Répondre aux questions visuelles à propos d'entités nommées

Paul Lerner

université PARIS-SACLAY

Répondre aux questions visuelles à propos d'entités nommées

Knowledge-based Visual Question Answering about Named Entities

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580,
Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Informatique
Graduate School : Informatique et sciences du numérique,
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Laboratoire interdisciplinaire des sciences du numérique** (Université Paris-Saclay, CNRS)
sous la direction d'**Olivier FERRET**, Directeur de recherche,
le co-encadrement de **Camille GUINAUDEAU**, Maîtresse de conférences

Thèse soutenue à Paris-Saclay, le 8 novembre 2023, par

Paul LERNER

Composition du jury

Membres du jury avec voix délibérative

Pierre ZWEIGENBAUM Directeur de recherche, Université Paris-Saclay, CNRS, LISN	Président
Josiane MOTHE Professeure, INSPE, IRIT UMR5505 CNRS, Université Toulouse Jean-Jaurès	Rapporteur & Examinatrice
Philippe MULHEM Chargé de recherches HDR, Université Grenoble Alpes, CNRS, Grenoble INP, LIG	Rapporteur & Examinateur
Michel CRUCIANU Professeur, CEDRIC-CNAM	Examinateur
Ewa KIJAK Maîtresse de conférences, Université de Rennes, Inria, IRISA	Examinatrice

THESE DE DOCTORAT

NNT : 2023UPASG074

Remerciements

Mes remerciements s'adressent évidemment en premier lieu à Olivier et Camille sans qui rien n'aurait été possible. Tous deux ont fait preuve d'une attention et d'une patience sans limite avec une particulière bienveillance durant ces trois années. J'en profite pour remercier Hervé Bredin qui m'a le premier engagé au LIMSI (futur LISN) comme ingénieur pour une première collaboration avec Camille. Quitte à remonter le temps, je remercie également mes précédentes encadrantes de stage Beatrice Biancardi et Catherine Pelachaud, et Laurence Likforman-Sulem, sans qui je n'aurais pas fait de thèse non plus. Je suis également reconnaissant aux membres du projet MEERQAT pour les discussions enrichissantes lors de nos réunions, en particulier à ceux qui ont contribué au jeu de données ViQuAE (chapitre 3) : Hervé Le Borgne, Romaric Besançon, Jose G Moreno et Jesús Lovón Melgarejo. Sans oublier Salem Messoud que j'ai encadré pendant son stage de Master, autre expérience enrichissante. C'est l'occasion de remercier officiellement l'ANR qui a financé ma thèse à travers ce projet ANR-19-CE23-0028 MEERQAT ainsi que l'IDRIS dont j'ai bénéficié d'un accès aux moyens de calcul du supercalculateur Jean Zay au travers de l'allocation de ressources 2021-AD011012846 attribuée par GENCI, renouvelée deux fois, en 2022 et 2023. Je tiens également à remercier Antoine Chaffin pour les interminables discussions sur la cross-modalité, les avantages et inconvénients des *cross-encoder* et *dual encoder*, qui ont contribué au chapitre 6. D'autre part, je remercie Frédéric Voisin, Joël Falcou, François Landes et Kim Gerdes qui m'ont fait confiance en tant que chargé de TD. J'adresse enfin un grand merci aux membres de mon jury de thèse : Pierre Zweigenbaum, Josiane Mothe, Philippe Mulhem, Michel Crucianu et Ewa Kijak pour leur interrogatoire de deux heures, dense mais très intéressant, ainsi que les corrections apportées à la première version de cette thèse ; et en particulier Josiane et Philippe pour leurs rapports.

Je m'adresse maintenant de façon plus personnelle aux membres du LISN et du département STL (ex-équipes ILES et TLP), en particulier les doctorantes et doctorants. Ce n'était pas évident de commencer ma thèse fin 2020, entre deux confinements et avant une année principalement télétravaillée, mais je pense que nous avons réussi à recréer une ambiance plus que conviviale dès la rentrée 2021-2022. Je suis très fier d'avoir fait ma thèse à vos côtés et il serait trop long d'énumérer tous les bons souvenirs. Citer des noms serait inévitablement injuste et je suis sûr que les concernés se reconnaîtront (si vous lisez ces lignes, c'est bon signe !)

J'abrège en remerciant ma famille et mes amis pour leur soutien, et Louise avec qui j'ai partagé ma vie pendant ces trois années.

Table des matières

Table des figures

Liste des tableaux

Glossaire

— Ajustement (*fine-tuning*) : mise à jour par descente de gradient des paramètres d'un modèle pré-entraîné

— Amorce (*prompt*) : préfixe utilisé pour interroger un gros modèle de langue

— Attention croisée (*cross-attention*) : mécanisme d'attention présent notamment dans le décodeur d'un *transformer*

— Auto-attention (*self-attention*) : mécanisme d'attention présent notamment dans l'encodeur d'un *transformer*

— BC : Base de Connaissances (structurée ou non)

— *Benchmark* : jeu de données utilisé pour évaluer et comparer différents systèmes et ainsi valider les hypothèses de recherche

— Cascade (*pipeline*) : enchaînement de plusieurs modèles dépendant les uns des autres de manière séquentielle (par exemple un système de question-réponse composé d'une RI et d'une extraction de réponse)

— DPR : *Dense Passage Retrieval*, modèle de recherche d'information fondé sur deux encodeurs BERT, proposé par Karpukhin et al. (2020)

— ECA : *Early Cross-Attention*, modèle de fusion multimodale fondée sur l'architecture *transformer* et le modèle pré-entraîné BERT

— Figer (*freeze*) : se dit de tout ou partie des paramètres d'un modèle qui ne sont pas mis à jour pendant l'ajustement

— ICT : *Inverse Cloze Task*, tâche de pré-entraînement pour la recherche d'information

— iid : (variables) indépendantes et identiquement distribuées

— ILF : *Intermediate Linear Fusion*, modèle de fusion multimodale standard, fondé sur des couches de projections linéaires

— Lot (*batch*) : collection d'exemples d'apprentissage. Chaque itération de la descente de gradient consomme un lot.

— GPU : *Graphics Processing Unit* (carte graphique)

— Hits@K : La proportion de questions où le modèle trouve *au moins un* document pertinent dans les K premiers ; équivalent au Rappel@K en considérant qu'il n'y a qu'*un seul* document pertinent

— KVQAE : *Knowledge-based Visual Question Answering about named Entities* (répondre aux questions visuelles à propos d'entités nommées)

— LLM : *Large Language Model* (gros modèle de langue)

— MIPS : *Maximum Inner-Product Search* (recherche du produit scalaire maximal)

— MRR : *Mean Reciprocal Rank* (rang réciproque moyen)

— PLM : *Pre-trained Language Model* (modèle de langue pré-entraîné)

— Recherche exhaustive (*grid search*) : utilisée pour optimiser les hyperparamètres d'un modèle parmi une plage de valeurs prédéfinie

— Référence (*baseline*) : système utilisé comme point de comparaison

— RI : Recherche d'Information

— TAL : Traitement Automatique des Langues

— VQA : *Visual Question Answering* (répondre aux questions visuelles), souvent utilisé pour désigner les questions visuelles classiques, ou dans un autre cadre que la KVQAE

Chapitre 1

Introduction

Le multimédia prend une part importante de nos vies quotidiennes et influence notre façon de communiquer, notamment à travers les réseaux sociaux et les smartphones (Ijaz et al., 2021; Sutisna et al., 2020). D'autre part, nous avons naturellement une communication multimodale qui combine parole, gestes et vision (Ekman et Friesen, 1969; Fröhlich et al., 2019). Par conséquent, la multimodalité a récemment pris une place importante dans la recherche académique, en particulier dans les domaines du Traitement Automatique des Langues (TAL), de la Recherche d'Information (RI), de la vision par ordinateur et de l'apprentissage automatique, ce dernier étant transverse aux autres domaines. Plusieurs tâches auparavant traitées de manière mono-modale ont ainsi désormais leur pendant multimodal, comme la traduction, la désambiguïsation d'entités nommées ou la recherche d'images (Specia et al., 2016; Adjali et al., 2020b; Rasiwasia et al., 2010). Cette multimodalité ambiante ouvre d'une part de nombreux défis scientifiques et problèmes de recherche, et, d'autre part, permet de fluidifier et rendre plus naturelle l'interaction entre l'utilisateur et la machine (Gurari et al., 2018; Yu et al., 2019).

Les systèmes de question-réponse ont longtemps intéressé chercheurs et industriels dans tous les champs de l'informatique mais plus particulièrement en TAL et en RI. La machine apparaît en effet comme ayant une mémoire virtuellement infinie, capable de stocker toutes les connaissances acquises depuis la nuit des temps et ce, de façon plus évidente depuis l'avènement du WWW. Bachelard (1938) nous dit que « *toute connaissance est une réponse à une question* ». Il est donc naturel pour l'utilisateur d'interroger une machine dotée de cette capacité. Les systèmes de question-réponse sont maintenant intégrés dans de nombreuses applications, en particulier dans les assistants virtuels (comme Siri) ou les moteurs de recherche. L'utilisateur peut donc interroger ces systèmes quotidiennement, avec des questions futiles, factuelles, philosophiques ou existentielles. Les entités nommées sont centrales dans les questions et les systèmes de question-réponse, qui vont souvent les désambiguïser, implicitement ou explicitement.

Dans ce contexte, nous proposons de définir une nouvelle tâche : répondre aux questions visuelles à propos d'entités nommées (KVQAE [1]). Nous estimons que cette tâche est bien définie et facilement évaluable et permet ainsi de suivre la progression de la qualité des représentations multimodales d'entités nommées.

1. *Knowledge-based Visual Question Answering about Named Entities.*

Question visuelle (entrée)	Passage visuel pertinent dans la base de connaissances
 *"Who succeeded him as president?"*	 De Gaulle resigned the presidency at noon, 28 April 1969 [...] Two months later Georges Pompidou was elected as his successor.
 *"How many avenues radiate from this building?"*	 The Arc de Triomphe is located on the right bank of the Seine at the centre of a dodecagonal configuration of twelve radiating avenues.

FIGURE 1.1 – Deux exemples de questions visuelles du jeu de données ViQuAE (Lerner et al., 2022) accompagnées de passages visuels pertinents issus de sa base de connaissances.

Ces représentations multimodales permettent de fluidifier les interactions homme-machine. Par exemple, en regardant un film, on peut se demander « *Où ai-je déjà vu cette actrice ?* » ou « *Est-ce qu'elle a déjà gagné un Oscar ?* » Un système de question-réponse multimodal nous éviterait alors la fastidieuse tâche de parcourir le générique du film et de chercher des informations à propos de ladite actrice.

1 Contexte et définitions

Le TAL, la RI, la vision par ordinateur et l'apprentissage automatique ont été bouleversés au cours de la dernière décennie par le succès de l'apprentissage profond (Krizhevsky et al., 2012; Mikolov et al., 2013; LeCun et al., 2015). Après un temps où les trois domaines applicatifs reposaient sur des technologies différentes, les méthodes ont convergé vers les mêmes paradigmes (réseau de neurones profond, pré-entraînement et apprentissage par transfert), jusqu'au point où l'architecture *transformer* (Vaswani et al., 2017) a été appliquée à une part importante des tâches de TAL (Devlin et al., 2019), de RI (Lin et al., 2021) et de vision par ordinateur (Khan et al., 2022). Les systèmes de question-réponse ou de compréhension de texte, autrefois composés de multiples systèmes en cascade (Ferret et al., 2001), ont été remplacés par un unique réseau de neurones, entraînable de bout en bout sous réserve d'une quantité suffisante de données annotées (Hermann et al., 2015; Rajpurkar et al., 2016). Même la décomposition de la tâche en deux étapes, RI et extraction de réponse, que nous suivons dans cette thèse, a été remise en cause par les gros modèles de langue (LLM ; Brown et al., 2020), capables de stocker des connaissances implicitement dans leurs paramètres et donc de répondre aux questions sans accès explicite à une BC. Cependant, cette unification des tâches relève de l'effet « boîte noire », ce qui n'est ni explicable ni interprétable, pour l'utilisateur comme pour le chercheur. Nous avons donc conservé la décomposition en deux étapes, en nous focalisant sur la partie RI. De plus, un courant de recherche récent vise à augmenter

les LLM avec une RI ou d'autres modules, afin de limiter leurs *hallucinations* (génération d'informations erronées ; Izacard et al., 2022; Shuster et al., 2021; Ji et al., 2023).

Dans ce contexte d'unification des méthodes, plusieurs tâches multimodales, plus précisément traitant des textes et des images, ont été introduites. Ce courant de recherche s'intéresse particulièrement aux interactions qui interviennent entre les modalités, lesquelles diffèrent selon que le texte est une légende de l'image, une question ancrée dans l'image ou que l'image sert de contexte distant (par exemple pour une traduction). Notre travail s'inscrit dans ce courant et étudie les interactions multimodales avec deux nouvelles perspectives : (i) les images représentent des entités nommées ; (ii) les interactions sont multiples et combinées puisqu'au lieu d'un texte et d'une image, nous étudions une question visuelle en interaction avec une base de connaissances multimodales.

Questions visuelles La figure 1.1 montre deux exemples de questions visuelles ainsi que des passages visuels pertinents correspondants, tirés du jeu de données ViQuAE et de sa Base de Connaissances (BC) multimodale [2]. Une question visuelle est constituée plus précisément d'une question textuelle contextualisée par une image, et symétriquement, un passage visuel est la réunion d'un passage textuel, issu d'un document lié à une entité, et d'une image associée. La question fait référence à l'entité dépeinte dans l'image contextuelle à travers une expression référentielle. Le besoin d'information exprimé dans la question est satisfait par le passage pertinent. L'image qui compose le passage visuel permet, elle, de reconnaître l'entité parmi toutes celles de la BC. La KVQAE est donc un problème à multiples niveaux : RI multimodale mais aussi compréhension du texte. Nous remarquons donc que l'interaction multimodale est différente au sein de la question ou du passage : le texte de la question est dépendant de l'image contextuelle tandis que le texte du passage a du sens indépendamment de l'image.

Entités nommées Contrairement aux questions visuelles classiques (VQA ; Antol et al., 2015), qui visent le contenu de l'image (par exemple « *De quelle couleur est la voiture ?* »), les questions en KVQAE visent des entités nommées et nécessitent donc de rechercher des informations dans une BC. Par conséquent, cette thèse se focalise sur une étape nécessaire pour répondre aux questions : la recherche d'information, qui vise à trouver un passage pertinent à partir de la question. Nous suivons la définition d'entité nommée standard en TAL, c'est-à-dire une entité qui a une identité propre et que l'on peut identifier à partir de son nom, comme une *personne*, un *lieu* ou une *organisation*, pour citer les types les plus courants en reconnaissance d'entités nommées. En revanche, il est à noter que le nom de l'entité ne fait justement pas partie des questions, où elle est alors désignée par une mention ambiguë, par exemple « *cette personne* ».

Images d'entités nommées Les représentations visuelles d'entités nommées sont diverses et multiples, comme en témoigne la figure 1.2. Une même image peut

2. Toutes les images utilisées dans les figures sont sous licence libre (par exemple Creative Commons). Leurs auteurs sont crédités à l'annexe D.

(a) Personne (Louis-Philippe I[er])

(b) Personne (Louis-Philippe I[er]) ou tableau (de Franz Xaver Winterhalter)

(c) Monument (Tour Eiffel)

(d) Site naturel (Mont Saint Helens)

(e) Entreprise (Apple) ou bâtiment (*Apple Fifth Avenue*)

(f) Entreprise (Apple) ou produit (iPhone 1)

FIGURE 1.2 – Quelques exemples d'images de différents types d'entités et différentes images du même type d'entité considérées dans notre travail

représenter plusieurs entités et une même entité a de multiples représentations. Par exemple, Louis-Philippe I[er] est dépeint à travers une photographie ou un tableau à la figure 1.2, et, réciproquement, son tableau peut le représenter mais est aussi une œuvre d'art à part entière. En dehors du support, les images varient selon la lumière, la pose ou l'âge du sujet, ce qui est bien documenté en reconnaissance faciale (Sim et al., 2002; Zou et al., 2007; Ding et Tao, 2016; Sawant et Bhurchandi, 2019). Mais Louis-Philippe I[er] reste néanmoins une personne et donc une entité concrète que l'on peut représenter directement. D'autres entités, comme les entreprises, sont abstraites et vont donc être représentées au travers d'entités concrètes. C'est le cas par exemple d'Apple, que l'on peut voir représentée à la figure 1.2 au travers d'un de ses magasins ou de ses produits phares. Ces subtilités distinguent clairement la KVQAE de la recherche d'image par le contenu, où la requête est une seule image. En KVQAE, l'image est conditionnée par la question et cette thèse s'intéresse particulièrement à ce type d'interactions multimodales.

Interactions mono- et cross-modales Nous définissons deux types d'interactions mono-modales, textuelle (TQTP) et visuelle (IQIP) entre question et passage, ainsi que trois cross-modales : au sein de la question visuelle (TQIQ), du passage visuel (TPIP) ou entre les deux (IQTP), comme illustré par la figure 1.3. Ces interactions interviennent dans toutes les étapes nécessaires pour arriver à la réponse mais seront particulièrement étudiées pour la RI, où elles seront alors modélisées comme des similarités entre les composantes des questions et des passages. Les interactions

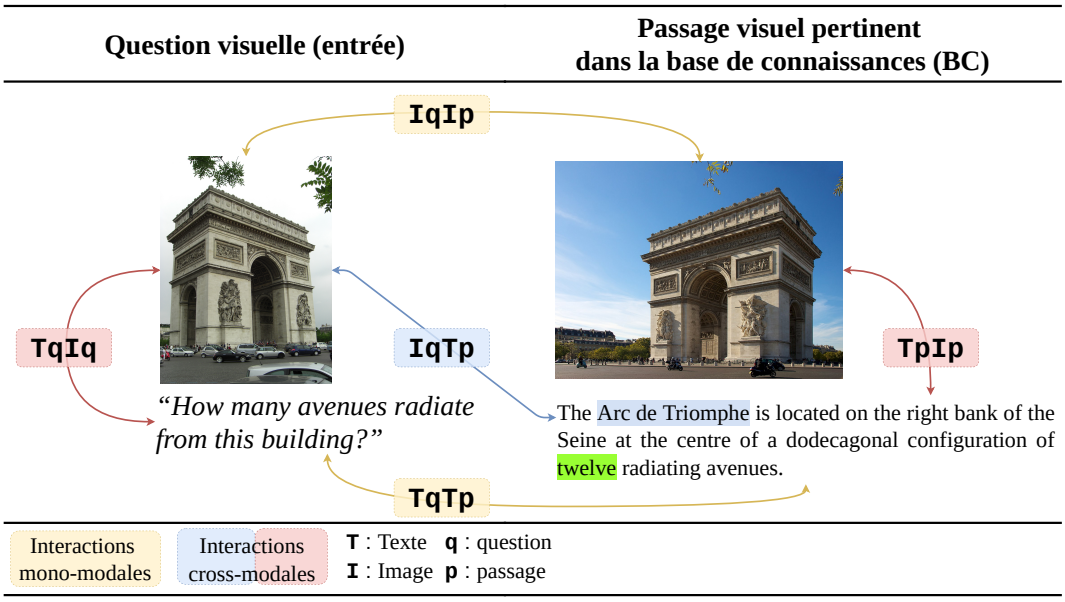


FIGURE 1.3 – Illustration des différents types d'interactions mono- et cross-modales étudiées

mono-modales ramènent notre travail à la RI et à la tâche de question-réponse textuel pour TQTP et à la recherche d'image par le contenu ou la reconnaissance faciale pour IQIP. L'interaction cross-modale TQIQ relève, elle, davantage de la fusion d'informations multimodales telle qu'étudiée notamment pour les questions visuelles classiques (Zhang et al., 2019), tandis qu'IQTP et TPIP s'apparentent à la recherche cross-modale ou la génération de légende (Vendrov et al., 2016; Vinyals et al., 2016).

Base de connaissances Nous utilisons le terme base de connaissances, ou BC, pour englober à la fois des bases de connaissances structurées, telles qu'elles sont considérées dans le champ de la représentation des connaissances en prenant souvent la forme de *graphe de connaissances*, mais aussi des collections de documents non-structurés, par exemple l'ensemble des articles Wikipédia. L'une ou l'autre de ces deux formes sont utilisées selon les travaux en KVQAE, ce qui distingue cette dernière de la VQA classique (Shah et al., 2019; Chen et al., 2023c). De plus, les images qui rendent ces BC multimodales sont pour le moment traitées de la même manière, que la BC soit structurée ou non (Shah et al., 2019; Chen et al., 2023c). Notre travail s'inscrit dans la lignée des travaux sur les BC non-structurées, bien qu'une approche hybride soit explorée dans le cadre du projet MEERQAT [3] (Adjali et al., 2023), lequel finance cette thèse.

3. *Multimedia Entity Representation and Question Answering Tasks*. Ce projet, coordonné par Hervé Le Borgne, vise à analyser des contenus textuels, visuels ou multimodaux ambigus en utilisant des connaissances à propos d'entités nommées (https://www.meerqat.fr/).

2 Questions de recherche et contributions

Cette thèse traite de plusieurs aspects relatifs aux questions visuelles, de la collecte et l'annotation de données aux représentations, recherches et connaissances multimodales.

Au chapitre 2, nous passons en revue l'état de l'art des systèmes de question-réponse et tâches connexes. En effet, notre travail se rapproche à la fois de la tâche de question-réponse textuel, où l'on cherche une information dans une BC non-structurée, mais aussi, à l'opposé, des questions visuelles où la RI se réduit à l'analyse d'une image et où les interactions cross-modales sont essentielles. Dans ce large spectre, plusieurs thématiques sont récurrentes :

— les modèles de langue pré-entraînés, qui sont parfois détournés pour devenir des modèles multimodaux ;

— l'apprentissage contrastif, utilisé de la même façon, jusqu'à la fonction de coût, à la fois pour la classification d'images, la détection de quasi-doublons, la RI visuelle, cross-modale ou textuelle ;

— la fusion d'informations multimodales.

Après avoir identifié un manque dans les jeux de données disponibles pour la KVQAE, nous présentons au chapitre 3 ViQuAE, une collection de 3 700 questions visuelles à propos de plus de 2 400 entités différentes, couvrant un large spectre de sujets, de types d'entités et de représentations visuelles. Ce jeu de données est accompagné d'une base de connaissances fondée sur Wikipédia qui permet de répondre à ses questions en combinant plusieurs modalités. ViQuAE nous permettra ensuite d'évaluer nos systèmes et ainsi valider nos méthodes et hypothèses de recherche. La collecte et l'annotation de ViQuAE a été semi-automatique car elle repose sur le principe d'une reformulation de questions textuelles, ce qui implique de détecter et désambiguïser leurs entités nommées et de trouver des images les dépeignant. Ce chapitre contribue à répondre à deux de nos questions de recherche :

— **Comment évaluer un système de KVQAE ?** Nous supposons qu'il est nécessaire de *collecter et annoter* un jeu de données à cet effet. De plus, nous supposons que la *présence de la réponse dans le passage* est indicative de sa pertinence.

— **Comment représenter visuellement une entité nommée ?** C'est-à-dire quelles images utiliser et comment les modéliser. Nous supposons que les *entités non-personnes* et *abstraites* ont de *multiples représentations* dont la modélisation serait bénéfique.

Dans cette thèse, nous avons choisi de distinguer deux étapes pour répondre aux questions : recherche d'information et extraction de réponse, qui s'articulent au sein d'un même système. Cette décomposition classique a été remise en cause au cours des dernières années par la multiplication des approches bout-en-bout. Néanmoins, après l'avènement des gros modèles de langue capables de stocker des connaissances implicitement dans leurs paramètres, et donc répondre aux questions sans BC, les travaux se réorientent vers des approches de combinaison de modèles, avec une RI réalisée en lien avec une BC explicite, pour contrer les hallucinations. Au chapitre 4,

nous proposons ainsi une première mise en œuvre de cette approche en deux étapes en nous appuyant sur des technologies bien établies et en privilégiant une certaine simplicité au travers de la fusion tardive de deux recherches mono-modales : textuelle et visuelle. Nous présenterons différentes manières de modéliser la recherche textuelle, en tenant compte des contraintes sur la faible quantité de données disponible. Nous verrons également comment modéliser la recherche visuelle, en distinguant plusieurs représentations selon le type d'entités nommées. Ensuite, l'extraction de la réponse à partir du passage visuel sera simplifiée au passage textuel, en faisant l'hypothèse que l'image n'est plus nécessaire *une fois* le passage pertinent retrouvé. Ce chapitre permettra d'identifier la RI comme principal verrou de la KVQAE. Le reste de la thèse y sera par conséquent principalement consacré.

La fusion tardive est limitée car elle néglige toutes les interactions cross-modales. Pour remédier à cette limitation, nous étudierons plusieurs méthodes de fusion précoce au chapitre 5 afin de répondre à notre troisième et principale question de recherche : **comment interagissent les modalités ?** Nous supposons qu'il est important de *modéliser les interactions cross-modales*, en particulier TQIQ, au sein de la question visuelle. Ces modélisations étant gourmandes en paramètres, nous aborderons le pré-entraînement des modèles dans ce même chapitre. Cela soulèvera plusieurs problématiques liées à l'apprentissage séquentiel et aux biais de l'apprentissage. Nous verrons également que toutes les représentations visuelles ne s'intègrent pas de la même façon dans une fusion tardive ou précoce.

Cette fusion précoce permettra une meilleure RI grâce aux interactions cross-modales considérées de façon assez holistique. Plus précisément, nous verrons que l'interaction IQTP, entre la question et le passage, est exploitée par cette fusion précoce et permet une meilleure RI. Cette piste sera davantage explorée au travers d'expériences avec le modèle CLIP (Radford et al., 2021), qui permet une recherche cross-modale, sans pré-entraînement supplémentaire (cf. chapitre 6). Cela renforcera les résultats précédents quant à l'importance de cette interaction. Ces résultats questionnent la façon d'effectuer une recherche visuelle : vaut-il mieux exploiter IQIP ou IQTP dans la figure 1.3 et comment ajuster un modèle cross-modal pré-entraîné comme CLIP ? Dans le même chapitre, nous présenterons également des différences importantes selon les métriques utilisées et discuterons de l'évaluation de la RI : intrinsèque en jugeant la pertinence de ses résultats ou extrinsèque via l'extraction des réponses.

Enfin, nous conclurons cette thèse par une discussion des résultats et des limitations de notre travail, ainsi que de nos perspectives pour les travaux futurs.

Cette thèse a donné lieu aux publications suivantes :

— Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno et Jesús Lovón Melgarejo. 2022. Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées en utilisant des bases de connaissances. Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN), Avignon, France. ATALA.

— Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno et Jesús Lovón Melgarejo. 2022. ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities. Proceedings of The 45th International ACM SIGIR Conference on Research

et Development in Information Retrieval, SIGIR '22, New York, NY, USA. Association for Computing Machinery.

— Paul Lerner, Olivier Ferret et Camille Guinaudeau. 2023. Multimodal Inverse Cloze Task for Knowledge-Based Visual Question Answering. Advances in Information Retrieval (ECIR), pages 569–587, Cham. Springer Nature Switzerland.

— Paul Lerner, Salem Messoud, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno et Jesús Lovón Melgarejo. 2023. Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées. Traitement Automatique des Langues (TAL), 2023, Intermodalité et multimodalité en traitement automatique des langues, 63 (2).

— Paul Lerner, Ferret Olivier et Camille Guinaudeau. 2023. Recherche cross-modale pour répondre à des questions visuelles. Actes de la 18e Conférence en Recherche d'Information et Applications (CORIA), pages 74–92, Paris, France. ATALA.

La contribution du jeu de données ViQuAE, que l'on retrouve au chapitre 3, a d'abord été présentée à SIGIR, avant d'être traduite et résumée pour TALN, puis enfin étendue pour la revue TAL, en intégrant le travail du stage de Salem Messoud (Messoud, 2022). Par ailleurs, l'ensemble du code (plus de 10 000 lignes), des modèles et des données produit est centralisé et rendu disponible librement via GitHub [4]. Le code est notamment fondé sur Lightning [5], PyTorch (Paszke et al., 2019) et Transformers (Wolf et al., 2020) pour l'entraînement des modèles, et Datasets (Lhoest et al., 2021), Faiss (Johnson et al., 2019) et Ranx (Bassani, 2022) pour la RI. Cela inclut également une documentation pour reproduire les expériences et une partie relative à l'annotation du jeu de données ViQuAE : le code pour la partie automatique mais aussi la plateforme d'annotation et les instructions destinées aux annotateurs.

4. https://github.com/PaulLerner/ViQuAE
5. https://www.pytorchlightning.ai/

Chapitre 2

État de l'art

1 Introduction

1.1 Contexte et positionnement

Répondre aux questions visuelles à propos d'entités nommées est un problème complexe qui se situe à l'intersection de plusieurs domaines de recherche : TAL, RI, vision par ordinateur et apprentissage automatique. Notre première question de recherche, *comment évaluer un système de KVQAE ?*, est davantage liée au TAL et à la RI. La deuxième, *comment représenter visuellement une entité nommée ?*, provient évidemment de la vision par ordinateur. Notre travail s'inscrit dans la lignée des travaux sur la multimodalité visant à fluidifier les interactions homme-machine et étudier les interactions cross-modales (Gurari et al., 2018; Yu et al., 2019; Specia et al., 2016; Adjali et al., 2020b) pour répondre à notre troisième et principale question de recherche : *comment interagissent les modalités ?*

Du point de vue de la tâche de question-réponse, nous nous situons, comme l'immense majorité des travaux sur les questions visuelles, ou plus généralement des tâches multimodales, dans ce que Rodriguez et Boyd-Graber (2021) appellent le paradigme de Manchester, par opposition au paradigme de Cranfield (Voorhees, 2019), c'est-à-dire que nous sommes plus intéressé par le test des capacités d'un modèle que par les bénéfices qu'il pourrait apporter aux utilisateurs [1]. Nous retrouvons cette idée de façon évidente au chapitre 3 où nous constituons le jeu de données ViQuAE de façon à évaluer des systèmes de KVQAE et suivre leur progrès. Cela s'oppose au paradigme de Cranfield, très usité en RI, où les questions proviennent souvent de l'historique des requêtes formulées à un moteur de recherche (Voorhees, 2001; Nguyen et al., 2016; Kwiatkowski et al., 2019) [2]. Notre travail est en cela similaire à l'immense majorité des travaux sur les questions visuelles (cf. section 6), à l'exception de Gurari et al. (2018) qui étudient des questions visuelles posées

1. On retrouve aussi cette idée dans la taxonomie de Rogers et al. (2021), qui distinguent les questions naturelles (*information-seeking*) des tests (*probes*).

2. Plus exactement, les évaluations de type Cranfield comme les campagnes TREC sont fondées sur un ensemble de requêtes associé à une collection de documents (une BC dans notre cadre) qui restent fixes (Cleverdon, 1967; Voorhees, 2019).

par des personnes aveugles. Cette asymétrie est causée par le manque d'interface utilisateur et de systèmes capables de répondre aux questions visuelles. Nous introduirons la tâche de question-réponse, textuel ou multimodal, et les tâches connexes à la section 1.2. Nous aborderons ensuite plus précisément la recherche d'image par le contenu à la section 3, la RI et la tâche de question-réponse textuel à la section 4, ce qui implique une certaine représentation des connaissances ; la recherche cross-modale fera l'objet de la section 5 et, enfin, les questions visuelles — et plus généralement la multimodalité — seront traitées à la section 6.

Notre recherche est également étroitement liée à l'apprentissage automatique. De ce point de vue, nous retrouverons deux leitmotivs au cours de cette thèse : l'apprentissage de représentations, plus précisément l'apprentissage contrastif, et le pré-entraînement dans un cadre d'apprentissage par transfert. Ce dernier est omniprésent dans la thèse, qui s'inscrit dans le paradigme *pré-entraînement et ajustement*, par opposition aux nouvelles approches d'amorçage (*prompting*) et aux méthodes d'apprentissage sans pré-entraînement. Nous présenterons différentes méthodes de pré-entraînement au sein des sections dédiées à la tâche visée en aval. L'apprentissage contrastif interviendra pour obtenir des représentations de texte, d'images ou de données multimodales aux chapitres 4 à 6, lesquelles seront utilisées pour rechercher des informations textuelles, visuelles, ou multimodales. Nous en discuterons à la section 2.2.

Ces techniques d'apprentissage ont largement été mises en œuvre en conjonction avec des modèles de langue que nous introduirons plus précisément à la section 2.1. Les modèles de langue modernes ont évolué à partir de l'architecture *transformer* (Vaswani et al., 2017) en trois familles : (i) encodeur (BERT ; Devlin et al., 2019) ; (ii) décodeur auto-régressif (GPT ; Radford et al., 2018) ; (iii) encodeur-décodeur (T5 ; Raffel et al., 2020), bien que les frontières entre ces catégories soient poreuses. Au-delà de ces différences architecturales, ces modèles ont été utilisés principalement dans deux cadres, ce dont témoigne le vocabulaire qui a évolué de *modèle de langue pré-entraîné* (PLM [3]) à *gros modèle de langue* (LLM [4]). Outre la taille croissante des modèles dont témoigne ce changement de vocabulaire, le premier usage, PLM, s'inscrit dans le paradigme *pré-entraînement et ajustement*, comme cette thèse et comme discuté dans les sections suivantes. Le pré-entraînement y est alors seulement une étape préliminaire en vue d'un apprentissage par transfert qui ajuste les paramètres du modèle ou en introduit de nouveaux pour traiter la tâche en aval. Le second usage, LLM, s'inscrit au contraire dans le paradigme *modèle de langue universel* où le « pré-entraînement » devient en fait le *seul* stade d'entraînement et où le modèle de langue est ensuite interrogé en complétant une amorce de texte, sans ajuster ses paramètres ou en introduire de nouveaux (Brown et al., 2020).

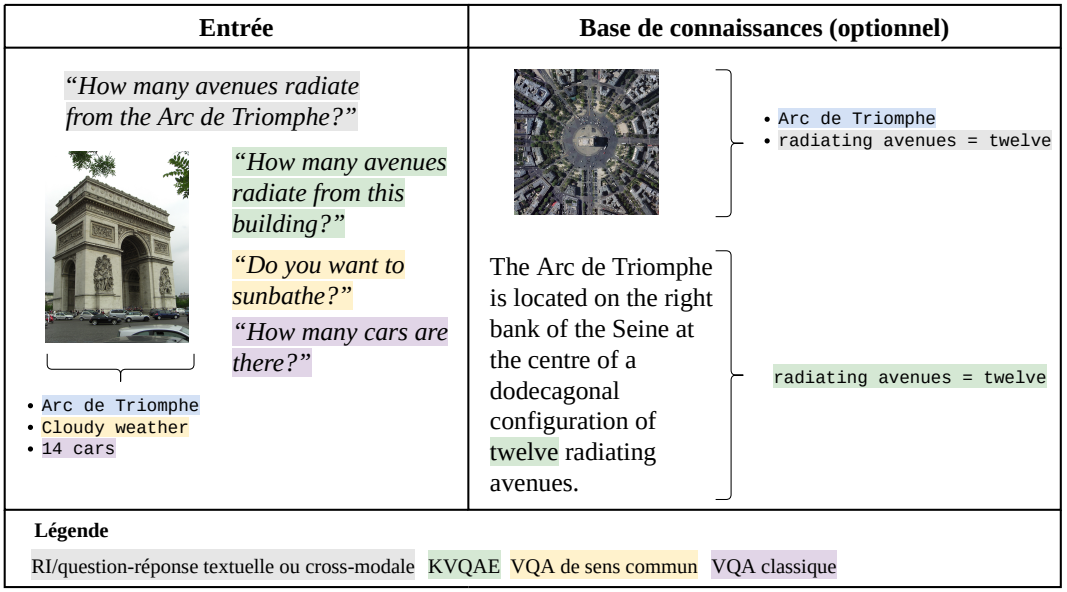


FIGURE 2.1 – Illustration de six tâches d'intérêt dans cette thèse et interactions multimodales connexes : 1. la recherche d'image par le contenu (IQIP) ; 2. la RI/question-réponse textuelle (TQTP) ; 3. la VQA classique (TQIQ) ; 4. la VQA de sens commun (TQIQ) ; 5. la RI/VQA cross-modale (TQIP) ; 6. et bien sûr, la KVQAE (IQIP, TQTP, TQIQ, TPIP et IQTP). Rappelons que les sigles des interactions sont composés des lettres T (Texte), I (Image), Q (Question) et P (Passage).

1.2 Recherche d'information et question-réponse

La KVQAE est liée à cinq autres tâches qui sont illustrées à la figure 2.1 :

1. la recherche d'image par le contenu (section 3), où en cherchant l'image de l'Arc de Triomphe à gauche, on souhaite retrouver celle à droite (texte inutilisé, on modélise l'interaction mono-modale IQIP) ;
2. la RI/question-réponse textuelle (section 4), où on souhaite avoir la réponse « *twelve* », ou la phrase complète, à la question « *How many avenues radiate from the Arc de Triomphe ?* », grâce au texte de la BC (image inutilisée, on modélise l'interaction mono-modale TQTP) ;
3. la VQA classique (section 6), où on cherche à extraire des informations de l'image contextuelle, le nombre de voitures dans l'exemple (BC inutilisée, on modélise l'interaction cross-modale TQIQ) ;
4. la VQA de sens commun (section 6), où les modalités sont similaires à la VQA classique mais où la réponse demande d'avoir du sens commun : *on ne peut pas bronzer s'il n'y a pas de soleil* (BC inutilisée également [5], on modélise la même interaction cross-modale TQIQ) ;

3. *Pre-trained Language Model*
4. *Large Language Model*
5. L'article présenté sur l'Arc de Triomphe n'est pas pertinent pour répondre à cette question mais on pourrait imaginer utiliser une autre BC pour cet usage, par exemple ConceptNet (Liu et Singh, 2004).

5. la RI/VQA cross-modale (section 5), où la requête est purement *textuelle*, comme pour la RI textuelle, mais où l'on cherche à analyser l'image de la BC. On peut compter, il y a bien *douze* avenues dans l'image à droite [6]. On peut aussi avoir une requête plus générale, par exemple « *Arc de Triomphe* », si l'on cherche des images, ou symétriquement, une requête purement visuelle pour chercher des textes relatifs à l'image, comme pour une génération de légende. Ce dernier usage est aussi proche de ce que Sun et al. (2022) appellent la *désambiguïsation visuelle d'entités nommées*. On modélise alors l'interaction cross-modale IQTP (ou symétriquement TQIP pour une requête textuelle).

Le lecteur aura bien sûr reconnu à la figure 2.1 l'exemple de la figure 1.1 : en KV-QAE (section 6), on veut répondre à la question « *How many avenues radiate from this building ?* », ancrée dans l'image de gauche, grâce au texte et à l'image de la BC [7]. Les six tâches que nous avons distinguées demandent d'analyser des données multimodales, d'en extraire des informations et de les mettre en correspondance. Nous remarquons que, dans la majorité des modèles étudiés, l'extraction d'information est implicite. Au contraire, d'autres tâches sont plus fondamentales, telles que la désambiguïsation multimodale d'entités nommées (« de quel *Arc de Triomphe* s'agit-il ? ») ou la résolution d'expressions référentielles (« à quel objet dans l'image *this building* fait-il référence ? »).

Toutes ces tâches sont confrontées au fossé sémantique (*semantic gap*) existant entre le besoin d'information/intention de l'utilisateur et les caractéristiques que l'on est capable de modéliser à partir des données multimodales (Smeulders et al., 2000). Ce fossé peut être réduit par des requêtes multimodales (Depeursinge et Müller, 2010). Le fossé sémantique est évident dans les images, où, au niveau du pixel, un changement de luminosité, par exemple, aura un fort impact sur la représentation des données tandis que le contenu sémantique restera inchangé (Smeulders et al., 2000). Pour les langues exprimées sous forme de texte, le fossé sémantique se retrouve :

— dans la synonymie (Krovetz, 1997 ; ou l'inadéquation du vocabulaire ; Furnas et al., 1987), lorsque l'utilisateur utilise un terme synonyme de celui présent dans le document pertinent, qui sont donc représentés par des chaînes de caractères tout à fait différentes mais au sens proche (*bureau* et *pupitre*) ;

— au contraire, dans la polysémie (Krovetz, 1997), lorsque l'utilisateur utilise un terme polysème ou du moins homographe d'un terme d'un document non pertinent, qui sont donc composés des mêmes caractères mais d'un sens différent (*bureau* a 11 sens selon Larousse, dont le meuble et la pièce).

Toutes les tâches que nous avons distinguées ont par ailleurs été bouleversées par l'apprentissage profond. La RI, de texte ou d'image, a évolué de représentations parcimonieuses à base de sac de mots (Deerwester et al., 1990; Robertson et al., 1995; Sivic et Zisserman, 2003) à des représentations denses (Karpukhin et al., 2020; Sharif Razavian et al., 2014) et les tâches multimodales ont largement émergé de

6. Souvent, l'image contextuelle est alors fournie avec la question pour évaluer les capacités de raisonnement plutôt que de RI (Chang et al., 2022).

7. Dans cette thèse, la réponse est systématiquement extraite depuis le texte de la BC mais on pourrait imaginer une approche hybride qui s'appuierait également sur l'image de la BC pour compter le nombre d'avenues.

cette révolution. Contrairement aux questions visuelles, où l'on doit *fusionner* les informations multimodales, la RI cross-modale et la génération de légende visent à *aligner* ou « traduire » d'une modalité à l'autre. Cette dernière problématique de recherche a été étudiée à travers différents points de vue : (i) naturellement, pour la RI cross-modale en tant que telle, par exemple pour la navigation Web ou dans un cadre de recherche sémantique, dans des collections d'images personnelles (Yang et al., 2023) ; (ii) pour l'apprentissage sans exemple, où l'appariement d'images et de textes permet d'extraire des attributs sémantiques de l'image (Frome et al., 2013; Radford et al., 2021) ; (iii) pour l'apprentissage de langue visuellement ancré (Lazaridou et al., 2015), plutôt fondé sur des théories cognitives (Harnad, 1990) en conjonction avec certains travaux qui montrent que les représentations purement textuelles (comme celles extraites de modèles de langue) manquent de sens commun (Gordon et Van Durme, 2013).

Toutes les tâches de RI, textuelle, visuelle ou cross-modale, doivent traiter de grandes collections de textes ou d'images (Radenović et al., 2018; Formal et al., 2021). La RI est alors souvent divisée en au moins deux étapes avec une recherche initiale très efficiente, parfois au prix d'une moindre précision, suivie d'un réordonnancement plus coûteux mais plus précis.

2 Cadre méthodologique des travaux réalisés

2.1 Modèles de langue : gros ou pré-entraînés, fondateurs ou spécifiques à une tâche ?

La révolution de l'apprentissage profond a largement été permise par la multiplication des données annotées, qui permettent d'entraîner des réseaux de neurones plus profonds donc plus expressifs (Russakovsky et al., 2015), efficaces sur des exemples issus de la même distribution mais également comme *pré-entraînement* dans un cadre d'*apprentissage par transfert* (Sharif Razavian et al., 2014; Girshick, 2015), bien que certaines études nuancent ses bénéfices (Kornblith et al., 2019; Lassance et al., 2023). En TAL, de gros jeux de données annotés ont également été motivés par, et ont motivé à leur tour, les approches neuronales (Socher et al., 2013; Rajpurkar et al., 2016; Williams et al., 2018) et été utilisés pour pré-entraîner des modèles (Min et al., 2017). Toutefois, les méthodes de pré-entraînement les plus emblématiques sont non supervisées. Celles-ci se sont longtemps concentrées sur les plongements lexicaux (*word embeddings* ; Mikolov et al., 2013; Pennington et al., 2014), utilisés pour obtenir une représentation vectorielle des mots sur laquelle on pouvait appliquer une méthode d'apprentissage standard pour traiter la tâche en aval (Chen et al., 2017), avant que les modèles de langue n'atténuent la frontière entre plongement lexical et modélisation de la tâche (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019).

Les modèles de langue ont d'abord été développés pour la reconnaissance de la parole (Bahl et al., 1983). En effet, ils permettent de lever l'ambiguïté en cas d'homophonie, par exemple « je mange des pâtes » vs. « je *mangent* des pâtes » (syntaxe) ou « je mange des *pattes* » (sémantique). Ces relations peuvent être partiellement

capturées avec des modèles statistiques très simples, comme des n-grammes (Brown et al., 1992). Un premier modèle de langue neuronal a été proposé dans l'article séminal de Bengio et al. (2000) mais a longtemps attendu le succès [8] et, après l'avènement de l'apprentissage profond, les méthodes sont revenues à des objectifs plus simples, fondés sur la co-occurrence de mots (Mikolov et al., 2013; Pennington et al., 2014). Les modèles de langue neuronaux sont seulement réapparus après des travaux sur la traduction automatique (Cho et al., 2014; Bahdanau et al., 2016), fondés sur les réseaux de neurones récurrents et le mécanisme d'attention, d'abord avec cette même architecture (Peters et al., 2017, 2018) puis avec l'architecture *transformer* que nous connaissons aujourd'hui (Radford et al., 2018; Devlin et al., 2019).

En augmentant la taille de ces modèles et la quantité de données utilisée pour les entraîner, ils développent des capacités de plus en plus poussées. Au-delà de la syntaxe et la sémantique partiellement capturées par les modèles statistiques, qu'ils maîtrisent évidemment bien mieux, ils développent également des connaissances à propos d'entités nommées, par exemple « Paris est la capitale de la France » vs. « Paris est la capitale de la *Belgique* » (Brown et al., 2020). Ces modèles peuvent donc être utilisés pour de nombreuses tâches, dont deux qui nous intéressent particulièrement : la RI et la tâche de question-réponse (section 4). Nous verrons qu'ils influencent également nos façons de représenter les connaissances. Ces modèles se regroupent en trois familles : (i) encodeur (BERT ; Devlin et al., 2019) ; (ii) décodeur auto-régressif (GPT ; Radford et al., 2018) ; (iii) encodeur-décodeur (T5 ; Raffel et al., 2020), qui diffèrent dans la façon de générer du texte.

