be useful to improve specific capabilities of vision-language models that are not learned during pre-training. For instance, it could concern their understanding of position. It should also limit the potential negative impact to other capabilities. These methods could alleviate the restrictions caused by the lack of resources.

Less resource-intensive methods could also help improve design of large vision-language models, especially in terms of visual input representations. Indeed, through this thesis, we have seen that the visual pre-training of vision-language models can be an important limiting factor for their multimodal capabilities.

Grounding with other modalities Research on vision-language transformers can transfer to other modalities. For instance, the use of large language models as a basis of a large vision-language model can be transferred to other modalities, such as audio. This would require work regarding the type of embeddings used to represent each modality, in order to extract relevant information from each modality. This would also require further work on the possible interactions between modalities. Indeed, interactions between images and texts are complex, and can range from redundancy, to complementarity and contradiction. This would also be applicable to other modalities, and could lead to new challenges in terms of multimodal learning. For instance, the use of the video modality would lead to the necessity of considering temporal continuity.

Vision-language models have difficulty understanding some multimodal concepts, such as position and size. The inclusion of information related to world knowledge, for example through model bases, could help models reach a better understanding of those concepts. However, the inclusion of such information could require significant changes in model architecture or pre-training protocol. Indeed, vision-language transformers are usually pre-trained on pairs of image-text data or multimodal documents, which may be too restrictive for the inclusion of other modalities.

More generally, the diversity of possible interactions between modalities could lead to the development of new types of architecture for multimodal models, and especially

in terms of multimodal fusion.

In the past few years, advances in machine learning have considerably changed the field of vision-language multimodality. This thesis aims to provide perspective on newer models, through the lens of vision-language transformers.

Bibliography

Abacha, Asma Ben et al. (2019). "VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019." In: *CLEF (working notes)* 2.6 (cit. on pp. 87, 89).

Agrawal, Aishwarya et al. (2018). "Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering". In: pp. 4971–4980. URL: https://openaccess.thecvf.com/content_cvpr_2018/html/Agrawal_Dont_Just_Assume_CVPR_2018_paper.html (visited on 05/13/2022) (cit. on pp. 67, 132, 134).

Agrawal, Harsh et al. (2019). "Nocaps: Novel object captioning at scale". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957 (cit. on p. 68).

Akula, Arjun et al. (July 2020). "Words Aren't Enough, Their Order Matters: On the Robustness of Grounding Visual Referring Expressions". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6555–6565. DOI: 10.18653/v1/2020.acl-main.586. URL: https://aclanthology.org/2020.acl-main.586 (visited on 05/13/2022) (cit. on pp. 68, 132).

Alayrac, Jean-Baptiste et al. (2022). "Flamingo: a Visual Language Model for Few-Shot Learning". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. URL: https://openreview.net/forum?id=EbMuimAbPbs (cit. on pp. 50, 51, 53, 54, 64, 76, 150, 178).

Allahyari, Mehdi et al. (2017). "Text summarization techniques: a brief survey". In: *arXiv preprint arXiv:1707.02268* (cit. on p. 44).

Antol, Stanislaw et al. (2015). "VQA: Visual Question Answering". In: pp. 2425–2433. URL: https://openaccess.thecvf.com/content_iccv_2015/html/Antol_VQA_Visual_Question_ICCV_2015_paper.html (visited on 04/03/2022) (cit. on pp. 67, 131, 144, 175).

Arkin, Ershat et al. (2023). "A survey: Object detection methods from CNN to transformer". In: *Multimedia Tools and Applications* 82.14, pp. 21353–21383 (cit. on p. 46).

Bahdanau, Dzmitry et al. (2014). "Neural machine translation by jointly learning to align and translate". In: *arXiv preprint arXiv:1409.0473* (cit. on p. 38).

Bai, Yutong et al. (2021). "Are transformers more robust than cnns?" In: *Advances in Neural Information Processing Systems* 34, pp. 26831–26843 (cit. on p. 48).

Baltrušaitis, Tadas et al. (2018). "Multimodal machine learning: A survey and taxonomy". In: *IEEE transactions on pattern analysis and machine intelligence* 41.2, pp. 423–443 (cit. on p. 38).

Bao, Hangbo, Li Dong, et al. (2021). "Beit: Bert pre-training of image transformers". In: *arXiv preprint arXiv:2106.08254* (cit. on pp. 47, 60).

Bao, Hangbo, Wenhui Wang, Li Dong, Qiang Liu, et al. (2021). "Vlmo: Unified vision-language pre-training with mixture-of-modality-experts". In: *arXiv preprint arXiv:2111.02358* (cit. on pp. 53, 64).

Bao, Hangbo, Wenhui Wang, Li Dong, and Furu Wei (2022). "Vl-beit: Generative vision-language pretraining". In: *arXiv preprint arXiv:2206.01127* (cit. on pp. 53, 60, 64).

Bardin, Laurence (1975). "Le texte et l'image". In: *Communication & Langages* 26.1, pp. 98–112 (cit. on p. 85).

Basaj, Dominika et al. (2021). "Explaining Self-Supervised Image Representations with Visual Probing". In: *International Joint Conference on Artificial Intelligence* (cit. on p. 82).

Bazi, Yakoub et al. (2023). "Vision–language model for visual question answering in medical imagery". In: *Bioengineering* 10.3, p. 380 (cit. on p. 29).

Bender, Emily M et al. (2020). "Climbing towards NLU: On meaning, form, and understanding in the age of data". In: *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5185–5198 (cit. on p. 46).

Bender, Emily M. et al. (Mar. 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*. FAccT '21. New York, NY, USA: Association for Computing Machinery, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922 (visited on 10/12/2022) (cit. on pp. 29, 76).

Birhane, Abeba et al. (2021). "Multimodal datasets: misogyny, pornography, and malignant stereotypes". In: *arXiv preprint arXiv:2110.01963* (cit. on p. 57).

Biten, Ali Furkan, Lluis Gomez, et al. (2019). "Good news, everyone! context driven entity-aware captioning for news images". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12466–12475 (cit. on pp. 94, 98).

Biten, Ali Furkan, Ruben Tito, et al. (2019). "Scene text visual question answering". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301 (cit. on p. 98).

Blodgett, Su Lin et al. (July 2020). "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. URL: https://aclanthology.org/2020.acl-main.485 (visited on 10/12/2022) (cit. on p. 83).

Boecking, Benedikt et al. (2022). "Making the most of text semantics to improve biomedical vision–language processing". In: *European conference on computer vision*. Springer, pp. 1–21 (cit. on pp. 31, 86).

Bogin, Ben et al. (2021). "COVR: A Test-Bed for Visually Grounded Compositional Generalization with Real Images". en. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 9824–9846. DOI: 10.18653/v1/2021.emnlp-main.774. URL: https://aclanthology.org/2021.emnlp-main.774 (visited on 05/07/2022) (cit. on p. 67).

Bommasani, Rishi et al. (2021). "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (cit. on p. 28).

Borji, Ali (2019). "Pros and cons of gan evaluation measures". In: *Computer Vision and Image Understanding* 179, pp. 41–65 (cit. on p. 79).

Bowman, Samuel R et al. (2015). "A large annotated corpus for learning natural language inference". In: *arXiv preprint arXiv:1508.05326* (cit. on pp. 91, 95).

Brown, Tom et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901 (cit. on pp. 29, 41, 45, 148).

Bugliarello, Emanuele, Ryan Cotterell, et al. (Sept. 2021). "Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs". en. In: *Transactions of the Association for Computational Linguistics* 9, pp. 978–994. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00408. URL: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00408/107279/Multimodal-Pretraining-Unmasked-A-Meta-Analysis (visited on 04/03/2022) (cit. on p. 140).

Bugliarello, Emanuele, Fangyu Liu, et al. (2022). "IGLUE: A benchmark for transfer learning across modalities, tasks, and languages". In: *International Conference on Machine Learning*. PMLR, pp. 2370–2392 (cit. on p. 80).

Burns, Andrea et al. (2022). "A dataset for interactive vision-language navigation with unknown command feasibility". In: *European Conference on Computer Vision*. Springer, pp. 312–328 (cit. on p. 86).

Cadene, Remi et al. (2019). "Rubi: Reducing unimodal biases for visual question answering". In: *Advances in neural information processing systems* 32 (cit. on p. 188).

Camburu, Oana-Maria et al. (2018). "e-snli: Natural language inference with natural language explanations". In: *Advances in Neural Information Processing Systems* 31 (cit. on pp. 91, 95).

Cao, Jize et al. (2020). "Behind the scene: Revealing the secrets of pre-trained vision-and-language models". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*. Springer, pp. 565–580 (cit. on pp. 82, 132, 134).

Cao, Yiyi et al. (2023). "Cucumber disease recognition with small samples using image-text-label-based multi-modal language model". In: *Computers and Electronics in Agriculture* 211, p. 107993 (cit. on p. 86).

Carion, Nicolas et al. (2020). "End-to-end object detection with transformers". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, pp. 213–229 (cit. on p. 46).

Caron, Mathilde et al. (2021). "Emerging properties in self-supervised vision transformers". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660 (cit. on p. 47).

Chang, Angel et al. (2017). "Matterport3D: Learning from RGB-D Data in Indoor Environments". In: *International Conference on 3D Vision (3DV)* (cit. on pp. 94, 98).

Changpinyo, Soravit et al. (2021). "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts". In: *CVPR* (cit. on pp. 55–57).

Chen, Mark et al. (2020). "Generative pretraining from pixels". In: *International conference on machine learning*. PMLR, pp. 1691–1703 (cit. on p. 47).

Chen, Ting et al. (2020). "A simple framework for contrastive learning of visual representations". In: *International conference on machine learning*. PMLR, pp. 1597–1607 (cit. on p. 42).

Chen, Yen-Chun et al. (2019). "Uniter: Learning universal image-text representations". In: (cit. on pp. 13, 15, 30, 49–51, 53, 59–61, 64, 67, 76, 109, 126, 138, 154).

Chen, Yuyang et al. (2022). "Multimodal detection of hateful memes by applying a vision-language pre-training model". In: *Plos one* 17.9, e0274300 (cit. on p. 87).

Chen, Zhenfang et al. (2020). "Cops-Ref: A New Dataset and Task on Compositional Referring Expression Comprehension". In: pp. 10086–10095. URL: https://openaccess.thecvf.com/content_CVPR_2020/html/Chen_Cops-Ref_A_New_Dataset_and_Task_on_Compositional_Referring_Expression_CVPR_2020_paper.html (visited on 03/15/2022) (cit. on pp. 68, 92, 98).

Cherian, Anoop et al. (2023). "Are deep neural networks SMARTer than second graders?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10834–10844 (cit. on pp. 95, 99).

Chiu, Tai-Yin et al. (2020). "Assessing image quality issues for real-world problems". In: *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3646–3656 (cit. on p. 99).

Cho, Jaemin et al. (2021). "Unifying vision-and-language tasks via text generation". In: *International Conference on Machine Learning*. PMLR, pp. 1931–1942 (cit. on pp. 53, 64).

Chollet, François (2019). "On the measure of intelligence". In: *arXiv preprint arXiv:1911.01547* (cit. on p. 78).

Chou, Shih-Han et al. (2020). "Visual question answering on 360 images". In: *2020 IEEE winter conference on applications of computer vision (WACV)*. IEEE, pp. 1596–1605 (cit. on pp. 94, 98).

Chowdhery, Aakanksha et al. (2022). "Palm: Scaling language modeling with pathways". In: *arXiv preprint arXiv:2204.02311* (cit. on p. 45).

Conneau, Alexis et al. (2018). "SentEval: An Evaluation Toolkit for Universal Sentence Representations". In: *arXiv preprint arXiv:1803.05449* (cit. on pp. 79, 82, 113).

Cubuk, Ekin D. et al. (2019). "RandAugment: Practical data augmentation with no separate search". In: *CoRR* abs/1909.13719. arXiv: 1909.13719. URL: http://arxiv.org/abs/1909.13719 (cit. on pp. 57, 155).

Das, Abhishek, Harsh Agrawal, et al. (2017). "Human attention in visual question answering: Do humans and deep networks look at the same regions?" In: *Computer Vision and Image Understanding* 163, pp. 90–100 (cit. on pp. 96, 99).

Das, Abhishek, Satwik Kottur, et al. (2017). "Visual dialog". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 326–335 (cit. on pp. 29, 86, 99).

De Vries, Harm et al. (2017). "Guesswhat?! visual object discovery through multi-modal dialogue". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5503–5512 (cit. on p. 99).

Devlin, Jacob et al. (June 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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423 (cit. on pp. 13, 38, 41–44, 51, 58, 61, 148, 155).

Díaz, Mark et al. (2022). "CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation". In: *2022 ACM Conference on Fairness, Accountability, and Transparency* (cit. on p. 57).

Diwan, Anuj et al. (Dec. 2022). "Why is Winoground Hard? Investigating Failures in Visuolinguistic Compositionality". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 2236–2250. URL: https://aclanthology.org/2022.emnlp-main.143 (cit. on pp. 81, 186).

Do, Virginie et al. (2020). "e-snli-ve: Corrected visual-textual entailment with natural language explanations". In: *arXiv preprint arXiv:2004.03744* (cit. on pp. 67, 95, 99).

Dong, Xiaoyi et al. (2022). "Cswin transformer: A general vision transformer backbone with cross-shaped windows". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12124–12134 (cit. on p. 48).

Dosovitskiy, Alexey et al. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=YicbFdNTTy (cit. on pp. 47, 50, 51, 119, 155).

Dou, Zi-Yi, Aishwarya Kamath, et al. (2022). "Coarse-to-Fine Vision-Language Pre-training with Fusion in the Backbone". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. URL: https://openreview.net/forum?id=o4neHaKMlse (cit. on pp. 51, 53, 64).

Dou, Zi-Yi, Yichong Xu, et al. (2022). "An empirical study of training end-to-end vision-and-language transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18166–18176 (cit. on pp. 52–54, 59, 60, 63, 64).

Edunov, Sergey et al. (2018). "Understanding back-translation at scale". In: *arXiv preprint arXiv:1808.09381* (cit. on p. 57).

Ettinger, Allyson (2020). "What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models". In: *Transactions of the Association for Computational Linguistics* 8, pp. 34–48 (cit. on pp. 46, 81).

Faghri, Fartash et al. (2018). "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives". In: URL: https://github.com/fartashf/vsepp (cit. on pp. 29, 37).

Fang, Zhiyuan et al. (2021). "Compressing visual-linguistic model via knowledge distillation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1428–1438 (cit. on p. 69).

Fenson, Larry et al. (2007). "MacArthur-Bates communicative development inventories". In: (cit. on p. 91).

Forbes, Maxwell et al. (2019). "Do neural language representations learn physical commonsense?" In: *arXiv preprint arXiv:1908.02899* (cit. on p. 46).

Frank, Michael C et al. (2017). "Wordbank: An open repository for developmental vocabulary data". In: *Journal of child language* 44.3, pp. 677–694 (cit. on p. 100).

Frank, Stella et al. (2021). "Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers". In: *CoRR* abs/2109.04448. arXiv: 2109.04448. URL: https://arxiv.org/abs/2109.04448 (cit. on p. 83).

Frome, Andrea et al. (2013). "DeViSE: A Deep Visual-Semantic Embedding Model". In: *NIPS* (cit. on p. 29).

Gan, Zhe et al. (2020). "Large-scale adversarial training for vision-and-language representation learning". In: *Advances in Neural Information Processing Systems* 33, pp. 6616–6628 (cit. on pp. 53, 63, 64).