Dans les premiers, la génération, ou décodage, est faite par un modèle linéaire, indépendamment pour chaque mot de la phrase [9]. Ils sont donc plutôt utilisés pour prédire un à quelque mots masqués d'une phrase au cours d'un *Cloze Test* (Taylor, 1953), renommé *Masked Language Modeling* par Devlin et al. (2019), ou sans génération, simplement pour leurs représentations latentes des mots ou de la phrase. Ce premier usage correspond au paradigme *modèle de langue universel* (Petroni et al., 2019) tandis que le second est plus souvent utilisé dans un cadre de *pré-entraînement et ajustement* (Devlin et al., 2019). Dans les deuxièmes, la génération est auto-régressive, c'est-à-dire qu'elle dépend des mots générés précédemment. On peut donc interroger le modèle en lui faisant compléter l'amorce d'une phrase, par exemple « Paris est la capitale de la . . . ». Ces modèles sont donc désormais le plus souvent utilisés dans le cadre *modèle de langue universel* (Radford et al., 2019; Brown et al., 2020) mais on peut également ajuster leurs paramètres pour traiter une tâche spécifique (Radford et al., 2018). Enfin, les troisièmes correspondent à l'architecture originellement proposée par Vaswani et al. (2017) pour la traduction, plus proche des travaux antérieurs fondés sur les réseaux de neurones récurrents (Cho et al., 2014; Bahdanau et al., 2016). La génération y est également auto-régressive [10] mais dépend additionnellement de la représentation latente de l'amorce (ou l'entrée) fournie par l'encodeur. Cette modélisation est utile lorsque l'espace d'entrée et de

8. L'article a seulement atteint son pic de citation en 2019 selon Semantic Scholar, soit près de 20 ans après sa première publication.

9. Nous utilisons les termes *mot* et *phrase* pour illustrer nos propos mais de façon plus générale, il s'agit d'une séquence de symboles discrets.

10. Nous notons des exceptions comme les travaux sur la traduction non-auto-régressive qui visent à accélérer la génération (Gu et al., 2018).

sortie sont différents, puisque l'encodeur et le décodeur partagent rarement leurs paramètres : par exemple si l'on traduit de l'anglais vers le français, si l'on génère la légende d'une image ou, dans une moindre mesure, la réponse à une question (Vaswani et al., 2017; Cho et al., 2021; Raffel et al., 2020). Ces modèles peuvent donc être utilisés dans les deux paradigmes (Raffel et al., 2020; Lester et al., 2021).

Nous évoquerons ces modèles *transformer* à maintes reprises au cours des sections suivantes, en particulier l'encodeur BERT, dont l'architecture permet d'obtenir une représentation des mots qui dépend de leur contexte passé et futur, contrairement aux décodeurs comme GPT qui appliquent un masque d'attention causal, qui limite la représentation des mots à leur contexte passé. BERT est pré-entraîné à accomplir deux objectifs de façon multi-tâche :

— MLM (*Masked Language Modeling*) : prédire les mots masqués dans l'entrée ;

— NSP (*Next Sentence Prediction*), une classification binaire « est-ce que les deux phrases en entrée sont consécutives ? »

Ces objectifs ont une vocation universelle et permettent de capturer des informations à la fois syntaxiques et sémantiques. Par-dessus tout, BERT a été appliqué à la quasi-totalité des tâches de TAL, ce pourquoi Bommasani et al. (2021) parlent de *modèle fondateur*. Au contraire, plusieurs tâches de pré-entraînement, souvent non supervisées ou faiblement [11] supervisées, ont été développées en visant une tâche en particulier. Citons, par exemple, l'*Inverse Cloze Task* de Lee et al. (2019), développée spécifiquement pour la RI. Ces pré-entraînements sont souvent eux-mêmes fondés sur un premier pré-entraînement universel tel que BERT. Lassance et al. (2023) parlent alors de *middle training*. Nous discuterons des pré-entraînements spécifiques à la RI et à la tâche de question-réponse textuel à la section 4 puis aux questions visuelles et tâches multimodales connexes à la section 6.

La RI, textuelle, visuelle ou multimodale, est maintenant étroitement liée à l'apprentissage contrastif de représentation, que ce soit pour le pré-entraînement ou l'ajustement des modèles. Nous introduisons donc ce concept à la section suivante, en comparant deux principales approches.

2.2 Apprentissage contrastif de représentation

L'apprentissage automatique a été bouleversé au cours de la dernière décennie par l'avènement de l'apprentissage profond (Krizhevsky et al., 2012; LeCun et al., 2015). Plutôt que d'apprendre à partir de caractéristiques extraites des données par des experts (*feature engineering*), l'apprentissage profond repose sur des réseaux de neurones profonds, capables d'apprendre une représentation des données de bout-en-bout en minimisant une fonction objectif par descente de gradient, les représentations étant plus abstraites au fur et à mesure de la profondeur du modèle (LeCun et al., 2015). Le jeu de données ImageNet (Deng et al., 2009), qui classe des millions d'images en milliers de catégories d'objets, a joué un rôle primordial dans cette évolution (Russakovsky et al., 2015), notamment à travers l'article séminal de Krizhevsky et al. (2012). En effet, puisque l'expressivité d'un réseau de neurones dépend de son nombre de paramètres (connexions entre les neurones), l'ensemble

11. En anglais on parle aussi de *webly supervised* lorsque les exemples sont récupérés sur le Web.

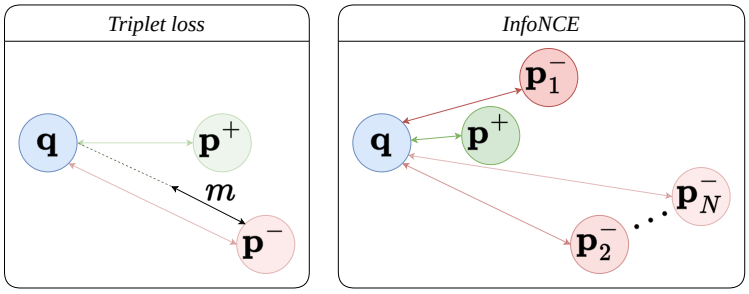


FIGURE 2.2 – Comparaison de deux fonctions objectifs populaires en apprentissage contrastif de représentations : la *triplet loss* et InfoNCE. Les deux visent à rendre la représentation de la question $\mathbf{q}$ proche de celle du passage pertinent $\mathbf{p}^+$ et éloignée des passages non pertinents $\mathbf{p}_j^-$. La *triplet loss* impose une marge m entre $\|\mathbf{q}-\mathbf{p}^+\|^2$ et $\|\mathbf{q}-\mathbf{p}^-\|^2$. Dans cet exemple, la contrainte est respectée, donc le gradient est nul. InfoNCE, au contraire, compare plusieurs exemples à la question $\mathbf{q}$, qui contribuent chacun au gradient selon leur distance.

d'apprentissage se doit d'être assez grand, faute de quoi, un grand nombre de paramètres conduit inévitablement au sur-apprentissage.

Bien qu'il soit possible d'utiliser ces représentations issues d'une classification dans un contexte plus générique [12], une part importante de l'apprentissage de représentations est contrastif, c'est-à-dire que la fonction objectif compare les exemples entre eux plutôt qu'à des classes (Le-Khac et al., 2020). L'objectif est donc d'obtenir un espace de représentation où une mesure de similarité, comme la similarité cosinus, est représentative de la similarité sémantique entre les exemples. Une fois cet espace à disposition, de multiples tâches peuvent être traitées : la vérification (« est-ce que ces deux exemples ont le même sens ? ») demande simplement de fixer un seuil de similarité et le clustering et la classification peuvent également être traités avec des modèles très simples tels que les K-moyennes ou les K plus proches voisins, respectivement. Quant à la RI, qui nous intéresse tout particulièrement, elle consiste alors simplement à ordonner les K plus proches voisins (Schroff et al., 2015; Johnson et al., 2019; Karpukhin et al., 2020).

Ces travaux utilisent souvent une fonction de classement à marge (*margin-based ranking loss*), qui impose une marge entre les exemples jugés différents, que l'on appelle *exemples négatifs* (Schroff et al., 2015; Reimers et Gurevych, 2019). Le succès d'une telle approche réside donc en partie dans la qualité de ces exemples négatifs, qui doivent être assez *difficiles* pour contribuer au gradient et entraîner le modèle (Xu et al., 2022). Il existe de nombreuses variantes de fonction de classement à marge mais la *triplet loss* est particulièrement populaire. Comme son nom l'indique, elle est fondée sur un triplet (ancre, exemple positif, exemple négatif), que nous notons $(\mathbf{q}, \mathbf{p}^+, \mathbf{p}^-)$ dans cette thèse pour représenter une question, un passage pertinent et un non pertinent :

$$\max\left(0, \|\mathbf{q}-\mathbf{p}^+\|^2 - \|\mathbf{q}-\mathbf{p}^-\|^2 + m\right) \tag{2.1}$$

12. Citons par exemple (Sharif Razavian et al., 2014), qui exploite des représentations apprises sur ImageNet pour la recherche d'image par le contenu, ce que nous explorerons également à partir du chapitre 4.

qui impose donc une marge m entre la distance euclidienne au carré entre $\mathbf{q}$ et $\mathbf{p}^+$ et entre $\mathbf{q}$ et $\mathbf{p}^-$:

$$\|\mathbf{q} - \mathbf{p}^+\|^2 + m < \|\mathbf{q} - \mathbf{p}^-\|^2 \tag{2.2}$$

Plus récemment, la fonction objectif proposée par Sohn (2016) et nommée InfoNCE par Oord et al. (2019), qui repose sur l'entropie croisée des similarités des différents exemples au sein d'un lot (*batch*), a été appliquée avec succès à de nombreuses tâches :

— RI et question-réponse textuel (section 4 ; Lee et al., 2019; Karpukhin et al., 2020; Xiong et al., 2020) ;

— pré-entraînement de CLIP (Radford et al., 2021), un modèle fondateur dont les représentations peuvent notamment servir à la classification d'image, la recherche visuelle ou la recherche cross-modale [13] (section 5) ;

— pré-entraînement d'une représentation visuelle pour la classification d'images (Chen et al., 2020a) ;

— détection de quasi-doublon (*near duplicate detection* ; Pizzi et al., 2022) ;

— recherche cross-lingue pour la traduction automatique (Cai et al., 2021).

Un des avantages d'InfoNCE par rapport aux fonctions de classement à marge est qu'il permet d'exploiter simultanément tous les exemples du lot comme exemples négatifs, à un coût virtuellement nul, ce qui est rendu très efficace par la parallélisation des calculs sur GPU. Formellement, étant donné la question $\mathbf{q}$, le passage pertinent $\mathbf{p}^+$ et un ensemble de N passages non pertinents $\mathbf{p}_j^-$, l'objectif est de minimiser la fonction suivante :

$$-\log \frac{\exp\left(\mathbf{q} \cdot \mathbf{p}^+\right)}{\exp\left(\mathbf{q} \cdot \mathbf{p}^+\right) + \sum_{j=1}^{N} \exp\left(\mathbf{q} \cdot \mathbf{p}_j^-\right)} \tag{2.3}$$

Cet objectif n'est donc plus contraint à une marge m fixe. Tous les exemples contribuent au gradient (cf. figure 2.2 et Chen et al., 2020a).

Au lieu d'étiquettes de pertinence binaire, il est également possible de distiller les connaissances d'un modèle plus complexe, typiquement un réordonnanceur dans un cadre de RI, qui appliquerait un mécanisme d'attention entre la question et le passage plutôt que d'encoder les deux indépendamment (Hofstätter et al., 2021; Menon et al., 2022). Cette méthode a également été appliquée à la recherche cross-modale par Miech et al. (2021). En effet, les encodeurs joints peuvent modéliser des interactions plus complexes entre la question et le passage et fournissent donc un score de similarité plus informatif que les jugements de pertinence, qui sont particulièrement parcimonieux (*sparse*) en RI (Craswell et al., 2021).

Dans la suite de ce chapitre, nous discutons plus spécifiquement des différentes tâches en lien avec nos travaux, tâches qui s'appuient sur les méthodes évoquées ici.

13. Le pré-entraînement de CLIP correspond à une recherche cross-modale avec des images $\mathbf{q}$ appariées à leur légende $\mathbf{p}^+$.

3 Recherche d'image par le contenu

La communauté de recherche d'image par le contenu en vision par ordinateur est divisée entre reconnaissance faciale et recherche d'entités non-personnes. Cela s'explique historiquement par des algorithmes de détection d'objets différents pour les visages et les autres objets (Sakai et al., 1972). Nous sommes au contraire intéressé par les deux méthodes afin de pouvoir répondre aux questions à propos de personnes ou d'autres types d'entités. Ces travaux sont également liés à l'étude de l'interaction mono-modale IQIP, entre les images des questions et des passages.

3.1 Recherche d'image pour des entités non-personnes

Cette branche de recherche est largement focalisée sur des méthodes non supervisées. Sivic et Zisserman (2003) se sont inspirés des méthodes de recherche lexicale pour proposer un index visuel inversé, où le vocabulaire est constitué de clusters de caractéristiques visuelles non supervisées telles que SIFT (Lowe, 2004). Cette méthode, bien qu'ancienne, est robuste et toujours compétitive aujourd'hui (Zheng et al., 2018).

Des méthodes plus modernes sont fondées sur un pré-entraînement supervisé. Sharif Razavian et al. (2014) montrent que les caractéristiques extraites par un modèle entraîné à la classification d'images sur ImageNet sont robustes, c'est-à-dire que les représentations qu'elles fournissent pour des images sont fidèles à leur contenu sémantique. La recherche peut alors être effectuée dans l'espace de représentation vectoriel, en choisissant une fonction de distance comme la distance euclidienne, ou de similarité comme la similarité cosinus. Remarquons que ces deux mesures sont monotones lorsque les vecteurs ont une norme unitaire, on a $\|\mathbf{q}-\mathbf{p}\|^2 = 2(1-\mathbf{q}\cdot\mathbf{p})$. Cette approche est ainsi devenue une référence privilégiée (Radenović et al., 2018).

La collection d'images utilisée pour chercher (analogue à la BC dans notre cadre) doit être grande et contenir de bons distracteurs, c'est-à-dire d'images non pertinentes mais qui sont semblables à la requête, même de façon surfacique. Radenović et al. (2018) proposent de redéfinir la BC utilisée pour les jeux de données Paris et Oxford (Philbin et al., 2007, 2008), où les résultats de méthodes concurrentes étaient saturés. Plus récemment, Weyand et al. (2020) ont proposé Google Landmarks v2, un grand jeu de données (5 millions d'images) centré sur les monuments et autres points d'intérêts touristiques (*landmarks*). La collection est fondée sur Wikimédia Commons et utilise des heuristiques sur ses catégories d'images. Nous utiliserons des heuristiques similaires au chapitre suivant pour collecter des images d'entités nommées.

Comme la RI textuelle, la recherche d'image se doit d'être très efficiente pour passer à l'échelle. Elle est donc également souvent traitée en deux étapes, où le réordonnancement suit la RI initiale (Siméoni et al., 2019). Pour les objets rigides, tels que les monuments, ce réordonnancement peut se faire à travers une vérification géométrique (Philbin et al., 2007). Il peut également servir à entraîner les modèles de RI initiale (Radenović et al., 2019). Tan et al. (2021) proposent d'utiliser un *transformer* pour réordonner des images en traitant la tâche comme une classification binaire, supervisée grâce à Google Landmarks v2 (Weyand et al., 2020).

3.2 Reconnaissance faciale

La reconnaissance faciale permet la recherche d'image pour des entités de type personne. Les progrès en la matière ont largement été permis par de grandes quantités de données annotées (Schroff et al., 2015; Guo et al., 2016). Beaucoup de travaux se sont donc concentrés sur l'apprentissage de représentations, que nous avons déjà évoqué à la section 2.2. Deng et al. (2019) ont proposé ArcFace, qui préfère imposer une marge autour de la classe plutôt qu'entre les exemples comme une *triplet loss* (Schroff et al., 2015).

Nous serons confrontés tout au cours de cette thèse à l'hétérogénéité des représentations visuelles d'entités nommées. Comme évoqué au chapitre 1, une personne, par exemple, peut être représentée à travers une photographie réaliste, un tableau ou une statue. Ces questions n'ont pas, à notre connaissance, été traitées par la littérature de la reconnaissance faciale. Nous notons cependant qu'il existe une littérature motivée par des applications de surveillance qui tentent de reconnaître des individus à partir d'images infrarouges ou de portraits robots (Huo et al., 2018; Peng et al., 2019; Shiri et al., 2018). Cependant, ce cadre est plus limité que le nôtre puisque l'image requête (infrarouge) et l'image référence (photographie standard) ont toujours le même format. Au contraire, nous traitons des images diverses à la fois dans les questions visuelles ou les passages visuels. Ainsi, nous n'avons pas utilisé ces méthodes, d'autant plus que nous sommes intéressé par les représentations visuelles de toutes les entités nommées et pas seulement des personnes.

4 Recherche d'information et question-réponse

Les méthodes pour la RI et la tâche de question-réponse, auparavant divisées entre les communautés RI et TAL, se sont unifiées au cours de ces dernières années, en particulier à partir des modèles de langue pré-entraînés tels que BERT. Ces recherches sont particulièrement connectées à nos travaux de par la composante textuelle des questions et des passages et l'étude de l'interaction mono-modale TQTP. L'articulation des différents modules du premier système de KVQAE proposé au chapitre 4 est largement inspirée par les travaux discutés dans cette section.

Nous évoquerons tout d'abord les différentes façons de représenter les connaissances avant de discuter des deux étapes cruciales d'un système de question-réponse : la recherche d'information et l'extraction (ou plus récemment, la génération) de réponse.

4.1 Représentation des connaissances

Un système de question-réponse a bien sûr besoin d'une certaine façon de représenter les connaissances. Nous en distinguons trois :

1. explicitement, à travers un corpus de textes non structurés, comme dans le cadre de cette thèse ;
2. explicitement, à travers une base de connaissances structurée, sans distinguer les graphes des bases tabulaires ;

3. implicitement, à travers les paramètres d'un gros modèle de langue ou plus généralement d'un réseau de neurones.

Dans les deux premiers cas, la RI peut se faire dans un espace latent en cherchant le document textuel ou le nœud du graphe le plus proche de la question (Karpukhin et al., 2020; Bordes et al., 2014). Alternativement, pour une BC structurée, la question peut être transformée en une représentation logique ou directement traduite en une requête SQL ou SPARQL (Berant et al., 2013; Sherborne et Lapata, 2022). Les modèles de langue peuvent, eux, directement générer les réponses aux questions à partir des connaissances stockées dans leurs paramètres. Nous discutons d'abord des ressources disponibles pour les BC avant d'évoquer quelques limites liées aux modèles de langue et comment les deux approches peuvent être complémentaires.

4.1.1 Ressources

La BC non structurée la plus répandue est sans conteste Wikipédia, qui a l'avantage d'être libre de droit et de couvrir plusieurs langues. Wikipédia évolue au fil du temps et quelques années de différence peuvent avoir un impact significatif sur la performance d'un système (Izacard et al., 2022). Au cours de cette thèse, nous utiliserons la version de KILT (Petroni et al., 2021), un *benchmark* qui regroupe plusieurs tâches de TAL qui demandent des connaissances, dont la tâche de question-réponse. La version de Wikipédia traitée par Karpukhin et al. (2020) a aussi fait autorité récemment. Tamber et al. (2023) l'étudient précisément et montrent que plusieurs choix arbitraires ont été faits, qui étaient souvent néfastes pour les performances du système.

Les graphes de connaissances ont beaucoup dérivés de Wikipédia en extrayant des informations du texte et des données semi-structurées telles que les *infobox* (Suchanek et al., 2007; Bollacker et al., 2008; Lehmann et al., 2015). La fondation Wikimédia a ensuite repris la main sur ses ressources en proposant Wikidata (Vrandečić et Krötzsch, 2014). Il s'agit du graphe de connaissances standard dans le domaine général aujourd'hui, bien que certains anciens graphes soient toujours utilisés puisqu'ils sont liés à des *benchmarks*. Nous utiliserons Wikidata au chapitre suivant pour annoter semi-automatiquement des questions. Notons également l'existence de graphes de connaissances dans des domaines spécialisés (comme ceux construits à partir du méta-thésaurus UMLS dans le domaine du médical et du biomédical ; Bodenreider, 2004) ainsi que des usages spécifiques, par exemple ConceptNet pour les connaissances de sens commun (Liu et Singh, 2004).

Les BC multimodales sont pour leur part assez limitées à l'heure actuelle. Bien que des données structurées aient été ajoutées à Wikimédia Commons au cours d'un projet de 2017 à 2020 [14], leur développement est encore balbutiant. Dans Wikidata, les images sont essentiellement représentées comme des chaînes de caractères, c'est-à-dire par leur URL. BabelNet (Navigli et Ponzetto, 2012) intègre des images d'articles Wikipédia à ses entités sans structurer plus précisément leurs relations [15].

14. https://meta.wikimedia.org/wiki/Structured_Data_on_Commons

15. Ce n'est pas discuté dans Navigli et Ponzetto (2012) qui présente une version antérieure de BabelNet mais plusieurs travaux sont fondés sur la multimodalité de BabelNet, par exemple (Calabrese et al., 2020).

Shah et al. (2019) et Chen et al. (2023c) utilisent une seule image pour représenter chaque entité de leur BC, issue de Wikidata ou de l'*infobox* Wikipédia. Nous avons principalement suivi la même stratégie au cours de cette thèse mais explorerons également une alternative : exploiter les multiples images des articles Wikipédia conjointement avec leur légende.

4.1.2 Modèles de langue

À la suite des résultats de Radford et al. (2019), qui montraient que les modèles de langue acquéraient des capacités bien plus complexes et diverses que la syntaxe d'une langue et que ce phénomène allait croissant avec leur nombre de paramètres, Brown et al. (2020) ont provoqué un changement de paradigme en poussant la taille de leur modèle, GPT-3, à 175 milliards de paramètres. GPT-3 a en effet balayé l'étape d'ajustement des modèles évoqués plus haut pour la remplacer par un amorçage (*prompting*), qui permet d'interroger le modèle sans ajuster ses paramètres ou en ajouter de nouveaux. GPT-3 a ensuite connu le succès grand public à travers ChatGPT, ce qui a encore accru et élargi l'intérêt des communautés académiques pour ces modèles.

Concernant la RI et la tâche de question-réponse, GPT-3 a donc court-circuité l'étape de RI puisque la génération de réponse est fondée seulement sur les connaissances stockées implicitement dans ses paramètres. Des travaux antérieurs avaient suggéré sa faisabilité (Petroni et al., 2019; Radford et al., 2019) mais restaient bien en deçà des méthodes standards, tandis que Brown et al. (2020) ont établi un nouvel état de l'art sur TriviaQA (Joshi et al., 2017) tout en étant compétitifs sur d'autres *benchmarks* [16].

Ces modèles peuvent en effet être interrogés de différentes façons pour répondre à une question, la plus naturelle étant le *cloze test* évoqué précédemment, par exemple « Paris est la capitale de la... ». Ainsi, ils n'ont pas besoin de BC et utilisent uniquement les connaissances stockées implicitement dans leurs paramètres. Cette approche a l'avantage de pouvoir condenser une grande quantité de texte, mais avec aussi plusieurs limites (Zamani et al., 2022), notamment :

— les hallucinations : les modèles de langue génèrent du texte plausible, même s'il contient des informations erronées à propos d'entités nommées réelles (Ji et al., 2023) ;

— l'interprétabilité et la confiance : à cause des hallucinations, l'utilisateur ne peut pas savoir si la réponse du modèle est fiable (AlKhamissi et al., 2022) ;

— la mise à jour et la généralisation : l'entraînement d'un modèle de langue est coûteux et son ajustement à de nouvelles données n'est pas pratique et sujet à l'oubli catastrophique (French, 1999; Tirumala et al., 2022).

Ainsi, de nombreux travaux ou applications récents conditionnent la génération du modèle de langue par le résultat d'une RI, par exemple `https://bing.com/chat` (Lewis et al., 2020; Zamani et al., 2022; Izacard et al., 2022; Schick et al., 2023). Il n'est pas trivial de marginaliser la probabilité de génération selon plusieurs passages

16. Les travaux sur la RI générative sont cependant encore à un stade embryonnaire (Metzler et al., 2021; Tay et al., 2022).

résultant de la RI et plusieurs méthodes ont été proposées (Lewis et al., 2020). Nous les décrirons à la section 4.3. Izacard et al. (2022) montrent qu'indépendamment de la RI, la performance du système va croissant avec la taille du modèle de langue générateur, ce qui suggérerait qu'un gros modèle de langue combine les informations issues de son pré-entraînement implicitement stockées dans ses paramètres et celles de la BC explicitement utilisées via la RI [17]. Ils montrent également que la BC peut être mise à jour sans ré-entraîner le modèle.

Les connaissances peuvent donc être représentées explicitement à travers une BC, structurée ou non, ou implicitement par un modèle de langue. Nous nous focaliserons dans la suite de cette section sur des systèmes modélisant explicitement des connaissances non-structurées, textuelles donc, pour une recherche d'information menée dans le contexte de l'extraction ou la génération d'une réponse à une question.

4.2 Recherche d'information

4.2.1 Recherche neuronale : dense ou parcimonieuse

La RI a évolué rapidement ces dernières années. En effet, les approches neuronales antérieures aux modèles de langue pré-entraînés tels que BERT produisaient des résultats mitigés (Lin et al., 2021). Pour la tâche de question-réponse, Lee et al. (2019) et Karpukhin et al. (2020) ont été les premiers à surpasser les méthodes lexicales telles que BM25. Nous commençons dans cette section par décrire plusieurs de ces modèles avant d'évoquer à la section suivante différentes méthodes pour les pré-entraîner.

DPR (Karpukhin et al., 2020) est fondé sur une architecture à encodeur double : un BERT pour représenter la question et un autre [18] pour le passage. La représentation finale de chaque texte est obtenue via le token spécial [CLS] qui fusionne donc les informations à travers le mécanisme d'auto-attention du *transformer*, comme pour une classification au niveau du texte entier (Devlin et al., 2019). Les deux encodeurs sont ensuite entraînés conjointement de manière contrastive comme introduit à la section 2.2.

Khattab et Zaharia (2020) ont proposé de rendre cette méthode plus robuste en conservant une représentation pour chaque token de la question et du passage. Néanmoins, cela entraîne une augmentation considérable de la taille de l'index [19]. Plusieurs travaux y font suite pour améliorer son efficience (Gao et al., 2021a; Fan et al., 2023). Formal et al. (2021), au contraire, visent à conserver des représentations parcimonieuses, en représentant les textes dans l'espace du vocabulaire de BERT, ce qui permet d'utiliser un index inversé. Néanmoins, cela entraîne un compromis entre efficacité et efficience et ne traite pas le problème de la *polysémie*, puisque les termes sont toujours représentés dans l'espace du vocabulaire. Ce travail a été étendu pour

17. Il s'agit de notre interprétation des résultats présentés mais les auteurs ne discutent pas de ce point. De futures études plus contrôlées seraient bienvenues.

18. Il est également possible de partager les paramètres des deux encodeurs, on parle alors de modèle *siamois* (Bromley et al., 1993; Fun et al., 2021).

19. La RI conserve le terme d'index même s'il s'agit de représentations denses pré-calculées, et non pas d'index inversé.

la recherche cross-modale par Chen et al. (2023a) et Luo et al. (2023). Nos travaux explorent la KVQAE, qui est encore loin des contraintes de production de la RI textuelle ou cross-modale. Ainsi, nous n'avons pas expérimenté avec ces méthodes car DPR permet de chercher dans une BC de dizaines de millions de documents très rapidement.

DPR surpasse BM25, certes, mais dans un cadre d'apprentissage supervisé où les données d'évaluation sont indépendantes et issues de la même distribution (iid) que celles d'entraînement. Thakur et al. (2021) s'intéressent au contraire à l'apprentissage sans exemple en proposant le *benchmark* BEIR. Ils trouvent que BM25 surpasse DPR et ANCE sur de nombreux jeux d'évaluation, ces modèles étant entraînés respectivement sur plusieurs jeux de données de question-réponse, comme décrit dans Karpukhin et al. (2020), et MS Marco (Nguyen et al., 2016). ANCE (Xiong et al., 2020) est un modèle très proche de DPR mais entraîné avec un meilleur minage d'exemples négatifs difficiles. En accord avec ces résultats, Sciavolino et al. (2021) montrent que DPR est sensible à la popularité des entités nommées et est incapable de trouver un passage pertinent, même pour les questions les plus simples telles que « *Où X est-il né ?* », si X n'est pas suffisamment connu. Pour y remédier, Sciavolino et al. (2021) montrent qu'il suffit d'ajuster les paramètres de l'encodeur du *passage*, ce qui peut sembler étrange puisque l'on veut répondre à de nouvelles *questions*. Néanmoins, cela rejoint les résultats de Rajapakse et de Rijke (2023), qui supposent que, lorsqu'un passage est lié à une seule question, l'encodeur du passage va « effacer » les autres informations du passage, qui seraient potentiellement pertinentes pour d'autres questions [20], donc mal généraliser. Leur solution est donc de générer plusieurs questions par passage. Nous verrons à la section suivante que les tâches de pré-entraînement proposées par Lee et al. (2019) et Ram et al. (2022) génèrent plusieurs pseudo-questions à partir du même passage, en accord avec ces résultats. Ces travaux sont également liés à ceux de Zhang et al. (2022) et Hong et al. (2022), qui produisent plusieurs représentations par passage. Hong et al. (2022), en particulier, démontrent une meilleure robustesse de leur modèle au changement de domaine.

Puisque ces modèles encodent la question et le passage indépendamment, ils sont tous relativement rapides. Une fois cette RI initiale effectuée, il est fréquent de réordonner un petit nombre de documents pour améliorer la précision de la RI (Liu et al., 2009). Cela peut se faire également avec un modèle de langue pré-entraîné tel que BERT (Wang et al., 2019; Nogueira et Cho, 2020), qui va cette fois encoder *conjointement* la question et le passage grâce au mécanisme d'auto-attention des *transformer*. On suppose souvent que ce mécanisme permet de modéliser des interactions plus complexes entre la question et le passage qu'un encodeur double qui modélise leur similarité avec un produit scalaire (Karpukhin et al., 2020).

20. Par exemple, le passage « Emmanuel Macron est né le *21 décembre 1977* à *Amiens*. » est potentiellement pertinent pour de nombreuses questions, parmi lesquelles « *Où* Emmanuel Macron est-il né ? » et « *Quand* Emmanuel Macron est-il né ? »

4.2.2 Pré-entraînement

Nous nous concentrons ici sur les méthodes de pré-entraînement, en mettant l'accent sur la RI initiale, qui sert à la fois aux moteurs de recherche Web et aux systèmes de question-réponse. En s'appuyant sur l'architecture de l'encodeur double (Bromley et al., 1993; Karpukhin et al., 2020), le but général est de générer des *pseudo-questions*, idéalement imitant le plus possible de vraies questions, tout en ayant le moins de contraintes possibles afin que l'approche soit robuste et extensible. Ces contraintes peuvent s'exercer sur les *données*, par exemple une approche fondée sur les hyperliens au sein des articles Wikipédias, ce qui limite la quantité de données utilisable, ou encore sur les *modèles*, par exemple en se reposant sur un détecteur d'entités nommées, qui sont disponibles pour un petit nombre de langues, ou, pire encore, sur un générateur de questions supervisé.

Évoquons tout d'abord l'*Inverse Cloze Task* (ICT) de Lee et al. (2019), qui sera généralisée à la multimodalité au chapitre 5. L'ICT consiste, étant donné une phrase $\mathbf{q}$, à retrouver son contexte $\mathbf{p}^+$, c'est-à-dire les phrases suivantes ou précédentes, parmi un ensemble de N distracteurs $\mathbf{p}_j^-$. Ce n'est pas sans rappeler le modèle Skip-gram de Mikolov et al. (2013), mais au niveau des phrases plutôt que des mots. Ce processus est très peu contraint et peut donc être appliqué à une quantité virtuellement illimitée de données, a priori agnostique par rapport à la langue. Cependant, il est assez coûteux puisqu'il n'y a pas de moyen évident de trouver des exemples négatifs difficiles, ce que Lee et al. (2019) contournent en utilisant un grand nombre d'exemples négatifs aléatoires ($N = 4096$). Ce modèle est ensuite ajusté de bout-en-bout avec un module d'extraction de réponse.

Ram et al. (2022) critiquent ce dernier point et se focalisent sur l'apprentissage sans ou avec peu d'exemples (*zero-shot* et *few-shot*). Contrairement à Guu et al. (2020) et Sachan et al. (2021), qui sont fondés sur un détecteur d'entités nommées, les auteurs s'appuient simplement sur la récurrence de n-grammes (entre 2 et 10 mots) au sein d'un même document et de quelques heuristiques développées dans leur travail précédent pour pré-entraîner un module d'extraction de réponse (Ram et al., 2021). Étant donné un article Wikipédia divisé en passages de texte dans lesquels on trouve un n-gramme récurrent, un de ces passages sert de pseudo-question $\mathbf{q}$ tandis qu'un autre va être le passage pertinent $\mathbf{p}^+$ à retrouver, ce qui est rendu plus *difficile* par un troisième passage du même article $\mathbf{p}^-$, *ne* contenant *pas* le n-gramme. Ram et al. (2022) démontrent empiriquement leur méthode comme efficace, même sans ou avec peu d'exemples supervisés pour ajuster le modèle par la suite, contrairement à l'ICT et CoCondenser qui, sans ajustement, tombent en dessous de BM25 (Robertson et al., 1995), un modèle de RI lexical non supervisé. CoCondenser (Gao et Callan, 2022) est très similaire à l'ICT mais utilise une architecture légèrement différente où le token $[\mathrm{CLS}]$ est contraint à *condenser* les informations des autres mots de la phrase sur la moitié des couches du modèle (Gao et Callan, 2021).

Ma et al. (2021b) et Zhou et al. (2022) utilisent tous deux des hyperliens Wikipédias pour pré-entraîner des modèles, ce qui limite l'extensibilité de leur approche. De plus, le premier vise plutôt la recherche web, ce qui s'éloigne de nos intérêts. L'approche de Zhou et al. (2022) est compétitive avec Ram et al. (2022) mais, puisque les deux articles sont concurrents, il faudrait de futures études pour les comparer équitablement.

4.3 Extraction ou génération de réponse

Une fois la RI effectuée, l'approche standard consiste tout naturellement à *extraire* la réponse du passage pertinent (Chen et al., 2017). Autrefois composé de multiples modules d'analyse sémantique (Ferret et al., 2001), la multiplication de gros jeux de données tels que SQuAD (Rajpurkar et al., 2016) ou TriviaQA (Joshi et al., 2017) a encouragé l'apprentissage de bout-en-bout de cette extraction (Hermann et al., 2015; Chen et al., 2017). On peut par exemple entraîner le modèle à prédire indépendamment si le l-ième token du passage est le début (resp. la fin) de la réponse puis fusionner les deux scores à l'inférence (Devlin et al., 2019).

Néanmoins, avec l'essor des modèles de langue génératifs introduits plus haut, on peut *générer* la réponse en *conditionnant* seulement cette génération par un ou plusieurs passages provenant de la RI (Lewis et al., 2020). Ces modèles sont entraînés comme un modèle de langue auto-régressif, c'est-à-dire à maximiser la vraisemblance des réponses générées. Ils diffèrent dans leur manière de conditionner la génération par les différents passages retrouvés par la RI, et la façon d'entraîner le modèle de RI et de génération : jointe ou disjointe (Lewis et al., 2020; Izacard et Grave, 2021; Izacard et al., 2022). Ces modèles ont également été étendus à des tâches multimodales par Hu et al. (2023c) et Chen et al. (2022). Les premiers s'intéressent aux questions visuelles de sens commun et les seconds aux questions cross-modales.

5 Recherche cross-modale

La KVQAE nécessite par essence d'aller au-delà de la tâche de question-réponse textuel et en particulier, de s'intéresser aux rapports cross-modaux entre texte et image dans un contexte de RI. La RI cross-modale, de même que la génération de légende, est permise par un *alignement* ou une « traduction » d'une modalité à l'autre, par opposition aux questions visuelles qui nécessitent de *fusionner* les deux modalités. Nous nous y intéressons toutefois de par sa proximité avec la question de la modélisation des interactions cross-modales (notamment IQTP, entre l'image de la question et le texte du passage). Par ailleurs, nous étudierons plus directement la recherche cross-modale au chapitre 6 comme une façon alternative de reconnaître l'entité dépeinte dans la question visuelle.

De par cet alignement des modalités, de nombreux travaux se sont fondés sur l'analyse canonique des corrélations, qui suppose que les deux modalités fournissent une représentation hétérogène des mêmes données et s'appuie sur leur matrice de covariance croisée pour effectuer un changement de base (Rasiwasia et al., 2010; Guo et al., 2019). Plus récemment, des approches neuronales fondées sur l'alignement d'images et de textes avec une fonction de classement à marge ont été proposées pour diverses applications. Tout d'abord, Vendrov et al. (2016) visaient la recherche cross-modale en tant que telle. Ensuite, Frome et al. (2013) étaient intéressés par la classification d'images sans exemple. En effet, une fois que les représentations visuelles et textuelles sont alignées, on peut espérer généraliser à de nouvelles images associées à des textes absents du jeu d'entraînement. Ce travail présageait CLIP, qui a augmenté drastiquement la complexité du modèle mais surtout la quantité de

données, en relâchant les contraintes sur l'alignement des textes et des images, et a ainsi permis d'apprendre des représentations uniquement à partir de cet entraînement cross-modal, plutôt que de les initialiser avec des pré-entraînements monomodaux. Enfin, Lazaridou et al. (2015) visaient au contraire à obtenir de meilleures représentations textuelles en s'appuyant sur des théories liées à l'apprentissage de langue visuellement ancré (Harnad, 1990). Cela vient également pallier le *reporting bias* auquel sont sujettes les représentations purement textuelles (Gordon et Van Durme, 2013). Nous retrouvons aussi des échos de ce travail plus récemment avec Wolfe et Caliskan (2022), qui trouvent que les représentations textuelles de CLIP sont meilleures que celles de GPT-2, son équivalent mono-modal. Ils supposent que la recherche cross-modale engendre une représentation plus sémantique du texte, moins syntaxique qu'un modèle de langue classique.

La recherche cross-modale a connu un regain d'intérêt avec les *transformer* multimodaux, qui l'utilisent comme pré-entraînement ou évaluation. Comme pour la RI textuelle, nous distinguons deux sortes de modèles :

1. les encodeurs doubles, comme CLIP, qui servent à la RI initiale ;
2. les encodeurs joints, comme UNITER (Chen et al., 2020b), qui servent au réordonnancement.

Nous remarquons toutefois que cette distinction a souvent été négligée dans la littérature sur la VQA et d'autres tâches multimodales, où l'évaluation de la recherche cross-modale est faite sur des jeux de données initialement conçus pour la génération de légendes, comme Flick30k et MS-COCO (Plummer et al., 2015; Lin et al., 2014). En effet, ces derniers ne fournissent pas de collection d'images (de BC dans notre cadre, qui contient des exemples distracteurs), ni de jugement de pertinence (de *qrels* dans un cadre de RI textuelle). Yang et al. (2023) cherchent à pallier ce problème en proposant une grande collection d'images et des jugements de pertinence dans le cadre de TREC 2023.

Nous expérimenterons avec CLIP tout au long de la thèse, d'abord pour une recherche d'image au chapitre 4, puis pour une recherche cross-modale au chapitre 6. Ce dernier point a en partie été motivé par le travail de Sun et al. (2022), qui emploient CLIP pour la recherche cross-modale dans le cadre d'une tâche connexe à la KVQAE : la désambiguïsation visuelle d'entités nommées, qui consiste à reconnaître les entités dépeintes dans une image de façon cross-modale, en exploitant donc l'interaction IQTP. Nous expérimenterons également avec un encodeur joint, similaire à UNITER, pour fusionner les modalités au sein d'une question visuelle, au chapitre 5.

La recherche cross-modale vise implicitement à obtenir une représentation des données invariante à la modalité. Les travaux de Quiroga (2012) suggèrent que de telles représentations existent dans le cerveau humain, où la même entité nommée, présentée sous la forme de son nom écrit ou d'une photographie, active les mêmes neurones. Goh et al. (2021) ont identifié des neurones similaires au sein de CLIP, ce qui pourrait expliquer pourquoi il est efficace pour reconnaître des entités nommées, comme nous le verrons par la suite. Aytar et al. (2017) visaient explicitement une représentation des données invariante à la modalité. Ils partagent ainsi les dernières couches de leur modèle à travers les modalités, en supposant qu'elles opèrent à un niveau plus abstrait, en conservant des premières couches spécifiques à chaque modalité. Nous étudierons également cette méthode au chapitre 5.

La RI multimodale, où à la fois la requête mais aussi la BC sont composées de paires (texte, image), comme pour la KVQAE, a été étudiée par le passé (Srihari et al., 2000; Clough et al., 2004; Depeursinge et Müller, 2010) mais nous n'avons pas connaissance de travaux plus modernes (hormis ceux sur la KVQAE bien entendu). Remarquons également que les modalités sont redondantes dans ces travaux sur la RI multimodale. Pour reprendre l'exemple de la figure 2.1, on pourrait utiliser l'image de l'Arc de Triomphe de façon redondante avec la question textuelle non-ambiguë « *How many avenues radiate from the Arc de Triomphe ?* » Au contraire, en ce qui concerne les questions visuelles, l'entité visée par la question est référencée par une expression ambiguë, « *this building* » pour reprendre l'exemple précédent.

6 Questions visuelles et multimodalité

De nombreux travaux sont connexes à la KVQAE, même en se restreignant à la multimodalité ou aux questions visuelles. Ces travaux sont, comme les nôtres, particulièrement intéressés par les interactions qui surviennent entre les modalités. Plus précisément, il s'agit souvent de *fusionner* les informations multimodales en étudiant l'interaction TQIQ au sein de la question visuelle.

Nous commencerons par évoquer deux tâches plus fondamentales qui composent parfois implicitement les systèmes de question-réponse multimodaux : la désambiguïsation multimodale d'entités nommées et la résolution d'expressions référentielles. Nous discuterons ensuite de deux différentes familles de modèles de langue multimodaux, les premiers, inspirés par BERT, s'inscrivant dans le paradigme *pré-entraînement et ajustement*, et les seconds, inspirés par GPT-3, s'inscrivant dans le paradigme *modèle de langue universel*. Nous distinguerons ensuite trois tâches assimilées aux questions visuelles :

1. la VQA classique, des premiers travaux à l'émergence des questions visuelles demandant des connaissances externes ;
2. la KVQAE, des travaux antérieurs aux nôtres, fondés sur une BC structurée et limités à un type d'entité, aux travaux plus récents qui exploitent d'autres sortes de BC et étudient divers types d'entités ;
3. et enfin, les questions cross-modales, dérivées des questions visuelles mais qui partagent des similarités avec la recherche cross-modale évoquée à la section précédente.

6.1 Désambiguïsation multimodale d'entités nommées

La désambiguïsation multimodale d'entités nommées (MEL [21]) a été introduite par Moon et al. (2018) dans un contexte applicatif : la désambiguïsation d'entités nommées dans les réseaux sociaux, où les publications sont quasi-systématiquement multimodales. La MEL repose sur la même hypothèse que la KVQAE, à savoir que les modalités prisent indépendamment sont souvent ambiguës mais peuvent être désambiguïsées conjointement en étudiant les interactions cross-modales. Cependant

21. *Multimodal Entity Linking*

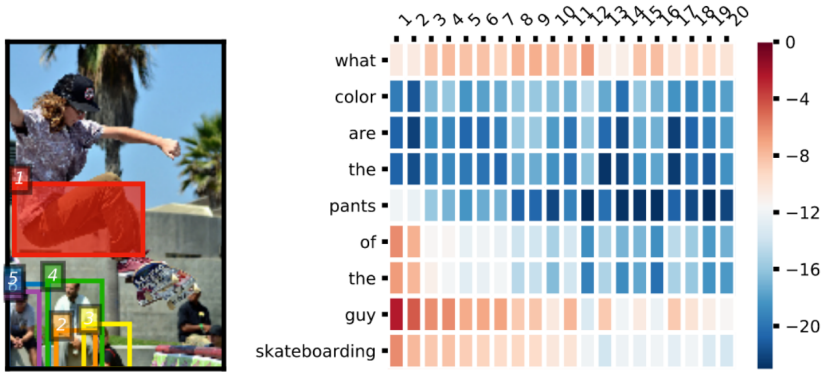


FIGURE 2.3 – Visualisation tirée de Kim et al. (2018) qui illustre comment les mécanismes d'attention permettent de fusionner les modalités pour traiter la VQA classique et la résolution d'expression référentielle. La question « *What color are* the pants *of the guy skateboarding ?* » est issue de VQA v2 (Goyal et al., 2017). Les numéros des colonnes énumèrent les régions de l'image les plus saillantes. La plus saillante correspond bien au pantalon désigné par l'expression référentielle.

Moon et al. (2018) peinent à surpasser leur référence textuelle. Puisque leurs données (publications Snapchat) sont privées, Adjali et al. (2020a) annotent automatiquement un jeu de données à partir de tweets, ce qui permet de reproduire les données sans les partager directement pour se conformer à la RGPD. Ils proposent de fusionner les modalités simplement et entraînent leur modèle avec une *triplet loss* mais les améliorations par rapport à la référence textuelle restent minimes (Adjali et al., 2020b). Ces travaux ont en partie motivé les membres du projet MEERQAT à travailler sur la KVQAE, où la modalité textuelle est moins prédominante.