Gao, Leo et al. (Dec. 2022). *EleutherAI/lm-evaluation-harness: v0.3.0*. Version v0.3.0. DOI: 10.5281/zenodo.7413426. URL: https://doi.org/10.5281/zenodo.7413426 (cit. on p. 79).

George, A Shaji et al. (2023). "The Environmental Impact of AI: A Case Study of Water Consumption by Chat GPT". In: *Partners Universal International Innovation Journal* 1.2, pp. 97–104 (cit. on p. 191).

Ghosal, Koustav et al. (2019). "Aesthetic Image Captioning From Weakly-Labelled Photographs". In: *arXiv preprint arXiv:1908.11310* (cit. on p. 99).

Goertzel, Ben (2014). "Artificial general intelligence: concept, state of the art, and future prospects". In: *Journal of Artificial General Intelligence* 5.1, p. 1 (cit. on p. 28).

Goodfellow, Ian et al. (2020). "Generative adversarial networks". In: *Communications of the ACM* 63.11, pp. 139–144 (cit. on p. 42).

Gordon, Daniel et al. (2018). "IQA: Visual Question Answering in Interactive Environments". In: *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE (cit. on pp. 94, 98).

Goyal, Yash et al. (2017). "Making the v in vqa matter: Elevating the role of image understanding in visual question answering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913 (cit. on pp. 61, 67, 80, 83, 98, 132, 134, 140, 175).

Gu, Jing et al. (2022). "Vision-and-language navigation: A survey of tasks, methods, and future directions". In: *arXiv preprint arXiv:2203.12667* (cit. on pp. 30, 68, 90).

Gu, Jiuxiang et al. (2018). "Recent advances in convolutional neural networks". In: *Pattern recognition* 77, pp. 354–377 (cit. on p. 37).

Gui, Liangke et al. (2022). "Training vision-language transformers from captions alone". In: *arXiv preprint arXiv:2205.09256* (cit. on pp. 52, 53, 60, 64).

He, Kaiming, Xinlei Chen, et al. (2022). "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009 (cit. on pp. 47, 60, 169).

He, Kaiming, Haoqi Fan, et al. (2020). "Momentum contrast for unsupervised visual representation learning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738 (cit. on p. 47).

He, Kaiming, Xiangyu Zhang, et al. (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (cit. on pp. 119, 129, 169).

He, Tianxing et al. (2022). "On the Blind Spots of Model-Based Evaluation Metrics for Text Generation". In: *arXiv preprint arXiv:2212.10020* (cit. on p. 79).

He, Xuehai et al. (2020). "PathVQA: 30000+ Questions for Medical Visual Question Answering". In: *arXiv preprint arXiv:2003.10286* (cit. on pp. 94, 98).

Hendricks, Lisa Anne, Kaylee Burns, et al. (2018). "Women also Snowboard: Overcoming Bias in Captioning Models". In: pp. 771–787. URL: https://openaccess.thecvf.com/content_ECCV_2018/html/Lisa_Anne_Hendricks_Women_also_Snowboard_ECCV_2018_paper.html (visited on 04/19/2022) (cit. on pp. 83, 132, 134, 175).

Hendricks, Lisa Anne, John F. J. Mellor, et al. (2021). "Decoupling the Role of Data, Attention, and Losses in Multimodal Transformers". In: *Transactions of the Association for Computational Linguistics* 9, pp. 570–585 (cit. on pp. 54, 58, 174, 175).

Hendricks, Lisa Anne and Aida Nematzadeh (2021). "Probing Image-Language Transformers for Verb Understanding". en. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Online: Association for Computational Linguistics, pp. 3635–3644. DOI: 10.18653/v1/2021.findings-acl.318. URL: https://aclanthology.org/2021.findings-acl.318 (visited on 04/21/2022) (cit. on pp. 81, 142, 145, 174).

Hendrycks, Dan et al. (2019). "Benchmarking neural network robustness to common corruptions and perturbations". In: *arXiv preprint arXiv:1903.12261* (cit. on p. 79).

Hewitt, John and Percy Liang (2019). "Designing and interpreting probes with control tasks". In: *arXiv preprint arXiv:1909.03368* (cit. on p. 108).

Hewitt, John and Christopher D Manning (2019). "A structural probe for finding syntax in word representations". 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)*, pp. 4129–4138 (cit. on pp. 82, 112).

Hochreiter, Sepp et al. (1997). "Long short-term memory". In: *Neural computation* 9.8, pp. 1735–1780 (cit. on pp. 36, 38).

Honnibal, Matthew et al. (2015). "An improved non-monotonic transition system for dependency parsing". In: *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 1373–1378 (cit. on p. 113).

Houlsby, Neil et al. (2019). "Parameter-efficient transfer learning for NLP". In: *International Conference on Machine Learning*. PMLR, pp. 2790–2799 (cit. on p. 42).

Hovy, Dirk et al. (2021). "Five sources of bias in natural language processing". In: *Language and Linguistics Compass* 15.8, e12432 (cit. on pp. 175, 187).

Hu, Edward J et al. (2021). "LoRA: Low-Rank Adaptation of Large Language Models". In: *International Conference on Learning Representations* (cit. on p. 194).

Hu, Hexiang et al. (Oct. 2019). "Evaluating Text-to-Image Matching using Binary Image Selection (BISON)". en. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Seoul, Korea (South): IEEE, pp. 1887–1890. ISBN: 978-1-72815-023-9. DOI: 10.1109/ICCVW.2019.00237. URL: https://ieeexplore.ieee.org/document/9022357/ (visited on 05/07/2022) (cit. on p. 67).

Hu, Xiaowei et al. (2021). "Scaling Up Vision-Language Pretraining for Image Captioning". In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17959–17968 (cit. on pp. 53, 64).

Huang, Shaohan et al. (2023). "Language Is Not All You Need: Aligning Perception with Language Models". In: *arXiv preprint arXiv:2302.14045* (cit. on pp. 52, 53, 59, 61, 64, 69, 76).

Huang, Po-Yao et al. (2021). "Multilingual Multimodal Pre-training for Zero-Shot Cross-Lingual Transfer of Vision-Language Models". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2443–2459. URL: https://arxiv.org/abs/2103.08849 (cit. on p. 69).

Huang, Zhicheng, Zhaoyang Zeng, Yupan Huang, et al. (2021). "Seeing out of the box: End-to-end pre-training for vision-language representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12976–12985 (cit. on pp. 53, 64).

Huang, Zhicheng, Zhaoyang Zeng, Bei Liu, et al. (2020). "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers". In: *arXiv preprint arXiv:2004.00849* (cit. on pp. 53, 64).

Huang, Zhongzhen et al. (2023). "KiUT: Knowledge-injected U-Transformer for Radiology Report Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19809–19818 (cit. on p. 29).

Hudson, Drew A et al. (2019). "Gqa: A new dataset for real-world visual reasoning and compositional question answering". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709 (cit. on pp. 67, 98, 131, 140, 175).

Ilinykh, Nikolai et al. (2021). "What does a language-and-vision transformer see: The impact of semantic information on visual representations". In: *Frontiers in Artificial Intelligence*, p. 182 (cit. on p. 82).

Jain, Aashi et al. (2021). "Mural: multimodal, multitask retrieval across languages". In: *arXiv preprint arXiv:2109.05125* (cit. on p. 69).

Jia, Chao et al. (2021). "Scaling up visual and vision-language representation learning with noisy text supervision". In: *International Conference on Machine Learning*. PMLR, pp. 4904–4916 (cit. on pp. 53, 56, 64).

Jin, Xin et al. (2019). "Aesthetic attributes assessment of images". In: *Proceedings of the 27th ACM international conference on multimedia*, pp. 311–319 (cit. on p. 99).

Jing, Quanliang et al. (2021). "TRANSFAKE: multi-task transformer for multimodal enhanced fake news detection". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8 (cit. on p. 86).

Johnson, Justin et al. (2017). "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910 (cit. on pp. 14, 91, 92, 98, 105, 126, 193).

Kafle, Kushal et al. (2017). "An Analysis of Visual Question Answering Algorithms". In: *ICCV* (cit. on pp. 91, 98).

Kakogeorgiou, Ioannis et al. (2022). "What to hide from your students: Attention-guided masked image modeling". In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXX*. Springer, pp. 300–318 (cit. on p. 47).

Kamath, Aishwarya et al. (2021). "Mdetr-modulated detection for end-to-end multi-modal understanding". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1780–1790 (cit. on pp. 53, 64).

Kannojia, Suresh Prasad et al. (2018). "Effects of varying resolution on performance of CNN based image classification: An experimental study". In: *Int. J. Comput. Sci. Eng* 6.9, pp. 451–456 (cit. on p. 54).

Karpathy, Andrej et al. (2015). "Deep visual-semantic alignments for generating image descriptions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137 (cit. on p. 37).

Kayser, Maxime et al. (2021). "e-vil: A dataset and benchmark for natural language explanations in vision-language tasks". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1244–1254 (cit. on pp. 96, 99).

Kazemzadeh, Sahar et al. (2014). "Referitgame: Referring to objects in photographs of natural scenes". In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798 (cit. on pp. 68, 98).

Ke, Junjie et al. (2023). "VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10041–10051 (cit. on p. 99).

Kedra, Joanna (2018). "What does it mean to be visually literate? Examination of visual literacy definitions in a context of higher education". In: *Journal of Visual Literacy* 37.2, pp. 67–84 (cit. on p. 85).

Kim, Wonjae et al. (2021). "Vilt: Vision-and-language transformer without convolution or region supervision". In: *International Conference on Machine Learning*. PMLR, pp. 5583–5594 (cit. on pp. 13, 15, 50, 53, 58, 59, 63, 64, 76, 109, 126, 138, 146, 154).

Kirkpatrick, James et al. (2017). "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the national academy of sciences* 114.13, pp. 3521–3526 (cit. on p. 42).

Kolve, Eric et al. (2017). "AI2-THOR: An Interactive 3D Environment for Visual AI". In: *ArXiv* abs/1712.05474 (cit. on pp. 94, 98).

Koubaa, Anis (2023). "GPT-4 vs. GPT-3.5: A concise showdown". In: (cit. on p. 41).

Krishna, Ranjay, Kenji Hata, et al. (2017). "Dense-Captioning Events in Videos". In: *International Conference on Computer Vision (ICCV)* (cit. on pp. 93, 98).

Krishna, Ranjay, Yuke Zhu, et al. (2016). "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations". In: URL: https://arxiv.org/abs/1602.07332 (cit. on pp. 55, 56, 67, 140, 155).

Krizhevsky, Alex et al. (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html (visited on 06/24/2022) (cit. on p. 148).

Kuznetsova, Alina et al. (July 2020). "The Open Images Dataset V4". en. In: *International Journal of Computer Vision* 128.7, pp. 1956–1981. ISSN: 1573-1405. DOI: 10.1007/s11263-020-01316-z. URL: https://doi.org/10.1007/s11263-020-01316-z (visited on 04/17/2022) (cit. on pp. 57, 136).

Lamiroy, Bart (2014). "Interpretation, Evaluation and the Semantic Gap... What if we were on a Side-Track?" In: *Graphics Recognition. Current Trends and Challenges: 10th International Workshop, GREC 2013, Bethlehem, PA, USA, August 20-21, 2013, Revised Selected Papers 10*. Springer, pp. 221–233 (cit. on p. 80).

Laurençon, Hugo et al. (2023). "OBELISC: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents". In: *arXiv preprint arXiv:2306.16527* (cit. on p. 178).

LeCun, Yann et al. (2015). "Deep learning". In: *nature* 521.7553, pp. 436–444 (cit. on p. 36).

Lei, Jie et al. (2021). "Less is more: Clipbert for video-and-language learning via sparse sampling". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7331–7341 (cit. on p. 70).

Li, Gen et al. (2020). "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07, pp. 11336–11344 (cit. on pp. 53, 64).

Li, Junnan, Dongxu Li, Silvio Savarese, et al. (2023). "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models". In: *arXiv preprint arXiv:2301.12597* (cit. on p. 76).

Li, Junnan, Dongxu Li, Caiming Xiong, et al. (2022). "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". In: *International Conference on Machine Learning* (cit. on pp. 52, 53, 57, 58, 64, 70, 156, 174, 177).

Li, Junnan, Ramprasaath Selvaraju, et al. (2021). "Align before fuse: Vision and language representation learning with momentum distillation". In: *Advances in neural information processing systems* 34, pp. 9694–9705 (cit. on pp. 49, 50, 53, 63, 64, 174).

Li, Linjie et al. (2020). "Hero: Hierarchical encoder for video+ language omni-representation pre-training". In: *arXiv preprint arXiv:2005.00200* (cit. on p. 70).

Li, Liunian Harold et al. (2019). "Visualbert: A simple and performant baseline for vision and language". In: *arXiv preprint arXiv:1908.03557* (cit. on pp. 53, 64, 138).

– (2020). "What Does BERT with Vision Look At?" en. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5265–5275. DOI: 10.18653/v1/2020.acl-

main.469. URL: https://www.aclweb.org/anthology/2020.acl-main.469 (visited on 04/03/2022) (cit. on pp. 82, 133).

Li, Manling et al. (2022). "CLIP-Event: Connecting Text and Images With Event Structures". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16420–16429 (cit. on pp. 36, 86).

Li, Teng et al. (2010). "Contextual bag-of-words for visual categorization". In: *IEEE Transactions on Circuits and Systems for Video Technology* 21.4, pp. 381–392 (cit. on p. 36).

Li, Wei et al. (2020). "UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning". In: *ArXiv* abs/2012.15409 (cit. on pp. 53, 64).

Li, Xiujun et al. (2020). "Oscar: Object-semantics aligned pre-training for vision-language tasks". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, pp. 121–137 (cit. on pp. 53, 60, 64, 138).

Li, Zihan et al. (2023). "Lvit: language meets vision transformer in medical image segmentation". In: *IEEE Transactions on Medical Imaging* (cit. on p. 29).

Liang, Percy et al. (2022). "Holistic evaluation of language models". In: *arXiv preprint arXiv:2211.09110* (cit. on pp. 77, 79, 85).

Lin, Tsung-Yi et al. (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*. Springer, pp. 740–755 (cit. on pp. 15, 55, 56, 68, 107, 114, 140, 144, 155, 175).

Lin, Xi Victoria et al. (2018). "Multi-Hop Knowledge Graph Reasoning with Reward Shaping". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3243–3253. DOI: 10.18653/v1/D18-1362. URL: https://aclanthology.org/D18-1362 (cit. on pp. 91, 95).

Lindström, Adam Dahlgren et al. (2021). "Probing Multimodal Embeddings for Linguistic Properties: the Visual-Semantic Case". In: *arXiv preprint arXiv:2102.11115* (cit. on p. 82).

Liu, Fangyu, Emanuele Bugliarello, et al. (Nov. 2021). "Visually Grounded Reasoning across Languages and Cultures". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10467–10485. DOI: 10.18653/v1/2021.emnlp-main.818. URL: https://aclanthology.org/2021.emnlp-main.818 (visited on 05/15/2022) (cit. on pp. 83, 150, 175).

Liu, Fangyu, Guy Edward Toh Emerson, et al. (2022). "Visual Spatial Reasoning". In: *ArXiv* abs/2205.00363 (cit. on p. 67).

Liu, Kuan et al. (2018). "Learn to combine modalities in multimodal deep learning". In: *arXiv preprint arXiv:1805.11730* (cit. on p. 83).

Liu, Runtao et al. (2019). "Clevr-ref+: Diagnosing visual reasoning with referring expressions". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4185–4194 (cit. on p. 98).

Liu, Xiao et al. (2021). "Self-supervised learning: Generative or contrastive". In: *IEEE Transactions on Knowledge and Data Engineering* 35.1, pp. 857–876 (cit. on p. 42).