6.2 Expressions référentielles

Les expressions référentielles ont été beaucoup étudiées d'un point de vue psycholinguistique (Krahmer, 2010). En effet, elles peuvent varier selon la position sociale du locuteur et son rapport à l'interlocuteur (Griffin, 2010).

Robert Dale, en s'appuyant sur ces travaux, a beaucoup contribué à la génération d'expression référentielle et y a notamment consacré sa thèse de doctorat (Dale, 1989). Dans le cas des questions visuelles, l'expression référentielle correspond au terme du texte qui fait référence à un objet dans l'image. Elle doit être suffisamment discriminante pour différencier l'objet des multiples objets dans l'image. Cela peut inclure des relations spatiales, mais aussi des relations de taille ou de couleur, par exemple « *le gros rond rouge sur la droite* » (Viethen et Dale, 2008; Mitchell et al., 2013). S'il y a un seul objet dans l'image, la génération est triviale et l'on peut simplement se référer au type de l'objet ou utiliser un pronom (Dale et Haddock, 1991).

La résolution d'expression référentielle [22] peut être traitée avec un mécanisme d'attention et les mêmes architectures que la VQA (Kim et al., 2018; Cho et al.,

22. Aussi connue sous le nom d'ancrage visuel (*visual grounding*).

2021 ; cf. figure 2.3). Dans notre cas, les expressions référentielles sont utiles dans deux cas bien distincts :

1. s'il y a plusieurs entités, le plus souvent des personnes, présentes dans l'image ; ce cas rejoint donc le cadre classique bien qu'on prendra garde à utiliser des attributs appropriés pour désigner des personnes ;
2. si l'image représente différentes entités selon le contexte ; comme discuté au chapitre 1, par exemple l'image (e) de la figure 1.2 peut servir à représenter l'entreprise *Apple* ou le magasin *Apple Fifth Avenue* ; ce type de résolution n'a pas été précédemment étudié à notre connaissance.

Nous ne traitons pas frontalement ce problème. Au chapitre 5, nous proposerons un modèle fondé sur un mécanisme d'attention entre le texte et l'image au sein de la question visuelle (interaction TQIQ) qui pourrait résoudre l'expression référentielle implicitement. Par ailleurs, l'interaction textuelle TQTP peut aider à résoudre ce problème, principalement dans le second cas.

6.3 Modèles de langue multimodaux

6.3.1 Pré-entraînement et ajustement

Les travaux modernes sur les tâches multimodales se sont en toute logique beaucoup concentrés sur la fusion des modalités. Cette fusion employant souvent un mécanisme d'attention, il est tout naturel que les travaux se soient emparés des *transformer* pré-entraînés peu après leur apparition.

Ainsi, de nombreux travaux ont concurremment proposé une extension de BERT à la multimodalité en s'appuyant sur l'appariement d'images et de leur légende textuelle (Tan et Bansal, 2019; Lu et al., 2019; Li et al., 2019; Su et al., 2020; Li et al., 2020; Chen et al., 2020b). Ces modèles sont souvent classés selon la façon dont ils intègrent l'image, elle-même initialement représentée par un détecteur d'objets lors de ces premiers travaux :

— à flux unique (*single stream*), lorsque l'image est projetée dans le même espace que les plongements lexicaux, donc prise en entrée par le *transformer* comme les mots du texte, espace où la fusion est effectuée à travers un mécanisme d'auto-attention (cf. figures 2.4a et 2.5a) ;

— à double flux (*dual stream*), lorsque l'image est d'abord traitée par un module *transformer* (figure 2.4b et 2.5b et c) et les modalités fusionnées à travers une attention croisée (figure 2.4c et 2.5d et e).

L'idée de considérer l'image comme un ou plusieurs plongements lexicaux, présente dans les modèles à flux unique, n'est pas nouvelle, et avait été explorée dès (Ren et al., 2015). Elle a été critiquée pour son aspect « pot pourri » qui traite deux modalités différentes de la même façon, avec les mêmes paramètres. Notamment, les relations spatiales entre les objets d'une image devraient être différenciées de la relation syntaxique entre les mots d'une phrase. Cependant, Bugliarello et al. (2021) replacent les modèles à flux unique ou double dans le même cadre et montrent qu'ils fonctionnent aussi bien, la différence rapportée dans les articles s'expliquant principalement par les données et les objectifs de pré-entraînement utilisés. Hessel

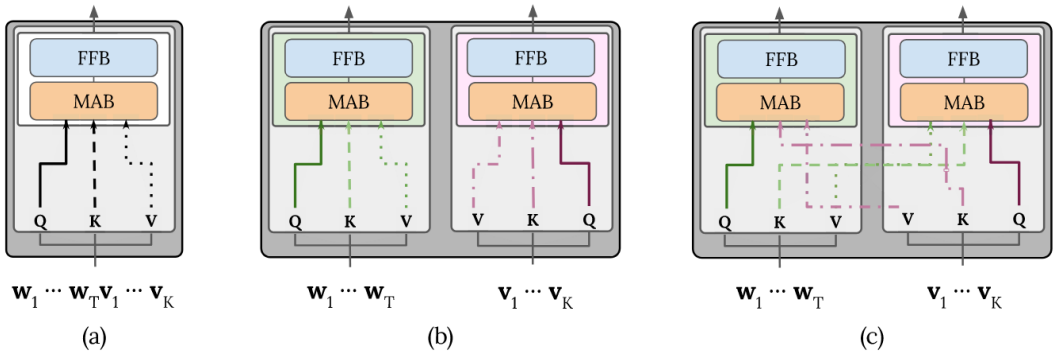


FIGURE 2.4 – *Transformer* multimodaux, à flux unique (a) ou double (b et c). $\mathbf{w}_i$ désigne les plongements lexicaux (texte) et $\mathbf{v}_j$ les représentations de l'image. MAB (*Multi-head Attention Block*) et FFB (*Feed-Forward Block*) désignent les modules composants du *transformer*. Adapté de Bugliarello et al. (2021).

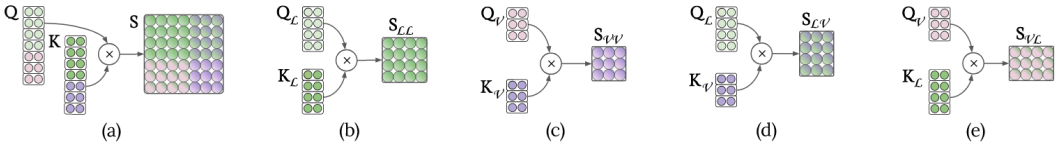


FIGURE 2.5 – Illustration tirée de Bugliarello et al. (2021) des différentes attentions au sein de *transformer* multimodaux selon les modalités : toutes les combinaisons mono- et cross-modales pour les modèles à flux unique (a) ; texte-texte (b) ; image-image (c) ; texte-image (d) ; et symétriquement, image-texte (e). Le texte est représenté par des teintes de vert et l'image par des teintes de rose.

et Lee (2020) montrent que ces modèles apprennent des interactions cross-modales non triviales pour la VQA.

Ces modèles sont pré-entraînés de différentes façons qui s'appuient toujours sur l'appariement d'une image et de sa légende (par exemple générer la légende de l'image ; Gan et al., 2022). Nous les décrirons plus précisément au chapitre 5.

6.3.2 Modèle de langue universel

Tout comme les travaux inspirés par BERT évoqués à la section précédente, ceux s'inscrivant dans le paradigme *modèle de langue universel* n'ont pas tardé, suite à la publication de GPT-3 (Brown et al., 2020). Comme les BERT multimodaux, ils s'appuient sur l'appariement d'une image et de sa légende pour rendre le modèle de langue multimodal. Comme évoqué à la section 4.1.2, puisque ces modèles sont enclins aux hallucinations et fonctionnent mieux une fois conditionnés par le résultat d'une RI, nous ne les avons pas utilisés et nous nous sommes focalisé sur la RI. Nous en discutons toutefois brièvement ici étant donné l'intérêt porté par la communauté, notamment les récents travaux de KVQAE décrits à la section suivante. Nous serons bref sur les modèles en cascade (*pipeline*), qui « traduisent » l'image en texte à l'aide d'un module dédié avant d'amorcer le modèle de langue avec le résultat (Yang et al., 2022).

Jin et al. (2022) entraînent un encodeur-décodeur à générer la légende d'une image. Ils peuvent ensuite amorcer le modèle sans ajuster ses paramètres pour répondre à des questions visuelles classiques.

Tsimpoukelli et al. (2021) ont proposé Frozen, qui est lui fondé sur Chinchilla-7B [23] (Hoffmann et al., 2022), un décodeur auto-régressif similaire à GPT mais dont la quantité de données d'entraînement a été augmentée proportionnellement à la taille du modèle. Frozen garde le modèle de langue figé, comme son nom l'indique, l'image y est donc intégrée en la projetant dans le même espace que les plongements lexicaux. Cette approche est ainsi similaire aux modèles multimodaux à flux unique évoqués plus haut, sauf qu'ici le *transformer* reste figé. Cela s'apparente alors plutôt à un ajustement de l'amorce (*prompt tuning* ; Lester et al., 2021). Frozen est ainsi « pré-entraîné » à générer des légendes d'images de manière auto-régressive. Ce modèle permet de traiter la VQA classique. Tsimpoukelli et al. (2021) l'amorcent sans ou avec peu d'exemples pour étudier son apprentissage *en contexte*.

Alayrac et al. (2022) étendent ce travail en considérant d'une part, des modèles de langue plus gros, avec Chinchilla-70B, mais surtout en ajoutant des couches d'attention croisée au sein du modèle de langue, ce qui permet une meilleure modélisation de l'image tout en laissant le modèle de langue figé pour éviter l'*oubli catastrophique*. Ce phénomène survient en effet lorsqu'on entraîne un réseau de neurones de façon séquentielle (French, 1999). Dans le cas présent, il mènerait Chinchilla à oublier ses connaissances acquises lors de son pré-entraînement textuel. En plus de l'habituelle génération de légendes d'images, le modèle est également entraîné sur des pages web multimodales. Ces pages web permettent au modèle d'acquérir un certain sens commun d'après leur étude d'ablation sur OK-VQA (Marino et al., 2019). Nous pouvons supposer qu'il acquiert également des connaissances à propos d'entités nommées et serait alors capable de traiter la KVQAE.

OpenAI a également proposé un GPT-4 multimodal mais sans donner d'information sur l'architecture ou les données d'entraînement ni d'évaluation sur des *benchmark* multimodaux (OpenAI, 2023). Nous évoquerons à la section 6.5 le travail de Li et al. (2023), qui applique un modèle similaire à la KVQAE.

6.4 Genèse des questions visuelles

Motivation La VQA a largement été étudiée après l'avènement de l'apprentissage profond en vision par ordinateur et en TAL (Krizhevsky et al., 2012; Mikolov et al., 2013) en présentant la tâche comme une généralisation de toutes les tâches de vision (Malinowski et Fritz, 2014; Antol et al., 2015; Kafle et Kanan, 2017). Il est donc naturel que la plupart des travaux proviennent de cette communauté (Wu et al., 2017; Bernardi et Pezzelle, 2021) [24]. La majorité des travaux s'inscrivent dans la perspective de tester les capacités (*to probe*) des modèles, ce que Rodriguez et Boyd-Graber (2021) appellent le paradigme de Manchester, en hommage à Alan Turing, avec des questions générées automatiquement ou collectées par des annotateurs qui

23. Vraisemblablement étant donné que Tsimpoukelli et al. (2021); Hoffmann et al. (2022); Alayrac et al. (2022) font partie d'une même équipe et qu'Alayrac et al. (2022) expérimentent avec Chinchilla-7B mais (Tsimpoukelli et al., 2021) est antérieur à (Hoffmann et al., 2022).

24. Les ateliers VQA étaient organisés annuellement à CVPR de 2016 à 2021 (`https://visualqa.org/workshop.html`).

connaissent la réponse (Ren et al., 2015; Antol et al., 2015). Les travaux de VQA sont souvent motivés par une assistance aux personnes malvoyantes mais seuls Gurari et al. (2018) ont proposé un jeu de questions posées par des personnes aveugles. Notre travail s'inscrit dans ce même paradigme de Manchester, puisque nous avons collecté des questions dans le but d'évaluer des systèmes de KVQAE (cf. chapitre 3). De plus, notre travail s'appuie sur TriviaQA, dont les questions sont également posées par des personnes qui connaissent la réponse, par opposition à une personne en quête d'information (Joshi et al., 2017).

Questions visuelles classiques Les travaux de VQA classique se sont largement concentrés sur les méthodes de fusion multimodale, étudiant l'interaction cross-modale TQIQ au sein de la question visuelle, d'abord en utilisant un mécanisme d'attention entre les mots représentés par un réseau de neurones récurrents et l'image représentée par réseau de convolutions (Fukui et al., 2016; Zhang et al., 2019), puis en exploitant le mécanisme d'auto-attention des modèles *transformer* (Li et al., 2019). Ces travaux traitent la tâche comme une classification parmi les réponses les plus fréquentes du jeu d'entraînement, les approches génératives sans contrainte étant étonnamment récentes (Cho et al., 2021). Puisque la majorité des *benchmarks*, dont le très compétitif VQA (v1 ou v2), proviennent des images de MS-COCO, les images ont longtemps été représentées par des détecteurs d'objet entraînés sur les mêmes données (Anderson et al., 2018; Chen et al., 2020b), l'apprentissage de bout-en-bout n'ayant finalement ressurgi qu'avec les méthodes de pré-entraînement (Kim et al., 2021).

Questions visuelles de sens commun La VQA a ensuite été étendue aux questions demandant des connaissances externes, souvent de sens commun, mais toujours à propos de catégories d'objets « gros grain », donc en traitant la tâche de manière similaire à la VQA classique (Wang et al., 2017, 2018; Marino et al., 2019; Gardères et Ziaeefard, 2020). Par la suite, Qu et al. (2021a) et Luo et al. (2021) ont proposé concurremment une approche assez similaire à celle du chapitre 5, puisqu'ils ont utilisé un BERT multimodal, LXMERT, pour représenter une question visuelle et chercher dans une BC textuelle. Bien que les deux expérimentent avec OK-VQA, leurs résultats sont contradictoires, ceux de Qu et al. (2021a) étant bien plus optimistes, ce qui peut s'expliquer par leurs différentes métriques et BC. Nous notons aussi que Luo et al. (2021) sont les premiers à traiter la VQA comme nous le verrons au chapitre 4, avec un module d'extraction de réponse plutôt qu'un classifieur des réponses les plus fréquentes.

De la difficulté d'un *benchmark* Le *benchmark* de sens commun le plus utilisé, OK-VQA (Marino et al., 2019), présente plusieurs problèmes d'annotation. En effet, Marino et al. (2019) ont collecté des questions en donnant aux annotateurs l'instruction de poser des questions qui pourraient « berner un robot intelligent [25] ». Cela a conduit les annotateurs à poser de nombreuses questions où l'on ne peut qu'essayer de *deviner* la réponse, par exemple « *How long does it take to make this*

25. « *fool a "smart robot"* »

food ? » D'autres annotateurs devaient ensuite répondre à ces questions, sans accès à une BC. Chen et al. (2023c) trouvent donc que seulement 29% des questions d'OK-VQA demandent des connaissances externes. Par conséquent, les réponses varient beaucoup selon les annotateurs [26]. Nous avons calculé que l'accord inter-annotateur est léger avec un Kappa de Fleiss de 0,358 et une exactitude [27] humaine de 64%, sur le jeu de validation, bien que Marino et al. (2019) aient filtré 73 000 des 87 000 questions collectées initialement.

6.5 Questions visuelles à propos d'entités nommées

Shah et al. (2019) étaient les premiers à traiter la KVQAE. Ils ont proposé KVQA, le premier jeu de données dédié à cette tâche mais généré automatiquement et limité aux entités de type personne. Nous discuterons des limitations que cela implique plus en détail au chapitre suivant. Shah et al. (2019) traitent la multimodalité de la tâche par une fusion tardive au niveau de la décision : les entités nommées sont détectées et désambiguïsées à la fois dans la question et l'image par des modules indépendants avant d'être regroupées. Un sous-graphe de Wikidata est ensuite construit à partir de ces entités et traité par un *memory network* (Weston et al., 2014). Ce jeu de données a malheureusement été assez peu utilisé, ce que nous ne changerons pas puisque nous utiliserons une BC non structurée, tandis que KVQA est assez spécifique à Wikidata. Seuls Vickers et al. (2021), Garcia-Olano et al. (2022) et Heo et al. (2022) ont proposé de nouvelles méthodes évaluées avec KVQA mais il est difficile de comparer leurs approches au reste de l'état de l'art car leurs systèmes prennent en entrée la légende de l'image, ce qui rend l'image elle-même redondante. Garcia-Olano et al. (2022) rapportent également un problème dans le protocole expérimental de Shah et al. (2019) et par conséquent de Vickers et al. (2021) et Heo et al. (2022) : 25% des questions à propos d'une entité donnée utilisent la même image que celle de référence pour cette entité dans leur BC.

Adjali et al. (2023) proposent une approche hybride exploitant à la fois une BC structurée et non structurée, en s'appuyant sur nos travaux (chapitre 4). Ils combinent la représentation textuelle du passage avec celles des plongements de graphes des entités qui y sont détectées mais sont ainsi principalement limités par le peu d'informations textuelles contenues dans les questions. Ils envisagent d'étendre leur méthode aux entités représentées dans les images.

Les travaux les plus proches du nôtre sont sans nul doute les très récents (Hu et al., 2023a) et (Chen et al., 2023c). Les premiers se focalisent sur la recherche d'entités. Les questions sont donc des variations de « *What is the name of* the model of this aircraft *?* », où seule l'expression référentielle (ici *the model of this aircraft*) est pertinente. Les seconds travaillent sur des questions visant différents attributs des entités, comme dans notre définition de la KVQAE. Leur jeu de données est divisé en deux sous-ensembles, avec un très grand généré automatiquement, et un de taille modeste annoté manuellement. Nous reviendrons sur ces données et les

26. On a par exemple les réponses divisées entre *Federer* et *Nadal* pour une question sur le joueur le mieux classé au tennis.

27. L'exactitude souple utilisée comme métrique officielle d'OK-VQA : « il faut qu'au moins 3 annotateurs sur 10 aient donné cette réponse pour qu'elle soit juste ».

comparerons au jeu de données ViQuAE au chapitre suivant. Les deux travaux étant menés conjointement, ils explorent les deux mêmes approches :

— un système en cascade où la recherche visuelle est faite avec CLIP ;

— un système bout-en-bout fondé sur un gros modèle de langue multimodal, PaLI (Chen et al., 2023b) ou OFA (Wang et al., 2022), qui n'a pas accès à une BC et stocke toutes ses connaissances dans ses paramètres.

Ces derniers modèles se révèlent inefficaces sans ajustement mais montrent des performances meilleures après ajustement, bien que très inférieures au système avec BC [28]. Les auteurs montrent que la RI est l'étape la plus cruciale dans le système en cascade et qu'une RI oracle améliore les performances de 126% à 151% selon les modèles. Mensink et al. (2023) arrivent à des conclusions similaires. Ils proposent eux aussi un nouveau *benchmark* mais limité aux *landmarks* (monuments et sites naturels) et aux espèces d'êtres vivants (animaux et végétaux).

Enfin, dernièrement, Li et al. (2023) ont poursuivi une approche sans BC mais sans ajustement non plus, en amorçant un gros modèle de langue avec différentes instructions.

6.6 Questions cross-modales

D'autres tâches se rapprochent de la KVQAE mais sont plus proches d'une RI ou d'une analyse cross-modale car elles étudient l'interaction IQTP entre l'image de la question et le texte du passage. Par exemple, Fu et al. (2022) proposent de déterminer le lieu et la date d'une image à partir de son seul contenu. Cette tâche très difficile demande un raisonnement en plusieurs étapes et peut être soutenue par des connaissances à propos d'entités nommées ou de sens commun. Les auteurs proposent une référence fondée sur CLIP qui fournit des résultats assez faibles, surtout pour la prédiction de la date, mais cette tâche n'a pas encore été davantage étudiée.

Comme évoqué dans l'introduction de ce chapitre, les questions cross-modales sont souvent dépendantes d'un contexte multimodal (Kembhavi et al., 2017; Sampat et al., 2020; Talmor et al., 2021; Chang et al., 2022; Reddy et al., 2022). Par exemple, Reddy et al. (2022) posent des questions sur des articles de presse où l'image sert à désambiguïser les différentes entités mentionnées dans le texte de l'article. Cela s'apparente donc davantage à une compréhension de texte (*reading comprehension*) qu'à une RI.

Néanmoins, certains de ces jeux de données ont été détournés. Liu et al. (2023) et Chen et al. (2022) travaillent tous deux sur WebQA (Chang et al., 2022) en traitant ses questions cross-modales sorties de leur contexte et étudient donc la partie RI, visant justement à retrouver le contexte pertinent. Liu et al. (2023) utilisent CLIP pour trouver des images pertinentes, ce qui s'apparente donc à ce que nous verrons au

28. Nous nous focalisons ici sur les résultats sur le sous-ensemble annoté manuellement, à propos d'entités absentes du sous-ensemble automatique qui sert à l'ajustement. Sur ce dernier, les approches sans BC sont compétitives vis-à-vis de celles avec BC. Nous remarquons aussi une contradiction entre Hu et al. (2023a) et Chen et al. (2023c) où les premiers trouvent que PaLI fonctionne d'autant mieux, par rapport à CLIP, que les entités sont davantage connues, tandis que les seconds trouvent exactement l'inverse.

chapitre 6 mais dans un sens inverse, puisque leur requête est textuelle et non visuelle. Ils trouvent aussi qu'il est préférable d'extraire les caractéristiques directement de l'image plutôt que d'utiliser un détecteur d'objets. Ces derniers ne seraient donc pas adaptés à la représentation d'entités nommées, comme nous le supposons tout au cours de cette thèse, ce qui rejoint également les résultats rapportés par Chang et Bisk (2022) autour du challenge WebQA. Chen et al. (2022), quant à eux, utilisent un modèle génératif conditionné par une RI (cf. section 4.3). Cette méthode est donc similaire à celle du chapitre 5 bien qu'à nouveau, la requête soit seulement textuelle et pas multimodale. Contrairement à l'ICT multimodale que nous proposons, Chen et al. (2022) pré-entraînent leur modèle en combinant différentes tâches : génération de légende d'images, de réponse à une question visuelle ou à une question textuelle. Cette approche reste aujourd'hui à appliquer à la KVQAE.

7 Conclusion

Notre travail s'inscrit dans différents domaines de recherche bien plus vastes. Seulement quelques travaux portent sur la KVQAE elle-même, mais de nombreux autres portent sur des tâches multimodales connexes ou sont fondés sur les mêmes méthodes d'apprentissage. Nous avons principalement discuté des méthodes appliquées à six tâches d'intérêt :

1. la recherche d'image par le contenu ;
2. la RI/question-réponse textuelle ;
3. la VQA classique ;
4. la VQA de sens commun ;
5. la RI/VQA cross-modale ;
6. et bien entendu, la KVQAE.

Toutes ces tâches sont liées au paradigme *pré-entraînement et ajustement* et la RI, mono- ou multimodale, est liée à l'apprentissage contrastif. De plus, la RI se doit d'être très efficiente pour passer à l'échelle et se divise souvent en deux étapes pour ce faire : RI initiale et réordonnancement, lesquelles sont désormais souvent traités par des *transformer*, qui encodent respectivement indépendamment ou de façon jointe la question et le passage. Les tâches multimodales ont en commun la fusion d'information multimodale, laquelle est désormais souvent effectuée à travers le mécanisme d'attention des *transformer*.

En ce qui concerne la KVQAE, nous avons identifié un manque de jeu de données, puisque seul KVQA de Shah et al. (2019) est disponible à ce jour [29] et qu'il est limité par sa diversité d'entités, de sujets, de lexique et de grammaire, ce que nous comblons au chapitre suivant avec le jeu de données ViQuAE. Étant les premiers à traiter la KVQAE avec une BC non structurée, nous avons dû définir un premier cadre expérimental fondé sur le jeu de données ViQuAE et cette BC. Notre

29. Bien qu'InfoSeek (Chen et al., 2023c) et OVEN (Hu et al., 2023a) aient été pré-publiés en Février 2023, leurs auteurs nous ont confidentiellement partagé les sous-ensembles annotés automatiquement le 27 Juin 2023 seulement. Les sous-ensembles annotés manuellement, plus intéressants pour l'évaluation, sont toujours indisponibles au moment de la rédaction.

premier système, présenté au chapitre 4, est fondé sur les travaux en reconnaissance faciale et recherche d'image générique discutés à la section 3, ainsi qu'en RI et question-réponse textuel (cf. section 4). Ayant dans ce chapitre identifié la RI comme principale composante de la KVQAE, en accord avec Chen et al. (2023c), nous y avons consacré la majeure partie du reste de la thèse. Au chapitre 5 nous viserons donc une meilleure RI à travers une meilleure fusion multimodale, en nous appuyant à la fois sur les travaux évoqués à la section 6 pour la modélisation multimodale mais aussi à la section 4.2.2 pour le pré-entraînement dans un cadre de RI. Enfin, d'après les résultats de ce chapitre 5 mais également des travaux sur la recherche cross-modale (cf. section 5), nous explorerons différentes manières de chercher avec et d'ajuster CLIP, de façon mono- ou cross-modale, au chapitre 6.

Chapitre 3

Jeu de données et base de connaissances

1 Introduction et motivation

Dans cette thèse, nous avons adopté le paradigme du *benchmark* qui consiste à évaluer la progression de nos recherches au travers de la performance croissante des systèmes sur des jeux d'évaluation. Cette approche se justifie dans notre cas où la KVQAE est facilement évaluable : une réponse à une question factuelle est vraie ou fausse. Nous pouvons donc raisonnablement estimer qu'un système répondant correctement à plus de questions qu'un autre est meilleur et valider ainsi les méthodes et hypothèses de recherche. En effet, nous nous intéressons aux questions factuelles, ou *factoïdes*, puisque leur factualité dépend de la BC utilisée. Nous renvoyons le lecteur à Bolotova et al. (2022) pour une taxonomie et un état de l'art des questions non-factoïdes, par exemple sujettes à débat « *Y a-t-il une vie extraterrestre ?* » Ces types de questions sont bien plus difficiles à évaluer automatiquement. Ce chapitre contribue à répondre à deux de nos questions de recherche :

— Comment évaluer un système de KVQAE ?

— Comment représenter visuellement une entité nommée ? Nous nous intéresserons ici aux images utilisées pour les représenter et étudierons différents modèles aux chapitres suivants.

Pour répondre à la première question, nous proposons ViQuAE, un jeu de données de 3 700 questions visuelles permettant de suivre la progression des systèmes de KVQAE. Ce jeu de données a notamment comblé un vide car il était seulement le *deuxième* de KVQAE après KVQA (Shah et al., 2019), ce dernier étant limité aux entités de type personne et généré automatiquement, comme discuté au chapitre précédent et approfondi à la section 1.1. On peut voir deux exemples issus du jeu de données ViQuAE à la figure 3.1. Nous avons choisi de travailler sur la langue anglaise pour permettre une meilleure dissémination des données et de nos travaux et, réciproquement, profiter plus facilement des ressources à disposition, qu'il s'agisse de modèles pré-entraînés ou de jeux de données. Nous avons décidé de fonder ViQuAE sur le jeu de données TriviaQA, comme expliqué à la section 1.2.

ViQuAE a été collecté et annoté semi-automatiquement en deux étapes décrites

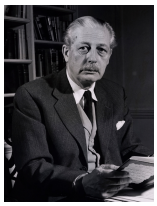
Question visuelle (entrée)	Passage visuel pertinent dans la base de connaissances
 *"Which constituency did this man represent when he was Prime Minister?"*	 Macmillan indeed lost Stockton in the landslide Labour victory of 1945, but returned to Parliament in the November 1945 by-election in Bromley.
 *"In which year did this ocean liner make her maiden voyage?"*	 Queen Elizabeth 2, often referred to simply as QE2, is a floating hotel and retired ocean liner built for the Cunard Line which was operated by Cunard as both a transatlantic liner and a cruise ship from 1969 to 2008.

FIGURE 3.1 – Deux exemples de questions visuelles du jeu de données ViQuAE accompagnées de passages visuels pertinents issus de sa base de connaissances.

dans les sections 2 et 3. L'annotation automatique s'est appuyée sur des jeux de données de question-réponse textuels existants ainsi que sur des outils de TAL standards et l'écosystème Wikimédia. Plus précisément, nous avons exploité TriviaQA (Joshi et al., 2017) comme source de questions textuelles (cf. section 1.2). Nous avons ensuite mené une analyse statistique des données. Enfin, nous présentons une base de connaissances fondée sur Wikipédia qui permet de répondre aux questions de ViQuAE avant de clore ce chapitre.

1.1 Limites de KVQA

KVQA est le premier jeu de données proposé pour la KVQAE, donc le seul existant au moment de la collecte de ViQuAE. Bien qu'il comprenne un grand nombre de questions, plus de 183 000, KVQA a deux principales limites :

— un seul type d'entité : *personne*, tandis que la diversité des types d'entités nommées, notamment en raison des représentations visuelles résultantes, est une question centrale dans cette thèse ;

— des questions générées automatiquement à partir de patrons et de Wikidata, ce qui limite leur sujet, leur lexique et leur grammaire.

Ainsi, 44% des questions de KVQA sont partagées quasi-équitablement entre les quatre questions textuelles suivantes (identiques au mot près), dans lesquelles l'image varie pour donner lieu à différentes questions visuelles :

— « *Who is the person in the image ?* »

— « *Was the person in the image born after the end of World War II ?* »

— « *In which country was the person in the image born ?* »

— « *In which continent was the person in the image born ?* »

En excluant les questions booléennes, 68% des questions de KVQA portent ainsi sur *trois* attributs d'entités de type personne : leur identité [1], leur pays ou leur continent

1. L'expression référentielle peut changer, par exemple « *Who is in* the left *?* » et pas seulement « *Who is the person in* the image *?* »

de naissance. Sachant que Shah et al. (2019) ont utilisé 18 relations Wikidata différentes, il est primordial de rééquilibrer KVQA avant de l'utiliser, ce qui réduirait considérablement sa taille. Hormis ce manque de diversité de sujets, nous remarquons évidemment la pauvreté linguistique du jeu de données d'un point de vue lexical ou grammatical. Il existe en effet des dizaines de façon de poser la même question tandis qu'ici, les questions sont toujours restreintes à la même chaîne de caractères. Pour remédier à ce problème, Shah et al. (2019) ont proposé de paraphraser automatiquement ces questions à l'aide de `https://paraphrasing-tool.com/`. Ces paraphrases automatiques ont ensuite été validées par des annotateurs mais même l'exemple cité dans l'article « *Were all the folks in the picture took birth in the same landmass ?* »[2] n'est ni grammaticalement correct, ni naturel. Nous verrons à la section 4 que le vocabulaire de KVQA est particulièrement limité, qu'il s'agisse des questions originales ou paraphrasées.

1.2 Quel jeu de question-réponse utiliser ?

Pour limiter les efforts d'annotation manuelle, nous nous sommes appuyé sur des jeux de données de question-réponse existants. Nous avions besoin d'un jeu de données assez grand, puisque toutes les questions textuelles ne sont pas reformulables en questions visuelles, comme nous le verrons par la suite, ce qui exclut des jeux de données tels que ceux fondés sur les campagnes d'évaluation TREC (Voorhees et Tice, 2000) tels que CuratedTREC (Baudiš et Šedivỳ, 2015). D'autre part, nous ne voulions pas de questions générées automatiquement afin que la langue soit riche et diverse, ce qui exclut des jeux de données comme WebQuestions (Berant et al., 2013) et ComplexWebQuestions (Talmor et Berant, 2018). Notre choix s'est donc limité à SQuAD (Rajpurkar et al., 2016), Natural Questions (Kwiatkowski et al., 2019) et TriviaQA (Joshi et al., 2017). SQuAD a rapidement été écarté car ses annotateurs ont écrit les questions à partir d'un article Wikipédia donné, ce qui favorise la paraphrase et les rend ambiguës en dehors de ce contexte (Joshi et al., 2017; Karpukhin et al., 2020). De plus, ses 100 000 questions portent sur seulement 500 entités différentes, ce qui biaise la distribution et rend le jeu de données inapproprié pour la RI (Lee et al., 2019; Karpukhin et al., 2020). Ce dernier point s'explique car SQuAD a été proposé comme un défi d'extraction de réponse conditionné par le contexte du passage Wikipédia pertinent et pas comme un problème de RI. Enfin, nous avons écarté Natural Questions car, bien que ses questions soient « naturelles », elles ont plusieurs limites. En effet, Natural Questions est issu de requêtes Google ; donc les questions sont souvent mal formées et portent fréquemment sur des entités difficilement représentables visuellement, comme par exemple avec la question « *how many episodes in season 2 breaking bad* ». La forme des questions, y compris le bas de casse systématique, aurait également nui à l'annotation automatique (cf. section suivante). Enfin, beaucoup de questions de Natural Questions sont ambiguës et dépendent du contexte temporel (Min et al., 2020; Izacard et al., 2022), à l'instar de la question « *when will the white house christmas tree be lit* ».

2. Qui paraphrase « *Were all the people in the image born in the same continent ?* »

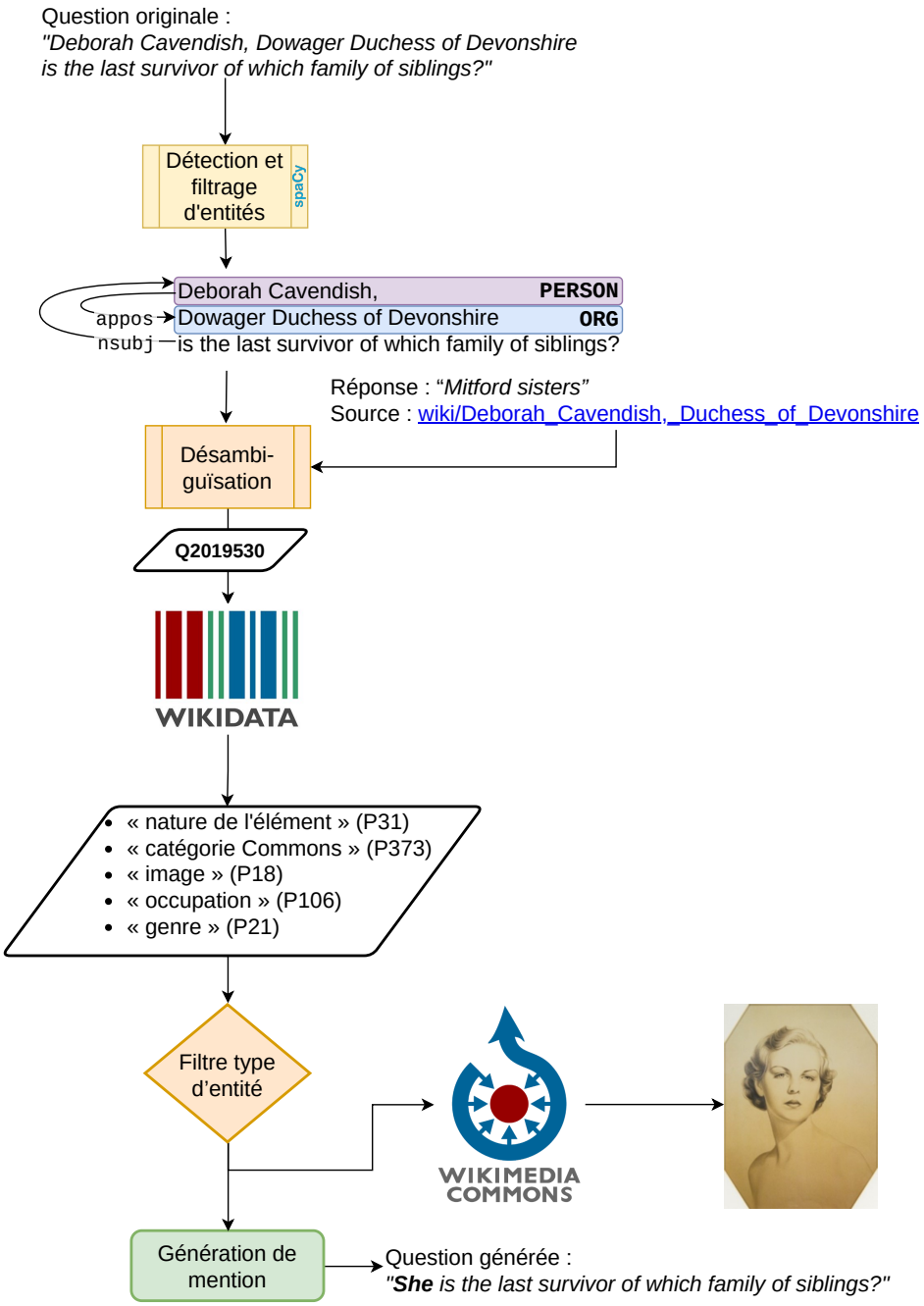


FIGURE 3.2 – Vue d'ensemble de l'annotation automatique. Remarquons que non seulement la mention d'entité (« Deborah Cavendish ») mais aussi ses enfants syntaxiques (« Dowager Duchess of Devonshire ») sont remplacés par la mention ambiguë.

2 Annotation automatique

L'idée principale du processus d'annotation est de remplacer la mention de l'entité dans la question par une représentation visuelle de l'entité. Celle-ci est alors référencée par une mention ambiguë (p. ex. « cet homme »). De cette façon, il n'est

pas possible de répondre à la question sans s'appuyer sur l'image contextuelle. La figure 3.2 illustre le processus d'annotation automatique complet, décrit en détail ci-après.

Données originales Nous utilisons la version KILT de TriviaQA (Petroni et al., 2021). Puisque son jeu de test sert à maintenir une compétition [3], ses annotations (les réponses aux questions) ne sont pas disponibles publiquement. Nous utilisons donc ses jeux d'entraînement et de validation, ce qui correspond au total à 67 000 questions.

Détection et filtrage d'entités nommées Cette détection est faite à l'aide de spaCy [4], qui utilise le schéma d'annotation OntoNotes (Weischedel et al., 2013). Nous mettons de côté les entités nommées qui sortent de notre définition et de nos intérêts, c'est-à-dire les entités numériques telles que les dates, les heures, etc. [5] De plus, nous filtrons les entités nommées selon leur rôle syntaxique dans la question. Nous ne conservons que les mentions ayant un des rôles suivants : *dobj, nsubj, pobj, obj, nsubjpass, poss, obl*. Par exemple, dans la figure 3.2, *Deborah Cavendish* est le sujet nominal (*nsubj*). Par ailleurs, l'analyse syntaxique permet de tronquer les enfants syntaxiques de l'entité, comme dans l'exemple de la figure 3.2. Cette étape produit ainsi 58 000 mentions valides, soit autant de potentielles questions visuelles, sachant qu'il peut y avoir plusieurs mentions valides dans la même question originale.

Désambiguïsation d'entités nommées Les entités nommées ont souvent de nombreux homonymes, ce qui rend leurs mentions ambiguës. Il convient de les désambiguïser en utilisant un système tel que TAGME (Ferragina et Scaiella, 2010). Cette tâche, difficile pour une machine, est ici facilitée car la réponse à la question est connue. Lorsqu'il existe plusieurs entités candidates (homonymes) pour une mention donnée, on suppose que la mention se réfère à l'entité dont l'article Wikipédia contient la réponse. C'est ainsi qu'ont procédé Joshi et al. (2017), car ils avaient initialement conçu TriviaQA pour l'extraction de réponse. Nous avons simplement fait correspondre nos mentions d'entités avec leurs entités désambiguïsées. Pour ce faire, nous calculons le taux d'erreur par mot (*Word Error Rate*, qui est équivalent à une distance de Levenshtein au niveau du mot) entre nos mentions et les noms et alias des entités désambiguïsées par Joshi et al. (2017) pour une question donnée. La mention est alors assignée à l'entité avec le plus faible taux d'erreur, sauf si celui-ci est supérieur à 0,5. 32 000 mentions d'entités sont ainsi désambiguïsées.

Informations structurées et génération de mention ambiguë Wikidata permet de recueillir des informations sur les entités désambiguïsées : leur type, leur profession, leur genre et leur catégorie Commons, laquelle contient plusieurs images ; les autres attributs sont nécessaires pour générer une mention ambiguë. Les personnes sont référencées par leur profession (p. ex. « cet écrivain ») et les autres entités par leur

3. https://eval.ai/web/challenges/challenge-page/689/leaderboard/1907
4. https://spacy.io/ avec le modèle `en_core_web_lg` fondé sur des plongements lexicaux statiques.
5. La liste complète est : *DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, CARDINAL.*

Étape	Nombre de questions
TriviaQA	67 000
Analyse syntaxique et détection d'entités nommées	58 000
Désambiguïsation d'entités nommées	32 000
Filtre type d'entité	14 000
Filtre images	12 000
Annotation manuelle	5 700
Questions validées (ViQuAE)	3 700

TABLEAU 3.1 – Récapitulatif des différentes étapes d'annotation qui permettent de passer de TriviaQA à ViQuAE. Le nombre de questions est arrondi au millier près pour les premières étapes, puis à la centaine près.

type (p. ex. « cette attraction touristique »). De plus, si le genre est disponible, nous utilisons également « *this man/woman* » et « *he-him-his/she-her-hers* » selon la dépendance syntaxique de la mention originale. Il convient de préciser que le genre n'est pas binaire dans Wikidata ; ainsi, les personnes transgenres et cisgenres sont mentionnées de la même façon et en pratique, il n'y a pas de personne identifiée comme intersexuée ou non-binaire dans TriviaQA. Étant donné que certaines entités abstraites, telles que les pays ou les nationalités, sont souvent mentionnées dans les questions mais ne sont pas pertinentes pour la KVQAE, le type d'entité est restreint à une liste de types et de sous-types construite manuellement, disponible avec le jeu de données. Puisqu'il est difficile de trouver un type d'entité dans Wikidata englobant tous les personnages fictifs ou mythologiques, nous utilisons une heuristique supplémentaire : l'entité est valide, peu importe son type, si elle a l'attribut « sexe ou genre » (P21) ou « occupation » (P106) afin de couvrir une grande partie des personnages fictifs ou mythologiques. De plus, pour se conformer à la RGPD [6], et étant donné que beaucoup de questions portent sur des personnes, nous conservons seulement les questions portant sur des personnes décédées qui ne sont pas concernées par la RGPD. Cette étape produit 14 000 questions avec mention ambiguë.

Images Les images sont récupérées à partir de la catégorie Commons de l'entité. Nous verrons par la suite que l'image est une source d'erreur importante. En effet, la catégorie Commons d'une entité liste de nombreuses images liées à celle-ci, de façon plus ou moins distante, mais qui ne la dépeignent pas forcément. Par exemple, dans celle de Barack Obama, on peut voir des cartes des résultats de vote en sa faveur, des manifestations ou des livres [7]. Par conséquent, nous utilisons plusieurs heuristiques pour trier les images par pertinence : tout d'abord, l'image doit être étiquetée comme dépeignant l'entité dans les données structurées de Commons [8] ; ensuite, le nom de l'entité doit être présent dans : (i) le titre de l'image ; (ii) la description de l'image ; (iii) le titre de toutes les catégories Commons auxquelles l'image est assignée. Par

6. https://eur-lex.europa.eu/eli/reg/2016/679/oj
7. https://commons.wikimedia.org/wiki/Category:Barack_Obama
8. https://commons.wikimedia.org/wiki/Commons:Structured_data

Question		Image
Originale (TriviaQA)	Which country's invasion of **Ethiopia** in **1935** forced *Haile Selassie* to flee ?	
Générée	Which country's invasion of **Ethiopia** in **1935** forced *this politician* to flee ?	
Corrigée (ViQuAE)	Which country's invasion forced *this politician* to flee ?	

FIGURE 3.3 – Exemple d'une question reformulée automatiquement puis corrigée manuellement. Parmi ses trois annotateurs, l'un l'a rejetée car il l'a jugée surspécifiée, l'autre l'a au contraire validée sans modification et le dernier l'a corrigée.

ailleurs, nous excluons l'image destinée à servir de référence dans la BC (section 5). Si plusieurs questions portent sur une même entité et qu'assez d'images sont disponibles, chaque question est contextualisée par une image différente. Certaines entités n'ont pas d'image ce qui réduit le nombre de questions visuelles à 12 000. Grâce aux contributeurs de Wikimédia Commons, toutes les images du jeu de données sont soit sous licence libre [9], soit dans le domaine public, ce qui nous permet de les redistribuer pour assurer la reproductibilité de notre travail. Nous décrivons comment filtrer cette annotation automatique dans la section suivante. Les différentes étapes de l'annotation sont résumées dans le tableau 3.1.