Liu, Yinhan et al. (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (cit. on pp. 41, 45).

Liu, Ze et al. (2021). "Swin transformer: Hierarchical vision transformer using shifted windows". In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022 (cit. on p. 48).

Lowe, David G (2004). "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60, pp. 91–110 (cit. on p. 37).

Lu, Jiasen, Dhruv Batra, et al. (2019a). "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html (visited on 04/03/2022) (cit. on p. 138).

– (2019b). "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks". In: *Advances in neural information processing systems* 32 (cit. on pp. 30, 51, 53, 64).

Lu, Jiasen, Vedanuj Goswami, et al. (2020). "12-in-1: Multi-task vision and language representation learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10437–10446 (cit. on pp. 53, 64).

Lu, Xiaoqiang et al. (n.d.). "Exploring Models and Data for Remote Sensing Image Caption Generation". In: *IEEE Transactions on Geoscience and Remote Sensing* 56.4 (), pp. 2183–2195. DOI: 10.1109/TGRS.2017.2776321 (cit. on p. 98).

Ma, Zixian et al. (2023). "CREPE: Can Vision-Language Foundation Models Reason Compositionally?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921 (cit. on pp. 77, 98).

MacCartney, Bill (2009). *Natural language inference*. Stanford University (cit. on pp. 91, 95).

Malinowski, Mateusz et al. (2014). "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., pp. 1682–1690. URL: http://papers.nips.cc/paper/5411-a-multi-world-approach-to-question-answering-about-real-world-scenes-based-on-uncertain-input.pdf (cit. on p. 98).

Marathe, Aboli et al. (2023). "WEDGE: A multi-weather autonomous driving dataset built from generative vision-language models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3317–3326 (cit. on p. 86).

Marino, Kenneth et al. (2019). "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge". In: *Conference on Computer Vision and Pattern Recognition (CVPR)* (cit. on p. 98).

Mathews, Alexander et al. (2016). "Senticap: Generating image descriptions with sentiments". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. 1 (cit. on pp. 96, 99).

Mikolov, Tomas et al. (2013). "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (cit. on p. 45).

Min, Bonan et al. (2021). "Recent advances in natural language processing via large pre-trained language models: A survey". In: *ACM Computing Surveys* (cit. on p. 46).

Naseer, Muhammad Muzammal et al. (2021). "Intriguing properties of vision transformers". In: *Advances in Neural Information Processing Systems* 34, pp. 23296–23308 (cit. on p. 48).

Nikolaus, Mitja and Abdellah Fourtassi (June 2021). "Evaluating the Acquisition of Semantic Knowledge from Cross-situational Learning in Artificial Neural Networks". In: *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*. Online: Association for Computational Linguistics, pp. 200–210. DOI: 10.18653/v1/2021.cmcl-1.24. URL: https://aclanthology.org/2021.cmcl-1.24 (visited on 02/11/2022) (cit. on pp. 134, 188).

Nikolaus, Mitja, Emmanuelle Salin, et al. (2022). "Do Vision-and-Language Transformers Learn Grounded Predicate-Noun Dependencies?" In: *arXiv preprint arXiv:2210.12079* (cit. on pp. 77, 92, 98).

Nilsback, Maria-Elena et al. (2008). "Automated flower classification over a large number of classes". In: *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, pp. 722–729 (cit. on pp. 43, 114, 115).

Oord, Aaron van den et al. (2018). "Representation learning with contrastive predictive coding". In: *arXiv preprint arXiv:1807.03748* (cit. on p. 61).

Ordonez, Vicente et al. (2011). "Im2text: Describing images using 1 million captioned photographs". In: *Advances in neural information processing systems* 24 (cit. on pp. 55, 56, 58, 140, 174).

Otto, Christian et al. (2019). "Understanding, categorizing and predicting semantic image-text relations". In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 168–176 (cit. on pp. 83, 88).

Parcalabescu, Letitia, Michele Cafagna, et al. (2021). "VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena". In: *Annual Meeting of the Association for Computational Linguistics* (cit. on pp. 77, 81, 91, 92, 98, 99).

Parcalabescu, Letitia, Albert Gatt, et al. (June 2021). "Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks". In: *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*. Groningen, Netherlands (Online): Association for Computational Linguistics, pp. 32–44. URL: https://aclanthology.org/2021.mmsr-1.4 (visited on 05/05/2022) (cit. on p. 81).

Peng, Zhiliang et al. (2023). "Kosmos-2: Grounding Multimodal Large Language Models to the World". In: *arXiv preprint arXiv:2306.14824* (cit. on pp. 52, 57, 62, 69, 172, 173).

Perronnin, Florent (2012). "AVA: A large-scale database for aesthetic visual analysis". In: *IEEE Conference on Computer Vision & Pattern Recognition* (cit. on pp. 96, 99).

Peters, Matthew E. et al. (2018). "Deep contextualized word representations". In: *CoRR* abs/1802.05365. arXiv: 1802.05365. URL: http://arxiv.org/abs/1802.05365 (cit. on pp. 43, 44).

Plummer, Bryan A, Matthew Brown, et al. (2017). "Enhancing video summarization via vision-language embedding". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5781–5789 (cit. on p. 86).

Plummer, Bryan A, Liwei Wang, et al. (2015). "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649 (cit. on p. 68).

Pratt, Sarah et al. (Mar. 2020). "Grounded Situation Recognition". In: *arXiv:2003.12058 [cs]*. arXiv: 2003.12058. URL: http://arxiv.org/abs/2003.12058 (visited on 03/15/2022) (cit. on p. 68).

Radev, Dragomir et al. (2015). "Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest". In: *arXiv preprint arXiv:1506.08126* (cit. on pp. 97, 99).

Radford, Alec, Jong Wook Kim, et al. (2021). "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR, pp. 8748–8763 (cit. on pp. 15, 51, 53, 64, 70, 79, 99, 138, 174).

Radford, Alec, Karthik Narasimhan, et al. (2018). "Improving language understanding by generative pre-training". In: (cit. on pp. 42–44).

Radford, Alec, Jeffrey Wu, et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9 (cit. on pp. 43, 45, 76).

Raganato, Alessandro et al. (2018). "An analysis of encoder representations in transformer-based machine translation". In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. The Association for Computational Linguistics (cit. on p. 41).

Raghu, Maithra et al. (2021). "Do vision transformers see like convolutional neural networks?" In: *Advances in Neural Information Processing Systems* 34, pp. 12116–12128 (cit. on pp. 48, 79).

Ramesh, Aditya et al. (2021). "Zero-shot text-to-image generation". In: *International Conference on Machine Learning*. PMLR, pp. 8821–8831 (cit. on p. 71).

Ramisa, Arnau et al. (2017). "Breakingnews: Article annotation by image and text processing". In: *IEEE transactions on pattern analysis and machine intelligence* 40.5, pp. 1072–1085 (cit. on pp. 94, 98).

Rasmy, Laila et al. (2021). "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction". In: *NPJ digital medicine* 4.1, p. 86 (cit. on p. 89).

Ravichander, Abhilasha et al. (2020). "On the systematicity of probing contextualized word representations: The case of hypernymy in BERT". In: *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pp. 88–102 (cit. on pp. 46, 79, 107, 137, 150).

Ren, Shaoqing et al. (2015). "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems* 28 (cit. on pp. 13, 50, 147).

Ribeiro, Marco Tulio et al. (July 2020). "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4902–4912. DOI: 10.18653/v1/2020.acl-main.442. URL: https://aclanthology.org/2020.acl-main.442 (cit. on pp. 46, 79, 189).

Rogers, Anna et al. (2021). "A primer in BERTology: What we know about how BERT works". In: *Transactions of the Association for Computational Linguistics* 8, pp. 842–866 (cit. on pp. 45, 79).

Rombach, Robin et al. (2022). "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695 (cit. on p. 71).