3 Annotation manuelle

Motivation L'annotation automatique décrite ci-dessus a des résultats imparfaits. Les deux principales sources d'erreurs sont : (i) l'image sélectionnée, qui peut être inappropriée ; (ii) la trop grande spécificité de la question, qui permet parfois de répondre sans avoir besoin de l'image. Dans le premier cas, l'image peut ne pas dépeindre l'entité correctement, comme expliqué ci-dessus. Elle peut aussi dépeindre plusieurs entités. Dans ce cas, il faut reformuler la question en choisissant une expression référentielle appropriée, par exemple « la personne sur la droite » (cf. chapitre 2). D'autre part, la question peut être surspécifiée : par exemple, à la figure 3.3, on peut voir que la question générée contient des termes discriminants tels qu'*Ethiopia* et *1935*, même si la mention originale de l'entité « Haile Selassie » a été remplacée automatiquement par l'ambiguë « this politician ». Dans certains cas, il peut être impossible de reformuler la question, notamment si la question porte sur une autre entité en plus de celle désambiguïsée, par exemple « *In which year was* ***the Boy Scout movement*** *founded by* `<entity>` » : la question porte principalement sur le mouvement Boy Scout et l'entité désambiguïsée apporte simplement une

9. https://freedomdefined.org/Definition

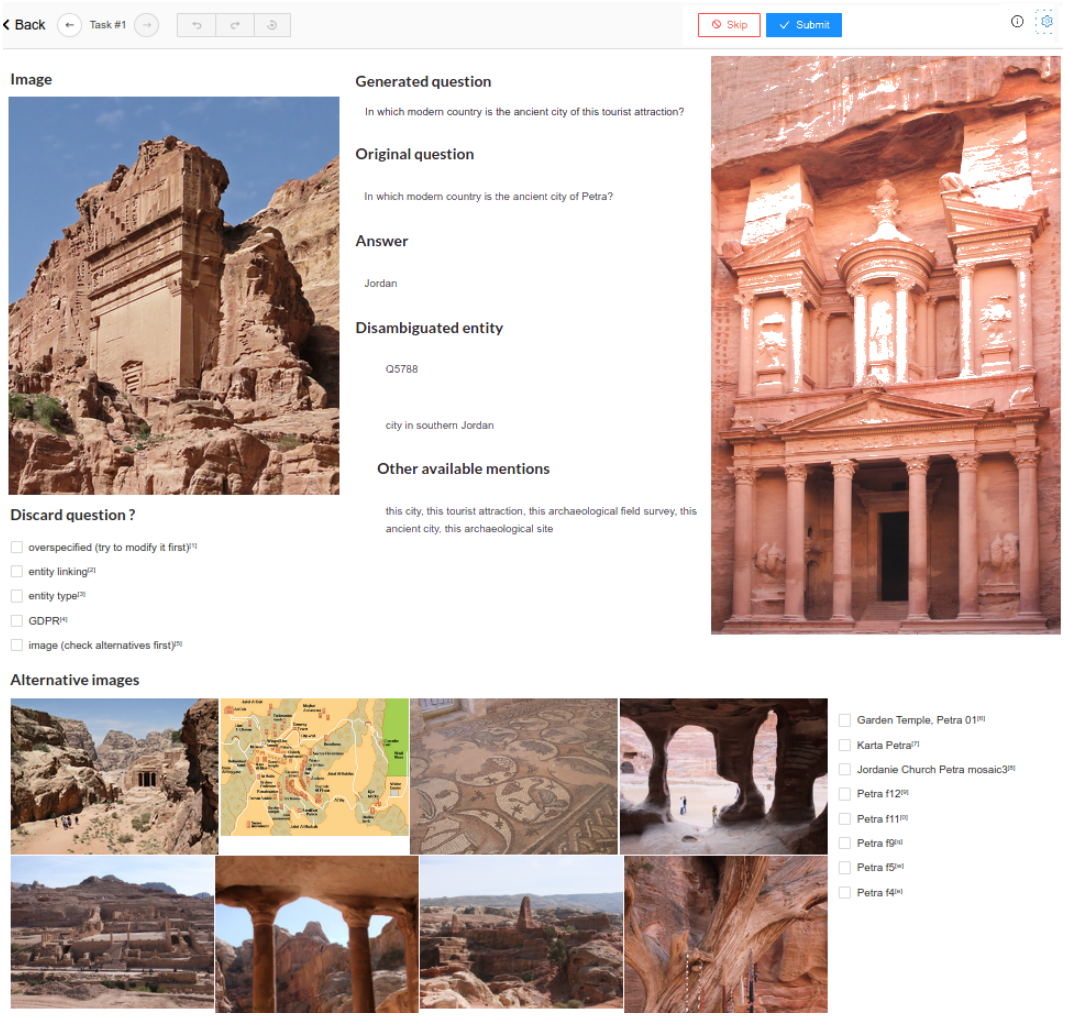


FIGURE 3.4 – Interface d'annotation pour corriger l'annotation automatique de ViQuAE.

information auxiliaire. Ce n'est donc pas possible de reformuler la question.

Interface Pour remédier à ce problème, une interface d'annotation, que l'on peut voir à la figure 3.4, a été conçue à l'aide de Label Studio [10]. L'annotateur peut reformuler librement la question (certaines mentions alternatives sont suggérées) tant que la réponse n'est pas modifiée. Ici, on voit que l'analyse syntaxique a échoué car la question générée est « *In which modern country is* the ancient city of *this tourist attraction* », que l'on peut reformuler en tronquant « *the ancient city of* ». L'annotateur doit également choisir parmi huit images candidates si celle sélectionnée (montrée en haut à gauche) n'est pas appropriée, en utilisant comme référence l'image de référence de la BC (montrée en haut à droite, cf. section 5) tout en s'assurant qu'il ne s'agit pas d'un quasi-doublon. En dernier recours, l'annotateur peut simplement rejeter la question. Les instructions d'annotation sont présentées à l'annexe B.

10. Une plateforme *open source* pour l'annotation (`https://labelstud.io/`)

Implémentation Cette annotation manuelle a été réalisée par sept annotateurs internes, les auteurs de Lerner et al. (2022). L'interface d'annotation permet de traiter environ 120 questions par heure. La proportion de questions à propos de personnes a été équilibrée pour assurer la diversité du jeu de données. Nous avons annoté 5 700 questions générées, parmi lesquelles 2 000 ont été écartées, principalement parce qu'elles étaient surspécifiées ou que l'image n'était pas pertinente. Par conséquent, le jeu de données ViQuAE est constitué de 3 700 questions, réparties aléatoirement en sous-ensembles de 1 190 questions pour l'entraînement, 1 250 questions pour la validation et 1 257 questions pour le test, sans recouvrement entre les images. La majorité (55%) des questions valides ont été éditées par les annotateurs, avec une distance de Levenshtein moyenne de 5 mots entre la version initiale et la version éditée.

Accord inter-annotateur Pour mesurer l'accord inter-annotateur, un sous-ensemble de 103 questions ont été annotées par au moins 3 annotateurs différents. L'accord a ensuite été calculé en utilisant le Kappa de Fleiss (Fleiss, 1971). Les annotateurs se sont mis d'accord pour rejeter ou non la question avec $\kappa = 0{,}33$, montrant un accord léger. Nous remarquons également que les annotateurs sont tous trois d'accord sur 56% des questions, ce qui est plus de deux fois supérieur à l'espérance d'un tirage aléatoire $\frac{1}{2^3} = 25\%$. En effet, déterminer si une question est surspécifiée ou non peut être assez subjectif. Par exemple « *This inner planet and which other planet in our solar system has no moon ?* » est valide *stricto sensu* mais on a une chance sur deux de trouver la bonne réponse sans regarder l'image en choisissant au hasard entre Vénus et Mercure. On peut donc la considérer comme surspécifiée. De plus, la reformulation de certaines questions surspécifiées peut être difficile. Par exemple, la question générée automatiquement de la figure 3.3 a été rejetée par un des trois annotateurs alors qu'il était possible de la reformuler. Enfin, ces désaccords ont été observés une fois l'annotation terminée. Pour arriver à un meilleur accord, nous aurions dû affiner le guide d'annotation itérativement, après discussion entre les annotateurs. Par exemple, certains annotateurs considéraient que les questions du type « *Qui est-ce ?* » étaient invalides. Cependant, il faut rappeler que, dans notre cas, les désaccords entre annotateurs ne concernent pas la *réponse* à la question mais seulement le filtrage du jeu de données généré automatiquement puisque les questions et les réponses sont définies dans TriviaQA et qu'un annotateur *ne peut pas changer la réponse*.

4 Analyse des données

Le jeu de données ViQuAE se compose finalement de 3 700 questions visuelles contextualisées par 3 300 images uniques, dont deux exemples sont présentés à la figure 3.1. La grande majorité des textes des questions sont uniques, ce qui témoigne de la diversité lexicale et syntaxique de ViQuAE. Les questions comportent en moyenne 12 mots, pour un vocabulaire de 4 700 mots. Pour étudier plus précisément leur diversité syntaxique, nous comptons 2 759 séquences de parties du discours (POS) différentes, dont la plus fréquente est « *PRON AUX DET NOUN ADP DET*

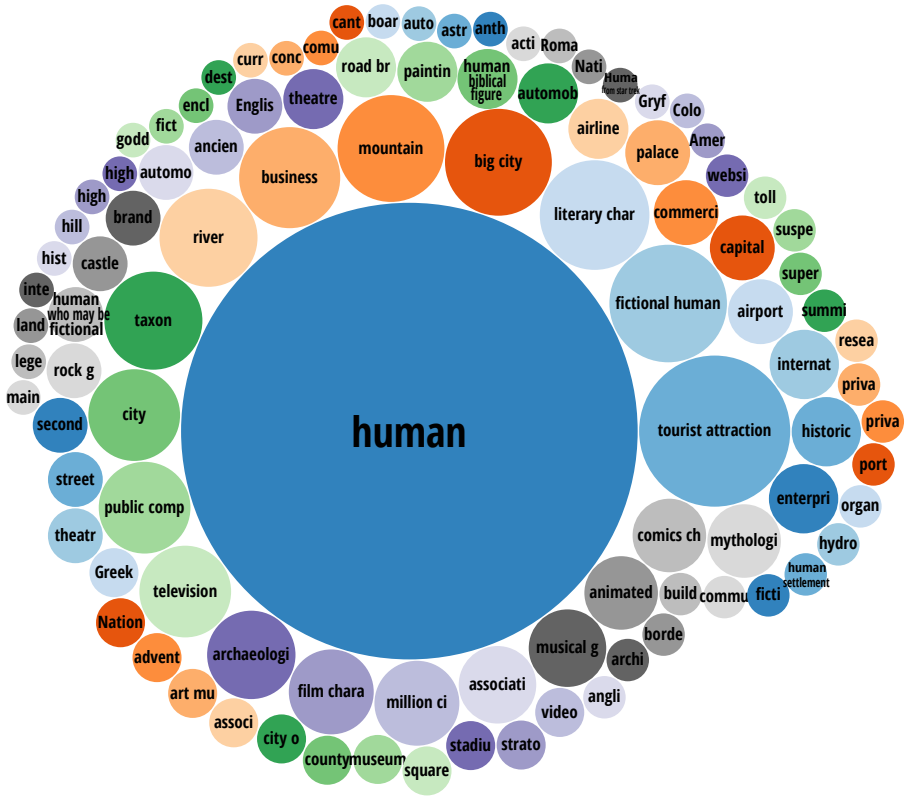


FIGURE 3.5 – Les 100 types d'entités (non-exclusifs) les plus fréquents dans le jeu de données.

NOUN PUNCT ». Sur les 3 700 réponses, les plus fréquentes, « France » et « Turquie », n'apparaissent que 13 fois, soit 0,3% du total, ce qui montre l'absence de biais a priori sur les réponses. De plus, il n'y a qu'un chevauchement de 25% des réponses et de 18% des entités entre les ensembles d'entraînement et de test. Ces trois points soulignent la différence entre la KVQAE et la VQA (fondée sur la connaissance ou pas) et démontrent que traiter la KVQAE comme une tâche de classification serait inefficace.

Une contribution importante du jeu de données est sa diversité de type d'entités, qui est l'un des principaux défis pour les représentations multimodales (cf. chapitre 1). ViQuAE comprend près de mille types d'entités différents (issus de l'ontologie Wikidata) parmi ses 2 400 entités uniques. Toutefois, ces types ne sont pas exclusifs : une moyenne de 1,6 type est attribuée à chaque entité. Le jeu de données comporte 43% de personnes, sans prendre en compte d'autres entités semblables, comme les personnages fictifs ou mythologiques ou des groupes de personnes, par exemple les groupes de musique. La figure 3.5 présente un graphique à bulles des 100 types d'entités les plus fréquents. Par ailleurs, Chen et al. (2023c) estiment que 95,2% des questions de ViQuAE demandent des connaissances externes, c'est-à-dire que les 4,8% restants peuvent être répondues « de tête » par l'annotateur moyen. Un résumé des statistiques comparées avec les jeux de données InfoSeek (Chen et al.,

	ViQuAE	ISM	ISA	EVQA	KVQA
# Questions visuelles	3 700	8 900	1 356 000	1 036 000	183 000
# Questions textuelles uniques	3 562	2 022	1 498	175 000	8 310
# Séquences de POS uniques	2 759	1 056	267	91 945	376
# Questions par image	1,1	1,0	1,4	2,0	7,4
Vocabulaire	4 700	1 307	725	40 787	8 400
Taille moyenne questions	12,4	7,8	8,9	11,6	10,1
Réponse prob. a priori	0,3%	–	0,6%	0,4%	15,9%
Chevauchement réponses	25,3%	–	48,1%	59,6%	89,4%
Chevauchement entités	18,1%	–	20,1%	82,0%	40,6%
# Questions par entité	1,5	11,0	117,6	62,5	9,7
# Types d'entités	980	527	2 739	–	1
Demande connaissances*	95,2%	95,6%	–	–	–

TABLEAU 3.2 – Statistiques de ViQuAE par rapport à InfoSeek (Chen et al., 2023c), Encyclopedic-VQA (EVQA ; Mensink et al., 2023) et KVQA (Shah et al., 2019). InfoSeek est divisé en deux sous-ensembles selon la méthode d'annotation : manuelle (ISM) ou automatique (ISA). *Selon l'étude menée par Chen et al. (2023c) sur un sous-ensemble de 500 questions.

2023c), Encyclopedic-VQA (Mensink et al., 2023) et KVQA [11] (Shah et al., 2019) est rapporté au tableau 3.2. Chen et al. (2023c) et Mensink et al. (2023) ont été pré-publiés très récemment et leurs données ne sont pas encore disponibles, ou seulement partiellement. Nous pouvons constater que, malgré sa petite taille, ViQuAE est plus diversifié sous certains aspects. Notamment, presque chaque question visuelle a une image unique, de même pour les entités nommées. Par conséquent, seul 18% des entités du jeu de test sont présentes dans le jeu d'entraînement.

Cependant, le jeu de données ViQuAE présente aussi certaines limites. L'un des inconvénients de notre processus d'annotation, et plus précisément de la désambiguïsation des entités nommées, est que les réponses sont systématiquement présentes dans la page Wikipédia de l'entité. Ainsi, les questions sont *mono-hop* au niveau du document. Bien sûr, la question peut toujours nécessiter un raisonnement sur plusieurs phrases ou paragraphes du document. En revanche, (Shah et al., 2019) comprend plusieurs questions *multi-hop* qui, même si elles ne semblent pas très naturelles, permettent d'évaluer les capacités de raisonnement du modèle.

Toujours dans la thématique du multi-hop, (Shah et al., 2019) comprend des images avec plusieurs personnes où les questions contiennent alors des expressions référentielles (par exemple « la personne sur la droite »). Dans le cadre de l'annotation automatique, nous avons au contraire visé à avoir une seule entité représentée de manière proéminente par image. Dans le cas où une telle image n'existait pas, l'annotateur a pu utiliser une expression référentielle en reformulant la question mais cela reste marginal par rapport à Shah et al. (2019).

Enfin, nous verrons aux chapitres suivants que ViQuAE a des biais textuels

11. Pour KVQA, il s'agit des statistiques des questions non-paraphrasées mais celles des questions paraphrasées automatiquement sont similaires. Le vocabulaire est notamment aussi pauvre dans les questions paraphrasées, ce qui soulève les problèmes de la paraphrase automatique évoqués à la section 1.1.

assez importants. Ces biais sont en partie inhérents à la KVQAE, voire à toute tâche multimodale tant qu'il reste une part de naturel à la langue, c'est-à-dire à moins de générer des exemples de façon contrôlée (Johnson et al., 2017). Cependant, ces biais sont accentués dans ViQuAE car il s'appuie sur TriviaQA, lequel est scrappé de 14 sites webs de *trivia*, donc de questions destinées à évaluer la culture générale des joueurs humains de manière ludique, comme au Trivial Pursuit. Ces questions comportent donc beaucoup d'éléments superflus destinés à aider et instruire les joueurs humains, comme dans l'exemple de la figure 3.3. Malheureusement, certains de ces éléments n'ont pas été retirés des questions par les annotateurs, comme pour l'exemple de la figure 3.3 où un annotateur a validé la question générée sans la modifier.

5 Base de connaissances

5.1 Collecte

La BC est construite à partir de la sauvegarde du 01/08/2019 de Wikipédia, disponible dans KILT (Petroni et al., 2021) et comprenant 5,9 millions d'articles. Chacun d'eux est associé à une entité Wikidata mais 11 000 de ces entités partagent le même article. Pour arriver à une correspondance 1-1, nous avons appliqué les heuristiques suivantes :

— garder seulement l'article qui permet de répondre à au moins une question de TriviaQA $\implies$ 262 désambiguïsations ;

— enlever les articles qui ont '*disambiguation*' dans leur titre afin de filtrer les pages de désambiguïsations $\implies$ 862 désambiguïsations supplémentaires ;

— garder l'article le plus long, toujours de manière à filtrer les pages de désambiguïsations.

Pour obtenir une représentation visuelle de l'entité, une image unique est extraite de Wikidata, dans l'ordre suivant de préférence des propriétés Wikidata : (i) P18 « image » ; (ii) P154 « image du logotype » ; (iii) P41 « image du drapeau » ; (iv) P94 « image du blason » ; (v) P2425 « ruban de médaille ». Les articles sans image sont écartés, ce qui aboutit à une BC d'1,5 million d'articles, dont 542 000 à propos de personnes, chacun associé à une image. La BC obtenue est donc cent et quinze fois plus grande que celle des expériences de Shah et al. (2019) et Chen et al. (2023c), respectivement. 95% des images de la base de connaissances sont uniques.

5.2 Analyse

Les questions dans ViQuAE sont ancrées dans une image, tout comme les articles de la BC. Une question sur une entité donnée utilise toujours une image différente de celle de la BC. Cependant, d'autres entités de la BC peuvent utiliser la même image qu'une question dans ViQuAE. Par exemple, une question sur *Odin* utilise la même image que *Hugin et Munin* dans la BC, ou une question sur *le pont sur la Severn* utilise la même image que *l'autoroute M48* dans la BC. Sur les 3 300 images de ViQuAE et les 1,4 million d'images de la BC, il y a un chevauchement de 98

images qui correspondent à 108 questions, soit 3% de ViQuAE. Cependant, il ne s'agit pas nécessairement d'un biais qui conduira à des résultats trop optimistes. En effet, seulement 54 de ces 108 questions ont une réponse dans l'article de la BC qui utilise la même image.

6 Conclusion

Nous avons présenté un nouveau jeu de données, ViQuAE, conçu comme un cadre d'évaluation pour suivre le progrès des systèmes de KVQAE. ViQuAE a été annoté selon une procédure semi-automatique que nous fournissons également. En particulier, une validation manuelle de questions générées automatiquement a été effectuée par sept annotateurs au cours d'environ 48 heures de travail au total. La génération automatique, elle, s'appuie sur le jeu de données de question-réponse textuel TriviaQA, ainsi que des outils de TAL standards pour l'analyse syntaxique et la détection et désambiguïsation d'entités nommées, en lien avec l'écosystème Wikimédia. La validation manuelle a néanmoins été nécessaire pour résoudre principalement deux problèmes : d'une part, la pertinence de l'image utilisée pour contextualiser la question visuelle ; d'autre part, la spécificité de la question, à laquelle on ne doit pas pouvoir répondre sans l'image.

On peut répondre aux questions de ViQuAE via une base de connaissances librement disponible d'1,5 million d'articles Wikipédia associés à des images. Nous démontrerons l'utilité pratique de cette BC aux chapitres suivants mais de futurs travaux pourraient y ajouter davantage d'images afin d'aider à capturer la diversité des représentations visuelles d'entités nommées, idéalement en modélisant le lien entre les images et les entités. Nous verrons au chapitre 6 que cet ajout doit se faire en contrôlant que les images ajoutées ne sont pas des quasi-doublons d'images utilisées pour les questions de ViQuAE.

Par rapport au jeu de données existant KVQA (Shah et al., 2019), ViQuAE couvre notamment différents types d'entités et de sujets. Cependant, il ne contient que des questions *mono-hop* au niveau du document et ses images représentent une seule et unique entité, sauf dans certains cas exceptionnels. Nous verrons que cette configuration fournit déjà de nombreux défis mais une future version du jeu de données pourrait introduire des questions *multi-hop* ou plusieurs entités par image.

La petite taille de ViQuAE a largement contraint nos choix de modélisation de la KVQAE. Dans les chapitres suivants, nous verrons d'une part comment exploiter des modèles probabilistes tels que BM25, qui ne nécessitent pas d'entraînement, mais surtout comment exploiter des modèles pré-entraînés, que ce soient des modèles visuels utilisables sans ajustement ou bien des modèles de langue ajustables avec quelques exemples. Nous verrons aussi différentes manières de pré-entraîner ces modèles pour arriver à une modélisation plus riche. Le chapitre suivant traite notamment de l'articulation du système de question-réponse avec la base de connaissances.

Chapitre 4

Recherche d'information et extraction de réponse

1 Introduction

La KVQAE s'apparente à un problème de recherche d'information multimodale, où le texte et l'image de la question visuelle peuvent servir à trouver un passage visuel pertinent. Une fois ce passage pertinent retrouvé, il peut être préférable de proposer une réponse concise à l'utilisateur, pour rendre la RI plus proche d'une recherche d'*information* que de document (Voorhees et Tice, 2000). Nous avons donc choisi de décomposer la tâche en deux étapes, où l'extraction de la réponse fait suite à la RI (cf. figure 4.1). Cette décomposition est standard en question-réponse textuel (Chen et al., 2017). Bien que les gros modèles de langue aient remis en question cette approche (Brown et al., 2020), ils ont certaines limites, comme discuté au chapitre 2. De plus, intégrer l'image dans un gros modèle de langue n'est pas trivial et de premières expériences ont seulement été menées récemment (Tsimpoukelli et al., 2021 ; cf. chapitre 2). Nous voulions au contraire, pour ce premier système, nous appuyer sur des technologies bien établies, telles que la recherche lexicale ou la reconnaissance faciale. Ce premier système se voulait également relativement simple afin d'explorer le jeu de données ViQuAE, introduit au chapitre précédent, et servir de référence pour la suite de cette thèse.

Nous avons donc suivi une approche de fusion tardive pour la RI, où les résultats de deux recherches mono-modales indépendantes, textuelle et visuelle, sont fusionnés au niveau du score. Pour la recherche textuelle, nous avons à la fois expérimenté avec une approche lexicale, BM25, ainsi qu'avec un modèle dense plus moderne, DPR, fondé sur BERT. Pour pré-entraîner ce dernier, étant donné la faible quantité de questions présentes dans ViQuAE, nous proposons d'utiliser TriviaQA, un jeu de question-réponse textuel (Joshi et al., 2017). La recherche visuelle combine quant à elle une reconnaissance faciale avec des représentations d'images plus génériques, obtenues à travers un pré-entraînement sur ImageNet, un jeu de classification d'images supervisée, ou sur un gros jeu de légendes d'images faiblement supervisé (CLIP). Ces représentations visuelles seront appliquées sans ajustement à ViQuAE et sont relatives à notre deuxième question de recherche : comment *représenter visuellement une entité nommée* ?

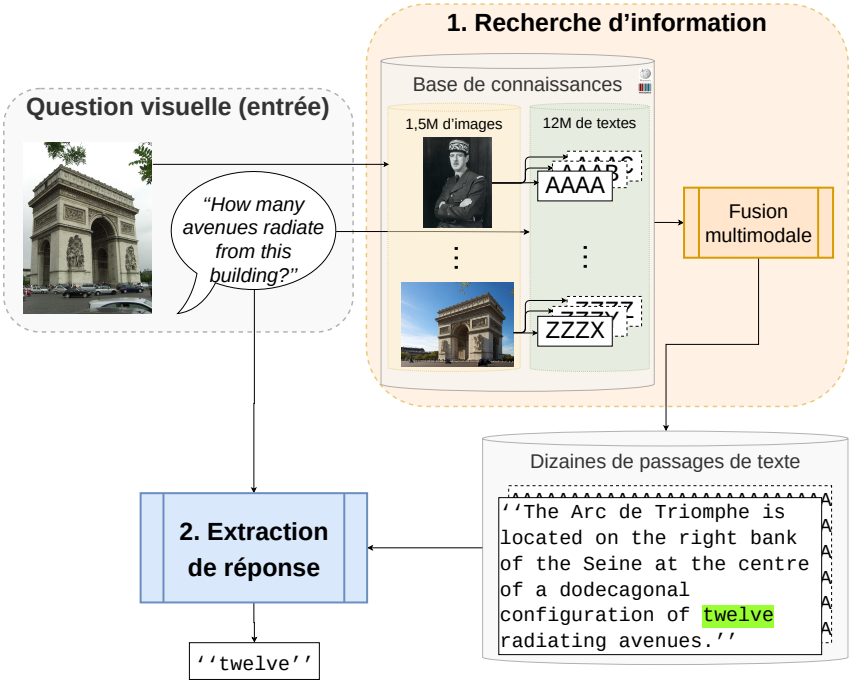


FIGURE 4.1 – Vue d'ensemble du système qui traite la KVQAE en deux étapes : (1) recherche d'information ; (2) extraction de réponse. La RI est faite indépendamment avec le texte et l'image avant de fusionner les résultats. La recherche d'image exploite l'interaction IQIP et la recherche de texte TQTP. Les 1,5 million d'articles appariés à des images sont prétraités en 12 millions de passages de 100 mots. Par conséquent, les résultats de la recherche d'images sont mis en correspondance avec les passages de texte pour permettre la fusion. L'extraction de réponse se fait ensuite à partir des résultats de la RI. On aura donc trois évaluations : au niveau de l'article ou du passage pour la RI puis l'exactitude de la réponse extraite.

Nous présenterons ces méthodes en détail et les résultats expérimentaux à la section 2.

Par la suite, nous proposons d'extraire la réponse des passages résultant de la RI avec BERT multi-passage (Wang et al., 2019). Ce modèle fondé sur BERT permet de traiter plusieurs passages par question en employant la normalisation globale de Clark et Gardner (2018). Ce modèle sera, comme DPR, pré-entraîné sur TriviaQA avant d'être ajusté sur ViQuAE. Ce modèle est purement textuel car nous supposons qu'*une fois le passage pertinent retrouvé*, on peut répondre à la question sans l'image (voir par exemple la figure 4.1). Nous mettrons cette hypothèse à l'épreuve et verrons que la qualité de l'extraction de réponse dépend grandement de la RI. Ainsi, nous étudierons une condition avec une RI *oracle* pour distinguer les erreurs commises par la RI de celles commises par l'extraction de réponse. Nous comparerons également nos approches aux méthodes de Adjali et al. (2023) et Chen et al. (2023c). Ce chapitre traite donc de notre première question de recherche : comment *évaluer* un système de KVQAE ?

Enfin, nous conclurons ce chapitre en résumant nos résultats principaux et nos perspectives.

2 Recherche d'information

Dans cette partie, nous nous intéressons à retrouver des passages visuels pertinents parmi les 1,5 million d'articles appariés à des images qui constituent la BC introduite au chapitre 3. Nous traitons la multimodalité de la tâche en décomposant la RI en deux recherches mono-modales :

— une recherche de texte, qui exploite l'interaction TQTP entre le texte de la question et du passage ;

— une recherche d'image, qui exploite l'interaction IQIP entre l'image de la question et du passage.

Notons que les autres interactions multimodales sont ici négligées et seront étudiées aux chapitres suivants. Les résultats de ces deux recherches sont ensuite fusionnés au niveau du score.

2.1 Recherche de texte

2.1.1 Prétraitement

Notre BC est donc constituée ici de 1,5 millions d'articles textuels. Les données semi-structurées, comme les tableaux et les listes, y sont filtrées, d'après Karpukhin et al. (2020). De plus, afin de traiter les articles avec des modèles de langue pré-entraînés, ici DPR et à la section suivante BERT multi-passage, ces articles sont divisés en passages disjoints de 100 mots tout en préservant les limites des phrases, ce qui produit 12 millions de passages (environ 8 passages par article). Le titre de l'article est concaténé au début de chaque passage en le séparant avec le token spécial [SEP], toujours d'après Karpukhin et al. (2020).

2.1.2 BM25

BM25, modèle probabiliste introduit par Robertson et al. (1995) et dérivé de TF-IDF (Sparck Jones, 1972), est, malgré son ancienneté, toujours une référence en RI (Thakur et al., 2021). De plus, puisqu'il repose sur le chevauchement lexical entre le texte de la question et du passage, il est très efficient car on peut calculer la similarité entre la question et tous les passages de la BC à travers un index inversé, en tenant compte donc *seulement des passages ayant un chevauchement lexical avec la question*. Formellement, le score de similarité entre une question $\mathbf{t_q}$ et un passage $\mathbf{t_p}$ se décompose comme une somme des scores de chaque terme $\mathbf{t_q}^{(j)}$ de la question :

$$\mathrm{BM25}(\mathbf{t_q}, \mathbf{t_p}) = \sum_{j=1}^{n} \mathrm{IDF}(\mathbf{t_q}^{(j)}) \frac{f(\mathbf{t_q}^{(j)}, \mathbf{t_p})(k_1 + 1)}{f(\mathbf{t_q}^{(j)}, \mathbf{t_p}) + k_1 \left(1 - b + b\frac{L}{\mu}\right)} \tag{4.1}$$

où IDF dénote la fréquence inverse en documents, f la fréquence du terme $\mathbf{t_q}^{(j)}$ dans le passage $\mathbf{t_p}$, L la longueur du passage, μ la longueur moyenne des passages et b et k_1, des hyperparamètres qui pondèrent l'importance de la longueur du passage et la saturation de la fréquence du terme, respectivement. Dans notre cas où tous les passages ont la même longueur, on peut s'attendre à ce que b ne soit pas important.

Les hyperparamètres sont déterminés sur le jeu de validation à travers une recherche exhaustive (*grid search*) pour maximiser le MRR. À part cela, ce modèle ne nécessite pas d'entraînement.

Avec BM25, les problématiques du fossé sémantique se retrouvent explicitement puisque :

— en cas de synonymie, $f(\mathbf{t_q}^{(j)}, \mathbf{t_p}) = 0$;

— au contraire, en cas de polysémie $f(\mathbf{t_q}^{(j)}, \mathbf{t_p})$ a la même valeur pour tous les sens de $\mathbf{t_q}^{(j)}$.

2.1.3 DPR

DPR a été introduit par Karpukhin et al. (2020) et a permis de surpasser BM25 pour la RI de question-réponse. Nous en avons déjà largement discuté au chapitre 2 mais nous rappelons brièvement qu'il est paramétré par deux encodeurs BERT [1], pour la question et le passage, que la représentation est obtenue à travers le token spécial [CLS] et que la similarité est calculée avec un produit scalaire :

$$\mathrm{DPR}(\mathbf{t_q}, \mathbf{t_p}) = \mathrm{BERT}_q(\mathbf{t_q})_{[\mathrm{CLS}]} \cdot \mathrm{BERT}_p(\mathbf{t_p})_{[\mathrm{CLS}]} \tag{4.2}$$

Puisque DPR s'appuie sur des représentations denses, contrairement à BM25, il est nécessaire de calculer ce produit avec chaque passage de la BC pour connaître la similarité exacte entre la question et les passages de la BC. Cependant, des techniques de recherche approximative du produit scalaire maximal (MIPS [2]) existent (Johnson et al., 2019) et, surtout, la représentation du passage peut être pré-calculée pour les 12 millions de passages de la BC. Les représentations contextuelles de BERT permettent en théorie de combler le fossé sémantique et, empiriquement, DPR surpasse souvent BM25.

Cependant, ce modèle doit être entraîné, malgré le pré-entraînement de BERT. Pour ce faire, nous exploitons le jeu de données TriviaQA, ViQuAE étant trop petit pour entraîner DPR [3]. Puisque certaines questions de TriviaQA ont été utilisées pour l'annotation de ViQuAE (cf. chapitre précédent), nous prenons garde de les filtrer. De plus, rappelons que le jeu de test de TriviaQA est privé. Parmi les 67 000 questions qui constituent ses jeux de validation et d'entraînement (dans la version KILT), nous en conservons donc 47 000 pour l'entraînement et 1 234 (resp. 1 247) pour la validation (resp. le test), qui correspondent aux questions utilisées pour le jeu de validation (resp. de test) de ViQuAE. Pour ce pré-entraînement, nous utilisons la BC de KILT complète, avant le filtre des articles sans image, soit 5,9 millions d'articles ou 32 millions de passages (prétraités comme précédemment). Enfin, pour l'entraînement contrastif de DPR, les exemples négatifs difficiles sont sélectionnés avec BM25, pour le pré-entraînement comme pour l'ajustement.

L'ajustement sur ViQuAE est effectué sur le texte de ses 1 190 questions du jeu d'entraînement, en conjonction avec les 12 millions de passages de notre BC.

1. Dans nos expériences, nous utilisons `bert-base-uncased` disponible dans la bibliothèque Transformers.
2. *Maximum Inner-Product Search*
3. Ce sera démontré empiriquement au chapitre suivant.

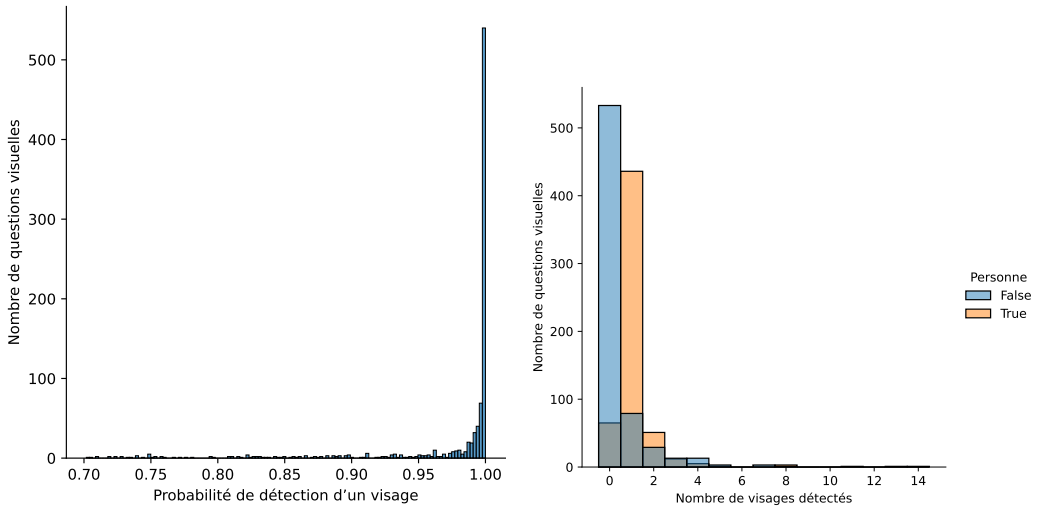


FIGURE 4.2 – Histogramme des probabilités de détection de visage (gauche) et du nombre de visage détecté selon le type d'entité (droite), sur l'ensemble de validation de ViQuAE.

Nous utilisons les mêmes hyperparamètres que Karpukhin et al. (2020), notamment 1 passage positif et 1 négatif difficile par question par lot, d'une taille totale de 128 questions (donc 256 passages) partagé sur 4 GPUs Nvidia V100. L'optimisation est faite avec Adam (Kingma et Ba, 2015) avec un taux d'apprentissage de 2×10^{-5} croissant linéairement pendant 4 époques puis décroissant pendant 40 époques, si l'entraînement n'est pas interrompu avant, ce qui est déterminé selon le MRR *au sein du lot* sur le jeu de validation. *Au sein du lot* signifie que l'on ordonne seulement les passages du lot, et pas de toute la BC, et que chaque question a un seul et unique passage pertinent (comme pour l'entraînement).

2.2 Recherche d'image

Nous avons évoqué au chapitre 2 différentes manières de représenter des images selon qu'il s'agit de l'image d'une personne, que l'on peut représenter à travers son visage, ou d'un autre type d'entité.

Nous avons donc naturellement utilisé ces différents types de représentations, en divisant la BC entre personnes et non-personnes. Les représentations de visage ou d'image complète sont alors utilisées de façon alternative selon qu'un visage est détecté ou non dans l'image de la question $\mathbf{i_q}$. Pour comparer plus équitablement ces différentes représentations, elles sont toutes obtenues à travers un réseau de convolution ResNet-50 (He et al., 2016), bien que l'architecture soit légèrement modifiée selon les cas.

2.2.1 Reconnaissance faciale

La reconnaissance faciale se fait en deux étapes : détection des visages puis représentation et recherche des visages les plus proches. Cette seconde étape est en cela proche de la RI textuelle ou de la recherche d'image générique.

Pour représenter les visages, nous utilisons ArcFace, proposé par Deng et al. (2019), qui a établi un état de l'art sur plusieurs *benchmarks*. Plutôt que d'imposer une marge entre les exemples dans une fonction de coût à triplet (Schroff et al., 2015), ArcFace impose une marge autour de la classe. À l'inférence, les représentations des visages sont comparées via une similarité cosinus. ArcFace est entraîné sur MS-Celeb (Guo et al., 2016), donc avec 100 000 classes (entités nommées). Ces entités ont un certain chevauchement avec ViQuAE, qui est analysé à la section 2.4. ArcFace est utilisé en conjonction avec le détecteur de visages MTCNN (Zhang et al., 2016). Les régions de l'image où sont détectés les visages sont prétraitées de telle façon que les 5 points de repère du visage (les yeux, le nez et les coins de la bouche) soient toujours à la même position, quelle que soit la pose originale de la personne.

On peut voir à la figure 4.2 (gauche) que les probabilités de détection se concentrent autour de la probabilité maximale. Nous conservons donc le seuil par défaut de MTCNN de 0,7. Si plusieurs visages sont détectés, seul celui associé à la plus forte probabilité est conservé. Cette stratégie peut paraître naïve mais on peut voir à la figure 4.2 (droite) qu'il n'y a souvent qu'un seul ou aucun visage détecté (moyenne $0{,}75 \pm 1{,}14$, premier quartile 0, médiane 1 et troisième quartile 1). 6,6% des personnes de la BC n'ont pas de visage détecté et ont donc été écartées.

2.2.2 Représentation d'image complète

Nous combinons deux représentations différentes pour l'image complète, que nous supposons assez génériques pour être appliquées à tout type d'entité. Premièrement, celle extraite de la dernière couche convolutive d'un ResNet entraîné à la classification d'images sur ImageNet. Ces représentations sont efficaces pour la recherche d'image par le contenu (Sharif Razavian et al., 2014) et celles de ResNet ont plus précisément été appliquées pour représenter des monuments (Radenović et al., 2018), un type d'entité qui nous intéresse particulièrement. Nous utilisons le *max-pooling* pour réduire la carte de caractéristiques (*feature map*), compte tenu des résultats rapportés dans (Radenović et al., 2018). D'autre part, nous exploitons une méthode plus récente, CLIP, qui est entraîné à faire correspondre une image et sa légende de façon contrastive. Cette faible supervision a permis d'entraîner le modèle sur plus de 400 millions de paires (texte, image). Son utilisation a été motivée par ses bons résultats pour la classification d'image et la recherche cross-modale (Radford et al., 2021) mais nous sommes les premiers à rapporter les résultats de son utilisation pour la recherche d'image par le contenu (mono-modale)[4]. CLIP et ImageNet-ResNet utilisent tous deux la similarité cosinus comme fonction de similarité.

2.3 Fusion tardive

Avant de pouvoir fusionner les résultats des recherches textuelles et visuelles, il est nécessaire de mettre en correspondance la recherche visuelle, effectuée au niveau de l'article (1,5 million dans la BC), et la recherche textuelle, effectuée au niveau du

4. Encore aujourd'hui, seulement quelques publications autour de la compétition GUIE existent à ce sujet (Conde et al., 2022; Huang et Li, 2022).

passage (environ 8 par article). Ainsi, la recherche visuelle n'est pas évaluée seule au niveau du passage mais seulement en combinaison avec la recherche textuelle (ou seule au niveau de l'article à titre indicatif).

Les résultats sont ensuite fusionnés au niveau du score par le biais d'une interpolation linéaire. Le score doit ainsi permettre de s'appuyer plus ou moins sur une modalité. Par exemple, pour une question textuelle contenant peu d'informations, comme « *Qui est-ce ?* », le terme IDF dans BM25 devrait donner des scores faibles à tous les passages, ce qui devrait conduire à s'appuyer davantage sur la recherche visuelle.

Formellement, en notant s_i les différentes fonctions de similarité et α_i les paramètres de l'interpolation linéaire, avec $i = A$, V et R pour respectivement ArcFace, CLIP et ImageNet-ResNet, $F \in \{0, 1\}$, pour indiquer la détection d'un visage dans $\mathbf{i_q}$ et $\mathbf{i_p}$ et $H \in \{0, 1\}$ si $\mathbf{i_p}$ correspond à une personne [5], nous définissons d'abord la similarité image s_I puis la similarité globale s de la façon suivante :

$$s_I(\mathbf{i_q}, \mathbf{i_p}) = FH\alpha_A s_A(\mathbf{i_q}, \mathbf{i_p}) + (1-F)(1-H)(\alpha_V s_V(\mathbf{i_q}, \mathbf{i_p}) + \alpha_R s_R(\mathbf{i_q}, \mathbf{i_p})) \quad (4.3)$$

$$s(\mathbf{t_q}, \mathbf{t_p}, \mathbf{i_q}, \mathbf{i_p}) = \alpha_B \mathrm{BM25}(\mathbf{t_q}, \mathbf{t_p}) + \alpha_D \mathrm{DPR}(\mathbf{t_q}, \mathbf{t_p}) + s_I(\mathbf{i_q}, \mathbf{i_p}) \quad (4.4)$$

Encore une fois, la confiance faite au détecteur de visage MTCNN (à travers F) peut paraître naïve mais nous trouvons sur le jeu de validation que ce détecteur prédit si l'entité-sujet est de type personne avec un rappel de 89% (précision 78%, score F1 83%).

Toutefois, ces systèmes fournissent des scores qui ne sont pas comparables : BM25 n'est pas borné, DPR non plus puisque les représentations ne sont pas normalisées avant le produit scalaire, contrairement à celles de la recherche visuelle. Par conséquent, les scores sont centrés-réduits avant la fusion. De plus, quand un passage n'est pas retrouvé par un système donné (mais par les autres, puisqu'on considère toujours le top-K d'un système), on lui assigne le score minimal des autres résultats de ce système (Ma et al., 2021a). Cette technique est essentielle en conjonction avec des scores centrés-réduits. Autrement, un passage classé à un rang supérieur à K aurait effectivement un score nul, donc moyen plutôt que mauvais. Les paramètres qui pondèrent l'interpolation linéaire sont déterminés sur le jeu de validation pour maximiser le MRR via une recherche exhaustive tout en contraignant leur somme à 1. De plus, nous utilisons une seule recherche textuelle à la fois, on a donc $\alpha_B = 0$ ou $\alpha_D = 0$.

2.4 Résultats

Jeu de données Nous n'avons pas expérimenté notre approche sur KVQA (Shah et al., 2019) puisque ce jeu de données ayant été généré automatiquement à partir de Wikidata, rien ne garantit que les réponses se trouvent dans notre BC [6]. De plus, il comprend 29% de questions booléennes (réponse oui/non) pour lesquelles nous ne pouvons pas évaluer la pertinence du passage automatiquement, ce qui est utile pour

5. On connaît seulement le type d'entité des images $\mathbf{i_p}$ de la BC, *pas* celles des questions $\mathbf{i_q}$.

6. Nous pouvons grossièrement estimer que 37% des questions (hors booléennes) de KVQA n'ont pas de réponse dans notre BC en vérifiant si la réponse est incluse dans l'article de l'entité-sujet.

#	Modèle	MRR@100	P@1	P@20	Hits@20
a	BM25 (texte seulement)	19,0	13,1	5,9	39,5
b	DPR non-ajusté (texte seulement)	30,5^{a}	21,2^{a}	16,2^{ac}	60,5^{ac}
c	BM25 + recherche visuelle	27,9^{a}	20,4^{a}	10,1^{a}	50,5^{a}
d	DPR non-ajusté + recherche visuelle	36,0^{abce}	26,7^{abce}	17,1^{ac}	65,2^{abce}
e	DPR ajusté (texte seulement)	32,8^{abc}	22,8^{a}	16,4^{ac}	61,2^{ac}
f	DPR ajusté + recherche visuelle	37,9^{abcde}	27,8^{abce}	**17,5**^{ac}	**65,7**^{abce}
*	DPR ajusté + KG + recherche visuelle	**38,3**	**29,0**	17,1	64,4

TABLEAU 4.1 – Résultats de la RI évaluée au niveau du passage avec les références textuelles et la fusion de la recherche multimodale, avec ou sans ajustement de DPR sur ViQuAE. Dans tous les cas, DPR est d'abord pré-entraîné sur TriviaQA. La recherche visuelle combine ArcFace, CLIP et ImageNet-ResNet. Les exposants dénotent des différences significatives selon le test de randomisation de Fisher avec $p \leq 0,01$. *Résultats rapportés par Adjali et al. (2023) en ajoutant des connaissances structurées (notées KG) pour lesquels nous ne calculons pas la significativité des différences.

entraîner DPR mais aussi pour l'évaluation intrinsèque de la RI. Nous présentons donc les résultats sur ViQuAE.

Métriques Puisqu'il est fondé sur TriviaQA (Joshi et al., 2017), ViQuAE n'est supervisé que de façon distante, c'est-à-dire qu'un passage est jugé pertinent s'il contient la réponse [7] après un prétraitement standard (insensibilité à la casse, aux déterminants et à la ponctuation).