Rösch, Philipp J. et al. (July 2022a). "Probing the Role of Positional Information in Vision-Language Models". In: *Findings of the Association for Computational Linguistics: NAACL 2022*. Seattle, United States: Association for Computational Linguistics, pp. 1031–1041. DOI: 10.18653/v1/2022.findings-naacl.77. URL: https://aclanthology.org/2022.findings-naacl.77 (cit. on pp. 11, 77).

– (2022b). "Probing the Role of Positional Information in Vision-Language Models". In: *NAACL-HLT* (cit. on p. 82).

Ruggeri, Gabriele et al. (2023). "A Multi-dimensional study on Bias in Vision-Language models". In: *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6445–6455 (cit. on p. 190).

Saharia, Chitwan et al. (2022). "Photorealistic text-to-image diffusion models with deep language understanding". In: *Advances in Neural Information Processing Systems* 35, pp. 36479–36494 (cit. on p. 71).

Saleh, Babak et al. (2015). "Large-scale classification of fine-art paintings: Learning the right metric on the right feature". In: *arXiv preprint arXiv:1505.00855* (cit. on p. 43).

Salin, Emmanuelle et al. (2022). "Are vision-language transformers learning multimodal representations? a probing perspective". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 10, pp. 11248–11257 (cit. on pp. 11, 77, 92, 98).

Salminen, Joni O. et al. (Oct. 2018). "Inter-Rater Agreement for Social Computing Studies". In: *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pp. 80–87. DOI: 10.1109/SNAMS.2018.8554744 (cit. on p. 138).

Schaeffer, Rylan et al. (2023). "Are emergent abilities of Large Language Models a mirage?" In: *arXiv preprint arXiv:2304.15004* (cit. on p. 153).

Schijndel, Marten van et al. (Nov. 2019). "Quantity doesn't buy quality syntax with neural language models". 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 5831–5837. DOI: 10.18653/v1/D19-1592. URL: https://aclanthology.org/D19-1592 (cit. on p. 46).

Schuhmann, Christoph, Romain Beaumont, et al. (2022). "LAION-5B: An open large-scale dataset for training next generation image-text models". In: *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. URL: https://openreview.net/forum?id=M3Y74vmsMcY (cit. on p. 56).

Schuhmann, Christoph, Richard Vencu, et al. (2021). "Laion-400m: Open dataset of clip-filtered 400 million image-text pairs". In: *arXiv preprint arXiv:2111.02114* (cit. on pp. 55–57).

Sebastiani, Fabrizio (2002). "Machine learning in automated text categorization". In: *ACM computing surveys (CSUR)* 34.1, pp. 1–47 (cit. on p. 36).

Selvaraju, Ramprasaath R et al. (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626 (cit. on p. 82).

Sharma, Piyush et al. (2018). "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning". In: *Proceedings of ACL* (cit. on pp. 55–57, 137, 140, 144, 155, 175).

Shekhar, Ravi et al. (July 2017). "FOIL it! Find One mismatch between Image and Language caption". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 255–265. DOI: 10.18653/v1/P17-1024. URL: https://aclanthology.org/P17-1024 (visited on 05/05/2022) (cit. on pp. 81, 91, 98, 99, 110).

Shen, Sheng et al. (2022). "How Much Can CLIP Benefit Vision-and-Language Tasks?" In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=zf_Ll3HZWgy (cit. on pp. 53, 64).

Shevchenko, Violetta et al. (2021). "Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge". In: *ArXiv* abs/2101.06013 (cit. on p. 70).

Shin, Taylor et al. (2020). "Autoprompt: Eliciting knowledge from language models with automatically generated prompts". In: *arXiv preprint arXiv:2010.15980* (cit. on p. 70).

Shin, Wonyoung et al. (2022). "e-clip: Large-scale vision-language representation learning in e-commerce". In: *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 3484–3494 (cit. on pp. 31, 86).

Shridhar, Mohit et al. (2020). "Alfred: A benchmark for interpreting grounded instructions for everyday tasks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10740–10749 (cit. on p. 87).

Shukor, Mustafa et al. (2022). "Structured Vision-Language Pretraining for Computational Cooking". In: *arXiv preprint arXiv:2212.04267* (cit. on p. 86).

Sidorov, Oleksii et al. (2020). "Textcaps: a dataset for image captioning with reading comprehension". In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, pp. 742–758 (cit. on p. 98).

Sikarwar, Ankur et al. (2022). "On the Efficacy of Co-Attention Transformer Layers in Visual Question Answering". In: *ArXiv* abs/2201.03965 (cit. on p. 82).

Simonyan, Karen et al. (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (cit. on p. 169).

Singh, Amanpreet, Vedanuj Goswami, et al. (2020). "Are we pretraining it right? digging deeper into visio-linguistic pretraining". In: *arXiv preprint arXiv:2004.08744* (cit. on pp. 58, 175).

Singh, Amanpreet, Ronghang Hu, et al. (2021). "FLAVA: A Foundational Language And Vision Alignment Model". In: *CoRR* abs/2112.04482. arXiv: 2112.04482. URL: https://arxiv.org/abs/2112.04482 (cit. on pp. 53, 64).

Singh, Amanpreet, Vivek Natarjan, et al. (2019). "Towards VQA Models That Can Read". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326 (cit. on p. 98).

Socher, Richard et al. (2013). "Recursive deep models for semantic compositionality over a sentiment treebank". In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642 (cit. on pp. 43, 44).

Srivastava, Aarohi et al. (2022). "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models". In: *arXiv preprint arXiv:2206.04615* (cit. on p. 79).

Steed, Ryan et al. (2021). "Image representations learned with unsupervised pretraining contain human-like biases". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 701–713 (cit. on p. 83).

Strubell, Emma et al. (2019). "Energy and policy considerations for deep learning in NLP". In: *arXiv preprint arXiv:1906.02243* (cit. on p. 191).

Su, Weijie et al. (2020). "VL-BERT: Pre-training of Generic Visual-Linguistic Representations". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SygXPaEYvH (cit. on pp. 53, 64, 138).

Suhr, Alane et al. (July 2019). "A Corpus for Reasoning about Natural Language Grounded in Photographs". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6418–6428. DOI: 10.18653/v1/P19-1644. URL: https://aclanthology.org/P19-1644 (visited on 03/15/2022) (cit. on pp. 66, 99, 146).

T. Levinboim A. Thapliyal, P. Sharma et al. (2019). "Quality Estimation for Image Captions Based on Large-scale Human Evaluations". In: *arXiv preprint arXiv:1909.03396* (cit. on pp. 97, 99).

Tan, Hao et al. (Nov. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 5100–5111. DOI: 10.18653/v1/D19-1514. URL: https://aclanthology.org/D19-1514 (cit. on pp. 13–15, 30, 49–51, 53, 61, 64, 67, 76, 109, 126, 138, 154, 175).

Taylor, Wilson L (1953). ""Cloze procedure": A new tool for measuring readability". In: *Journalism quarterly* 30.4, pp. 415–433 (cit. on p. 44).

Thrush, Tristan et al. (2022). "Winoground: Probing vision and language models for visio-linguistic compositionality". In: *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pp. 5238–5248 (cit. on pp. 11, 77, 81, 92, 98, 132, 188).

Touvron, Hugo, Matthieu Cord, Matthijs Douze, et al. (2021). "Training data-efficient image transformers & distillation through attention". In: *International conference on machine learning*. PMLR, pp. 10347–10357 (cit. on pp. 48, 169).

Touvron, Hugo, Matthieu Cord, Alaaeldin El-Nouby, et al. (2022). "Three things everyone should know about vision transformers". In: *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*. Springer, pp. 497–515 (cit. on p. 54).

Touvron, Hugo, Thibaut Lavril, et al. (2023). "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971* (cit. on pp. 41, 45).

Vaswani, Ashish et al. (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30 (cit. on pp. 29, 34, 38–40, 138).

Vedantam, Ramakrishna et al. (2015). "Cider: Consensus-based image description evaluation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575 (cit. on p. 68).

Verma, Prateek et al. (2021). "Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions". In: *arXiv preprint arXiv:2105.00335* (cit. on p. 70).

Wallace, Eric et al. (2019). "Do NLP models know numbers? probing numeracy in embeddings". In: *arXiv preprint arXiv:1909.07940* (cit. on p. 46).

Wang, Alex, Yada Pruksachatkun, et al. (2019). "Superglue: A stickier benchmark for general-purpose language understanding systems". In: *Advances in neural information processing systems* 32 (cit. on pp. 44, 65, 78).

Wang, Alex, Amanpreet Singh, et al. (2019). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rJ4km2R5t7 (cit. on pp. 44, 45, 65, 77, 78).

Wang, Boxin et al. (2021). "Adversarial glue: A multi-task benchmark for robustness evaluation of language models". In: *arXiv preprint arXiv:2111.02840* (cit. on p. 189).

Wang, Peng et al. (2022). "OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework". In: *CoRR* abs/2202.03052 (cit. on pp. 53, 64).

Wang, Wenhui et al. (2023). "Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186 (cit. on p. 76).

Wang, Xiaosong et al. (2017). "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106 (cit. on pp. 94, 98).

Wang, Yizhong et al. (2022). "Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks". In: *Conference on Empirical Methods in Natural Language Processing* (cit. on p. 79).

Wang, Zeyu et al. (2020). "Towards fairness in visual recognition: Effective strategies for bias mitigation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8919–8928 (cit. on p. 79).

Wang, Zirui et al. (2022). "SimVLM: Simple Visual Language Model Pretraining with Weak Supervision". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=GUrhfTuf_3 (cit. on pp. 50, 51, 53, 58, 64).

Warstadt, Alex et al. (2019). "Neural network acceptability judgments". In: *Transactions of the Association for Computational Linguistics* 7, pp. 625–641 (cit. on p. 45).

Wei, Jason et al. (2022). "Emergent abilities of large language models". In: *arXiv preprint arXiv:2206.07682* (cit. on p. 153).

Wen, Congcong et al. (2023). "Vision-Language Models in Remote Sensing: Current Progress and Future Trends". In: *arXiv preprint arXiv:2305.05726* (cit. on pp. 36, 87).

Williams, Adina et al. (2017). "A broad-coverage challenge corpus for sentence understanding through inference". In: *arXiv preprint arXiv:1704.05426* (cit. on pp. 43–45).

Wimmer, Christopher et al. (2023). "Leveraging vision-language models for granular market change prediction". In: *arXiv preprint arXiv:2301.10166* (cit. on p. 87).

Wolf, Thomas et al. (Oct. 2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: https://aclanthology.org/2020.emnlp-demos.6 (cit. on p. 119).

Wu, Hui et al. (2021). "The Fashion IQ Dataset: Retrieving Images by Combining Side Information and Relative Natural Language Feedback". In: *CVPR* (cit. on p. 99).

Wu, Qi et al. (2017). "Visual question answering: A survey of methods and datasets". In: *Computer Vision and Image Understanding* 163, pp. 21–40 (cit. on p. 29).

Xie, Ning et al. (2019). "Visual entailment: A novel task for fine-grained image understanding". In: *arXiv preprint arXiv:1901.06706* (cit. on p. 67).

Xie, Yujia et al. (2020). "A fast proximal point method for computing exact wasserstein distance". In: *Uncertainty in artificial intelligence*. PMLR, pp. 433–453 (cit. on p. 61).

Xue, Hongwei et al. (2021). "Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training". In: *Advances in Neural Information Processing Systems* 34, pp. 4514–4528 (cit. on pp. 53, 63, 64, 82).

Yagcioglu, Semih et al. (2018). "Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes". In: *arXiv preprint arXiv:1809.00812* (cit. on pp. 93, 98).

Yang, Jinyu et al. (2022). "Vision-language pre-training with triple contrastive learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15671–15680 (cit. on pp. 50, 53, 59–64, 174).

Yang, Zhengyuan et al. (2021). "Crossing the Format Boundary of Text and Boxes: Towards Unified Vision-Language Modeling". In: *CoRR* abs/2111.12085. arXiv: 2111.12085. URL: https://arxiv.org/abs/2111.12085 (cit. on pp. 50, 53, 64).

Yang, Zhilin, Zihang Dai, et al. (2019). "Xlnet: Generalized autoregressive pretraining for language understanding". In: *Advances in neural information processing systems* 32 (cit. on pp. 43, 45).

Yang, Zhilin, Peng Qi, et al. (2018). "HotpotQA: A dataset for diverse, explainable multi-hop question answering". In: *arXiv preprint arXiv:1809.09600* (cit. on pp. 91, 95).

Yang, Zichao et al. (2016). "Stacked attention networks for image question answering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29 (cit. on p. 37).

Yang, Zongming et al. (2022). "SeeWay: Vision-Language Assistive Navigation for the Visually Impaired". In: *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, pp. 52–58 (cit. on pp. 36, 86).

Ye, Keren et al. (2021). "A case study of the shortcut effects in visual commonsense reasoning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 35. 4, pp. 3181–3189 (cit. on p. 67).

Young, Peter et al. (2014a). "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions". In: *TACL* 2, pp. 67–78 (cit. on pp. 68, 112).

– (2014b). "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions". In: *Transactions of the Association for Computational Linguistics* 2, pp. 67–78 (cit. on p. 140).

Yu, Fei et al. (2021). "Ernie-vil: Knowledge enhanced vision-language representations through scene graphs". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 4, pp. 3208–3216 (cit. on pp. 53, 60, 64).

Yu, Jiahui et al. (2022). "Coca: Contrastive captioners are image-text foundation models". In: *arXiv preprint arXiv:2205.01917* (cit. on pp. 53, 64–66).

Yu, Licheng et al. (2016). "Modeling context in referring expressions". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, pp. 69–85 (cit. on p. 68).

Yu, Yong et al. (2019). "A review of recurrent neural networks: LSTM cells and network architectures". In: *Neural computation* 31.7, pp. 1235–1270 (cit. on p. 37).

Yuan, Lu et al. (2021). "Florence: A new foundation model for computer vision". In: *arXiv preprint arXiv:2111.11432* (cit. on pp. 53, 64, 76, 79).

Yuksekgonul, Mert et al. (2022). "When and why vision-language models behave like bags-of-words, and what to do about it?" In: *arXiv e-prints*, arXiv–2210 (cit. on pp. 77, 98).

Zellers, Rowan et al. (2019). "From recognition to cognition: Visual commonsense reasoning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731 (cit. on pp. 67, 96, 99).

Zeng, Yan et al. (2021). "Multi-grained vision language pre-training: Aligning texts with visual concepts". In: *arXiv preprint arXiv:2111.08276* (cit. on pp. 53, 64).

Zhai, Xiaohua et al. (2019). "A Large-scale Study of Representation Learning with the Visual Task Adaptation Benchmark". In: *arXiv: Computer Vision and Pattern Recognition* (cit. on pp. 77, 78).

Zhang, Jiajun et al. (2015). "Deep Neural Networks in Machine Translation: An Overview." In: *IEEE Intell. Syst.* 30.5, pp. 16–25 (cit. on p. 44).

Zhang, Pengchuan et al. (2021). "Vinvl: Revisiting visual representations in vision-language models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588 (cit. on pp. 15, 50, 53, 57, 64, 123, 138, 154).

Zhao, Cairong et al. (2022). "Learning Domain Invariant Prompt for Vision-Language Models". In: *ArXiv* abs/2212.04196 (cit. on p. 70).

Zhao, Dora et al. (2021). "Understanding and evaluating racial biases in image captioning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14830–14840 (cit. on p. 57).

Zhao, Jieyu et al. (2017). "Men also like shopping: Reducing gender bias amplification using corpus-level constraints". In: *arXiv preprint arXiv:1707.09457* (cit. on p. 175).