Étant donné les K premiers passages retrouvés par un système pour une question visuelle, nous les évaluons comme suit, avec $R@K = |\text{résultats@}K \cap \text{pertinents}|$:

— $\text{Hits@}K = \min(1, R@K)$;

— $\text{Précision@}K (P@K) = \frac{R@K}{K}$;

— $\text{MRR@}K = \frac{1}{\text{rang}}$, le rang réciproque moyen (MRR [8]), où rang est le rang du *premier* document pertinent retrouvé pour la requête ou $\text{rang} = \infty$ si aucun document pertinent n'est dans $\text{résultats@}K$.

Il est à noter que ces métriques sont toutes équivalentes pour $K = 1$. Elles sont moyennées pour toutes les questions visuelles du jeu de test, ce dont témoigne le *moyen* dans rang réciproque *moyen*, qui est éludé pour les autres métriques. Hits@K nous donne une indication du rappel [9], que nous ne pouvons pas calculer faute d'estimer le nombre total de passages pertinents par question. Les tests de significativité statistique sont effectués à l'aide du test de randomisation de Fisher (Fisher, 1937; Smucker et al., 2007).

DPR ou BM25 ? Les résultats principaux sont rapportés au tableau 4.1. On voit que la fusion multimodale apporte des améliorations importantes et significatives

7. Ou plutôt *les* réponses car les alias Wikipédia d'une entité constituent une réponse valide.
8. *Mean Reciprocal Rank*
9. $\text{rappel@}K = \frac{R@K}{|\text{pertinents}|}$

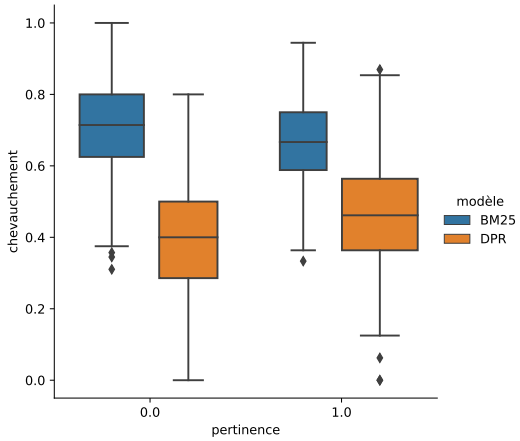


FIGURE 4.3 – Chevauchement entre les lemmes de la question et du premier passage retourné par BM25 et DPR (non-ajusté) en fonction de la pertinence du passage. Chaque boîte montre les quartiles tandis que ses moustaches s'étendent pour montrer le reste de la distribution, à l'exception des valeurs extrêmes. La largeur de la boîte est proportionnelle au nombre de passages concernés, ce qui rappelle que DPR trouve davantage de passages pertinents que BM25.

Sous-ensemble	#	Modèle	MRR@100	P@1	P@20	Hits@20
Visage détecté	-	ArcFace	54,3	50,2	5,5	65,3
Pas de visage détecté	a	ImageNet	17,5	11,9	4,9	36,1
	b	CLIP	**27,5**^{a}	**20,5**^{a}	**9,5**^{a}	**53,1**^{a}

TABLEAU 4.2 – Résultats de la RI évaluée au niveau de l'article sur deux sous-ensembles selon la détection d'un visage.

bien que les références textuelles soient assez fortes. Nous distinguons deux versions de DPR : avec ou sans ajustement sur ViQuAE (mais toujours pré-entraîné sur TriviaQA). Même la version sans ajustement surpasse BM25, y compris la fusion multimodale fondée sur BM25, bien que ce ne soit pas significatif pour le MRR et la P@1. DPR comble le fossé sémantique auquel est soumis BM25 grâce aux représentations contextuelles de BERT. Il est alors capable de trouver des passages pertinents, même s'ils ont peu de chevauchement lexical avec la question, comme on le voit à la figure 4.3. La méthode de Adjali et al. (2023), fondée sur la fusion tardive de DPR et de la recherche visuelle, mais qui ajoute des connaissances structurées dans la représentation des passages de DPR, améliore légèrement le MRR et la P@1. D'autre part, la relative bonne performance de BM25 et de DPR, qui utilisent seulement le texte, peut partiellement s'expliquer par les biais textuels de ViQuAE, dont nous discutons davantage dans les sections suivantes. Ces biais peuvent être accentués par DPR, qui est capable de capturer les biais de la distribution du jeu d'entraînement.

Recherche visuelle Les performances des modèles purement textuels ne doivent pas conduire à négliger l'apport de l'image dans la KVQAE, comme en témoignent

#	Modèle	MRR	P@1	P@20	Hits@20
a	BM25 (texte seulement)	19,8	14,4	6,1	37,6
b	DPR non-ajusté (texte seulement)	28,0^{a}	19,2^{a}	14,4^{a}	57,9^{a}
c	BM25 + recherche visuelle	32,4^{a}	24,4^{a}	11,9^{a}	56,0^{a}
d	DPR non-ajusté + recherche visuelle	37,9^{abce}	28,9^{abe}	17,4^{abce}	67,4^{abce}
e	DPR ajusté (texte seulement)	31,1^{ab}	21,7^{a}	15,2^{ac}	57,5^{a}
f	DPR ajusté + recherche visuelle	**40,4**^{abce}	**29,8**^{abce}	**18,4**^{abcde}	**67,8**^{abce}
a	BM25 (texte seulement)	18,3	12,1	5,8	41,0
b	DPR non-ajusté (texte seulement)	32,7^{ac}	22,9^{ac}	**17,7**^{ac}	62,6^{ac}
c	BM25 + recherche visuelle	24,1^{a}	17,1^{a}	8,5^{a}	45,9^{a}
d	DPR non-ajusté + recherche visuelle	34,3^{ac}	24,7^{ac}	16,9^{ac}	63,4^{ac}
e	DPR ajusté (texte seulement)	34,1^{ac}	23,8^{ac}	17,4^{ac}	**64,3**^{ac}
f	DPR ajusté + recherche visuelle	**35,7**^{abc}	**26,0**^{ac}	16,8^{ac}	64,0^{ac}

TABLEAU 4.3 – Résultats de la RI évaluée au niveau du passage pour les questions à propos de personnes (partie supérieure) et de non-personnes (partie inférieure). La recherche visuelle combine ArcFace, CLIP et ImageNet-ResNet.

les résultats de la fusion multimodale, qui améliore significativement les résultats de la recherche textuelle. L'évaluation directe de la recherche visuelle se fait quant à elle au niveau de l'article (cf. tableau 4.2). Remarquons notamment qu'ArcFace est très précis mais a un rappel relativement faible (Hits@K croît lentement avec K), contrairement à DPR qui peut facilement accroître Hits@K avec K : pour reprendre l'exemple de la figure 3.1, « *Which constituency did this man represent when he was Prime Minister ?* », un modèle textuel peut proposer K circonscriptions différentes (une par passage), tandis qu'un modèle visuel aura une chance infime qu'une personne ressemblant à Harold Macmillan ait été également élue à *Bromley*. Par ailleurs, CLIP surpasse largement ImageNet-ResNet quand aucun visage n'est détecté. Ces résultats concordent avec ceux de Radford et al. (2021) et pourraient motiver de futurs travaux sur l'utilisation de CLIP pour la recherche d'image par le contenu.

Fusion multimodale Enfin, nous étudions l'apport de la fusion multimodale selon le type de l'entité-sujet. On voit au tableau 4.3 que, pour les questions à propos de personnes, la P@1 passe de 14,4 avec BM25 seul à 24,4 en fusionnant BM25 et la recherche visuelle, soit une amélioration de 70%. En comparaison, l'amélioration est plus faible, seulement 41%, en termes de P@1 pour les questions sur les non-personnes. Cela s'explique par la bonne performance d'ArcFace, qui est fondé sur un grand jeu de données annoté, MS-Celeb. Ceci est d'autant plus vrai qu'environ un quart des questions de ViQuAE portent sur une personne présente dans MS-Celeb et, sur cette intersection, lorsqu'un visage est détecté, la P@1 d'ArcFace au niveau de l'article atteint 68,3% (contre 50,2 pour tous les visages ; cf. tableau 4.2) [10]. Plus généralement, les représentations visuelles des personnes, par leur visage, sont

10. Il ne faudrait pas surinterpréter cette différence de performance. Le chevauchement entre ViQuAE et MS-Celeb cache également la date de naissance des personnes présentes dans ce jeu de données, elle-même corrélée avec le format de l'image (tableau, statue, photographie ancienne ou

intrinsèquement mieux définies que celles d'autres entités, qui sont plus hétérogènes et abstraites.

3 Extraction de réponse

3.1 Méthodes

Une fois la RI effectuée, on cherche à extraire une réponse concise du passage pertinent. Pour cela, nous exploitons un modèle de l'état de l'art, qui prend en entrée la concaténation du texte de la question et du passage. En effet, nous supposons qu'*une fois le passage pertinent retrouvé*, l'image n'est plus nécessaire pour répondre à la question (voir par exemple la figure 3.1). Nous verrons à la section suivante que cette hypothèse peut être nuancée lorsque le module d'extraction de réponse prend en entrée plusieurs passages candidats, dont des non-pertinents qui peuvent le distraire.

L'extraction de réponse est faite par BERT multi-passage (Wang et al., 2019), un modèle d'extraction de réponse de l'état de l'art (Devlin et al., 2019), fondé sur BERT, à la subtilité près que la normalisation softmax est *globale* (Clark et Gardner, 2018), c'est-à-dire que tous les passages liés à une question donnée partagent la même normalisation softmax afin que les scores soient comparables à travers les passages, bien qu'ils soient encodés indépendamment. Plus précisément, le modèle est entraîné à prédire la position de début et de fin de la réponse dans le passage, ce qui est paramétré par deux perceptrons indépendants. Si le passage n'est pas pertinent (s'il ne contient pas la réponse), le modèle doit prédire la position 0, celle associée au token spécial [CLS]. À l'inférence, la probabilité que la réponse se trouve entre les tokens i et j est le produit des probabilités de début à i et de fin à j. Nous expérimenterons également la pondération de cette probabilité par le score de la RI afin de tenir compte du score du passage. Formellement, pour prédire le début de la réponse (processus analogue pour la fin de la réponse), on a :

$$\forall k \in [\![1, K]\!], a_k = \mathrm{BERT}([\mathbf{t_q}; \mathbf{t_p}^{(k)}]) \cdot \mathbf{W} \tag{4.5}$$

avec K, le nombre de passages en entrée (24 dans nos expériences) et $\mathbf{W} \in \mathbb{R}^d$ donc $a_k \in \mathbb{R}^L$ où d dénote la dimension du modèle et L la longueur des passages. $\mathbf{W}$ et BERT sont entraînés conjointement à minimiser la log-vraisemblance négative du début de la réponse dans les K passages. La réponse peut apparaître jusqu'à R fois dans un passage ($R = 10$ dans toutes nos expériences, d'après Karpukhin et al.).

$$-\sum_{r=1}^{R}\sum_{k=1}^{K}\sum_{l=1}^{L} y_{rkl} \log \frac{\exp\left(a_{kl}\right)}{\sum_{k'=1}^{K}\sum_{l'=1}^{L}\exp\left(a_{k'l'}\right)} \tag{4.6}$$

où $y_{rkl} \in \{0, 1\}$ dénote la vérité terrain. La normalisation globale (Clark et Gardner, 2018) apparaît au dénominateur où l'on somme sur les K passages. Toutefois, nous

moderne). L'année de naissance médiane pour les personnes de ViQuAE est 1909 si ces personnes sont présentes dans MS-Celeb, 1812 sinon.

avons en premier lieu implémenté la fonction objectif suivante, d'après (Karpukhin et al., 2020) :

$$-\sum_{r=1}^{R}\max_{k=1}^{K}\sum_{l=1}^{L} y_{rkl}\log\frac{\exp\left(a_{kl}\right)}{\sum_{k'=1}^{K}\sum_{l'=1}^{L}\exp\left(a_{k'l'}\right)} \tag{4.7}$$

Les raisons qui ont poussé Karpukhin et al. (2020) à implémenter ce *max-pooling* ne sont pas claires [11] mais nous avons par la suite trouvé cette méthode moins efficace que l'équation 4.6. C'est pourquoi les résultats présentés à la section suivante sont différents de ceux rapportés dans les articles publiés à SIGIR, TALN, ECIR, TAL et CORIA, bien que les conclusions restent inchangées.

Comme pour DPR, le modèle est d'abord pré-entraîné sur TriviaQA avant d'être ajusté sur ViQuAE. Pour TriviaQA, la BC KILT de 32 millions de passages fournit les passages en conjonction avec BM25. Pour ViQuAE, la BC de 12 millions de passages fournit les passages en conjonction avec le meilleur modèle de RI, c'est-à-dire la fusion de DPR ajusté et de la recherche visuelle. Les mêmes hyperparamètres que Karpukhin et al. (2020) sont utilisés, à l'exception du ratio de passages pertinents et non pertinents par question, qui est fixé à $\frac{8}{16}$. Nous utilisons notamment l'optimiseur Adam avec un taux d'apprentissage constant de 1×10^{-5} et un lot de 4 questions et 96 passages (24 par question) sur une seule GPU V100. L'entraînement est interrompu pour sélectionner le modèle avec le meilleur score F1 sur le jeu de validation.

3.2 Résultats

Nous utilisons des métriques standards pour évaluer l'extraction de réponse : la correspondance exacte (EM pour *exact match*) et le score F1 (au niveau des sacs de mots) entre la réponse extraite et la vérité terrain, avec le même prétraitement qu'à la section précédente.

Les résultats sont présentés dans le tableau 4.4. En ce qui concerne l'ajustement sur ViQuAE, ils suivent la même tendance que pour la RI, c'est-à-dire que l'ajustement est bénéfique mais que le modèle pré-entraîné sur TriviaQA est robuste [12]. Nous supposons aussi que cet ajustement est limité par la petite taille de ViQuAE, sur lequel le modèle sur-apprend rapidement.

Ces résultats sont cependant bien en deçà des performances habituelles en question-réponse textuel. Nous pouvons les comparer aux performances de l'extraction de réponse sur le sous-ensemble de TriviaQA qui a servi a générer le test de ViQuAE : 72,8 de F1 et 67,8 d'EM en prenant en entrée le top-24 de BM25 [13], ce qui est du même ordre de grandeur que les résultats obtenus par Wang et al. (2019) et Karpukhin et al. (2020) sur les sous-ensembles officiels de validation et de test respectivement.

Pour mieux comprendre ces résultats, et distinguer les erreurs commises par le module de RI de celles de l'extraction, nous avons étudié deux configurations

11. Voir https://github.com/facebookresearch/DPR/issues/244.

12. Ce pré-entraînement sur TriviaQA est également nécessaire. Sans lui, le modèle sur-apprend rapidement et atteint seulement 16,4 de F1 et 13,4 d'EM.

13. 74,2 de F1 et 69,1 d'EM en pondérant avec le score de BM25. BM25 a un MRR de 70,6 et une P@1 de 60,2 sur ce sous-ensemble.

Ajustement sur ViQuAE	Entrée	F1	EM
✗	Top-24 RI	26,2	24,1
	+ pondération	26,4	24,3
✓	Top-24 RI	32,2	29,4
	+ pondération	32,5	29,8
	Mi-oracle	49,1	46,1
	Oracle complet	72,7	68,3

TABLEAU 4.4 – Résultats de l'extraction de réponses sur l'ensemble de test de ViQuAE, avec ou sans ajustement du modèle d'extraction.

oracles différentes. Premièrement, *mi-oracle*, où les 24 passages résultant de la RI sont filtrés pour ne contenir que des passages pertinents (s'il y en a ; sinon, la réponse extraite sera fausse). Cette configuration se traduit par une amélioration significative de 57% d'EM par rapport à la référence et montre ainsi que le modèle ne fait pas bien la distinction entre un passage pertinent et non pertinent. Par exemple, à la figure 4.4, deux passages sur trois ne sont pas pertinents mais fournissent une réponse plausible à la question. En comparaison, l'amélioration par la pondération du score de la RI est insignifiante. Cela ouvre la voie à une meilleure intégration de l'image pour l'extraction de réponse. Enfin, nous avons considéré la configuration *oracle complet*, où le modèle ne reçoit que des passages pertinents [14]. L'écart de performance continue de se creuser : +48% d'EM relativement à la condition *mi-oracle*, donc +132% relativement à la référence. Ce constat corrobore les résultats de la section précédente : la KVQAE est très difficile pour les représentations d'images actuelles et de futurs travaux devraient porter sur une meilleure fusion des informations multimodales, ce qui sera l'objet des chapitres 5 et 6. De plus, ces chiffres assez élevés, comparables aux résultats sur TriviaQA, confirment notre hypothèse : *une fois que le passage pertinent a été retrouvé*, il est possible de répondre à la question sans regarder l'image. Ces résultats oracles sont destinés à servir de référence haute aux futures études.

4 Discussion et conclusion

Ce chapitre fixe un premier cadre d'expérimentation pour la KVQAE ainsi qu'une référence sur le *benchmark* ViQuAE présenté au chapitre précédent. Nous avons traité la KVQAE en deux étapes : (i) recherche d'information ; (ii) extraction de réponse. La RI combine plusieurs modalités de manière tardive, avec des représentations visuelles dédiées selon le type d'entité ainsi que deux variantes pour la recherche textuelle : lexicale (BM25) ou dense (DPR).

Nous trouvons que la fusion multimodale apporte des améliorations significatives à la recherche textuelle, en particulier lorsque la question porte sur une personne, qui peut bénéficier des représentations des visages d'ArcFace. En effet, d'une part

14. En leur absence dans les résultats de la RI, on utilise ceux liés à l'article Wikipédia de l'entité-sujet.

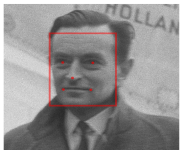
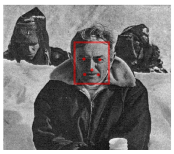
Question visuelle	1er résultat	2e résultat	3e résultat
			
« *This arch bridge spans what river ?* »	« Marlow Bridge [SEP] [...] The Széchenyi Chain Bridge, spanning the River *Danube* in Budapest, was also designed by William Clark and it is a larger scale version of Marlow bridge. »	« Hudson River [SEP] The width of the Lower *Hudson River* required major feats of engineering to cross [...]	« Pont de la Tournelle [SEP] [...] This bridge connected the Eastern bank of the **Seine** (le quai Saint-Bernard) to l'île Saint-Louis. [...]
			
« *What was the last film directed by this film producer ?* »	« David Lean [SEP] Sir David Lean (25 March 1908-16 April 1991) was [...] responsible for large-scale epics such as "The Bridge on the River Kwai" (1957), "Lawrence of Arabia" (1962), "Doctor Zhivago" (1965) and "**A Passage To India**" (1984). »	« Bernard Herrmann [SEP] [...] is particularly known for his collaborations with director Alfred Hitchcock, most famously "*Psycho*", "*North by Northwest*", "*The Man Who Knew Too Much*", and "*Vertigo*". »	« David Lean [SEP] [...] Lean recruited long-time collaborators for the cast and crew, [...] John Box, the production designer for "*Dr. Zhivago*". »

FIGURE 4.4 – Questions visuelles accompagnées des trois premiers résultats de la RI multimodale (DPR ajusté + recherche visuelle). La réponse (dans le passage pertinent) est imprimée en caractères gras et les réponses plausibles dans les passages non pertinents sont imprimées en italique. Les visages détectés sont indiqués en rouge. Le passage de texte a été raccourci pour la mise en page.

ArcFace est fondé sur un grand jeu de données, MS-Celeb, qui contient des millions d'images de 100 000 entités de type personne dont certaines apparaissent dans ViQuAE ; mais d'autre part, les personnes sont intrinsèquement mieux représentées visuellement que les autres entités, de par leur visage. Pour reprendre l'exemple évoqué au chapitre 1, à la figure 1.2, Louis-Philippe Ier est représenté de diverses façons mais on peut toujours reconnaître son *visage*. D'autres types d'entités ont des représentations bien plus diverses : par exemple un monument peut être photographié depuis différents points de vue, de l'extérieur ou de l'intérieur, et pire encore, les entités abstraites telles que les entreprises sont représentées à travers d'autres entités. Nous tenterons de proposer des solutions à ce problème aux chapitres suivants en nous appuyant sur une représentation multimodale de l'entité.

Bien que la fusion multimodale bénéficie à la RI, la marge d'amélioration reste très importante. Nous supposons que les interactions cross-modales, négligées par la fusion tardive, permettraient d'améliorer la RI. Nous étudierons donc une méthode de fusion précoce au chapitre suivant ainsi qu'une recherche cross-modale au chapitre 6.

Pour ce qui est de l'extraction de réponse, bien qu'elle soit réalisable avec un

modèle textuel, nous avons vu que ce dernier pouvait être distrait par les passages candidats non pertinents résultant de la RI, ce qui persiste malgré la pondération du score de la RI. De futurs travaux pourraient donc traiter l'extraction de réponse avec un modèle multimodal. Un autre axe d'amélioration est la RI elle-même, comme discuté précédemment, puisqu'avec une RI oracle, l'extraction de réponse atteint des résultats similaires aux *benchmarks* de question-réponse textuel.

D'autre part, nous avons vu que la référence de recherche textuelle était relativement forte, ce qui met en évidence les biais textuels inhérents à la KVQAE et plus particulièrement ceux du jeu de données ViQuAE. En effet, à moins de retirer tout naturel à la langue, il reste forcément des informations dans le texte d'une question visuelle, indépendamment de son image. Par exemple, pour « *Dans quel pays est-il né ?* », on peut répondre un pays au hasard sans regarder l'image. Ce biais peut être accentué par le jeu de données étudié, ici ViQuAE, dont la distribution n'est pas uniforme pour tous les attributs de toutes les entités. Les biais textuels de ViQuAE ont également été étudiés par Chen et al. (2023c), qui obtiennent un EM de 31,5 avec PaLM (Chowdhery et al., 2022), un gros modèle de langue (textuel), amorcé avec cinq exemples, atteignant ainsi une performance similaire à notre meilleur modèle multimodal. Cependant, PaLM a 540 milliards de paramètres, soit environ 1 000 fois plus que nos modèles (en tenant compte à la fois des modèles de RI et d'extraction de réponse). De plus, environ 20% de TriviaQA, dont ViQuAE est dérivé, se chevauche avec le jeu de pré-entraînement de PaLM selon Chowdhery et al. (2022). Par ailleurs, PaLM atteint 76,9 d'EM sur TriviaQA en l'amorçant sans exemple et 81,4, en l'amorçant avec un exemple. La marge d'amélioration par rapport à ViQuAE est donc très importante. Nous pouvons faire la même observation d'après nos expériences oracles, où BERT multi-passage atteint 68,3 d'EM sur ViQuAE.

Mentionnons aussi le travail de Li et al. (2023), qui emploient un gros modèle de langue multimodal capable de traiter la KVQAE. Toutefois, leur modèle génère de longues réponses explicatives. Ils l'évaluent donc avec ROUGE-L (Lin, 2004), avec un score de 29,6 sur ViQuAE, qui n'est malheureusement pas comparable à nos résultats.

Chapitre 5

Fusion précoce et *Inverse Cloze Task* multimodale

1 Introduction

Le chapitre précédent présente une décomposition de la KVQAE en deux étapes : Recherche d'Information (RI) et extraction de réponse. Il démontre que l'extraction de réponse fonctionne bien, grâce au pré-entraînement sur TriviaQA, si un passage pertinent est fourni par la RI. Cependant, la RI fournit une référence raisonnable, mais la marge d'amélioration est importante. En effet, le rappel est modeste, c'est-à-dire qu'il n'y a parfois aucun passage pertinent dans ses résultats, mais la précision aussi, c'est-à-dire que les résultats contiennent beaucoup de passages non pertinents qui distraient par la suite le module d'extraction de réponse. Ceci est particulièrement vrai pour les entités de type autre que *personne*, qui bénéficient d'une bonne représentation visuelle par leur visage, lesquels ont des représentations vectorielles robustes.

Dans ce chapitre, nous nous intéressons donc à améliorer la composante RI d'un système de KVQAE. La RI de notre premier système suit une approche de fusion tardive, où la recherche est faite indépendamment avec le texte et l'image avant de fusionner les résultats au niveau du score. Au contraire, une fusion précoce des modalités permettrait de modéliser les interactions multimodales introduites au chapitre 1. Notamment, l'interaction TQIQ entre l'image et le texte au sein de la question visuelle est illustrée à la figure 5.1 : la même image peut représenter plusieurs entités, ici l'entreprise *Apple* ou le magasin *Apple Fifth Avenue*. Une recherche purement visuelle serait ainsi susceptible de manquer l'entreprise *Apple* ici puisqu'elle est représentée à travers son magasin. Plus simplement, plusieurs entités peuvent aussi être dépeintes dans la même image, par exemple deux personnes : une sur la gauche et une sur la droite. L'interaction multimodale peut aider même si les modalités sont redondantes : par exemple, le texte de la première question de la figure 5.1 spécifie que le magasin considéré est de la marque *Apple*, ce qui peut être aussi utile pour filtrer les résultats de la recherche visuelle. Plus globalement, l'étude présentée dans ce chapitre tente ainsi de répondre à notre troisième et principale question de recherche : *comment interagissent les modalités ?*

Question visuelle		Passage visuel pertinent	
	« When did *this Apple store* open ? » « When was *this company* founded ? »	 	The store opened on **May 19, 2006**, as Apple's 147th store. Apple was founded as Apple Computer Company on **April 1, 1976** [...]

FIGURE 5.1 – Illustration du besoin de la fusion précoce pour modéliser l'interaction TQIQ entre l'image et le texte au sein de la question visuelle : la même image peut représenter plusieurs entités, ici l'entreprise *Apple* ou le magasin *Apple Fifth Avenue*.

La fusion d'informations est un sujet ancien qui nous ramène aux premiers travaux sur les réseaux de neurones (Nilsson, 1965; Sharkey, 1999). Ces méthodes ont naturellement été appliquées à de nombreuses tâches multimodales, pour une fusion précoce (au niveau des données ou des *features*) ou tardive (au niveau du score ou de la décision ; Kludas et al., 2007). La fusion d'informations multimodales a été très étudiée pour les questions visuelles classiques avec beaucoup de méthodes centrées autour d'un mécanisme d'attention (Zhang et al., 2019). Plus récemment, inspiré par le succès de BERT et du paradigme *pré-entraînement et ajustement*, de nombreuses méthodes ont, de manière concurrente, émergé en s'appuyant sur l'architecture *transformer*, le plus souvent sur un modèle de langue pré-entraîné tel que BERT (Tan et Bansal, 2019; Lu et al., 2019; Li et al., 2019; Su et al., 2020; Li et al., 2020; Chen et al., 2020b). En effet, le mécanisme d'attention du *transformer* permet de fusionner les modalités en modélisant des interactions non-triviales (Hessel et Lee, 2020). Nous avons donc suivi cette approche dans ce chapitre, tout en la comparant avec une méthode de fusion linéaire plus classique.

Un autre point saillant du travail que nous présentons est la possibilité d'entraîner un tel modèle de fusion précoce, en particulier avec peu de données annotées. Dans le cas des tâches telles que la VQA par exemple, la possibilité d'exploiter de grands jeux de données annotées (de l'ordre du million de questions visuelles ; Goyal et al., 2017) réduit fortement l'intérêt des méthodes de pré-entraînement, qui ne produisent que des améliorations marginales. Au contraire, nous travaillons ici à l'échelle beaucoup plus réduite du jeu de données ViQuAE (3 700 questions visuelles). Nous avons donc proposé une tâche de pré-entraînement, l'*Inverse Cloze Task* (ICT) multimodale, puisque les pré-entraînements proposés pour les questions visuelles classiques, tels que la modélisation de la langue, ne sont pas adaptés à la RI multimodale, comme nous le discuterons à la section 2.

Après avoir discuté des méthodes de pré-entraînement existantes, nous formaliserons notre cadre de recherche d'information multimodale et décrirons les modèles de fusion à la section 3. Puis, nous aborderons à la section 4 plus directement les différentes tâches d'entraînement proposées, en particulier l'ICT multimodale. Nous analyserons ensuite les résultats des différentes expériences menées à la section 5 et en discuterons à la section 6 avant de clore ce chapitre.

2 Pré-entraînements multimodaux existants

Comme discuté au chapitre 2, les modèles multimodaux pré-entraînés sont largement inspirés par BERT. Leurs tâches de pré-entraînement ne font pas exception. Nous discutons ici des plus fréquentes selon Gan et al. (2022) et citons quelques articles les utilisant à titre d'exemples :

— Le MLM (*Masked Language Modeling*) a été popularisé par BERT. Il consiste à prédire les mots masqués dans l'entrée. Toutefois, ici le modèle est conditionné à la fois par le texte et l'image (Tan et Bansal, 2019; Chen et al., 2020b; Cho et al., 2021; Kim et al., 2021).

— L'ITM (*Image-Text Matching*) imite également la *Next Sentence Prediction* de BERT mais cherche à déterminer si l'image correspond bien au texte (la légende de l'image). La tâche est donc similaire au pré-entraînement de CLIP, sauf que le texte et l'image sont encodés conjointement, tandis que CLIP est un encodeur double qui encode le texte et l'image indépendamment. Le modèle peut donc servir à réordonner des images ou des textes ; mais il est bien trop coûteux pour être appliqué en RI initiale (Tan et Bansal, 2019; Chen et al., 2020b; Cho et al., 2021; Kim et al., 2021).

— Le MIM (*Masked Image Modeling*), qui symétriquement au MLM, consiste à prédire des parties de l'image masquées. Pour les modèles fondés sur des détecteurs d'objet, il s'agit donc de prédire l'objet détecté par ce module externe (Lu et al., 2019; Tan et Bansal, 2019; Chen et al., 2020b) tandis que pour les modèles plus récents qui prennent directement en entrée les pixels de l'image, une formulation plus générique a été proposée (Bao et al., 2022; Singh et al., 2022).

Ces objectifs, bien que variés et originaux, reposent tous sur l'appariement d'une image et de sa légende. Nous supposons donc qu'ils ne sont pas adaptés au pré-entraînement d'un système de question-réponse car les légendes contiennent peu d'informations factuelles. D'autre part, certains objectifs, comme MIM, ne sont pas applicables aux images d'entités nommées puisqu'ils reposent largement sur des détecteurs d'objets. Par-dessus tout, ces objectifs sont fondés sur *une* paire (texte, image) tandis que nous formulons la KVQAE comme un problème de RI où l'on doit estimer la pertinence d'une paire (texte, image), en l'occurrence le passage visuel, selon une autre, la question visuelle.

Pour confirmer nos intuitions, nous avons mené des expériences avec le modèle ViLT (Kim et al., 2021), entraîné pour le MLM et l'ITM, ainsi que qu'avec VL-T5 (Cho et al., 2021), entraîné pour ces mêmes tâches, en parallèle des tâches cibles en aval, telles que la VQA classique. Les expériences avec ViLT ont été menées dans le même cadre de RI multimodale initiale que celui de ce chapitre mais ont fourni de très faibles résultats, bien en deçà des références mono-modales. Les expériences menées dans le même cadre avec VL-T5 par Paul Grimal durant son stage au CEA-List ont également conduit à des résultats décevants (Grimal, 2022). D'autres expériences ont été menées avec ViLT pour réordonner les résultats de la RI initiale mais ont été également infructueuses (Messoud, 2022).

Pour ces raisons, nous proposons un nouvel objectif de pré-entraînement, l'*Inverse Cloze Task* multimodale, décrit dans les sections suivantes.

3 Modélisation

3.1 Cadre de recherche d'information multimodale

Nous généralisons le cadre du chapitre précédent, pour l'entraînement de DPR, en incluant cette fois l'image de la question visuelle et du passage visuel. Les deux peuvent donc être représentés par une paire (texte, image). Notre but est d'optimiser les paramètres de l'encodeur E afin qu'il produise des représentations adéquates de $\mathbf{q} = E(\mathbf{t_q}, \mathbf{i_q})$ et $\mathbf{p} = E(\mathbf{t_p}, \mathbf{i_p})$, c'est-à-dire que $\mathbf{q}$ soit proche de $\mathbf{p}$ s'il est pertinent et éloigné sinon (ce que l'on note avec les exposants $(^+)$ et $(^-)$). Une fois ces représentations calculées, la RI se résume à un problème de recherche des plus proches voisins. Comme pour DPR, notons que la représentation du passage visuel $\mathbf{p} = E(\mathbf{t_p}, \mathbf{i_p})$ est indépendante de celle de la question visuelle. Cela permet de pré-calculer les représentations de toute la BC, puis d'utiliser une méthode de MIPS, comme au chapitre précédent. Il incombe donc au modèle E de tenir compte ou pas des informations contenues dans $(\mathbf{t_q}, \mathbf{i_q})$ et $(\mathbf{t_p}, \mathbf{i_p})$. Selon le modèle, on peut revenir à DPR (recherche purement textuelle) ou bien au contraire à une recherche purement visuelle. Les interactions TQTP, IQIP et IQTP, entre la question et le passage se retrouvent dans la similarité entre les vecteurs $\mathbf{q}$ et $\mathbf{p}$, tandis que l'interaction TQIQ (resp. TPIP) au sein de la question visuelle (resp. passage visuel) sera modélisée, ou pas, par E.

En calculant la similarité entre deux vecteurs comme leur produit scalaire, l'objectif utilisé pour entraîner E est de minimiser la log-vraisemblance négative suivante pour toutes les questions visuelles du jeu de données :

$$-\log \frac{\exp\left(\mathbf{q} \cdot \mathbf{p}^+\right)}{\exp\left(\mathbf{q} \cdot \mathbf{p}^+\right) + \sum_j \exp\left(\mathbf{q} \cdot \mathbf{p}_j^-\right)} \tag{5.1}$$

Les exemples négatifs $\mathbf{p}_j^-$ proviennent soit des passages pertinents pour les autres questions du lot, auquel cas ils sont tirés de façon aléatoire, soit ils sont sélectionnés spécifiquement pour la question $\mathbf{q}$. On dit alors que ce sont des exemples négatifs *difficiles*.

Nous avons étudié en particulier deux modèles E que nous présentons à la section suivante, auxquels s'ajoute la référence textuelle BERT (que l'on désigne *DPR* quand il est utilisé dans ce cadre d'encodeur double).

3.2 Modèles

Les modèles sont construits avec les mêmes briques qu'au chapitre précédent, c'est-à-dire BERT pour représenter le texte, CLIP et ImageNet-ResNet pour représenter l'image de façon générique et ArcFace pour représenter les visages détectés avec MTCNN. Ces modèles visuels sont figés pendant l'entraînement pour éviter l'oubli catastrophique. Les représentations visuelles résultantes sont fusionnées avec les représentations textuelles par deux méthodes différentes : ECA et ILF.

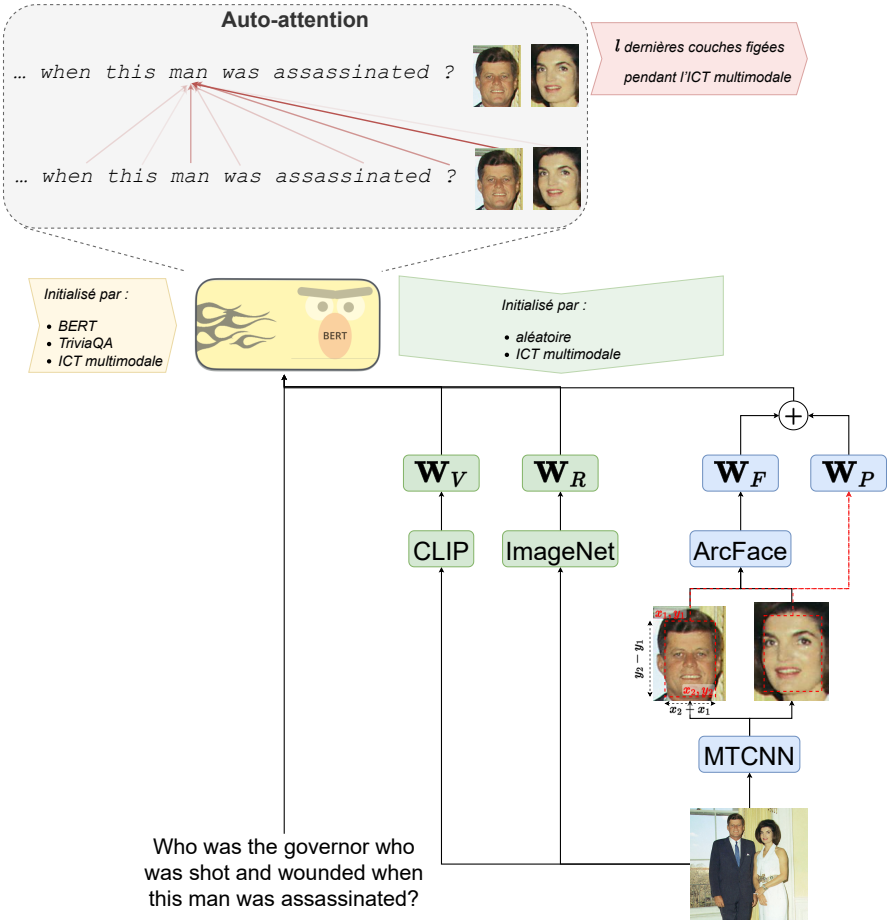


FIGURE 5.2 – Illustration de l'architecture ECA (*Early Cross-Attention*) pour la fusion multimodale. Le modèle est décrit à la section 3.2. Il est entraîné en trois phases : (i) Question-réponse textuel « TriviaQA » ici ; (ii) *Inverse Cloze Task* (ICT) multimodale ; (iii) KVQAE ; cf. section 4.1.

3.2.1 ECA

ECA (*Early Cross-Attention*) est un BERT multimodal à flux unique, comme par exemple (Chen et al., 2020b), sauf que l'image est représentée par CLIP, ImageNet-ResNet et ArcFace et non avec un détecteur d'objets comme Faster R-CNN (on suppose qu'il n'est pas adapté pour représenter des entités nommées). L'intuition est de représenter l'image dans le même espace que les plongements lexicaux du texte afin que le modèle puisse prendre en entrée l'image à la suite des tokens. Cette approche est donc liée au *prompt tuning*. Les informations sont ensuite naturellement fusionnées à travers le mécanisme d'auto-attention du *transformer* (cf. figure 5.2).

Formellement, chaque représentation de l'image est projetée linéairement dans un espace de dimension d, avec des paramètres dédiés pour chaque représentation $\mathbf{W}_{\{V,R,A,P\}} \in \mathbb{R}^{d_{\{V,R,A,P\}} \times d}$, où V, R, A notent CLIP, ImageNet-ResNet, ArcFace et MTCNN, respectivement.

Une image peut contenir plusieurs visages. Nous en représentons N_A au maximum, les plus probables selon MTCNN. Le texte peut contenir une expression référentielle relative à sa position, par exemple « la personne sur la droite ». Par

conséquent, à l'instar de (Chen et al., 2020b), nous représentons les coordonnées du visage dans le même espace que le visage lui-même et sommons leurs représentations de manière analogue à un plongement positionnel dans un *transformer*. Les positions des visages sont représentées par les coordonnées des points en haut à gauche et en bas à droite du visage, ainsi que la largeur, la hauteur et l'aire du rectangle ainsi délimité. Formellement, on a :

$$\forall j \in [\![1, N_A]\!], \mathbf{e}_{Aj} = \mathrm{ArcFace}(\mathbf{i})_j \cdot \mathbf{W}_A + \mathrm{MTCNN}(\mathbf{i})_j \cdot \mathbf{W}_P \tag{5.2}$$

où $\mathbf{W}_P \in \mathbb{R}^{d_P \times d}$.

Enfin, toutes ces représentations sont concaténées à la suite de la séquence des plongements lexicaux avant d'être données en entrée au *transformer*. La représentation finale est obtenue à travers le token spécial $[\mathrm{CLS}]$:

$$\mathrm{ECA}(\mathbf{t}, \mathbf{i}) = \mathrm{BERT}([\mathbf{t}_1; \mathbf{t}_2; \ldots; \mathbf{t}_L; \mathbf{e}_V; \mathbf{e}_R; \mathbf{e}_{A_1}; \mathbf{e}_{A_2}; \ldots; \mathbf{e}_{A_{N_A}}])_{[\mathrm{CLS}]} \tag{5.3}$$

Le *transformer* est initialisé à partir de BERT. Ainsi, l'interaction TQIQ (resp. TPIP) au sein de la question visuelle (resp. passage visuel) est modélisée par le mécanisme d'auto-attention du *transformer*.

3.2.2 ILF

ILF (*Intermediate Linear Fusion*) introduit de façon plus standard les paramètres supplémentaires $\mathbf{W}_T \in \mathbb{R}^{d \times d}$ afin de projeter la représentation textuelle dans le même espace que les images :

$$\mathrm{ILF}(\mathbf{t}, \mathbf{i}) = \mathrm{BERT}(\mathbf{t})_{[\mathrm{CLS}]} \cdot \mathbf{W}_T + \mathbf{e}_V + \mathbf{e}_R + \mathbf{e}_{A_1} + \mathbf{e}_{A_2} + \ldots + \mathbf{e}_{A_{N_A}} \tag{5.4}$$

Il faut remarquer que cette approche équivaut à concaténer $[\mathrm{BERT}(\mathbf{t})_{[\mathrm{CLS}]}; \mathbf{e}_V; \mathbf{e}_R; \mathbf{e}_{Aj}]$ et réaliser la projection du vecteur issu de cette concaténation dans l'espace commun texte-image en utilisant la matrice concaténant $[\mathbf{W}_T; \mathbf{W}_V; \mathbf{W}_R; \mathbf{W}_A]$. Il n'y a donc *pas* de prise en compte des interactions TQIQ (resp. TPIP) au sein de la question visuelle (resp. du passage visuel). ILF nous permet donc de vérifier si les améliorations d'ECA par rapport à la fusion tardive s'expliquent par le pré-entraînement à l'ICT multimodale ou par une meilleure fusion des modalités.

3.2.3 DPR

Il faut par ailleurs rappeler que, dans ce cadre, DPR représente le texte avec $\mathrm{BERT}(\mathbf{t})_{[\mathrm{CLS}]}$ et ne modélise donc pas d'interaction cross-modale. Nous distinguerons une version des modèles fondée sur toutes les représentations visuelles (comme au chapitre précédent), notée ECA_{V+R+A} et ILF_{V+R+A}, et une version fondée seulement sur CLIP, notée ECA_V et ILF_V.

4 Pré-entraînement et ajustement

Malgré l'utilisation de modèles fondateurs comme BERT, dont le pré-entraînement a une vocation universelle, de nombreux travaux portent sur une seconde phase de

pré-entraînement, avant l'ajustement final sur la tâche d'intérêt (Gao et Callan, 2022). Notre travail s'inscrit dans cette lignée.

Nous étudions deux pré-entraînements différents, qui peuvent être combinés : un pré-entraînement purement textuel sur TriviaQA, comme au chapitre précédent ; l'autre multimodal, que nous décrivons dans cette section. Les deux peuvent être combinés de façon séquentielle : le question-réponse textuel d'abord, l'ICT multimodale ensuite. Nous sommes alors confronté à un risque d'oubli catastrophique. Pour y remédier, nous avons gardé certaines couches du *transformer* figées afin qu'elles conservent les aptitudes apprises en question-réponse textuel. Pour ILF, il est possible de figer l'intégralité des 12 couches du *transformer* puisque la fusion se fait à travers la projection $\mathbf{W}_T$. Pour ECA, nous avons expérimenté avec la méthode d'Aytar et al. (2017), qui consiste à figer les l dernières couches du modèle. Cette méthode est contraire à l'apprentissage par transfert classique où le modèle pré-entraîné sert d'extracteur de caractéristiques et où seul le classifieur est ajusté. L'intuition est que les l dernières couches opèrent avec des représentations plus abstraites, plus proches de la tâche, et que seules les premières couches doivent être ajustées à la multimodalité. Cela rejoint parfaitement la philosophie de l'architecture ECA : les entités nommées textuelles apprises pendant la première phase sont en quelque sorte remplacées par leurs représentations visuelles.

Une fois le modèle pré-entraîné, il peut être ajusté sur ViQuAE sans crainte d'oubli catastrophique, en laissant donc toutes les couches entraînables.

Dans cette section, nous décrivons plus formellement cet apprentissage séquentiel (section 4.1) ainsi que la tâche de pré-entraînement proposée : l'ICT multimodale (section 4.2).

4.1 Apprentissage séquentiel en trois phases

(i) Question-réponse textuel Cette phase est en tout point identique à l'entraînement de DPR décrit au chapitre précédent, à la section 2.1.3. Pour mémoire, DPR est fondé sur deux encodeurs BERT, paramétrés différemment, qui serviront à initialiser ECA et ILF. Les données utilisées sont celles de TriviaQA, filtré de toutes les questions utilisées dans ViQuAE, accompagné de la BC textuelle KILT de 32 millions de passages. Les passages *négatifs difficiles* sont sélectionnés avec BM25 en s'assurant qu'ils *ne* sont *pas* pertinents, ce qui est réalisé en vérifiant qu'ils *ne* contiennent *pas* la réponse à la question.

(ii) *Inverse Cloze Task* multimodale La génération d'exemples d'entraînement est détaillée à la section suivante. Pour ce pré-entraînement multimodal, les *transformer* d'ECA et ILF sont initialisés à partir de la phase d'entraînement précédente (ou bien de BERT pour l'étude d'ablation). Par ailleurs, nous expérimenterons avec la méthode d'Aytar et al. (2017) en figeant les $l = 6$ dernières couches d'ECA, soit la moitié, et $l = 12$ couches d'ILF. Les couches de projection $\mathbf{W}$ sont, quant à elles, initialisées aléatoirement. Nous expérimenterons à la fois avec des exemples négatifs aléatoires et des exemples négatifs difficiles.

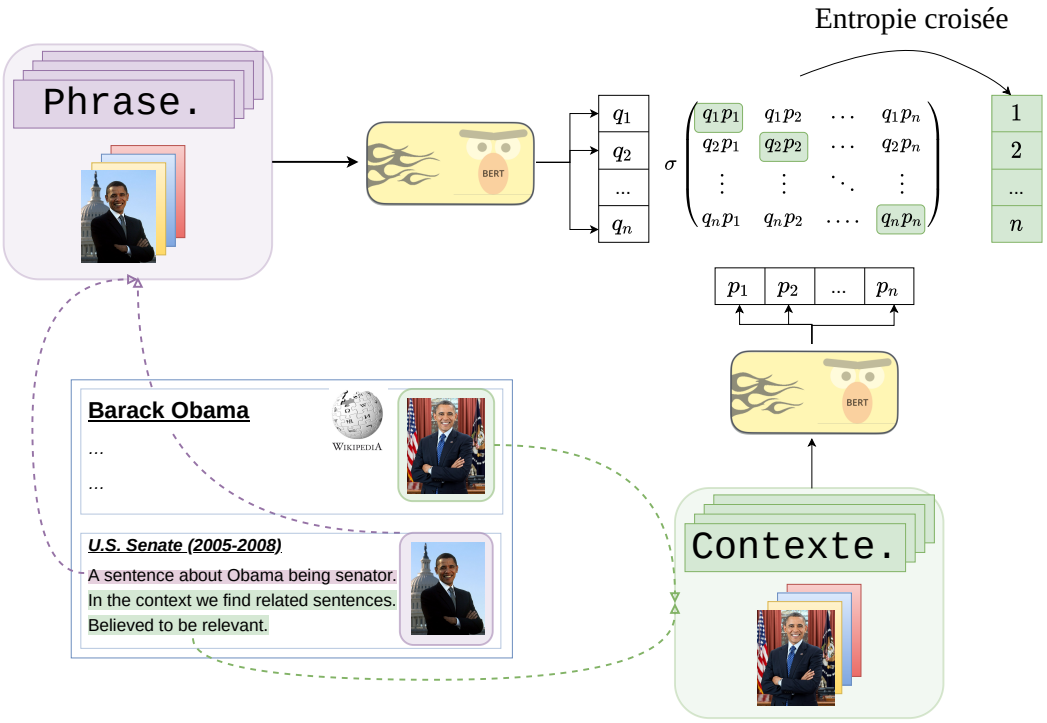


FIGURE 5.3 – Aperçu de l'*Inverse Cloze Task* multimodale via Wikipédia/WIT.

(iii) KVQAE Cette phase consiste à ajuster les modèles sur ViQuAE, de façon similaire à DPR au chapitre précédent sauf que les questions et les passages sont multimodaux. Les *transformer* sont initialisés à partir de la deuxième phase d'entraînement (ou bien de BERT ou de la première phase pour l'ablation). Les couches de projection $\mathbf{W}$ ne peuvent être pré-entraînées qu'à partir de la deuxième phase d'entraînement ; sinon, elles sont initialisées aléatoirement (cf. figure 5.2). Toutes les couches des *transformer* d'ECA et ILF sont ajustées pendant cette phase puisque l'on n'a plus à craindre d'oubli catastrophique. Contrairement au chapitre précédent, où les exemples négatifs étaient sélectionnés avec BM25, nous avons décidé ici de les sélectionner avec la RI multimodale (fusion tardive) du chapitre précédent, comme pour le module d'extraction de réponse. Ce choix a été fait car nous supposons que BM25 sélectionne des exemples difficiles seulement d'un point de vue textuel [1] et donc, non visuel [2]. Il serait donc sous-optimal d'entraîner ECA et ILF avec ces passages. Nous vérifierons que, pour DPR, cela n'entraîne pas de différence significative [3].

4.2 *Inverse Cloze Task* multimodale

L'ICT textuelle a été proposée par Lee et al. (2019). Elle consiste, étant donné une phrase, à chercher son contexte et peut ainsi être vue comme un Skip-gram

1. Par exemple, un passage avec un fort chevauchement lexical avec la question, ce qui pourrait distraire un modèle textuel comme DPR.
2. Par exemple, une image d'une personne ressemblant à l'entité-sujet.
3. Puisque la RI multimodale du chapitre précédent est fondée sur DPR, il serait possible que cette deuxième itération soit renforcée, qu'elle « apprenne de ses erreurs ».

au niveau des phrases. Il s'agit d'une tâche auto-supervisée car on peut aisément masquer une phrase de son contexte pour créer des exemples d'entraînement. L'ICT multimodale est plutôt *faiblement* supervisée car elle s'appuie sur des documents multimodaux semi-structurés, en l'occurrence, des articles Wikipédia. En effet, nous tirons avantage des images contextuelles associées aux sections des articles Wikipédia, en considérant qu'elles sont pertinentes pour le texte de ces sections, ainsi que de l'image de l'*infobox*, en considérant qu'elle est pertinente pour l'ensemble du texte de l'article. Cette dernière supposition se retrouve également dans la BC ViQuAE, où une seule image est associée à l'ensemble du texte de l'article. L'image de la section sert donc à illustrer la phrase masquée pour former une pseudo-question visuelle tandis que l'image de l'*infobox* illustre les phrases environnantes pour former un passage visuel pertinent (cf. figure 5.3).

Ce processus est mis en œuvre grâce au corpus WIT (Srinivasan et al., 2021). WIT associe images et textes de l'écosystème Wikimédia de plusieurs manières : légendes d'images Commons, légendes Wikipédia et section Wikipédia appariée. Nous ne conservons que cette dernière afin de mimer la BC ViQuAE et l'ICT textuelle de Lee et al. (2019). WIT est également multilingue, mais nous nous concentrons sur la langue anglaise, comme pour l'ensemble de la thèse. Ce sous-ensemble de WIT consiste en 400 000 articles, appariés à des images d'*infobox*. Les articles sont eux-mêmes composés de sections, 1,2 million au total, elles-mêmes appariées à des images. Ces 1,2 million de sections se composent de 13,6 millions de phrases, soit autant de potentielles pseudo-questions (cf. figure 5.3). Les phrases ont une longueur moyenne de 26 mots. Par conséquent, pour imiter les passages des BC KILT et ViQuAE, où les passages sont limités à 100 mots, nous limitons ici les passages à quatre phrases [4]. Comme pour les BC, le titre de l'article est concaténé au début du passage et séparé par le token spécial [SEP]. Ces quatre phrases peuvent se trouver avant et/ou après la pseudo-question dans la section. Deux exemples de pseudo-questions visuelles générées de cette façon sont montrés à la figure 5.4. On peut voir que, pour les deux pseudo-questions visuelles, le passage visuel pertinent peut être retrouvé à partir du texte ou de l'image. Cela motive l'utilisation de cette tâche pour pré-entraîner des modèles de RI multimodale.

Après avoir filtré les images dont les fichiers sont corrompus ou celles dont le format n'est pas approprié (par exemple .svg) ainsi que les sections contenant une seule phrase, nous obtenons 975 000 sections appariées à des images. Ce jeu de données est simplement baptisé WIT-ICT. Il est divisé en sous-ensembles d'entraînement (878 000 sections), de validation (48 000 sections, pour ajuster les hyperparamètres) et de test (48 000, comme *sanity-check*) de manière à ce qu'il n'y ait pas de chevauchement entre les articles.

4.2.1 Exemples négatifs difficiles

Les exemples négatifs difficiles font partie intégrante de tout apprentissage contrastif (Xu et al., 2022). Dans notre cadre, ils ont été empiriquement démontrés comme essentiels à DPR par Karpukhin et al. (2020) et d'autres à la suite (Qu

4. Cette approche diffère légèrement de celle de Lee et al. (2019), qui considèrent des passages contenant jusqu'à 288 sous-mots, avant le masquage des pseudo-questions.

Pseudo-question visuelle	Passage visuel pertinent
 *[CLS] deutsch was born in alencon to a hungarian - jewish father and a romanian mother. [SEP]*	 [CLS] lorant deutsch [SEP] lorant deutsch is a french actor and writer. an ardent catholic, deutsch claims to be a royalist. in 2005, deutsch met actress marie - julie baup when they worked together during amadeus. after working together for several more years while cast in the importance of being earnest, they married in 2009, on 3 october, and now have three children. [SEP]
 *[CLS] several bay platforms are arranged in an elevated position between the running lines from the north and south hall. [SEP]*	 [CLS] dresden hauptbahnhof [SEP] these are mainly used for stabling short sets. impressive entrances to the station building were built not only from the east, but also from the north and the south. additionally from these sides there are direct entrances to the central train shed under the elevated through tracks. the entrance from wiener platz to the station hall [...] [SEP]

FIGURE 5.4 – Exemples de pseudo-questions visuelles accompagnées de leurs passages visuels pertinents, générés à partir de WIT-ICT. Le deuxième passage a été raccourci pour des raisons de mise en page.

et al., 2021b). En effet, si les exemples $\mathbf{p}_j^-$ sont trop éloignés de $\mathbf{q}$, l'objectif de l'équation 5.1 devient trivial. Cependant, ils ne sont pas utilisés pour l'ICT textuelle proposée par Lee et al. (2019). À la place, les auteurs utilisent une très grande taille de lot (4 096 exemples), donc un grand nombre d'exemples négatifs aléatoires, ce qui contourne le problème (Qu et al., 2021b).

Contrairement au question-réponse, où l'on peut vérifier la pertinence d'un passage par la présence de la réponse, nous n'avons pas ici de façon évidente de juger la pertinence d'un passage pour une pseudo-question visuelle. Nous avons donc expérimenté avec une permutation des images des passages visuels. Ainsi, dans le cas de la figure 5.4, nous créons un nouveau passage avec l'image du premier exemple et le texte du second. Cet exemple est alors *non-pertinent* pour toutes les pseudo-questions du lot. En particulier, il est *difficile* pour la deuxième question visuelle puisque *le texte est pertinent*, mais pas l'image et, symétriquement, il est *difficile* pour la première question visuelle puisque *l'image est pertinente*, mais pas le texte. Nous espérons ainsi contraindre le modèle à utiliser les deux modalités et modéliser leurs interactions. Le nombre de passages visuels dans le lot est par ailleurs doublé [5] car chaque texte du passage visuel n est associé à l'image du passage visuel $n-1$.

5. Pour un lot de taille N, nous pourrions ainsi générer $N(N-1)$ exemples, mais nos expériences se limitent à générer N exemples. Cette configuration est ainsi similaire au question-réponse où un seul exemple négatif difficile est utilisé pour chaque question du lot.

4.2.2 Probabilité de masquage

Lee et al. (2019) démontrent empiriquement qu'il est bon de ne pas systématiquement masquer la pseudo-question du passage mais de la laisser dans 10% des cas. Cela permet au modèle d'apprendre que le chevauchement lexical est une caractéristique importante pour juger de la pertinence d'un passage. Cependant, nous avons mené la plupart de nos expériences avec un taux de masquage de 100% car nous supposons que laisser la pseudo-question dans 10% des cas n'est ni nécessaire, étant donné que le modèle est déjà pré-entraîné au question-réponse sur TriviaQA et devrait donc déjà avoir appris à utiliser le chevauchement lexical, ni bénéfique, car il pourrait alors ignorer l'image et utiliser seulement le texte.

4.2.3 Chevauchement entre WIT-ICT et ViQuAE

Puisque les images de WIT-ICT et ViQuAE proviennent toutes deux de Wikimédia Commons, nous pouvons estimer un chevauchement de 14% des images de ViQuAE avec WIT-ICT selon leur URL. Nous vérifions que cela n'entraîne pas un biais à la section suivante.

4.3 Implémentation

Pour éviter l'oubli catastrophique, les représentations d'images (CLIP V, ImageNet-ResNet R et ArcFace A) restent figées pendant l'entraînement. Les dimensions des espaces de représentations sont $d_V = 1024, d_R = 2048, d_A = 512, d_P = 7, d = 768$. Sauf indication contraire, nous utilisons au plus *quatre* représentations de visages par image (les plus probables) pour ECA et ILF. Ce choix a été motivé par la détection en moyenne de 1 ± 3 visages (médiane 0, troisième quartile 1) sur les images liées aux sections de WIT-ICT et 1 ± 2 (médiane 0, troisième quartile 1) sur les images des *infobox*, soit des statistiques similaires au jeu de données ViQuAE.

Les encodeurs de la question visuelle $(\mathbf{t_q}, \mathbf{i_q})$ et du passage visuel $(\mathbf{t_p}, \mathbf{i_p})$ ne partagent pas leurs paramètres.

Nous utilisons les mêmes hyperparamètres que pour l'entraînement de DPR au chapitre précédent. En particulier, l'optimisation est faite avec Adam avec un taux d'apprentissage de 2×10^{-5} croissant pendant 100 et 4 itérations puis décroissant pendant 8 000 et 80 itérations aux phases d'entraînement 2 et 3, respectivement. Cependant, nous avons constaté, sur le sous-ensemble de validation, qu'ILF convergeait plus rapidement avec un taux d'apprentissage de 2×10^{-3}, constant au cours de la 2e phase. Nous expliquons cela par le fait qu'ILF fige BERT dans cette phase et n'a donc pas besoin d'un faible taux d'apprentissage.

Nous appliquons un *dropout* après avoir projeté les couches de projection linéaire $\mathbf{W}$ avec une probabilité de 0,1. La *layer normalization* (Ba et al., 2016) est appliquée après les additions élément-par-élément de vecteurs dans ILF et dans ECA pour les visages. Les gradients sont normalisés pour avoir une norme maximale de 2.

Les modèles des phases 2 et 3 sont entraînés avec un lot de 512 et 298 questions visuelles, respectivement [6]. Ces relativement grandes tailles de lot permettent

6. 256 pour le 2e phase en cas de génération d'exemples négatifs difficiles, ce qui produit un total de 512 passages visuels. De la même façon, pour la 3e phase, le lot compte 596 passages visuels, soit

d'obtenir beaucoup d'exemples négatifs aléatoires et assurent donc le succès de l'apprentissage contrastif. Nous avons constaté que le *gradient checkpointing* (Chen et al., 2016) permettait d'économiser énormément de mémoire, et donc d'augmenter la taille du lot, au prix d'une rétro-propagation plus longue. Ainsi, contrairement au chapitre précédent, où nous utilisions quatre GPU Nvidia V100 avec 32 Go de RAM chacune pour une taille de lot totale de 128 questions, nous sommes en mesure ici de tenir en mémoire un lot de 298 questions (comme indiqué ci-dessus) sur une seule GPU V100. Cette différence d'implémentation a été contrôlée pour DPR en même temps que la sélection d'exemples négatifs.

La 2e phase est responsable de la majeure partie du temps de calcul, dans ce chapitre comme pour l'ensemble de la thèse, la plupart des modèles convergeant après environ 8 000 itérations, soit environ trois jours de calcul. Nous discutons du bilan carbone associé à l'annexe A. Dans toutes les phases, l'entraînement est interrompu selon le rang réciproque moyen *au sein du lot*, sur le sous-ensemble de validation, comme au chapitre précédent. *Au sein du lot* signifie que l'on ordonne seulement les passages visuels du lot, et pas de toute la BC, et que chaque question a un seul et unique passage pertinent (comme pour l'entraînement).

5 Résultats

Cette section présente les résultats, qui varient particulièrement selon les phases d'entraînement et les représentations visuelles utilisées. Nous commencerons par présenter les résultats en amont, sur la tâche de pré-entraînement proposée : l'ICT multimodale (section 5.1). Ces résultats ne nous intéressent pas directement mais permettent d'analyser les résultats en aval sur notre tâche d'intérêt, la KVQAE, qui sont présentés par la suite (section 5.2). Nos principaux résultats (§5.2.1) suivent les trois phases d'entraînement décrites plus haut, mais nous avons également fait une étude d'ablation du pré-entraînement pour le question-réponse (la 1re phase), qui est présenté dans une sous-section à part (§5.2.2). Au sein de ces sous-sections, nous comparons également des modèles avec ou sans pré-entraînement pour l'ICT multimodale (la 2e phase) ainsi que sans ajustement sur ViQuAE (la 3e phase). Par ailleurs, les résultats sont présentés sur le jeu de test, sachant que les différentes variantes d'ECA ont été validées en premier lieu sur le jeu de validation (cf. annexe C). ILF tient également le rôle d'une ablation d'ECA. Nous ne présentons donc que ses résultats pour la variante $\mathrm{ILF}_V(l = 12)$, fondée seulement sur CLIP et avec l'intégralité des 12 couches du *transformer* figées pendant l'ICT multimodale.

5.1 En amont, ICT multimodale sur WIT-ICT

Comme discuté plus haut, nous ne disposons pas de jugement de pertinence pour chaque passage selon chaque pseudo-question, ni de façon évidente d'en obtenir. L'évaluation se fait donc ici toujours *au sein du lot*, comme pour l'interruption de l'entraînement. En plus du rang réciproque moyen (MRR), nous rapportons

deux par question.

#	Modèle	Initialisation	l	Négatifs difficiles	MRR	P@1
a	ILF_V	Question-réponse	12	✗	87,1	79,9
b	ECA_{V+R+A}	Question-réponse	6	✓	91,4	86,1
c	ECA_{V+R+A}	Question-réponse	6	✗	91,9	87,0
d	ECA_V	Question-réponse	6	✗	91,6	86,6
e	ECA_V	Question-réponse	0	✗	**92,9**	88,3
f	DPR	BERT	0	✗	89,4	83,9
g	ECA_V	BERT	0	✗	92,4	87,6
h	ECA_{V+R+A}	BERT	0	✗	**92,9**	**88,4**

TABLEAU 5.1 – Évaluation *standard* en amont, ICT multimodale sur WIT-ICT. Les lettres en indice du nom des modèles désignent les représentations visuelles : CLIP V, ImageNet-ResNet R et ArcFace A. Question-réponse est la 1re phase d'entraînement, qui peut être optionnellement omise en initialisant les modèles directement à partir de BERT. l correspond au nombre des dernières couches figées dans le modèle pour préserver le pré-entraînement question-réponse, donc 0 si la 1re phase d'entraînement est omise. Les exemples négatifs difficiles sont utilisés ici seulement pendant l'entraînement, pas l'évaluation.

#	Modèle	Initialisation	l	Négatifs difficiles	MRR	P@1
a	ECA_{V+R+A}	Question-réponse	6	✓	**93,8**	**89,8**
b	ECA_{V+R+A}	Question-réponse	6	✗	75,5	55,3
c	DPR	BERT	0	✗	68,7	44,2

TABLEAU 5.2 – Évaluation *difficile* en amont, ICT multimodale sur WIT-ICT. La colonne « négatifs difficiles » indique leur utilisation pendant l'entraînement. Ils sont toujours utilisés pour l'évaluation des deux modèles.

également la précision@1 (P@1). Pour comparer deux modèles, il est donc crucial que la taille du lot soit constante. Nous distinguons deux évaluations :

— *standard* (tableau 5.1), avec 1 024 pseudo-questions associées à autant de passages visuels qui sont donc des exemples négatifs aléatoires ;

— *difficile* (tableau 5.2), avec 512 pseudo-questions associées à 1 024 passages visuels, soit un exemple négatif difficile par question.

En premier lieu, on peut remarquer au niveau du tableau 5.1 que DPR, initialisé seulement à partir de BERT, fournit une très bonne référence bien qu'il prenne seulement le texte en entrée. Cependant, la multimodalité apporte une amélioration importante (f vs. g et h). En comparant ECA_{V+R+A} et ECA_V, nous remarquons que les différences sont très légères, avec ou sans pré-entraînement question-réponse, ce qui suggérerait qu'ECA n'exploite pas bien les représentations des visages fournies par ArcFace (g vs. h et c vs. d).

Concernant les exemples négatifs difficiles, on peut voir, en comparant les tableaux 5.1 et 5.2, que, s'ils ne semblent pas bénéfiques pour l'évaluation standard, ils sont au contraire essentiels à l'évaluation difficile. Cela suggère que, sans exemples

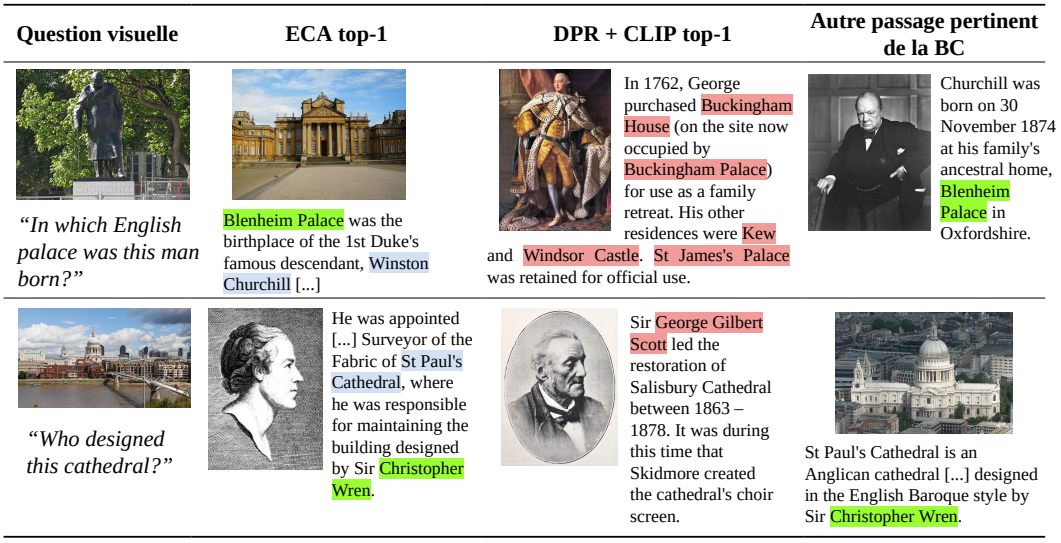


FIGURE 5.5 – Exemples où $\mathrm{ECA}_V(l = 6)$, pré-entraîné pour le question-réponse sur TriviaQA puis pour l'ICT multimodale, classe un passage visuel pertinent en premier, contrairement à la fusion tardive (DPR + CLIP). Ces exemples suggèrent qu'ECA exploite l'interaction IQTP, entre l'image de la question et le texte du passage, ce qui permet de traiter l'hétérogénéité des représentations visuelles. Cette dernière est illustrée par la dernière colonne qui montrent d'autres passages pertinents, présents dans la BC mais absents des résultats des deux modèles.

négatifs difficiles pendant l'entraînement, ECA n'exploite pas bien les interactions entre les modalités. Il surpasse tout de même la référence textuelle DPR, qui sert ici de point de référence entre les deux conditions.

Lorsqu'il est entièrement entraînable ($l = 0$), ECA obtient des résultats très similaires, avec ou sans pré-entraînement question-réponse, ce qui est sans doute le signe de son oubli catastrophique de la 1re phase d'entraînement (d vs. f).

Nous observons également que ILF_V obtient les moins bons résultats car le *transformer* reste figé. Il en va de même pour ECA_V, qui fonctionne mieux avec $l = 0$ que $l = 6$ (c vs. d au tableau 5.1). Ces résultats ne sont pas intéressants en eux-mêmes mais sont utiles pour analyser les résultats en aval sur ViQuAE (section suivante).

5.2 En aval, KVQAE sur ViQuAE

5.2.1 Principaux résultats : avec pré-entraînement au question-réponse sur TriviaQA

Les résultats principaux sont présentés au tableau 5.3. Nous les présentons selon les différentes dimensions importantes pour la KVQAE puis selon plusieurs phénomènes liés à nos pré-entraînements.

Interactions cross-modales Lorsqu'ils sont fondés seulement sur CLIP, ECA et ILF surpassent la fusion tardive pour toutes les métriques (d et f vs. b). ECA se distinguant d'ILF par la prise en compte des interactions TQIQ et TPIP, nous

#	Modèle	ICT	MRR	P@1	P@20	Hits@20
a	DPR	NA	32,8	22,8	16,4	61,2
b	DPR_V	NA	34,5^{a}	24,8^{a}	15,8	61,8
c	$\text{ECA}_V(l = \text{NA})$	✗	34,6	25,9^{a}	17,2^{ab}	61,6
d	$\text{ECA}_V(l = 6)$	✓	**37,8**^{abce}	26,7^{a}	**19,5**^{abce}	**67,6**^{abce}
e	$\text{ECA}_V(l = 0)$	✓	35,1	24,7	17,6^{b}	63,7
f	$\text{ILF}_V(l = 12)$	✓	37,3^{a}	**26,8**^{a}	19,1^{abce}	66,9^{abc}
g	DPR_{V+R+A}	NA	**37,9**	**27,8**	17,5	65,7
h	$\text{ECA}_{V+R+A}(l = 6)$	✓	37,7	27,2	**18,8**	**66,7**

TABLEAU 5.3 – Évaluation en aval, KVQAE sur ViQuAE, des modèles pré-entraînés pour le question-réponse sur TriviaQA et ajustés sur ViQuAE. l correspond au nombre de dernières couches figées dans le modèle pendant l'ICT multimodale, si applicable. Les lettres en indice du nom des modèles désignent les représentations visuelles : CLIP V, ImageNet-ResNet R et ArcFace A. Les exposants dénotent des différences significatives selon le test de randomisation de Fisher avec $p \leq 0,01$.

supposons que cette supériorité résulte de leur exploitation de l'interaction IQTP entre l'image de la question et le texte du passage. Deux exemples allant dans ce sens sont montrés à la figure 5.5. Dans le premier exemple, l'image du passage proposé par ECA est ainsi très différente de celle de la question, mais le texte mentionne Winston Churchill, qui est bien l'entité-sujet dépeinte dans l'image de la question. On peut dire la même chose du second exemple, où la cathédrale Saint-Paul est seulement mentionnée dans le texte du passage mais pas dépeinte dans l'image. Cette interaction, si elle est effectivement prise en compte, permet ainsi de traiter l'hétérogénéité des représentations visuelles des entités nommées. En effet, on peut répondre à la question sur Churchill à partir de son propre article Wikipédia : il est né au *Palais de Blenheim*. Toutefois, Churchill est représenté à travers une *statue* dans cette question visuelle tandis que, dans la BC, il est représenté par un *portrait photographique*.

Enfin, on voit que les deux modèles atteignent des performances similaires (d vs. f), contrairement à ce que suggèrent les travaux connexes (chapitre 2). D'une part, cela démontre la robustesse de l'ICT multimodale et d'autre part, cela laisse à penser que l'interaction IQTP est plus importante que les interactions TQIQ et TPIP, qu'ILF ne peut modéliser, à moins que ces dernières ne soient pas bien modélisées par ECA.

Représentation des visages Malgré les bons résultats avec CLIP seulement, on voit que les représentations d'ImageNet-ResNet, et surtout d'ArcFace pour les visages, n'améliorent pas les résultats d'ECA (d vs. h), contrairement à la fusion tardive (b vs. g), et donc que les deux arrivent au même niveau (g vs. h, sans aucune différence significative dans toutes les métriques). Remarquons qu'ECA est alors meilleur pour les questions à propos de non-personnes (39,3 de MRR vs. 35,7 pour la fusion tardive) mais moins bon pour les questions à propos de personnes (35,8 de MRR vs. 40,4), ce qui suggère à nouveau qu'il est incapable d'exploiter

#	Modèle	ICT	(-) diff.	MRR	P@1	P@20	Hits@20
a	DPR	NA	NA	**30,5**^{bcfg}	**21,2**^{bcfg}	**16,2**^{bcdfg}	**60,5**^{bcfg}
b	$\text{ECA}_V(l=6)$	✓	✗	22,8^{c}	13,5^{c}	9,4^{cf}	51,1^{c}
c	$\text{ECA}_V(l=0)$	✓	✗	18,3	9,9	8,0	47,5
d	$\text{ILF}_V(l=12)$	✓	✗	30,0^{bcfg}	19,6^{bcfg}	13,8^{bcfg}	60,4^{bcfg}
e	DPR_{V+R+A}	NA	NA	**36,0**^{abcdfg}	**26,7**^{abcdfg}	**17,1**^{bcdfg}	**65,2**^{abcdfg}
f	$\text{ECA}_{V+R+A}(l=6)$	✓	✓	21,1^{c}	11,9	8,7^{c}	50,8^{c}
g	$\text{ECA}_{V+R+A}(l=6)$	✓	✗	23,8^{cf}	14,6^{cf}	9,7^{cf}	52,7^{c}

TABLEAU 5.4 – Évaluation en aval, KVQAE sur ViQuAE, des modèles pré-entraînés au question-réponse sur TriviaQA mais *sans* ajustement sur ViQuAE. L'utilisation d'exemples négatifs difficiles pendant l'ICT multimodale est notée dans la colonne « (-) diff. »

les représentations d'ArcFace. Intuitivement, nous pensons que les représentations d'ArcFace sont diluées avec les autres dans ECA, ce qui n'est pas bénéfique pour la prise en compte des représentations plus spécialisées comme celles d'ArcFace. Nous avons mené plusieurs études d'ablation dont nous discutons à la section 6.1.

Sur-apprentissage et oubli catastrophique Discutons à présent des différents phénomènes liés à nos pré-entraînements : sur-apprentissage, oubli catastrophique, exemples négatifs difficiles et chevauchement entre les jeux d'entraînement et d'évaluation.

Il apparaît au tableau 5.3 que l'ICT multimodale est essentielle à ECA. Sans elle, les performances retombent au même niveau que la fusion tardive à cause du sur-apprentissage sur le petit jeu de données de ViQuAE (c vs. d). Cependant, l'ajustement sur ViQuAE est non moins essentiel à ECA. Sans lui, ECA tombe bien en dessous de la référence textuelle, ce qui témoigne du phénomène d'oubli catastrophique, s'appliquant ici au pré-entraînement sur TriviaQA (cf. tableau 5.4, b vs. a). Cet oubli semble atténué par le figement des dernières couches du modèle, comme le montre la différence entre les lignes b et c, mais l'ajustement sur ViQuAE reste nécessaire. Plus globalement, en amont sur WIT-ICT (tableau 5.1), nous avions l'ordonnancement suivant des modèles : $\text{ILF}_V(l=12) < \text{ECA}_V(l=6) < \text{ECA}_V(l=0)$. Ici, l'ordre tend à s'inverser : une meilleure ICT multimodale induit davantage d'oubli catastrophique.

Exemples négatifs difficiles Le tableau 5.4 montre que les exemples négatifs difficiles, générés par permutation des images pendant l'ICT multimodale, ne sont pas bénéfiques à la KVQAE (f vs. g). Puisque la section 5.1 suggérait qu'au contraire, ces exemples négatifs *difficiles* l'étaient effectivement pour plusieurs versions d'ECA, ces résultats viennent troubler le reste de nos conclusions : vaut-il mieux ignorer les interactions cross-modales pour traiter la KVQAE ? En tout cas, nous pouvons conclure que ces exemples négatifs « difficiles » ne sont pas une bonne approximation des vrais passages visuels de la BC.

#	Modèle	ICT	MRR	P@1	P@20	Hits@20
a	DPR	✗	22,2	13,4	10,5	46,9
b	$\mathrm{DPR}(l=0)$	T	23,5	14,7	10,6	$50{,}4^{a}$
c	$\mathrm{DPR}_V(l=0)$	T	$27{,}3^{ab}$	$18{,}7^{ab}$	11,4	$52{,}7^{a}$
d	$\mathrm{ECA}_V(l=0)$	MM	$28{,}3^{ab}$	$17{,}9^{ab}$	$13{,}1^{abc}$	$\mathbf{59{,}2}^{abc}$
e	$\mathrm{ILF}_V(l=12)$	MM	$\mathbf{29{,}4}^{ab}$	$\mathbf{18{,}9}^{ab}$	$\mathbf{13{,}4}^{abc}$	$58{,}9^{abc}$
f	$\mathrm{DPR}_{V+R+A}(l=0)$	T	$\mathbf{32{,}7}^{abcdeg}$	$\mathbf{23{,}3}^{abcdeg}$	$13{,}3^{abc}$	$\mathbf{58{,}8}^{abc}$
g	$\mathrm{ECA}_{V+R+A}(l=0)$	MM	$27{,}8^{ab}$	$16{,}9^{a}$	$\mathbf{13{,}4}^{abc}$	$58{,}3^{abc}$

TABLEAU 5.5 – Évaluation en aval, KVQAE sur ViQuAE, des modèles sans pré-entraînement au question-réponse sur TriviaQA mais *avec* ajustement sur ViQuAE. Ici, l'ICT peut être textuelle (T) ou multimodale (MM). Aucune couche du modèle n'est figée ($l = 0$) puisque l'on n'a pas à se soucier de l'oubli catastrophique. $l = 12$ pour ILF dans l'ICT multimodale, mais ILF est d'abord entraîné avec $l = 0$ à l'ICT textuelle.

Chevauchement entre WIT-ICT et ViQuAE Rien n'indique que le chevauchement de 14% d'images de ViQuAE avec WIT-ICT biaise les résultats. ECA fonctionne mieux sur le sous-ensemble en dehors de WIT-ICT (38,0 vs. 36,5 MRR pour $\mathrm{ECA}_V(l=6)$), mais c'est l'inverse pour DPR et la fusion tardive.

5.2.2 Ablation : sans pré-entraînement au question-réponse sur TriviaQA

La section précédente rapporte une amélioration des performances grâce à l'ICT multimodale mais notre stratégie de pré-entraînement en trois phases est complexe et engendre un oubli catastrophique que nous nous sommes efforcé de pallier en figeant une partie du modèle. D'autre part, de par la similarité entre TriviaQA — utilisé pour la première phase d'entraînement — et ViQuAE, la référence textuelle DPR est très efficace. Nous expérimentons donc ici dans un cadre plus simple sans pré-entraînement au question-réponse sur TriviaQA, donc avec une seule phase de pré-entraînement : l'ICT multimodale. Dans cette condition, DPR peut être entraîné pour l'ICT textuelle. ILF est alors initialisé à partir de ce DPR avant de poursuivre une ICT multimodale comme à la section précédente.

Nous observons au tableau 5.5 que la référence textuelle chute par rapport au pré-entraînement sur TriviaQA. De plus, l'ICT textuelle apporte un gain seulement léger par rapport au pré-entraînement de BERT seul (b vs. a). Ces résultats rejoignent (Ram et al., 2022) et (Zhou et al., 2022). On pourrait se demander si la probabilité de masquage (section 4.2.2) de 100% au lieu de 90% en est la cause, mais, empiriquement, cela ne change pas les résultats avec l'ICT textuelle.

Par ailleurs, ECA atteint dans cette condition des résultats comparables à la fusion tardive pour les modèles fondés sur CLIP (d vs. c), voire moins bons pour ceux fondés également sur ArcFace (g vs. f).

Dans cette condition, il est également clair qu'ILF surpasse ou du moins égale ECA (e vs. d), sans être remis en cause par un éventuel oubli catastrophique comme à la section précédente, puisque le pré-entraînement sur TriviaQA n'a pas lieu et n'est donc pas oublié. Cela renforce nos conclusions précédentes : l'interaction IQTP

est la plus importante.

Nous pouvons conclure de cette expérience que l'ICT n'est pas suffisante pour pré-entraîner nos modèles et que l'ICT multimodale interagit avec TriviaQA.

6 Discussion

6.1 Représentations des visages

Les résultats rapportés à la section précédente montrent que les représentations des images par ImageNet-ResNet, et surtout les représentations des visages par ArcFace, ne sont pas bénéfiques aux modèles de fusion précoce, notamment ECA, ou seulement de façon marginale pour l'ICT multimodale en amont, contrairement à la fusion tardive où ces représentations permettent une précision accrue. Intuitivement, nous pensons que les représentations d'ArcFace sont diluées avec les autres dans ECA et ne lui apportent pas de gain d'information. En effet, si l'on reconnaît l'entité-sujet dans l'image de la question, les autres modalités risquent davantage de brouiller cette reconnaissance que de la préciser, du moins pour ViQuAE. Nous avons démontré cela en fusionnant ECA et les représentations purement visuelles de manière tardive : le gain de performance est important, 43,3 de MRR vs. 37,7 pour la fusion précoce seulement.

Nos constats sur la prise en compte de représentations telles qu'Arface au niveau d'ECA s'appuient par ailleurs sur des expérimentations concernant plusieurs façons d'intégrer les visages ou d'entraîner ECA, sans obtenir néanmoins de meilleurs résultats que la version standard.

Tout d'abord, nous avons étudié si ECA était capable de traiter plusieurs visages (*quatre* dans nos expériences) en parallèle. La fusion tardive profite d'une heuristique où la représentation du visage est utilisée alternativement aux représentations de CLIP et ImageNet. Pour obtenir une version plus comparable, nous avons essayé une version d'ECA utilisant *un seul* visage (le plus probable), alternativement aux représentations de CLIP et ImageNet. Néanmoins, cette version détériore les résultats (34,8 de MRR vs. 37,7). On peut aussi noter que, bien que la version standard d'ECA soit inférieure à la fusion tardive quand *un seul et unique* visage est détecté dans l'image de la question (36,1 de MRR vs. 39,7), elle est meilleure lorsque plusieurs visages sont détectés (38,7 vs. 35,9), ce qui suggérerait qu'ECA est capable de traiter plusieurs visages en parallèle.

Enfin, nous avons vu que seul un quart des images de WIT-ICT comportaient un visage détecté. De plus, un visage peut être détecté dans l'image de la question mais pas dans celle du passage, ou vice-versa. Cela pourrait perturber le modèle, bien qu'il devrait être idéalement robuste à ces perturbations. Néanmoins, pré-entraîner ECA sur un sous-ensemble de WIT-ICT dans lequel au moins un visage est détecté des deux côtés (un cinquième de WIT-ICT complet) n'apporte pas de changement significatif (37,2 de MRR vs. 37,7).

6.2 Interaction TQIQ au sein de la question visuelle

Les résultats de la section précédente suggèrent qu'ECA surpasse la fusion tardive principalement grâce à l'interaction IQTP entre l'image de la question et le texte du passage. Ce résultat est assez surprenant et dévie de notre motivation première : modéliser l'interaction TQIQ au sein de la question visuelle.

Nous avons expérimenté différentes architectures et conditions d'entraînement, mais, plus fondamentalement, on peut s'interroger sur l'*Inverse Cloze Task* multimodale. En effet, en utilisant une phrase sortie de son contexte comme pseudo-question, l'image contextuelle, bien que souvent pertinente, n'a pas la même relation avec le texte que dans une vraie question visuelle. Reprenons l'exemple de la figure 5.4 : le texte a du sens indépendamment de l'image, bien que les deux soient liés. Au contraire, à la figure 5.1, le texte est ancré dans l'image : il y fait référence et est vide de sens sans elle. Ainsi, l'interaction TQIQ pour une pseudo-question visuelle ressemble plutôt à TPIP : l'interaction au sein du passage visuel.

De façon concurrente à notre travail, Hu et al. (2023c) ont proposé une tâche de pré-entraînement très similaire à l'ICT multimodale et aussi fondée sur WIT. Étant donné une image et sa légende, l'objectif est de retrouver le texte de la section appariée. Ainsi, le texte de la pseudo-question (la légende) est bien ancrée dans l'image, donc pourrait permettre une meilleure modélisation de TQIQ.

On pourrait aller plus loin et critiquer les données d'évaluation : ViQuAE. En effet, le jeu de données a été construit en visant à ce que chaque image représente une seule entité. Il y a donc peu d'images représentant plusieurs entités, par exemple plusieurs personnes, ou, comme à la figure 5.1, où la résolution d'expression référentielle ou plus généralement l'interaction TQIQ sont essentielles. On pourrait envisager d'ajouter des exemples adverses à ViQuAE, où, pour chaque question visuelle, on poserait une question à propos d'une autre entité également représentée dans l'image.

7 Conclusion

Nous avons contribué à une nouvelle méthode de pré-entraînement, l'*Inverse Cloze Task* multimodale, appliquée à des questions visuelles à propos d'entités nommées. Cette tâche se repose sur des documents multimodaux, en l'occurrence des articles Wikipédia, pour générer des pseudo-questions visuelles. Elle permet d'utiliser des modèles de fusion multimodale plus complexes que la fusion tardive proposée précédemment, ce qui mène à une amélioration de 10% relativement au MRR sur le jeu de données ViQuAE. Nous expliquons cette amélioration par la modélisation d'interactions cross-modales, négligées par la fusion tardive. Plus précisément, nos résultats suggèrent que l'interaction IQTP, entre l'image de la question et le texte du passage, est la plus importante et permettrait de prendre en compte l'hétérogénéité des représentations visuelles des entités nommées.

Nous avons mené nos expériences avec deux architectures différentes : ECA, qui fusionne les modalités de manière précoce à l'aide du mécanisme d'auto-attention des *transformer* ; et ILF, un modèle plus simple qui fusionne les modalités de façon moins précoce à travers des projections linéaires. Contrairement aux travaux connexes (cf.

chapitre 2), nous trouvons que les deux obtiennent des résultats comparables. Cela démontre d'une part, la robustesse de l'ICT multimodale et, d'autre part, confirme que l'interaction IQTP est la plus importante, puisqu'ILF ne peut modéliser les autres interactions cross-modales. Nous avons également fait face à un problème d'apprentissage séquentiel, où l'oubli catastrophique a partiellement été contourné en figeant les *dernières* couches d'ECA, contrairement à la pratique habituelle, ce qui renforce l'hypothèse d'Aytar et al. (2017).

Bien que le pré-entraînement soit essentiel pour ces modèles de fusion, nous avons vu que l'ajustement sur la tâche finale n'était pas à négliger par ailleurs. En cela, notre travail s'inscrit clairement dans le paradigme *pré-entraînement et ajustement* plutôt que *zero-shot*.

Ces résultats sont limités par leur manque de généralisation à de multiples représentations visuelles, notamment les représentations de visages par ArcFace, qui restent meilleures en étant fusionnées tardivement que de façon précoce. De ce point de vue, nous supposons que les représentations d'ArcFace sont diluées avec les autres dans ECA. Cela pourrait être causé par l'architecture d'ECA, l'ICT multimodale ou les données d'évaluation.

Nos expériences nous ont également montré que nous ne pouvons pas nous abstraire du pré-entraînement sur TriviaQA pour le question-réponse textuel dans la mesure où l'ICT, multimodale ou textuelle, s'est avérée insuffisante pour pré-entraîner un modèle de RI, multimodal ou textuel. Ces résultats rejoignent (Ram et al., 2022) et (Zhou et al., 2022), qui montrent que l'ICT textuelle apporte un gain seulement léger par rapport au pré-entraînement de BERT seul.

L'ICT multimodale pourrait être par ailleurs améliorée, par exemple en exploitant les légendes liées aux images, comme discuté à la section précédente. Une autre piste d'amélioration réside dans les données d'évaluation. En effet, beaucoup d'images de ViQuAE dépeignent une seule et unique entité, ce qui limite l'étude des interactions cross-modales.

Dans le prochain chapitre, nous proposons une nouvelle façon d'exploiter le modèle CLIP, jusqu'ici utilisé seulement pour ses représentations visuelles. En effet, motivés par les résultats sur l'interaction IQTP, nous proposons de la modéliser explicitement avec CLIP en utilisant ses représentations textuelles pour effectuer une recherche cross-modale.

Chapitre 6

Recherche cross-modale

1 Introduction

La recherche cross-modale vise à comparer le contenu sémantique de textes et d'images pour trouver des images pertinentes selon une requête textuelle, ou inversement, des textes pertinents selon une requête visuelle. Nous l'emploierons dans cette dernière direction puisque ce chapitre s'intéresse principalement à reconnaître l'entité présente dans l'image de la question $\mathbf{i_q}$. Il s'agit d'un problème essentiel pour la KVQAE puisqu'on ne peut répondre à une question visuelle sans reconnaître l'entité sur laquelle elle porte.

La recherche cross-modale a été étudiée selon différentes perspectives, comme évoqué au chapitre 2 :

— en tant que telle, puisque la RI cross-modale a de nombreuses applications, par exemple pour la navigation Web ou dans un cadre de recherche sémantique, dans des collections d'images personnelles (Yang et al., 2023) ;

— dans le cadre de l'apprentissage sans exemple (Frome et al., 2013; Radford et al., 2021) ;

— dans le cadre de l'apprentissage de langue visuellement ancré (Lazaridou et al., 2015).

Aux chapitres précédents, nous nous sommes intéressé à la recherche cross-modale comme moyen d'entraîner des représentations visuelles robustes et faiblement supervisées à travers le modèle CLIP (Radford et al., 2021). Nous nous intéresserons ici directement à la recherche cross-modale, motivée par les résultats du chapitre précédent, qui montrent que les modèles de fusion multimodale apprennent à exploiter l'interaction IQTP entre l'image de la question et le texte du passage. En particulier, certains exemples suggéraient que cette interaction permettait de traiter l'hétérogénéité des représentations visuelles d'entités nommées, comme dans le cas d'une statue de Churchill, jugée plus proche dans l'espace multimodal du texte « *Churchill* » que de sa photographie. Nous poursuivons cette étude à travers le modèle CLIP, pré-entraîné de façon cross-modale sur des légendes d'images et qui peut donc servir directement à la recherche cross-modale, contrairement à ECA et ILF qui devaient être pré-entraînés et ajustés. Nous traitons donc toujours de notre troisième et principale question de recherche : comment *interagissent les modalités* ?