Zhao, Tiancheng et al. (2022). "An Explainable Toolbox for Evaluating Pre-trained Vision-Language Models". In: *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 30–37 (cit. on pp. 81, 98).

Zhong, Wanjun et al. (2023). "Agieval: A human-centric benchmark for evaluating foundation models". In: *arXiv preprint arXiv:2304.06364* (cit. on p. 79).

Zhou, Kaiyang et al. (2022). "Learning to prompt for vision-language models". In: *International Journal of Computer Vision* 130.9, pp. 2337–2348 (cit. on p. 70).

Zhou, Luowei et al. (2020). "Unified vision-language pre-training for image captioning and vqa". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 07, pp. 13041–13049 (cit. on pp. 53, 64).

Zhou, Wangchunshu et al. (2022). "VLUE: A Multi-Task Benchmark for Evaluating Vision-Language Models". In: *arXiv preprint arXiv:2205.15237* (cit. on p. 80).

Zhu, Yuke et al. (2016). "Visual7W: Grounded Question Answering in Images". In: pp. 4995–5004. URL: https://openaccess.thecvf.com/content_cvpr_2016/html/Zhu_Visual7W_Grounded_Question_CVPR_2016_paper.html (visited on 05/04/2022) (cit. on pp. 131, 140, 175).

Zhu, Yukun et al. (2015). "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books". In: *Proceedings of the IEEE international conference on computer vision*, pp. 19–27 (cit. on p. 45).

Zitnick, C Lawrence et al. (2013). "Bringing semantics into focus using visual abstraction". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3009–3016 (cit. on p. 98).

A. Taxonomy: Details regarding News-related Study

We detail in this Appendix the examples chosen for the study of news-related data. We select 5 online news sources from several countries and varying demographics. We restrict ourselves to English language newspapers.

— The New York Times (NYT), a daily American newspaper [1]
— Daily Mail, a daily British tabloid [2]
— The Wall Street Journal (WSJ), a daily American business newspaper [3]
— France 24, a French international news network [4]
— Al Jazeera, a Qatari international news network [5]
— Global Times, a daily Chinese English-language newspaper. [6]

We select three dates and study a captioned image from those newspapers for each of those dates, selecting a topic at random for each example. These examples vary across topics: ranging from business to culture.

The image/caption pairs are detailed in Table A.1. Those examples show that captions and images can sometimes be unrelated. The use of multimodal documents to process news-related data could help provide more context.

1. https://www.nytimes.com/
2. https://www.dailymail.co.uk/
3. https://www.wsj.com/
4. https://www.france24.com/en/
5. https://www.aljazeera.com/en
6. https://www.globaltimes.cn

Source	Caption	Image URL
NYT	The Rijksmuseum show gathers 28 Vermeer works, including some of the artist's most famous, from left: "Girl Reading a Letter at an Open Window," "Girl With a Pearl Earring" and "The Milkmaid."	https://static01.nyt.com/images/2023/02/05/multimedia/05vermeer-01-zvhm/05vermeer-01-zvhm-superJumbo.jpg?quality=75&auto=webp
Daily Mail	Each monarch has a Throne Chair – unique to them – for the enthronement part of the ancient ritual	https://i.dailymail.co.uk/1s/2023/02/04/23/67329673-0-image-m-3_1675555191629.jpg
Global Times	Three helicopters attached to a ship-borne helicopter regiment of the navy under the PLA Southern Theater Command hover in formation during a round-the-clock training exercise in late January, 2023.	https://www.globaltimes.cn/Portals/0/attachment/2023/2023-01-05/cfb51357-b82b-4e79-85df-af82eee7696a.jpeg
Al Jazeera	A taxi stand, seen shut here in Srinagar, Indian-controlled Kashmir on July 28, 2020. Despite claims by the Indian government that a controversial abrogation of the region's special status would help its economy, investments have dried up	https://www.aljazeera.com/wp-content/uploads/2023/02/AP20212272338222.jpg?resize=770%2C513&quality=80
WSJ	Bulgari silver watch, $5,050, gold watch, $31,400, and coin cuff, $33,900, Bulgari.com, Prasi bracelet, $6,600, PrasiOfficial.com, Celine by Hedi Slimane shirt, price upon request	https://images.wsj.net/im-706327?width=700&size=0.7498535442296427&pixel_ratio=2
France 24	Ukrainian servicemen stand on a tank in the eastern Donetsk region on February 4, 2023	https://s.france24.com/media/display/7eef2608-a4ca-11ed-87b3-005056bf30b7/w:980/p:16x9/2023-02-04T113816Z_204102323_RC294Z9GRX89_RTRMADP_3_UKRAINE-CRISIS.webp
NYT	Sam Bankman-Fried faces investigations by the Securities and Exchange Commission and the Justice Department	https://static01.nyt.com/images/2022/11/28/business/28FTX-Catchup/28FTX-Catchup-superJumbo.jpg?quality=75&auto=webp
Daily Mail	Liverpool interviewed 12 candidates in their extensive search for a new club doctor	https://i.dailymail.co.uk/1s/2022/11/28/07/65000557-11476479-image-a-28_1669619915043.jpg
Global Times	A soccer production company in Yiwu, East China's Zhejiang Province, has supplied 100,000 commemorative balls for the Qatar 2022 FIFA World Cup, featuring logos of the Qatar World Cup, mascots and national ags of the 32 countries that entered the competition	https://www.globaltimes.cn/Portals/0/attachment/2022/2022-11-13/28390bfc-d779-4d11-abb4-6df353dfc332.jpeg
Al Jazeera	3 million people need help	https://liberties.aljazeera.com/wp-content/uploads/2022/11/1669624309.jpg
WSJ	Monkeypox was first identified in a group of laboratory monkeys in Denmark. It is rare for the WHO to rename an old disease.	https://archive.is/NR2Hg/6b31461f786674bdfd93365251c01461fe8de4f5.jpg
France 24	Naxos graviera is one of Greece's most popular cheeses.	https://s.france24.com/media/display/27745d68-6ee6-11ed-b5d1-005056a90284/9b7b7df69ad49d06d994bb4b06ada0a6129bf068.webp
NYT	WHAT THE FUTURE HOLDS FOR UNDOCUMENTED IMMIGRANTS The fate of our democracy is tied to theirs.	https://static01.nyt.com/images/2020/12/01/opinion/01tobar-01/01tobar-01-superJumbo.jpg?quality=75&auto=webp
Daily Mail	In episode seven, set in the 1980s, Princess Margaret (played by Helena Bonham Carter, pictured above) stumbles across the Queen Mother's nieces', Nerissa and Katherine Bowes-Lyon, existence and is appalled by their treatment	https://i.dailymail.co.uk/1s/2020/12/02/01/36336334-9008071-image-a-63_1606873687384.jpg
Global Times	French writer Herve le Tellier and his book L'Anomalie	https://www.globaltimes.cn/Portals/0/attachment/2020/2020-12-01/ee9e5b59-e2f0-4be0-9336-c7d7988fd176.jpeg
Al Jazeera	The Orange Line Metro Train on its first test-run, travels along a track in a neighbourhood in Lahore, Pakistan on May 16, 2018	https://www.aljazeera.com/wp-content/uploads/2020/12/lahore-orange-train.jpg?resize=770%2C513&quality=80
WSJ	Big airlines face a major change in their clientele if business travel is reduced. A new analysis of travel purposes estimates a 19%-36% permanent loss of corporate trips as a result of more meetings shifting to screens.	https://archive.is/WDEuY/ff8c6a1068b34614345a273ec200d1341dbc842b.jpg
France 24	This photo, said to show Brazilian football legend Pelé at the grave of Argentine legend Diego Maradona, has been circulating online since November 27. It turns out, it was photoshopped.	https://s.observers.france24.com/media/display/dc0d6c38-3ae5-11eb-bacb-005056a98db9/w:980/p:16x9/Pele_maradona_debunked.webp

Table A.1. – Details of the image/caption pairs used for the study of news-related data