En dehors des résultats présentés au chapitre précédent, notre étude a également été motivée par Sun et al. (2022), qui emploient CLIP pour la recherche cross-modale dans le cadre d'une tâche connexe à la KVQAE : la désambiguïsation visuelle d'entités nommées. Ils montrent ainsi que CLIP surpasse un modèle de reconnaissance faciale, même sans ajustement. Nos résultats suggèrent le contraire mais Sun et al. (2022) utilisent un modèle de reconnaissance faciale différent du nôtre et ne précisent ni la version ni la taille du modèle CLIP avec lequel ils expérimentent. Leur travail se focalise sur le jeu de données qu'ils proposent et n'emploie ainsi CLIP que de manière cross-modale, avec un ajustement laissant les encodeurs figés et donc, en ajoutant un perceptron multi-couches. Il reprend en cela le modèle CLIP-Adapter de Gao et al. (2021b), qui supposent que l'ajustement de l'ensemble des paramètres de CLIP conduiraient inévitablement au sur-apprentissage. Toutefois, Gao et al. (2021b) ne vérifient pas cette hypothèse et se comparent principalement aux approches de *prompting*. Au contraire de Sun et al. (2022), nous utilisons CLIP à la fois pour la recherche mono- et cross-modale et ajustons l'ensemble de ses paramètres sans en introduire de nouveaux, comme décrit à la section suivante.

Plus précisément, nous étudions comment traiter la KVQAE en faisant intervenir une recherche cross-modale dans la recherche visuelle, en plus de l'habituelle recherche mono-modale dans le cadre décrit au chapitre 4. Le modèle ECA du précédent chapitre offre également un autre point de vue sur ce travail : plutôt que de modéliser implicitement toutes les interactions cross-modales comme ce dernier, nous décomposons explicitement la similarité entre une question visuelle et un passage visuel en trois termes : deux termes monomodaux (texte-texte et image-image) et un terme cross-modal (image-texte), cf. section 2. Nous verrons les avantages et les inconvénients des recherches mono- ou cross-modales pour la recherche visuelle à partir de l'image, qui peuvent s'expliquer par le pré-entraînement cross-modal de CLIP. Nous étudierons également l'ajustement de CLIP, contraint, comme toujours, par la petite taille du jeu de données ViQuAE.

Ainsi, le modèle du présent chapitre s'appuie seulement sur BERT (ou DPR) et CLIP. Il est donc plus simple que celui du chapitre 4, qui utilise ArcFace pour représenter le visage des personnes, lesquelles bénéficient alors d'une meilleure représentation que les 979 autres types d'entités de ViQuAE, et que le système du chapitre précédent, qui demande un pré-entraînement coûteux. D'autre part, nous verrons que les différences de performance entre les systèmes varient selon que l'on évalue intrinsèquement la pertinence des passages résultant de la RI ou extrinsèquement, à travers l'extraction de réponse à partir de ces passages. En cela, ce chapitre traite également de notre première question de recherche : comment *évaluer* un système de KVQAE ? Nous discutons des différences entre ces métriques et des implications pour l'évaluation de la KVQAE à la section 6.

Par ailleurs, pour explorer l'impact sur la recherche cross-modale de la variété des représentations visuelles des entités nommées (deuxième question de recherche), nous avons mené des expérimentations visant à changer les images de notre BC afin d'inclure plusieurs images par entité et éventuellement la légende de ces images. Nous supposons que ces paires (image, légende) permettent d'obtenir une meilleure représentation multimodale des entités nommées. De plus, ces paires sont largement disponibles sur le Web, dans de nombreuses langues, ce qui s'avérerait utile pour

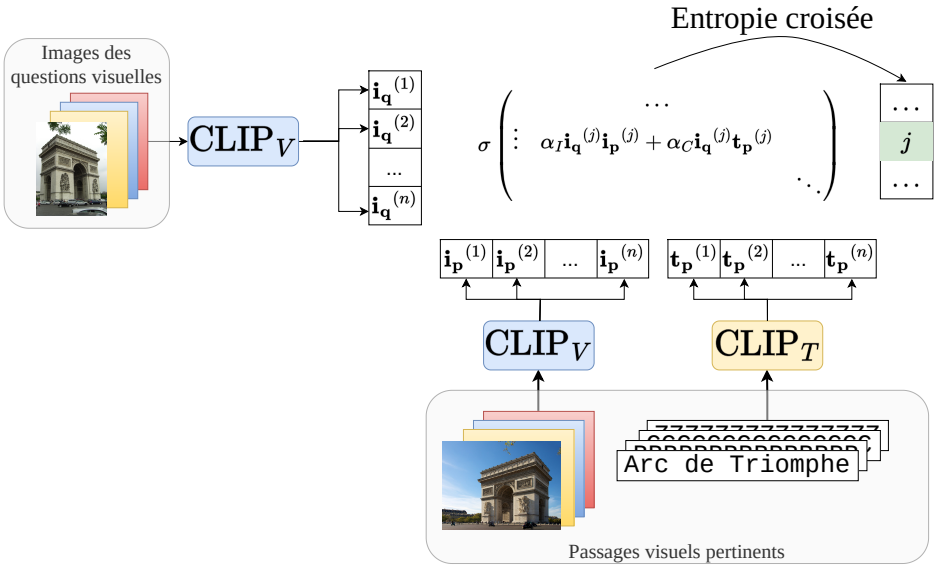


FIGURE 6.1 – Aperçu de l'apprentissage conjointement mono- et cross-modal de CLIP pour un lot de n questions visuelles $(\mathbf{t_q}, \mathbf{i_q})$ appariées à des passages visuels pertinents $(\mathbf{t_p}, \mathbf{i_p})$. Le terme $\mathbf{i_q}^{(j)}\mathbf{i_p}^{(j)}$ correspond à la similarité mono-modale qui modélise l'interaction IQIP tandis que $\mathbf{i_q}^{(j)}\mathbf{t_p}^{(j)}$ correspond à la similarité cross-modale qui modélise IQTP. La fusion des deux similarités est régie par les paramètres scalaires α_I et α_C.

entraîner un modèle et étudier la KVQAE dans une autre langue que l'anglais ou dans un cadre multilingue. Cependant, nous verrons à la section 5 que ces expériences sont rendues difficiles par la présence de quasi-doublons d'images de ViQuAE dans ces nouvelles images, phénomène que nous avions pu contrôler lors de l'annotation du jeu de données (cf. chapitre 3) et qui peut fausser les résultats.

2 Méthodes

Nous nous concentrons en premier lieu sur l'étape de recherche d'information avant de considérer l'extraction des réponses à sa suite à la section 4.3.

Étant donné une question visuelle $(\mathbf{t_q}, \mathbf{i_q})$ et une BC consistant en une collection de passages visuels $(\mathbf{t_p}, \mathbf{i_p})$, nous cherchons à trouver des passages pertinents, c'est-à-dire permettant de répondre à la question. Nous nous concentrons ici sur les interactions cross-modales entre les questions et les passages. Nous laissons donc de côté les interactions cross-modales au sein des questions (TQIQ) et des passages (TPIP). Par ailleurs, nous ne considérons pas non plus la similarité TQIP entre la question et l'image du passage dans ce cadre car nous jugeons que la spécification de l'entité par le biais de la seule partie textuelle de la question est très peu discriminante du point de vue de l'entité référencée. Par conséquent, nous nous focalisons sur la recherche visuelle à partir de l'image $\mathbf{i_q}$. Pour ce faire, nous définissons la fonction de similarité suivante, qui combine similarités mono- et cross-modale :

$$s(\mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = \alpha_I s_I(\mathbf{i_q}, \mathbf{i_p}) + \alpha_C s_C(\mathbf{i_q}, \mathbf{t_p}) \tag{6.1}$$

où les paramètres $\alpha_{\{I,C\}}$ pondèrent chaque similarité. Cette décomposition nous permet d'exploiter directement des modèles pré-entraînés. Plus précisément, dans ce chapitre, nous nous focalisons sur l'ajustement de CLIP pour implémenter $s_I(\mathbf{i_q}, \mathbf{i_p})$ et $s_C(\mathbf{i_q}, \mathbf{t_p})$. L'objectif est donc de rapprocher l'image de la question de l'image de cette entité dans la BC (*apprentissage mono-modal*) ou bien de son nom (*apprentissage cross-modal*) ou les deux de façon jointe (cf. figure 6.1).

2.1 Objectif d'apprentissage et modèles

Plus formellement, l'objectif sous-tendant notre modèle de RI est de maximiser $s(\mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p})$ si les deux images $\mathbf{i_q}$ et $\mathbf{i_p}^{(+)}$ représentent la même entité, nommée sous la forme textuelle $\mathbf{t_p}^{(+)}$, et de la minimiser sinon. Les données étant traitées par lot, ces entités négatives, pour lesquelles les représentations textuelles et visuelles sont notées respectivement $\mathbf{t_p}^{(j)}$ et $\mathbf{i_p}^{(j)}$, sont, dans une telle approche contrastive, constituées des autres entités du lot. Pour mettre en œuvre cette approche, nous optimisons de façon jointe $s_I(\mathbf{i_q}, \mathbf{i_p})$ et $s_C(\mathbf{i_q}, \mathbf{t_p})$ pour chaque image $\mathbf{i_q}$ du lot en minimisant l'objectif suivant [1], étant donné τ la température :

$$-\log \frac{\exp\left(s(\mathbf{i_q}, \mathbf{t_p}^{(+)}, \mathbf{i_p}^{(+)})e^{\tau}\right)}{\exp\left(s(\mathbf{i_q}, \mathbf{t_p}^{(+)}, \mathbf{i_p}^{(+)})e^{\tau}\right) + \sum_j \exp\left(s(\mathbf{i_q}, \mathbf{t_p}^{(j)}, \mathbf{i_p}^{(j)})e^{\tau}\right)} \tag{6.2}$$

Puisque nous implémentons $s_C(\mathbf{i_q}, \mathbf{t_p})$ avec CLIP, en notant son encodeur visuel CLIP_V et textuel CLIP_T, nous obtenons :

$$s_C(\mathbf{i_q}, \mathbf{t_p}) = \cos\left(\mathrm{CLIP}_V(\mathbf{i_q}), \mathrm{CLIP}_T(\mathbf{t_p})\right) \tag{6.3}$$

Cet objectif correspond à celui utilisé pour le pré-entraînement de CLIP si $\alpha_I = 0$ et $\alpha_C = 1$ (apprentissage cross-modal seulement), sauf qu'il est asymétrique (la fonction softmax exprime les probabilités selon $\mathbf{i_q}$ et pas selon $\mathbf{t_p}$). Puisque $\mathbf{i_q}$, $\mathbf{t_p}$ et $\mathbf{i_p}$ sont encodés indépendamment, cet objectif permet d'exploiter toutes les autres images et textes du lot de manière très efficace (il suffit d'un produit matriciel pour calculer le dénominateur de l'équation 6.2). Nous implémentons $s_I(\mathbf{i_q}, \mathbf{i_p})$ de manière similaire : $\cos\left(\mathrm{CLIP}_V(\mathbf{i_q}), \mathrm{CLIP}_V(\mathbf{i_p})\right)$, comme au chapitre 4.

Les résultats de cette recherche visuelle peuvent être combinés avec la recherche textuelle $s_T(\mathbf{t_q}, \mathbf{t_p})$, en redéfinissant s de la façon suivante :

$$s(\mathbf{t_q}, \mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = \alpha_T s_T(\mathbf{t_q}, \mathbf{t_p}) + \alpha_I s_I(\mathbf{i_q}, \mathbf{i_p}) + \alpha_C s_C(\mathbf{i_q}, \mathbf{t_p}) \tag{6.4}$$

Nous discutons des difficultés à optimiser ces trois similarités de façon jointe à la section 3. De ce fait, $s_T(\mathbf{t_q}, \mathbf{t_p})$ est implémenté par un modèle entraîné indépendamment et les poids $\alpha_{\{T,I,C\}}$ sont déterminés par une recherche exhaustive sur le jeu de validation pour maximiser le rang réciproque moyen en contraignant leur somme à 1. Nous notons aussi cette fonction de similarité DPR_{V+T} car elle est fondée sur DPR, CLIP_V et CLIP_T, ou $\mathrm{DPR}_{\mathbf{V+T}}$ (en gras) lorsque CLIP est ajusté.

1. Déjà utilisé aux chapitres précédents pour DPR, ILF et ECA, sauf que s y était défini différemment et que nous avions $\tau = 0$.

2.2 Systèmes de références

Le premier système du chapitre 4, DPR_{V+R+A}, combine DPR, CLIP_V, ArcFace et un modèle ResNet entraîné sur ImageNet. Rappelons brièvement que DPR est fondé sur deux encodeurs BERT : un pour la question et un pour le passage. Dans notre cas, il implémente $s_T(\mathbf{t_q}, \mathbf{t_p}) = \mathrm{BERT}_q(\mathbf{t_q})_{[\mathrm{CLS}]} \cdot \mathrm{BERT}_p(\mathbf{t_p})_{[\mathrm{CLS}]}$. Les résultats des quatre modèles sont combinés de la même façon qu'à l'équation 6.4, où DPR implémente $s_T(\mathbf{t_q}, \mathbf{t_p})$; CLIP_V, ArcFace et ImageNet composent $s_I(\mathbf{i_q}, \mathbf{i_p})$ et il n'y a pas de similarité cross-modale ; donc $s_C(\mathbf{i_q}, \mathbf{t_p}) = 0$. Plus précisément, ArcFace est utilisé de manière alternative à CLIP_V et ImageNet, seulement lorsqu'un visage est détecté. La recherche est alors effectuée sur les seules entités nommées de type personne de la BC, en supposant que les visages sont pertinents seulement pour les personnes. Formellement, en notant ArcFace A, CLIP_V V, ImageNet R, $F \in \{0, 1\}$ la détection d'un visage dans $\mathbf{i_q}$ et $\mathbf{i_p}$ et $H \in \{0, 1\}$ si $\mathbf{i_p}$ correspond à une personne, nous obtenons :

$$s_I(\mathbf{i_q}, \mathbf{i_p}) = FH\alpha_A s_A(\mathbf{i_q}, \mathbf{i_p}) + (1-F)(1-H)(\alpha_V s_V(\mathbf{i_q}, \mathbf{i_p}) + \alpha_R s_R(\mathbf{i_q}, \mathbf{i_p})) \quad (6.5)$$

Nous considérons également les modèles $\mathrm{ECA}_V(l = 6)$ et $\mathrm{ILF}_V(l = 12)$ du chapitre précédent (fondés seulement sur CLIP pour la représentation visuelle), que nous notons plus simplement ECA_V et ILF_V ici. Rappelons qu'ECA fusionne les modalités de manière précoce à l'aide d'un mécanisme d'attention. La similarité est donc calculée en s'appuyant sur la représentation de la question et du passage suivant $s(\mathbf{t_q}, \mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = \mathrm{ECA}(\mathbf{t_q}, \mathbf{i_q}) \cdot \mathrm{ECA}(\mathbf{t_p}, \mathbf{i_p})$ et combine ainsi toutes les interactions multimodales étudiées. ILF fusionne les modalités avec une simple projection linéaire et n'a donc, comme notre méthode, ni interaction TQIQ ni interaction TPIP puisque la similarité s'y réduit à :

$$s(\mathbf{t_q}, \mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = s_T(\mathbf{t_q}, \mathbf{t_p}) + s_{C'}(\mathbf{t_q}, \mathbf{i_p}) + s_I(\mathbf{i_q}, \mathbf{i_p}) + s_C(\mathbf{i_q}, \mathbf{t_p}) \quad (6.6)$$

Les différentes interactions mono- et cross-modales utilisées par les modèles étudiés sont résumées au tableau 6.1.

Il est à noter qu'aux chapitres précédents, nous utilisions CLIP_V avec l'architecture ResNet tandis que nous avons opté ici pour ViT (Dosovitskiy et al., 2021) pour des raisons d'implémentation (mais nous comparons les deux à la section 4 sans trouver de différence significative) [2].

3 Implémentation

3.1 Données

Nous utilisons principalement la BC décrite au chapitre 3, qui consiste en 1,5 million d'articles Wikipédia et images des entités Wikidata correspondantes. Nous étudierons une autre BC à la section 5. Les articles sont divisés en 12 millions de passages de 100 mots. Par conséquent, tous les passages d'un même article partagent

2. Plus précisément, il s'agit de `RN50` et `ViT-B/32`, disponibles à https://github.com/openai/CLIP

Modèle	Mono-modal		Cross-modal		
	TQTP	IQIP	TQIQ	TPIP	IQTP
DPR (référence)	✓				
DPR_V (référence)	✓	✓			
DPR_{V+R+A} (chapitre 4)	✓	✓			
ECA_V (chapitre 5)	✓	✓	✓	✓	✓
ILF_V (chapitre 5)	✓	✓			✓
DPR_{V+T} (ce chapitre)	✓	✓			✓

TABLEAU 6.1 – Récapitulatif des différentes interactions mono- et cross-modales utilisées par les modèles étudiés.

la même image. La section 4 présente donc une évaluation de la RI à deux niveaux : article et passage.

Pour mémoire, le recouvrement entre les entités du jeu d'entraînement et de test de ViQuAE est très faible, seulement de 18%. Nos modèles doivent donc apprendre à généraliser non seulement à de nouvelles images mais aussi à de nouvelles entités.

3.2 Problème de l'annotation de référence

Comme nous l'avons indiqué à la section 2.1, l'apprentissage joint de DPR et CLIP (équation 6.4) s'est avérée difficile et nous avons finalement opté pour un apprentissage disjoint des deux modèles. Nous expliquons cette difficulté par l'annotation des passages visuels pertinents, trop bruitée pour entraîner un modèle visuel. Plus précisément, tous les passages visuels du même article partageant *la même image*, celle-ci peut être considérée comme pertinente ou non pertinente selon *le texte qui lui est associé*. De plus, la même image (ou deux images de la même entité) peut illustrer deux articles différents, donc encore une fois avoir une pertinence variable selon le texte associé. À l'inverse, on peut trouver un passage pertinent dans un autre article que celui de l'entité-sujet, donc illustré par une image très différente, mais qui sera alors considérée comme pertinente. À cause de ces difficultés, nous avons opté pour une autre annotation, indépendante du passage. Il est néanmoins intéressant de constater que nous avons réussi, au chapitre précédent, à entraîner ECA et ILF avec cette même annotation bruitée des passages. Ce succès pourrait être expliqué par la représentation jointe d'ECA (qui modélise TQIQ et TPIP) ou par l'expressivité d'ECA et d'ILF. Une autre explication, pas nécessairement incompatible, serait liée à l'interaction IQTP, car ECA et ILF considèrent le *passage entier* tandis que CLIP n'est appliqué qu'au *titre* de l'article. Pour reprendre notre exemple motivant ce travail : le passage mentionnant Churchill provient justement de l'article *Bleinheim Palace*. CLIP devrait donc le juger dissimilaire à la photographie de Churchill tandis qu'ECA et ILF peuvent exploiter le texte entier du passage mentionnant *Churchill*.

À la place de cette annotation au niveau du passage, nous utilisons l'annotation au niveau de l'entité fournie dans ViQuAE car chaque question visuelle porte sur une seule et unique entité. Pour ce faire, nous retirons 25 questions visuelles du jeu

d'entraînement de ViQuAE pour le réduire à 1 165 car les entités correspondantes sont absentes de la BC, faute d'images libres de droit.

3.3 Hyperparamètres

Pour profiter au mieux des entités $\mathbf{t_p}^{(j)}$ et $\mathbf{i_p}^{(j)}$ associées aux autres images des questions visuelles du lot, nous utilisons un lot de la plus grande taille possible, ici 1 165 triplets $(\mathbf{i_q}, \mathbf{t_p}^{(+)}, \mathbf{i_p}^{(+)})$, soit l'intégralité du jeu d'entraînement. Nous utilisons une seule GPU Nvidia V100 avec 32 Go de mémoire vive. La grande taille de lot est en partie permise par le *gradient checkpointing*.

Puisque le jeu d'entraînement est petit, l'entraînement est très peu coûteux : notre meilleur modèle converge [3] au bout de 11 époques/itérations, en moins de 15 minutes, ce qui est négligeable par rapport au pré-entraînement de 8 000 itérations en trois jours présenté au chapitre précédent.

Nous utilisons un taux d'apprentissage très faible, de 2×10^{-6}, croissant linéairement pendant 4 époques puis décroissant pendant 46 époques, si l'entraînement n'est pas interrompu avant. L'optimisation est faite avec AdamW (Loshchilov et Hutter, 2019), avec $\lambda = 0{,}1$. Pour l'apprentissage joint, nous initialisons $\alpha_I = \alpha_C = 0{,}5$ et leur assignons un taux d'apprentissage de $0{,}02$, beaucoup plus grand que le reste du modèle. À l'instar de Radford et al. (2021), la température τ reste entraînable mais, étant donné le faible taux d'apprentissage, elle reste proche de sa valeur initiale, soit $4{,}6$ [4]. Ces hyperparamètres ont été déterminés manuellement sur le jeu de validation.

L'entraînement est interrompu et le meilleur modèle sélectionné selon le meilleur rang réciproque moyen au sein du lot sur le jeu de validation, c'est-à-dire en réordonnant les images ou textes du lot selon le score de similarité s, pour éviter de calculer les représentations de toute la BC à chaque époque.

4 Résultats

Nous évaluons la RI à deux niveaux :

— article (qui contient plusieurs passages), à la section 4.1 ;

— passage, à la section 4.2, afin de pouvoir nous comparer aux méthodes proposées précédemment.

Une fois la RI effectuée au niveau du passage visuel, les réponses sont extraites à l'aide du modèle BERT multi-passage (cf. chapitre 4). Ces résultats sont présentés à la section 4.3.

4.1 Recherche d'information au niveau de l'article

Nous explorons dans un premier temps trois modes d'apprentissage et trois manières d'utiliser CLIP au travers d'expériences menées sur le jeu de validation. Ces trois modes peuvent être décrits à partir de l'équation 6.1 :

3. C'est-à-dire commence à sur-apprendre.

4. Nous avons gardé la formulation de Radford et al. (2021) mais la température est habituellement exprimée sous la forme $\frac{1}{\tau'}$ et non pas e^{τ}, ce qui équivaudrait à $\tau' = \frac{1}{100}$ ici.

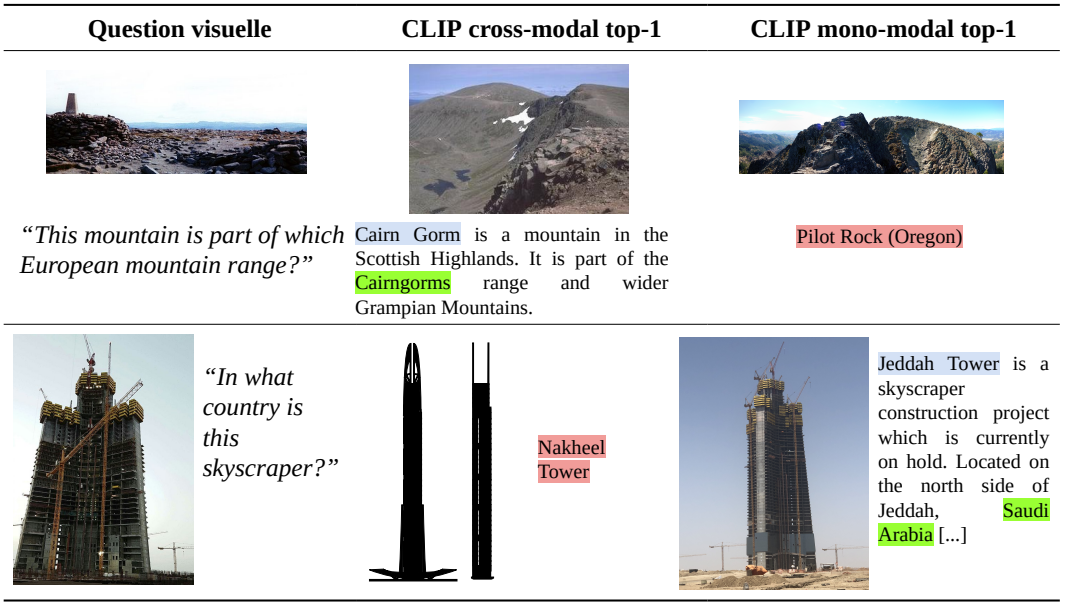


FIGURE 6.2 – Exemples des forces et faiblesses des recherches mono- et cross-modales. Les questions sont issues du jeu de validation de ViQuAE. Le texte de la question est montré pour référence mais n'est pas exploité par ces modèles. De la même façon, l'image et le texte du passage visuel (ou le titre de l'article) sont systématiquement montrés mais CLIP cross-modal n'exploite que le titre de l'article et CLIP mono-modal que son image. Il s'agit ici des versions non-ajustées de CLIP.

— recherche/apprentissage mono-modal entre les deux images, soit $\alpha_I = 1, \alpha_C = 0$;

— recherche/apprentissage cross-modal entre l'image et le nom de l'entité, soit $\alpha_I = 0, \alpha_C = 1$;

— recherche hybride ou apprentissage joint, soit $\alpha_I > 0, \alpha_C > 0$.

Rappelons que le mode d'apprentissage ne restreint pas le mode de recherche, comme en témoigne le tableau 6.2. Pour mémoire, CLIP est pré-entraîné de manière cross-modale uniquement (Radford et al., 2021).

Recherche mono- ou cross-modale Avant de comparer les différentes méthodes d'apprentissage, nous pouvons d'ores et déjà remarquer au niveau du tableau 6.2 que la RI cross-modale est systématiquement supérieure [5] à la RI mono-modale, notamment sans ajustement, ce qui peut paraître surprenant puisque les noms propres portent a priori peu de sémantique. On s'étonne donc que CLIP parvienne à généraliser [6] la représentation d'entités à partir de leurs seuls noms. Néanmoins, certains noms sont tout de même porteurs de sens. Par exemple, un nom peut indiquer le genre d'une personne ou suggérer sa nationalité [7]. De plus, nous travaillons ici avec

5. Significativement selon le test de randomisation de Fisher avec $p \leq 0{,}01$.

6. À moins que son jeu de pré-entraînement ne contienne suffisamment d'entités de ViQuAE pour que ce ne soit pas nécessaire.

7. Une visualisation interactive est disponible à l'adresse suivante : https://PaulLerner.github.io/ViQuAE/#image.

Recherche	Apprentissage	MRR	P@1	P@20	Hits@20
Mono-modale (IQIP)	✗	29,4	21,8	9,1	53,4
	Mono-modal	30,0	21,8	9,2	55,7
	Cross-modal	29,8	21,4	**9,5**	54,7
	Joint	**30,4**	**22,0**	**9,5**	**55,8**
Cross-modale (IQTP)	✗	32,7	23,1	10,9	60,6
	Mono-modal	31,6	21,9	10,9	59,6
	Cross-modal	**37,1**	**26,9**	**11,9**	**67,8**
	Joint	30,8	21,3	10,4	59,5
Hybride	✗	39,6	30,6	11,8	63,9
	Mono-modal	40,1	31,8	11,6	63,6
	Cross-modal	***44,1***	***34,9***	***12,7***	***69,9***
	Joint	41,0	32,6	11,6	64,9
	Disjoint	43,7	34,5	***12,7***	***69,9***

TABLEAU 6.2 – Évaluation des différentes méthodes d'ajustement de CLIP (ainsi que la version non-ajustée pour référence) sur le sous-ensemble de validation de ViQuAE pour la recherche visuelle (à partir de l'image de la question $\mathbf{i}_q$). L'évaluation est faite ici au niveau de l'article. Pour chaque recherche (mono- ou cross-modale), les meilleurs résultats sont marqués en gras. Les meilleurs résultats au total sont obtenus par la fusion des deux types de recherche et sont marqués en gras italique. La condition apprentissage disjoint pour la recherche hybride dénote la fusion d'une recherche mono-modale apprise de manière mono-modale et, réciproquement, d'une recherche cross-modale apprise de manière cross-modale.

les titres des articles Wikipédia, qui sont également susceptibles de contenir la nature de l'entité (par exemple la profession d'une personne ou le type de monument). Ces caractéristiques peuvent ainsi être mises en correspondance avec des attributs visuels. Enfin, nous attribuons principalement le succès de la RI cross-modale à son adéquation avec le pré-entraînement de CLIP : l'espace de représentation de CLIP est organisé pour rapprocher textes et images similaires ; la proximité mono-modale des images n'en est qu'une conséquence indirecte. Nous montrons des exemples de réussite et d'échec à la figure 6.2. En accord avec les résultats du chapitre précédent, nous constatons que la recherche mono-modale est plus sensible aux détails superficiels des images (photographie en couleur ou noir et blanc, pose du sujet…). Ici, les deux photographies au sommet de deux montagnes, montrant l'horizon, sont jugées similaires bien qu'il s'agisse de montagnes différentes. La recherche mono-modale est en revanche plus efficace dans le second exemple, où les deux photographies de la Tour de Djeddah sont prisent de points de vue semblables. Ces exemples renforcent notre hypothèse selon laquelle la recherche cross-modale permet de traiter l'hétérogénéité des représentations visuelles des entités nommées.

Complémentarité des deux recherches La condition *Hybride* montre, avec une simple combinaison de leurs résultats au niveau du score (comme à l'équation 6.1),

que les recherches mono- et cross-modales sont complémentaires. Pour l'apprentissage joint, nous pourrions utiliser directement les poids α appris par descente de gradient avec le reste du modèle sur le jeu d'entraînement, mais cela détériore légèrement les résultats. La comparaison avec les autres méthodes d'apprentissage serait alors injuste. Ainsi, sans ajustement, la fusion des deux recherches apporte une amélioration relative de 32% en P@1 par rapport à la recherche cross-modale seulement (significatif avec $p \leq 0{,}01$). Il serait intéressant d'étudier si ces résultats se généralisent à d'autres tâches. Cette méthode pourrait par exemple bénéficier à la recherche visuelle par le contenu dans un contexte de navigation Web.

Différents modes d'apprentissage Nous étudions ici l'amélioration des différents modes d'apprentissage par rapport à la version non-ajustée de CLIP. On peut voir que l'apprentissage cross-modal améliore légèrement la recherche mono-modale mais pas l'inverse. Dans les trois cas, l'apprentissage dans un mode améliore au moins la recherche dans le même mode. Il est intéressant de noter que l'apprentissage joint détériore la RI cross-modale mais améliore la fusion. Mais au total, l'apprentissage cross-modal semble être la meilleure option, surpassant significativement l'apprentissage mono-modal. Nous l'expliquons encore une fois largement par son adéquation avec le pré-entraînement de CLIP : nous manquons probablement de données pour réorganiser l'espace de représentations. Conséquemment, nous supposons que l'optimisation de la similarité mono-modale s_I pénalise l'apprentissage joint (équation 6.1). On voit également que les différences entre les modes d'apprentissage de la recherche mono-modale sont très faibles. Par conséquent, il n'est pas bénéfique de combiner la recherche mono-modale apprise de manière mono-modale et la recherche cross-modale apprise de manière cross-modale (« apprentissage disjoint » dans le tableau 6.2), puisque l'apprentissage cross-modal du même modèle fonctionne mieux. Ainsi, dans la suite du chapitre nous utilisons le modèle entraîné de manière cross-modale et présentons les résultats sur le jeu de test.

4.2 Recherche d'information au niveau du passage visuel

Les résultats sont présentés au tableau 6.3. Nous utilisons comme référence DPR (recherche avec le texte seulement) ainsi que sa fusion avec la recherche mono-modale de CLIP non-ajusté, notée DPR_V. Puisque ces résultats, ainsi que ceux des autres modèles présentés précédemment, sont fondés sur l'architecture ResNet pour CLIP, nous ajoutons également les résultats obtenus avec l'architecture ViT, utilisée dans le reste de nos expériences. Les deux architectures fournissent des résultats similaires.

Notre méthode améliore la P@1 de 3% relativement à DPR_{V+R+A}, sans utiliser ArcFace, ni ImageNet, ni l'heuristique de la division de la BC entre personnes et non-personnes, et de 7% relativement aux modèles du chapitre 5, sans pré-entraînement supplémentaire. Les différences de MRR avec ces modèles sont très faibles, mais ECA_V et ILF_V surpassent notre méthode en P@20 et Hits@20, ce qui suggérerait un avantage de la représentation jointe d'ECA_V et de l'expressivité d'ECA_V et d'ILF_V.

On peut voir que les améliorations par rapport à la référence DPR_V sont assez

Modèle	MRR	P@1	P@20	Hits@20
DPR (référence)	32,8	22,8	16,4	61,2
DPR_V (référence)	34,5	24,8	15,8	61,8
DPR_V* (référence)	34,7	24,3	16,0	62,8
DPR_{V+R+A} (chapitre 4)	**37,9**	27,8	17,5	65,7
ECA_V (chapitre 5)	37,8	26,7	**19,5**	**67,6**
ILF_V (chapitre 5)	37,3	26,8	19,1	66,9
$\mathrm{DPR}_{\mathbf{V+T}}$* (ce chapitre)	37,6	**28,6**	16,3	63,6

TABLEAU 6.3 – Résultats de la RI évaluée au niveau du passage visuel sur le jeu de test de ViQuAE. *CLIP_V fondé sur l'architecture ViT au lieu de ResNet.

Recherche d'information	Correspondance exacte	F1
DPR (référence)	23,1	25,6
DPR_V (référence)	24,3	27,1
DPR_V* (référence)	26,4	29,1
DPR_{V+R+A} (chapitre 4)	29,4	32,2
ECA_V (chapitre 5)	26,0	29,2
ILF_V (chapitre 5)	28,0	31,3
$\mathrm{DPR}_{\mathbf{V+T}}$* (ce chapitre)	**30,9**	**34,3**

TABLEAU 6.4 – Résultats de l'extraction des réponses sur le jeu de test de ViQuAE. Le modèle d'extraction prend en entrée le top-24 des différents systèmes de RI. *CLIP_V fondé sur l'architecture ViT au lieu de ResNet.

modestes, beaucoup plus faibles qu'à la section précédente où nous étudiions les résultats au niveau de l'article et avant la fusion avec DPR. Nous verrons à la section suivante que l'impact sur l'extraction des réponses est, lui, plus important, et démontre la supériorité de notre approche.

4.3 Extraction de réponse

Nous conservons le modèle du chapitre 4 pour extraire la réponse des passages résultant de la RI. Pour rappel, ce modèle est purement textuel et utilise l'architecture BERT multi-passage. Il est pré-entraîné sur TriviaQA puis ajusté sur ViQuAE de façon analogue à DPR. Il prend 24 passages en entrée, lesquels varient selon les systèmes de RI. Ainsi, on peut le considérer comme une évaluation extrinsèque des systèmes de RI.

Les résultats sont présentés au tableau 6.4. On voit que la recherche cross-modale et l'ajustement de CLIP apportent 23% à 25% d'amélioration relativement à la référence DPR_V selon les métriques, ce qui est plus cohérent avec les résultats de la RI au niveau de l'article (section 4.1) où nous avions alors 31% à 60% d'amélioration relative selon les métriques (avant la fusion avec DPR) [8]. Ainsi, notre méthode ap-

8. La section 4.1 présente les résultats sur le jeu de validation mais nous obtenons des résultats similaires sur le jeu de test.

porte des améliorations appréciables, de 12% à 20% selon les métriques et modèles : par rapport à DPR_{V+R+A}, sans utiliser ArcFace, ni ImageNet, ni l'heuristique de la division de la BC entre personnes et non-personnes, et par rapport aux modèles du chapitre 5, sans pré-entraînement supplémentaire.

5 Représentations visuelles multiples

5.1 Introduction

Les sections précédentes présentent une méthode simple et efficace pour traiter la KVQAE mais qui s'appuie sur une seule représentation visuelle par entité. Nous menons ici une étude exploratoire complémentaire pour enrichir la BC avec plusieurs représentations visuelles par entité. De plus, dans notre cadre de recherche cross-modale, nous étudierons l'utilisation de la légende de ces images, pouvant être utilisée pour représenter l'entité plus précisément que son seul nom et surtout, de façon plus proche du pré-entraînement de CLIP. Par ailleurs, ces légendes peuvent être utilisées pour ajuster un modèle cross-modal tel que CLIP afin de le rendre plus sensible à la représentation des entités nommées.

5.2 Méthodes et implémentation

Pluralité des représentations par entité Pour la RI, les méthodes sont identiques au reste du chapitre, mis à part le traitement de plusieurs images et légendes par article de la BC WIT (section 5.3). Dans ce cas, le score assigné à l'article est le score maximal de similarité entre la question visuelle et les différentes images (ou légendes) de l'article.

Ajustement de CLIP Pour l'ajustement sur WIT-légendes (présenté à la section 5.4), qui est donc seulement cross-modal, nous gardons le même objectif que pour le pré-entraînement de CLIP, c'est-à-dire comme l'équation 6.2 mais avec une normalisation softmax symétrique.

Le modèle est ajusté pendant 45 000 itérations avec un lot de 320 paires (image, légende), soit un peu plus de 2 époques, avec les mêmes hyperparamètres que précédemment, mis à part le *gradient checkpointing*, qui est désactivé.

5.3 Base de connaissances WIT

Nous exploitons encore une fois le jeu de données WIT, cette fois pour récupérer les paires (image, légende) au sein d'un article Wikipédia[9], tout en conservant le texte complet de l'article provenant de KILT, tandis que notre BC habituelle comprend une seule image par article : celle associée à l'entité correspondante dans Wikidata. Le sous-ensemble en langue anglaise de WIT contient 8 millions de paires (image, légende) provenant de 2 millions d'articles Wikipédia. Nous filtrons 29 000 paires dupliquées au sein de ces mêmes articles.

9. WIT contient également la légende de l'image sur Commons que l'on conserve si elle n'est pas identique à celle de l'article Wikipédia.

Puisque les images de ViQuAE proviennent également de Wikimédia Commons (cf. chapitre 3), il est important de retirer les quasi-doublons présents dans la BC, faute de quoi la recherche visuelle serait rendue triviale. Par exemple, l'article du Queen Elizabeth 2 [10] contient la même image que celle utilisée pour l'exemple de la figure 3.1, avec la légende « *QE2 in Dubai with Cunard titles removed from her superstructure* ». Les quasi-doublons ont de multiples définitions. Nous suivons celle de Pizzi et al. (2022) : une image est le quasi-doublon d'une autre si les deux proviennent de la même image source en deux dimensions.

Pour filtrer les quasi-doublons, nous utilisons le système de Gadeski et al. (2017), fondé sur une méthode de *locality sensitive hashing*. Nous fixons un seuil de 168 pour la distance de Hamming de façon à n'avoir aucun faux négatif dans le top-5 de 50 images du jeu de test de ViQuAE. Ce seuil filtre 67 000 images uniques, soit 142 000 paires (image, légende) de la BC WIT dont nous avions au préalable filtré quelques images illisibles et 4 000 URLs chevauchant celles de ViQuAE. Ces images proviennent de 1 960 000 articles Wikipédia dont 1 870 000 sont présents dans la base textuelle KILT. Cette intersection sert donc finalement de BC, que l'on nomme simplement BC WIT par la suite, et compte 7 550 000 paires (image, légende).

5.4 Jeu de légendes d'images

En parallèle de la BC WIT, nous avons constitué un jeu de données appariant des images à leur légende, sans nous soucier de l'intersection avec KILT, afin d'ajuster CLIP. Une autre différence importante avec la BC WIT est que nous dédupliquons ici les paires (image, légende) pour en garder 6 580 000 uniques, alors que certaines paires (image, légende) se répètent *dans différents articles* de la BC. Ce jeu de données, que l'on nomme WIT-légendes, est divisé aléatoirement en trois sous-ensembles, avec 90% de données pour l'entraînement et 5% chacun pour la validation et le test. Comme au chapitre précédent, le test sert seulement à mieux comprendre les résultats sur ViQuAE mais n'est pas notre intérêt premier.

5.5 Résultats

En amont, sur WIT-légendes Comme au chapitre précédent, faute de jugement de pertinence nous évaluons les résultats en amont avec les métriques *au sein du lot*. Les lots en question sont constitués de 512 paires (image, légende). Nous remarquons que la direction de la recherche cross-modale (image vers texte ou texte vers image) n'a pas d'importance ici puisque les similarités fournies par CLIP sont symétriques. Le tableau 6.5 montre simplement que l'ajustement est efficace, bien que CLIP sans ajustement fournisse une bonne performance.

Protocole expérimental Comme à la section 4.1, nous présentons aux tableaux 6.6 et 6.7 les résultats sur le jeu de validation de ViQuAE afin d'explorer les différents hyperparamètres du modèle sans compromettre le jeu de test.

10. https://en.wikipedia.org/w/index.php?title=Queen_Elizabeth_2&oldid=1154536883

Ajustement	MRR	P@1
✗	74,4	65,6
✓	86,8	80,1

TABLEAU 6.5 – Résultats de CLIP en amont, recherche cross-modale sur WIT-légendes, avec ou sans ajustement.

Avec la BC habituelle avec une image par entité Les résultats sont présentés au tableau 6.6. On voit que l'ajustement de CLIP sur WIT améliore la recherche cross-modale mais détériore la mono-modale, par rapport à CLIP non-ajusté [11]. L'amélioration finale pour la recherche hybride est donc décevante. Cette méthode d'ajustement semble donc moins intéressante que celle présentée précédemment, d'autant plus qu'elle est plus coûteuse (45 000 itérations, soit 30h avec une GPU V100). Ces résultats sont présentés pour avoir un cadre d'expérimentation proche de la section 4 mais nous étions en premier lieu intéressés par l'enrichissement de la BC avec davantage de représentations visuelles.

Images de la BC WIT Afin de distinguer les différences dues au changement d'images et celles liées à l'ajout de nouvelles, nous menons une étude d'ablation avec un sous-ensemble de la BC WIT qui contient une seule paire (image, légende) par article, en priorité celle de l'*infobox*. Cela n'a pas d'effet pour le titre qui est toujours unique à chaque article. On peut donc déjà observer une baisse de performance entre la BC habituelle et ce sous-ensemble de la BC WIT, en comparant les tableaux 6.7 et 6.6. La différence pour la recherche image-titre est minime car la BC WIT ne rajoute que quelques centaines de milliers de titres distracteurs, mais la recherche mono-modale chute de façon importante. Nous en concluons que les images de la BC WIT, qui proviennent des *infobox* des articles Wikipédia, seraient moins représentatives des entités nommées que celles de la BC habituelle, qui proviennent de Wikidata.

Représentations textuelles pour la recherche cross-modale : titre ou légende ? Sur le sous-ensemble avec une seule image, nous observons qu'il est préférable d'utiliser le titre de l'article plutôt que la légende de l'image. Toutefois, les résultats principaux du tableau 6.7, avec plusieurs images par article, montrent qu'il est bénéfique de combiner plusieurs représentations visuelles et textuelles par entité. Cependant, ces résultats sont optimistes, au moins pour la recherche mono-modale, à cause des quasi-doublons toujours présents dans la BC WIT malgré les filtres appliqués. En effet, en inspectant le top-1 de la recherche mono-modale non-ajustée, pour 50 questions visuelles où ce top-1 est pertinent, on trouve 7 quasi-doublons, des rognages, fréquents sur Wikipédia mais difficiles à détecter, surtout par une méthode de *hashing* comme celle de Gadeski et al. (2017).

Ajustement sur WIT-légendes Comme avec la BC habituelle (cf. tableau 6.6), on trouve que l'ajustement améliore la recherche cross-modale de l'image vers le titre de l'article Wikipédia (cf. tableau 6.7). Cependant, la recherche mono-modale est

11. Les résultats de CLIP non-ajusté étaient également présentés au tableau 6.2 de la section 4.1.

Recherche	Ajustement sur WIT	MRR	P@1	P@20	Hits@20
Mono-modale (IQIP)	✗	29,4	21,8	9,1	53,4
	✓	28,0	20,6	9,1	51,4
Cross-modale (IQTP)	✗	32,7	23,1	10,9	60,6
	✓	36,3	26,2	11,7	64,7
Hybride	✗	39,6	30,6	11,8	63,9
	✓	42,2	33,0	12,2	68,0

TABLEAU 6.6 – Résultats de CLIP pour la recherche visuelle (à partir de l'image de la question $\mathbf{i_q}$) sur le jeu de validation de ViQuAE avec la BC habituelle (une image par entité), avec ou sans ajustement sur WIT-légendes. L'évaluation est faite ici au niveau de l'article.

Recherche	Ajustement sur WIT	MRR	P@1	P@20	Hits@20
Mono-modale (IQIP)	✗*	23,0	14,8	8,5	49,4
	✗	28,9	18,7	11,7	58,4
	✓	28,5	18,2	12,2	59,4
Image-titre (IQTP)	✗	32,0	22,8	12,0	59,7
	✓	36,2	26,0	12,7	64,6
Image-légende (IQTP)	✗*	23,9	14,5	10,1	52,3
	✗	31,8	20,1	14,7	66,0
	✓	32,1	21,3	14,8	66,1

TABLEAU 6.7 – Résultats de CLIP pour la recherche visuelle (à partir de l'image de la question $\mathbf{i_q}$) sur le jeu de validation de ViQuAE avec la BC WIT (1,9 million d'articles/7,6 millions d'images et légendes), avec ou sans ajustement sur WIT-légendes. La recherche cross-modale peut ici être faite entre l'image et la légende et pas seulement entre l'image et le titre. L'évaluation est faite ici au niveau de l'article. *Évaluation avec un sous-ensemble de la BC WIT, avec une seule paire (image, légende) par article (pas applicable pour le titre qui est déjà unique par article).

également détériorée selon les métriques et la recherche cross-modale de l'image vers la légende est seulement légèrement améliorée. Ce dernier point est assez surprenant puisque le modèle est précisément ajusté avec des paires (image, légende), mais l'amélioration est bien plus forte pour la recherche image-titre.

6 Discussion

La section 4 rapporte des différences importantes entre les métriques de RI au niveau du passage visuel et de l'extraction des réponses. Notre système peut être replacé dans le cadre de l'apprentissage augmenté par RI défini par Zamani et al. (2022). Les métriques de RI constituent alors une évaluation intrinsèque des modèles de RI tandis que l'extraction de réponse en représente une évaluation extrinsèque. D'autre part, les métriques d'extraction de réponse évaluent la KVQAE de façon plus

Question visuelle (entrée)	$\text{DPR}_{\text{V+T}}$	ECA_V
 *"In which state of the USA would you find this National Park?*	 Yosemite National Park is located in the central Sierra Nevada of California [...]	 Udall oversaw the addition of four national parks [...] including Canyonlands National Park in Utah, North Cascades National Park in Washington, Redwood National Park in California, the Great Swamp National Wildlife Refuge in New Jersey [...]
 *"This municipality is a ski resort in which European country?"*	 Zermatt and Saas-Fee have both summer ski areas. [...] the majority of ski resorts in Switzerland tend to open in December and run through to April.	 Major ski resorts are located mostly in the various European countries (e.g. Andorra, Austria, Bulgaria, Bosnia-Herzegovina, Croatia, Czech Republic, [...], Poland, Romania, Serbia, Sweden, Slovakia, Slovenia, Spain, Switzerland, Turkey) [...]

FIGURE 6.3 – Exemples où BERT multi-passage parvient à extraire la réponse du passage pertinent fourni par la RI cross-modale tandis qu'il est distrait par le passage fourni par ECA_V (chapitre 5), qui contient beaucoup de réponses plausibles mais n'est pas vraiment pertinent (tout en étant considéré comme tel car il contient la réponse).

complète. Il est intéressant de noter que les métriques de RI que nous utilisons ont été conçues pour des utilisateurs humains et que les modèles d'apprentissage exploitent les résultats de la RI de manière assez différente. Par exemple, le modèle BERT multi-passage ne tient pas compte du rang du passage visuel, les top-K passages étant traités en parallèle.

De plus, les métriques d'extraction de réponse sont moins sensibles au biais textuel inhérent à la KVQAE. Pour reprendre le premier exemple de la figure 6.2, « *This mountain is part of which European mountain range ?* », on peut citer les différentes chaînes de montagnes européennes, sans regarder l'image, et obtenir ainsi un $\text{Hits@20} = 1$. Néanmoins, un modèle d'extraction de réponse tel que BERT multi-passage aura seulement une chance sur 20 environ d'extraire la bonne réponse car elles sont toutes plus ou moins plausibles, comme discuté au chapitre 4. Au contraire, avec une RI purement visuelle, donc exempte de biais textuel, on récupère les passages de l'article Wikipédia de Cairn Gorm ou d'autres montagnes ressemblantes. Les métriques de RI sont alors mauvaises, car la plupart des passages ne sont pas pertinents, mais il suffit d'un seul passage pertinent dans le top-24 pour que BERT multi-passage puisse extraire la réponse sans ambiguïté car seuls les passages pertinents fournissent alors des réponses plausibles.

Nous observons ce phénomène quantitativement : entre la recherche purement textuelle (DPR) et la référence multimodale (DPR_V), il n'y a presque pas de différence pour les métriques de RI au niveau du passage (cf. tableau 6.3), voire une détérioration de la P@20, alors que les métriques d'extraction de réponse (cf. tableau 6.4) montrent 16% à 17% d'amélioration relative pour DPR_V. De la même manière, les modèles de fusion précoce du chapitre 5 sont plus à même d'exploiter

les biais textuels que CLIP, qui cherche seulement à partir de l'image. Deux exemples sont montrés à la figure 6.3.

Bien que cette évaluation extrinsèque corrige certains biais de l'évaluation intrinsèque, puisqu'elle repose sur un modèle externe, elle ajoute aussi plusieurs facteurs, notamment :

— l'architecture du modèle d'extraction (ici BERT multi-passage) ;

— son entraînement, notamment les données et le système de RI utilisés (ici DPR_{V+R+A}) ;

— le top-K (ici 24).

Ces questions sont interdépendantes. Par exemple, le top-K à l'inférence pourrait dépendre du nombre de passages utilisés pendant l'entraînement (24 aussi ici) ou de l'architecture : nous avons tenté de fusionner le score d'extraction de réponse et de RI sans obtenir d'amélioration significative (cf. chapitre 4). L'étude de ces facteurs sort du cadre de ce chapitre mais devrait être réalisée dans de futurs travaux.

Par ailleurs, nous avons étudié l'enrichissement de la BC avec de multiples images associées à des légendes afin d'obtenir plusieurs représentations multimodales d'entité nommée. Cependant, ces expériences ont été rendues difficiles par la présence de quasi-doublons d'images de ViQuAE dans cette BC. En effet, une part non négligeable de ces quasi-doublons subsistent malgré le filtre du système de Gadeski et al. (2017), qui n'est pas en mesure de filtrer les rognages, lesquels sont pourtant fréquents dans Wikimédia Commons. Nous avons également expérimenté avec un système de détection de quasi-doublons neuronal plus coûteux, celui de Pizzi et al. (2022), mais nous avons alors rencontré le problème inverse, puisque son apprentissage contrastif auto-supervisé, proche de SimCLR (Chen et al., 2020a), mène à une représentation partiellement sémantique, qui filtre des images d'une même entité nommée, bien que ce ne soit pas des quasi-doublons. D'autre part, nos résultats suggèrent que ces images de WIT seraient moins représentatives des entités nommées que celles de Wikidata, utilisées dans la BC habituelle, la recherche mono-modale étant en conséquence détériorée. Enfin, l'ajustement de CLIP sur ces paires (image, légende) a fourni des gains de performance variables, tout en étant plus coûteux que l'ajustement sur ViQuAE proposé précédemment. Nous jugeons donc finalement cet ajustement non nécessaire.

7 Conclusion

Dans ce chapitre, nous avons étudié la recherche cross-modale et sa combinaison avec la recherche mono-modale pour répondre aux questions visuelles à propos d'entités nommées, en nous focalisant sur le modèle CLIP. Nos résultats démontrent la supériorité de la recherche cross-modale, mais aussi la complémentarité des recherches mono- et cross-modale, qui peuvent être combinées facilement. Il renforce ainsi les résultats du chapitre précédent concernant la relation entre la recherche cross-modale et l'hétérogénéité des représentations visuelles des entités nommées. Ces résultats se sont avérés robustes à un changement de BC : même en ajoutant plusieurs images par entité, dont des quasi-doublons difficiles à filtrer, la recherche cross-modale reste supérieure à la mono-modale. Il serait intéressant d'étudier si

ces résultats se généralisent à d'autres tâches. Cette méthode pourrait par exemple bénéficier à la recherche visuelle par le contenu dans un contexte de navigation Web. Par ailleurs, l'abondance de données cross-modales a permis d'entraîner un modèle avec la capacité de CLIP, ce qui aurait été difficile avec une annotation mono-modale, mais limite aussi la portée de nos résultats car il est difficile de contrôler une telle masse de données et donc d'estimer les capacités de généralisation de CLIP. Nous avons également étudié différentes manières d'ajuster CLIP et trouvé que l'apprentissage cross-modal est la meilleure solution, encore une fois grâce à son adéquation avec son pré-entraînement. Cette conclusion pourrait changer si nous disposions de suffisamment de données pour réorganiser l'espace de représentation. Nous avons également vu que cet ajustement ne demandait pas de pré-entraînement ou *middle training* supplémentaire, ou du moins que les paires (image, légende) de WIT ne s'y prêtaient pas.

Notre méthode surpasse la référence (recherche mono-modale) mais aussi les méthodes décrites dans les chapitres précédents, tout en étant plus simple et moins coûteuse. Nos résultats questionnent toutefois les métriques utilisées pour évaluer les modèles de RI, notamment l'évaluation intrinsèque des passages, qui est sujette aux biais textuels. Nous préconisons donc de comparer prudemment des modèles fondés sur les mêmes données, comme à la section 4.1, ou bien d'évaluer extrinsèquement les résultats via un modèle d'extraction de réponse (section 4.3). Toutefois, l'utilisation d'un modèle d'apprentissage pour évaluer les résultats est source de variabilité et pourrait donc changer les conclusions d'une étude. Nous avons notamment identifié trois facteurs importants dans la section 6 qui devraient être étudiés dans de futurs travaux. Il serait également intéressant d'étudier l'apprentissage joint de la RI et de l'extraction/génération de réponse. De récents travaux ont montré sa faisabilité pour des tâches connexes à la KVQAE (Chen et al., 2022; Hu et al., 2023c).

Chapitre 7

Conclusion et perspectives

Cette thèse a défini une nouvelle tâche multimodale, répondre aux questions visuelles à propos d'entités nommées (KVQAE), et a contribué à l'évaluer et la traiter. Elle s'inscrit ainsi dans le courant des approches multimodales, qui font suite à l'avènement de l'apprentissage profond mais se trouvent à l'intersection de plusieurs domaines et visent à fluidifier les interactions homme-machine. Nos principales questions de recherche, dans ce cadre de système de question-réponse et de recherche d'information multimodale étaient :

— Comment *évaluer* un système de KVQAE ?

— Comment *représenter visuellement une entité nommée* ?

— Comment *interagissent les modalités* ?

Notre travail fait partie d'une recherche très active à l'intersection du TAL, de la RI, de la vision par ordinateur et de l'apprentissage automatique, dont nous avons donné un aperçu au chapitre 2. Nous avons proposé d'évaluer les systèmes de KVQAE à travers le jeu de données ViQuAE (chapitre 3) et différentes métriques, en supposant que la présence de la réponse dans un passage était indicative de sa pertinence (chapitre 4). Nous avons ensuite remis en cause cette hypothèse au chapitre 6 en privilégiant l'extraction de réponse comme évaluation extrinsèque des systèmes de RI. Concernant les interactions cross-modales, nous avons vu qu'elles pouvaient être : ignorées (chapitre 4), modélisées implicitement à travers un pré-entraînement (chapitre 5), ou explicitement grâce à un modèle entraîné de façon cross-modale sur des images appariées avec leur légende textuelle (chapitre 6). Pour ce qui est enfin de la représentation des entités nommées, diverses questions se sont posées. Tout d'abord, aux chapitres 3 et 6, celle des images les représentant puis, dans le même chapitre 3, celle de leur détection et de leur désambiguïsation pour annoter les questions. Ensuite, au chapitre 4, s'est posée la question de l'utilisation d'une représentation dédiée pour les personnes à travers leur visage. Enfin, aux chapitres 5 et 6, l'étude des représentations des entités nommées s'est inscrite dans celle des interactions cross-modales.

Nous concluons cette thèse en synthétisant nos résultats (section 1) avant de discuter de leurs limites (section 2) et de nos perspectives en conséquence (section 3). Par ailleurs, nous dressons un bilan carbone partiel de nos activités à l'annexe A.

1 Synthèse

Nous avons défini et étudié une nouvelle tâche multimodale, la KVQAE, en nous focalisant sur les données et les représentations multimodales. Ainsi, un des principaux résultats apparaissant aux chapitres 5 et 6 est l'importance de l'interaction cross-modale IQTP, entre l'image de la question et le texte du passage, que nous n'avions pas considéré à l'origine.

Nous avons collecté et annoté ViQuAE, un jeu de données de 3 700 questions visuelles à propos de 2 400 entités nommées, qui nous a servi à évaluer nos méthodes au cours de cette thèse, étant donné les limites du seul autre jeu de données existant, KVQA (Shah et al., 2019 ; cf. chapitre 3). ViQuAE a été annoté semi-automatiquement et la même méthode pourrait être appliquée à de nouveaux domaines ou langues, ou simplement pour collecter davantage de données. Son annotation manuelle a demandé environ 48 heures de travail au total, par sept annotateurs différents. ViQuAE est associé à une base de connaissances non structurée fondée sur 1,5 million d'articles Wikipédia et d'images des entités associées dans Wikidata. Cette BC est simple mais permet de traiter la KVQAE, de répondre aux questions de ViQuAE, et surtout de comparer équitablement différents systèmes.

Nous étions les premiers à traiter la KVQAE avec une BC non-structurée. Ainsi, nous avons proposé un cadre où la tâche est traitée en deux étapes : recherche d'information et extraction de réponse (cf. chapitre 4). Nous avons alors trouvé que la RI était le principal verrou de la KVQAE et nous nous sommes concentré sur cette étape par la suite.

Dans le même chapitre, nous avons étudié différentes manières de représenter visuellement les entités nommées selon leur type, en utilisant une reconnaissance faciale pour les personnes et des représentations issues d'ImageNet et de CLIP pour les autres entités. Nous avons alors trouvé une supériorité de la reconnaissance faciale, que nous expliquons par la qualité des données utilisées pour entraîner le système mais aussi par une meilleure définition de la représentation visuelle intrinsèque aux visages. Par ailleurs, nous étions également les premiers à rapporter les résultats de CLIP dans un cadre de recherche d'image par le contenu, trouvant qu'il surpassait ImageNet, en accord avec ses performances supérieures sur d'autres tâches (Radford et al., 2021).

Dans ce premier système, les interactions cross-modales étaient négligées et la fusion multimodale tardive. De façon complémentaire, nous avons ensuite étudié la fusion précoce au chapitre 5, avec pour objectif initial de modéliser l'interaction TQIQ au sein de la question visuelle. Cependant, nos résultats suggèrent finalement que c'est l'interaction IQTP, entre la question et le passage, qui est principalement utilisée. Nous supposons que cette interaction permet notamment de traiter l'hétérogénéité des représentations visuelles des entités nommées. Ces résultats ont été permis par une nouvelle méthode de pré-entraînement, l'*Inverse Cloze Task* multimodale.

Nous avons donc ensuite naturellement poursuivi l'étude de cette interaction IQTP à travers le modèle CLIP (cf. chapitre 6). Nous avons trouvé que la recherche visuelle cross-modale (qui modélise IQTP) était supérieure à la mono-modale (qui modélise IQIP) mais aussi que les deux étaient complémentaires et pouvaient être

```python
from datasets import load_dataset
from meerqat.models.mm import ECAEncoder

dataset = load_dataset("PaulLerner/viquae_dataset")
model = ECAEncoder.from_pretrained("PaulLerner/question_eca_l6_wit_mict")
```

Algorithme 1: Comment accéder aux données et aux modèles pré-entraînés grâce à quelques lignes de code Python et les bibliothèques Datasets et Transformers (dont dépend `meerqat`).

facilement combinées, sans ajustement supplémentaire. Ces résultats avaient été suggérés par d'autres travaux connexes mais nous sommes les premiers à l'avoir étudié spécifiquement. Cette recherche visuelle hybride pourrait être étudiée dans d'autres contextes où les deux modalités sont disponibles, par exemple au cours d'une navigation Web. Par ailleurs, dans notre cadre où peu de données sont disponibles, nous avons trouvé qu'il était préférable d'ajuster CLIP de façon cross-modale, comme son pré-entraînement, plutôt que de façon mono-modale ou hybride.

Au même chapitre, nous avons étudié différentes méthodes d'évaluation pour les systèmes de RI, et avons trouvé qu'une évaluation extrinsèque à travers une extraction de réponse était préférable à une évaluation intrinsèque de la pertinence des passages, plus sujette aux biais textuels (comme discuté également à la section 2.2).

Tous nos résultats peuvent être reproduits en utilisant notre code (plus de 10 000 lignes) librement disponible à l'adresse suivante `https://github.com/PaulLerner/ViQuAE`. Nous montrons à l'algorithme 1 comment charger le jeu de données ViQuAE ou le modèle ECA pré-entraîné sur WIT à l'*Inverse Cloze Task* multimodale en quelques lignes de code Python grâce aux bibliothèques Datasets et Transformers.

2 Discussion

2.1 Interaction TQIQ au sein de la question visuelle

Nous nous sommes particulièrement intéressé dans cette thèse aux interactions cross-modales. Nous étions à l'origine focalisé sur l'interaction TQIQ, en supposant que le texte et l'image de la question étaient ambigus seuls mais pouvaient être désambiguïsés conjointement.

Cependant, l'approche proposée au chapitre 5 a plusieurs limites. Notamment, l'*Inverse Cloze Task* multimodale, qui sert à pré-entraîner les modèles, pourrait ne pas susciter d'interaction TQIQ au sein de la question visuelle puisque le texte généré de la pseudo-question n'est pas ancré dans l'image et qu'il a du sens indépendamment d'elle. Pour remédier à ce problème, nous pourrions exploiter la légende de l'image comme pseudo-question.

D'autre part, ces résultats proviennent d'une évaluation sur le jeu de données ViQuAE, où la majorité des images représentent une seule et unique entité, rendant ainsi l'interaction TQIQ superflue. Pour mieux comprendre ce phénomène, une méta-

annotation de ViQuAE ou d'un autre jeu de KVQAE serait appréciable et pourrait mener à une nouvelle collecte de données.

2.2 Biais textuels et *benchmark*

Comme toute tâche multimodale, la KVQAE est confrontée aux biais induits par l'une ou l'autre de ses modalités (Yuan, 2021; Dancette et al., 2021). Ici, la modalité textuelle a été la plus problématique et évoquée à chaque chapitre. Plus précisément, un tel biais est présent si répondre à la question visuelle peut se faire à partir de sa seule composante textuelle, donc en ignorant l'image, ou du moins lorsqu'une évaluation automatique juge la réponse comme correcte (si la chaîne de caractères correspond à la vérité-terrain) ou le passage la contenant comme pertinent. Nous avons identifié cinq facteurs de biais textuels principaux. Il existe premièrement un biais textuel *intrinsèque à la KVQAE* : à moins d'enlever tout naturel à la langue, le texte de la question contient des informations permettant d'inférer la réponse. Par exemple, pour une question telle que « *Dans quel pays est-il né ?* », on peut citer un pays au hasard. Deuxièmement, *TriviaQA*, duquel les questions de ViQuAE ont été reformulées, est également source de biais puisque, de par son origine ludique, ses questions sont souvent *surspécifiées*, à l'image de la question « *Dans quel* pays fondateur de l'Union Européenne *est-il né ?* » pour reprendre l'exemple ci-dessus. Troisièmement, *l'annotation de ViQuAE* pourrait être révisée, puisqu'il est évident que le deuxième exemple peut être reformulé en limitant le biais textuel. Quatrièmement, ces biais peuvent être renforcés par les *modèles* capables de les *apprendre automatiquement*, par exemple si la majorité des personnes du jeu de données sont nées en France. Finalement, la *méthode d'évaluation automatique* exacerbe ces biais, comme discuté au chapitre 6, puisqu'un document est jugé pertinent s'il contient la réponse, même s'il ne traite pas de l'entité-sujet ou du sujet visé par la question. Ce biais peut également être exploité par les méthodes d'apprentissage automatique : nous avons vu aux chapitres 5 et 6 des exemples où DPR et ECA retournaient des passages contenant de nombreuses entités nommées (comme les six pays fondateurs de l'UE pour reprendre l'exemple précédent).

Nous avons largement discuté de ce dernier point au chapitre 6, où nous recommandons d'évaluer la RI extrinsèquement à travers l'extraction de réponse, moins sujette aux biais textuels.

Le quatrième point lié à l'apprentissage automatique est peut-être le plus important car il a conditionné nos approches et directions de recherche. Par exemple, on a vu au chapitre 4 que le pré-entraînement de DPR sur TriviaQA était efficace. Par conséquent, nous l'avons conservé comme premier stade de pré-entraînement avant l'*Inverse Cloze Task* multimodale au chapitre 5. Mais quelle part de cette efficacité est-elle due aux biais textuels de TriviaQA et ViQuAE ? Plus généralement, les biais textuels remettent en cause l'approche *benchmark* : en essayant de surpasser un modèle biaisé comme DPR, on risque de se contraindre à exploiter les mêmes biais. Les représentations visuelles sont confrontées au même problème : au chapitre 4 nous pouvions facilement analyser le chevauchement entre les entités du jeu d'entraînement d'ArcFace et de ViQuAE, ce qui n'est pas le cas de CLIP, dont le jeu

d'entraînement n'est pas disponible mais qui, même s'il l'était [1], pourrait difficilement fournir une annotation sur les entités présentes, puisqu'il a été collecté à grande échelle en scrappant le Web. Ces questions dépassent la KVQAE et concernent toutes les communautés de recherche faisant usage de modèles pré-entraînés ou d'apprentissage automatique [2]. Une alternative, ou plutôt une variante du paradigme du *benchmark*, répandu en RI mais peu usité dans nos autres domaines d'intérêt, est l'étude de satisfaction des utilisateurs (Voorhees, 2002; Buchanan et Mckay, 2022). Toutefois, elle est très coûteuse. Une solution moins drastique serait une continuelle mise à jour des *benchmarks* ; on peut par exemple citer l'initiative LongEval dans le cadre de CLEF 2023 (Alkhalifa et al., 2023). Nous discutons de quelques alternatives aux modèles pré-entraînés à la section suivante.

3 Perspectives

Malgré l'ensemble de nos travaux et un regain d'intérêt très récent (Hu et al., 2023a; Chen et al., 2023c; Hu et al., 2023b; Mensink et al., 2023; Li et al., 2023), la KVQAE reste encore largement à explorer. Nous proposons ici quelques pistes pour de futurs travaux, dont certaines viennent pallier les limites évoquées à la section précédente. En lien avec le chapitre 6, nous commencerons par discuter de façons d'obtenir de meilleurs jugements de pertinence des passages afin d'évaluer intrinsèquement la RI. Par ailleurs, ce même chapitre motive l'entraînement joint des modèles de RI et d'extraction de réponse. Nous discuterons ensuite d'alternatives au paradigme *pré-entraînement et ajustement* pour l'apprentissage avec peu d'exemples ainsi qu'un ajustement dans un cadre non-iid, c'est-à-dire de changement de distribution ou de domaine pour mettre à l'épreuve la robustesse des méthodes proposées jusqu'à présent. Nous conclurons nos perspectives en discutant d'entités nommées : comment représenter leurs liens, leurs images et résoudre les expressions référentielles les concernant ?

3.1 De meilleurs jugements de pertinence ou *qrels*

Avec de meilleurs *qrels*, les métriques de RI évaluant intrinsèquement la pertinence des passages retournés ne seraient plus biaisées, comme évoqué à la section précédente et au chapitre 6.

Nous avons expérimenté avec deux méthodes différentes pour obtenir automatiquement de meilleurs *qrels*, que nous n'avons finalement pas jugées satisfaisantes mais qui permettraient d'assister une annotation manuelle.

En premier lieu, nous avons essayé de filtrer les *qrels* en ne conservant que les passages correspondant à l'article de l'entité-sujet. Ainsi, pour que le passage soit pertinent, il faut qu'il contienne la réponse *et* provienne de l'article de l'entité-sujet.

1. On peut noter l'existence d'approches plus transparentes telles qu'OpenCLIP, associé au jeu de données LAION (Schuhmann et al., 2022).

2. Rappelons que les biais des jeux de données ne sont pas spécifiques aux tâches multimodales mais sont présents dans tous les domaines (Belinkov et al., 2019; Kementchedjhieva et al., 2019; Ghaddar et al., 2021; Zhan et al., 2022; Ferré et Langlais, 2023). Certaines de ces références sont tirées de la présentation invitée de Philippe Langlais à CORIA-TALN 2023.

Nous n'avons pas jugé cette solution satisfaisante car elle produit des faux négatifs, puisque la réponse peut se trouver dans un autre article étant donné la redondance des informations dans Wikipédia [3], tout en conservant une partie des faux positifs. Par exemple, il y a 220 occurrences du terme « France » dans l'article lié à Emmanuel Macron alors que seulement quelques-unes permettent de dire qu'il y est né.

Dans un second temps, nous avons envisagé d'utiliser un modèle de RI tel que DPR pour déterminer les *qrels* de TriviaQA automatiquement, avec notre méthode habituelle (passage pertinent s'il contient la réponse), puis de les transposer aux questions correspondantes de ViQuAE. Néanmoins, se reposer sur un tel modèle impliquerait autant de variabilité que pour le modèle d'extraction de réponse. Par ailleurs, en concordance avec notre point précédent concernant la redondance des informations dans Wikipédia, nous trouvons, même en nous limitant au top-10 de DPR, un grand nombre de passages pertinents par question : $5{,}4 \pm 3{,}4$ en moyenne (quartiles : 2, 5 et 9), provenant de presque autant d'articles différents ($4{,}5 \pm 3{,}0$ en moyenne, quartiles : 2, 4 et 7). Cette diversité d'articles, chacun lié à une image différente, motive une modélisation des liens entres les entités, dont nous discuterons à la section 3.5.

Pour conclure à propos des jugements de pertinence, nous pensons qu'une annotation manuelle, d'après un ensemble de passages candidats obtenus par *pooling*, sur le modèle des campagnes TREC, est la meilleure solution, bien qu'elle soit coûteuse. Pour réduire son coût, on pourrait se contenter d'annoter seulement un sous-ensemble de ViQuAE. D'autre part, cela impliquerait de fixer la BC, ce qui pourrait être bénéfique par ailleurs afin d'obtenir des résultats pleinement comparables entre les différentes équipes de recherche. Pour le moment, nous réitérons nos recommandations du chapitre 6 : utiliser les métriques de RI pour comparer prudemment des modèles qui prennent la même entrée et se fier davantage aux métriques d'extraction de réponse.

3.2 Entraînement joint de la recherche d'information et de l'extraction de réponse

Nous avons pour le moment entraîné les modèles de RI et d'extraction de réponse indépendamment. Suite aux problèmes liés à la pertinence des passages discuté ci-dessus, qui servent pour l'évaluation mais aussi l'entraînement de la RI, il est tout naturel de proposer un entraînement joint de la RI et de l'extraction ou génération de réponse, d'autant plus que cela a été éprouvé pour le question-réponse textuel (Lewis et al., 2020; Izacard et al., 2022) mais aussi pour les questions cross-modales (Chen et al., 2022) ou visuelles de sens commun (Hu et al., 2023c).

Cette approche pourrait être rendue difficile sur ViQuAE faute d'un jeu d'entraînement assez grand ou d'un pré-entraînement adéquat. Le grand jeu de données annoté automatiquement par Chen et al. (2023c) pourrait être exploité dans ce sens, d'autant que les auteurs démontrent qu'un gros modèle de langue multimodal et génératif, sans accès à une BC, peut être entraîné de cette façon et fournit une référence raisonnable.

3. cf. Alshomary et al. (2019) ; voir par exemple la figure 5.5 et la discussion associée au chapitre 5.

3.3 Apprentissage avec peu d'exemples

Nous nous sommes principalement concentré, tout comme Chen et al. (2023c) et la majorité des travaux relevant de l'apprentissage profond, sur le paradigme *pré-entraînement et ajustement*, permettant ainsi de traiter la KVQAE dans le cadre du peu de données de ViQuAE.

Ces approches sont toutefois coûteuses et néfastes pour l'environnement (Luccioni et al., 2022). De futurs travaux pourraient au contraire utiliser d'autres méthodes d'apprentissage avec peu d'exemples, comme le méta-apprentissage (Hospedales et al., 2022) ou le *parameter-efficient fine-tuning* (Ding et al., 2023).

D'autre part, nous pourrions nous appuyer davantage sur des méthodes sans apprentissage, telles que BM25, comme au chapitre 4. Par exemple, selon le type d'entité présent dans l'image, une recherche pourrait être effectuée grâce aux descripteurs SIFT ou une vérification géométrique. Ces approches sans apprentissage peuvent s'avérer plus robustes que des méthodes d'apprentissage dans un cadre sans ou avec peu d'exemples (Thakur et al., 2021; Zheng et al., 2018).

3.4 Robustesse et généralisation

L'ajustement, dans le cadre du paradigme *pré-entraînement et ajustement*, a systématiquement été réalisé dans cette thèse dans un cadre iid, bien que nous ayons par ailleurs exploité des modèles non supervisés ou des modèles fondateurs sans les ajuster (cf. chapitres 4 et 6) et que seulement 18% des entités du jeu d'entraînement de ViQuAE soient présentes dans le jeu de test. Nous avons vu au chapitre 2 que les méthodes de recherche dense telles que DPR étaient particulièrement efficaces dans un cadre iid, mais pouvaient être surpassées par d'anciens modèles tels que BM25 dans un cadre adverse ou sans ajustement, ce que nous avons également observé pour ECA au chapitre 5.

Il serait donc intéressant d'étudier la performance de nos méthodes sur des exemples adverses, tels qu'évoqués au chapitre 5 : deux questions ancrées dans la même image mais visant deux entités différentes (toutes deux représentées dans l'image). Nous pourrions également étudier la performance de nos méthodes à travers différents jeux de données, tels que ceux de Hu et al. (2023a) et Chen et al. (2023c), malheureusement indisponibles à ce jour. Nous avons évoqué plusieurs méthodes améliorant la robustesse de DPR au chapitre 2, dont celle de Hong et al. (2022), qui consiste à considérer de multiples représentations par passage. Il serait intéressant d'étendre leur méthode aux passages visuels afin de prendre en compte la multiplicité des représentations des entités nommées, sujet dont nous discutons par ailleurs aux sections suivantes.

3.5 Liens entre les entités

Les entités nommées sont liées par des relations, ce qui influe également sur leurs représentations visuelles, notamment pour les entités abstraites comme les entreprises. Pour reprendre un de nos exemples phares, Apple peut-être représentée visuellement par une photographie d'un de ses magasins (l'*Apple Fifth Avenue*) ou de ses produits emblématiques (l'iPhone). Cet aspect n'est pas traité explicitement dans

cette thèse, où les connaissances sont modélisées à travers une collection de passages visuels, c'est-à-dire de paires (image, texte) où l'image représente une entité et le texte exprime des connaissances en anglais. Nous avons cependant expérimenté l'utilisation de plusieurs représentations visuelles par entité au chapitre 6, ce qui peut contourner le problème (on peut trouver des images de plusieurs magasins et produits d'Apple au sein de son article Wikipédia et, inversement, la même image dans plusieurs articles).

Adjali et al. (2023) proposent d'intégrer des connaissances structurées dans DPR pour traiter la KVQAE mais se focalisent ainsi davantage sur la partie textuelle. D'autre part, Shah et al. (2019) exploitent également un graphe de connaissance mais seulement un sous-graphe autour de la personne reconnue dans l'image par un module externe. Nous envisageons de modéliser explicitement le lien entre les entités nommées et les images les représentant, en les considérant toutes deux comme des nœuds du graphe de connaissance, en exploitant par exemple les données structurées de Wikidata et Wikimédia Commons. Quelques travaux explorent des approches similaires mais sont limités à de petits graphes de connaissance et une évaluation intrinsèque des représentations en complétant le même graphe (Xie et al., 2017; Pezeshkpour et al., 2018; Wilcke et al., 2020; Alberts et al., 2021).

Une autre approche serait fondée sur le pseudo-retour de pertinence (*pseudo-relevance feedback*), une méthode de RI liée à l'expansion de requête, qui s'appuie sur l'hypothèse que les premiers résultats de la RI sont assez pertinents pour les utiliser en enrichissant la requête afin d'effectuer une seconde recherche ou réordonner les résultats initiaux (Xu et Croft, 1996; Chatterjee et Dietz, 2022). Dans notre cas, on pourrait alors faire une première requête à partir de l'image de la question puis utiliser les premiers résultats pour enrichir le texte de la question. Pour reprendre l'exemple précédent et la figure 5.1 où Apple est représentée à travers son magasin : même si la recherche visuelle trouve l'article lié au magasin et pas à l'entreprise, on peut supposer que l'on parvienne à enrichir la question « *When was this company founded ?* », par exemple en extrayant une mention de l'entité nommée *Apple*. Cette extraction pourrait être faite simplement en considérant l'entité la plus fréquemment détectée dans l'article (Chatterjee et Dietz, 2022). Des expériences préliminaires que nous avons réalisées suggèrent qu'elle pourrait également être conditionnée par la question, avec un modèle tel que BERT multi-passage, habituellement utilisé pour l'extraction de réponse. Par ailleurs, d'autres de nos expériences préliminaires indiquent que l'expansion de requête pourrait être effectuée en concaténant directement la mention de l'entité ainsi extraite au texte de la question.

3.6 Représentations visuelles d'entités non-personne

Nous avons évoqué tout au cours de cette thèse la difficulté de représenter des entités non-personne, telles que l'entreprise Apple, par opposition aux personnes qui peuvent être représentées par leur visage, bien que le support (photographie ou tableau) et d'autres facteurs (l'âge de la personne) soient vecteurs d'une variabilité. Nos résultats (voir en particulier le chapitre 4) vont dans ce sens, bien qu'il soit difficile de dissocier l'impact des images de celui des modèles et des données utilisées pour les entraîner.

On peut dénoter trois types de défis pour les représentations d'entités non-personne, peu étudiés jusqu'à présent :

— les entités abstraites qui n'ont pas de représentation propre mais sont représentées à travers d'autres entités concrètes, comme discuté ci-dessus ;

— les entités spatiales, telles que les monuments, qui peuvent être vus de l'intérieur ou de l'extérieur, ou les chaînes de montagne, qui ont plusieurs sommets ;

— les entités fictives, telles que Sherlock Holmes, qui possèdent des attributs caractéristiques dont on souhaite généralement ne pas tenir compte pour les personnes, comme leur tenue, mais qui vont ici définir l'entité (la pipe de Sherlock Holmes).

Nous remarquons que, pour les entités spatiales comme les monuments, la recherche d'image par le contenu standard ne considère pas une image de l'intérieur d'un monument comme pertinente pour une image de l'extérieur du même monument (Radenović et al., 2018). Weyand et al. (2020) filtrent leur jeu de données pour traiter ce problème.

Nous avons particulièrement cherché à traiter l'hétérogénéité des représentations visuelles au chapitre 6 en nous appuyant sur une recherche cross-modale ou de multiples représentations par entité. Alternativement, nous pourrions exploiter des modèles et des données spécifiques à chaque type d'entité, par exemple une vérification géométrique pour les monuments, ce qui rejoint notre discussion de la section 3.3. Ces modèles seraient alors régis par un détecteur de type d'entité. Il faudrait également prêter attention à la propagation des erreurs, comme pour tout système en cascade. Par ailleurs, pour bien analyser ce phénomène, il sera nécessaire de définir une taxonomie des différentes représentations visuelles d'entités nommées, au-delà des trois défis évoqués plus haut, ainsi qu'annoter en conséquence ViQuAE, ou un autre jeu de KVQAE.

3.7 Résolution d'expression référentielle

Pour conclure nos perspectives, nous voulions aborder un autre point que nous n'avons pas traité explicitement : la résolution d'expression référentielle. Il s'agit de déterminer à quelle entité représentée dans l'image on se réfère dans la question. Comme discuté au chapitre 2, on peut décliner cette résolution selon deux cas pour la KVQAE :

1. s'il y a plusieurs entités, le plus souvent des personnes (« *la personne sur la gauche* »), présentes dans l'image, ce qui rejoint la définition standard de la tâche (Plummer et al., 2015) ;
2. si l'image représente différentes entités selon le contexte, par exemple une entreprise ou son magasin, comme discuté plus haut ; ce type de résolution n'a pas été précédemment étudiée à notre connaissance.

Nous ne traitons pas cette résolution explicitement bien qu'ECA, le modèle proposé au chapitre 5, devrait être capable d'apprendre implicitement cette résolution grâce à son mécanisme d'attention (Kim et al., 2018; Cho et al., 2021).

Cependant, nous ne sommes pas en mesure d'étudier la performance de nos modèles pour les questions demandant une résolution d'expression référentielle car

cette méta-annotation n'est pas présente dans ViQuAE. La taxonomie des représentations visuelles pourrait être utile pour étudier le second type de résolution (entités abstraites). Une annotation de ViQuAE ou d'un autre jeu de KVQAE sera alors nécessaire. Pour le premier type de résolution (plusieurs personnes), on peut noter l'utilité du jeu KVQA (Shah et al., 2019), où la position relative des personnes (de gauche à droite dans l'image) est annotée. Heo et al. (2022) traitent spécifiquement ce problème en représentant les positions relatives des personnes dans les images de KVQA dans le même espace que les connaissances structurées puis appliquent un mécanisme d'attention entre la question et ce sous-graphe de connaissances.

Bibliographie

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, et Brigitte Grau. 2020a. Building a multimodal entity linking dataset from tweets. *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4285–4292.

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, et Brigitte Grau. 2020b. Multimodal Entity Linking for Tweets. *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 463–478, Cham. Springer International Publishing.

Omar Adjali, Paul Grimal, Olivier Ferret, Sahar Ghannay, et Hervé Le Borgne. 2023. Explicit knowledge integration for knowledge-aware visual question answering about named entities. *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, page 29–38, New York, NY, USA. Association for Computing Machinery.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo : a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35 :23716–23736.

Houda Alberts, Ningyuan Huang, Yash Deshpande, Yibo Liu, Kyunghyun Cho, Clara Vania, et Iacer Calixto. 2021. VisualSem: a high-quality knowledge graph for vision and language. *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 138–152, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rabab Alkhalifa, Iman Bilal, Hsuvas Borkakoty, Jose Camacho-Collados, Romain Deveaud, Alaa El-Ebshihy, Luis Espinosa-Anke, Gabriela Gonzalez-Saez, Petra Galuščáková, Lorraine Goeuriot, Elena Kochkina, Maria Liakata, Daniel Loureiro, Harish Tayyar Madabushi, Philippe Mulhem, Florina Piroi, Martin Popel, Christophe Servan, et Arkaitz Zubiaga. 2023. Longeval : Longitudinal evaluation of model performance at clef 2023. *Advances in Information Retrieval*, pages 499–505, Cham. Springer Nature Switzerland.

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, et Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv :2204.06031*.

Milad Alshomary, Michael Völske, Tristan Licht, Henning Wachsmuth, Benno Stein, Matthias Hagen, et Martin Potthast. 2019. Wikipedia text reuse : Within and

without. *Advances in Information Retrieval : 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 747–754. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, et Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, et Devi Parikh. 2015. VQA: Visual Question Answering. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile. IEEE.

Yusuf Aytar, Lluis Castrejon, Carl Vondrick, Hamed Pirsiavash, et Antonio Torralba. 2017. Cross-modal scene networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(10) :2303–2314. Publisher : IEEE.

Jimmy Lei Ba, Jamie Ryan Kiros, et Geoffrey E. Hinton. 2016. Layer Normalization. *arXiv :1607.06450 [cs, stat]*. ArXiv : 1607.06450.

Gaston Bachelard. 1938. *La formation de l'esprit scientifique : contribution à une psychanalyse de la connaissance*. Vrin. Google-Books-ID : E1iPyMlagS8C.

Dzmitry Bahdanau, Kyunghyun Cho, et Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv :1409.0473 [cs, stat]*. ArXiv : 1409.0473.

Lalit R. Bahl, Frederick Jelinek, et Robert L. Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2) :179–190.

Hangbo Bao, Wenhui Wang, Li Dong, et Furu Wei. 2022. VL-BEiT: Generative Vision-Language Pretraining. Number : arXiv :2206.01127 arXiv :2206.01127 [cs].

Elias Bassani. 2022. ranx: A Blazing-Fast Python Library for Ranking Evaluation and Comparison. *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 259–264, Cham. Springer International Publishing.

Petr Baudiš et Jan Šedivỳ. 2015. Modeling of the question answering task in the yodaqa system. *Experimental IR Meets Multilinguality, Multimodality, and Interaction : 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings 6*, pages 222–228. Springer.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, et Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Yoshua Bengio, Réjean Ducharme, et Pascal Vincent. 2000. A Neural Probabilistic Language Model. *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Jonathan Berant, Andrew Chou, Roy Frostig, et Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Raffaella Bernardi et Sandro Pezzelle. 2021. Linguistic issues behind visual question answering. *Language and Linguistics Compass*, 15(6).

Olivier Bodenreider. 2004. The unified medical language system (umls) : integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1) :D267–D270.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, et Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, et Mark Sanderson. 2022. A Non-Factoid Question-Answering Taxonomy. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 1196–1207, New York, NY, USA. Association for Computing Machinery.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia

Zheng, Kaitlyn Zhou, et Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv :2108.07258 [cs]*. ArXiv : 2108.07258.

Antoine Bordes, Jason Weston, et Nicolas Usunier. 2014. Open Question Answering with Weakly Supervised Embedding Models. *Machine Learning and Knowledge Discovery in Databases*, Lecture Notes in Computer Science, pages 165–180, Berlin, Heidelberg. Springer.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, et Roopak Shah. 1993. Signature Verification using a "Siamese" Time Delay Neural Network. *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann.

Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, et Robert L. Mercer. 1992. Class-based *n*-gram models of natural language. *Computational Linguistics*, 18(4) :467–480.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, et Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

George Buchanan et Dana Mckay. 2022. What the actual... examining user behaviour in information retrieval. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3444–3446.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, et Desmond Elliott. 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, 9 :978–994.

Deng Cai, Yan Wang, Huayang Li, Wai Lam, et Lemao Liu. 2021. Neural Machine Translation with Monolingual Translation Memory. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 7307–7318, Online. Association for Computational Linguistics.

Agostina Calabrese, Michele Bevilacqua, et Roberto Navigli. 2020. Fatality Killed the Cat or: BabelPic, a Multimodal Dataset for Non-Concrete Concepts. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4680–4686, Online. Association for Computational Linguistics.

Yingshan Chang et Yonatan Bisk. 2022. WebQA: A Multimodal Multihop NeurIPS Challenge. *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*, pages 232–245. PMLR. ISSN : 2640-3498.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, et Yonatan Bisk. 2022. Webqa : Multihop and multimodal qa. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16495–16504.

Shubham Chatterjee et Laura Dietz. 2022. BERT-ER: Query-specific BERT Entity Representations for Entity Ranking. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 1466–1477, New York, NY, USA. Association for Computing Machinery.

Chen Chen, Bowen Zhang, Liangliang Cao, Jiguang Shen, Tom Gunter, Albin Madappally Jose, Alexander Toshev, Jonathon Shlens, Ruoming Pang, et Yinfei Yang. 2023a. STAIR: Learning Sparse Text and Image Representation in Grounded Tokens. ArXiv :2301.13081 [cs].

Danqi Chen, Adam Fisch, Jason Weston, et Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1870–1879.

Tianqi Chen, Bing Xu, Chiyuan Zhang, et Carlos Guestrin. 2016. Training Deep Nets with Sublinear Memory Cost. Number : arXiv :1604.06174 arXiv :1604.06174 [cs].

Ting Chen, Simon Kornblith, Mohammad Norouzi, et Geoffrey Hinton. 2020a. A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR. ISSN : 2640-3498.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, et William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, et Radu Soricut. 2023b. PaLI: A Jointly-Scaled Multilingual Language-Image Model. *Proceedings of the International Conference on Learning Representations*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, et Ming-Wei Chang. 2023c. Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? ArXiv :2302.11713 [cs].

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, et Jingjing Liu. 2020b. Uniter : Universal image-text representation learning. *European Conference on Computer Vision*, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, et Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. *Proceedings of the 38th International Conference on Machine Learning*, pages 1931–1942. PMLR. ISSN : 2640-3498.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, et Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, et Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. ArXiv :2204.02311 [cs].

Christopher Clark et Matt Gardner. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

Cyril Cleverdon. 1967. The cranfield tests on index language devices. *Aslib proceedings*, volume 19, pages 173–194. MCB UP Ltd.

Paul Clough, Mark Sanderson, et Henning Müller. 2004. The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. *Image and Video Retrieval*, Lecture Notes in Computer Science, pages 243–251, Berlin, Heidelberg. Springer.

Marcos V. Conde, Ivan Aerlic, et Simon Jégou. 2022. General Image Descriptors for Open World Image Retrieval using ViT CLIP. ArXiv :2210.11141 [cs].

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Ellen M. Voorhees, et Ian Soboroff. 2021. Trec deep learning track: Reusable test collections in the large data regime. *Proceedings of the 44th International ACM SIGIR Conference on*

Research and Development in Information Retrieval, SIGIR '21, page 2369–2375, New York, NY, USA. Association for Computing Machinery.

Robert Dale. 1989. *Generating referring expressions in a domain of objects and processes*. Ph.D. thesis, University of Edinburgh.

Robert Dale et Nicholas Haddock. 1991. Generating Referring Expressions Involving Relations. *Fifth Conference of the European Chapter of the Association for Computational Linguistics*, Berlin, Germany. Association for Computational Linguistics.

Corentin Dancette, Remi Cadene, Damien Teney, et Matthieu Cord. 2021. Beyond question-based biases : Assessing multimodal shortcut learning in visual question answering. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1574–1583.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, et Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6) :391–407.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, et Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. ISSN : 1063-6919.

Jiankang Deng, Jia Guo, Niannan Xue, et Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Adrien Depeursinge et Henning Müller. 2010. Fusion Techniques for Combining Textual and Visual Information Retrieval. Henning Müller, Paul Clough, Thomas Deselaers, et Barbara Caputo, editors, *ImageCLEF : Experimental Evaluation in Visual Information Retrieval*, The Information Retrieval Series, pages 95–114. Springer, Berlin, Heidelberg.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, et Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Changxing Ding et Dacheng Tao. 2016. A comprehensive survey on pose-invariant face recognition. *ACM Trans. Intell. Syst. Technol.*, 7(3).

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu, Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, et Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence*, 5(3) :220–235. Number : 3 Publisher : Nature Publishing Group.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, et Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.

Paul Ekman et Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1) :49–98.

Zhen Fan, Luyu Gao, Rohan Jha, et Jamie Callan. 2023. Coilcr : Efficient semantic matching in contextualized exact match retrieval. *Advances in Information Retrieval*, pages 298–312, Cham. Springer Nature Switzerland.

Paolo Ferragina et Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*, pages 1625–1628, New York, NY, USA. Association for Computing Machinery.

Arnaud Ferré et Philippe Langlais. 2023. An analysis of entity normalization evaluation biases in specialized domains. *BMC bioinformatics*, 24(1) :227.

Olivier Ferret, Brigitte Grau, Martine Hurault-Plantet, Gabriel Illouz, Laura Monceaux, Isabelle Robba, et Anne Vilnat. 2001. Finding An Answer Based on the Recognition of the Question Focus. *Proceedings of The Tenth Text REtrieval Conference, TREC*, Gaithersburg - USA, Unknown Region.

R. A. Fisher. 1937. The design of experiments. *The design of experiments.*, (2nd Ed). Publisher : Oliver & Boyd, Edinburgh & London.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5) :378–382.

Thibault Formal, Benjamin Piwowarski, et Stéphane Clinchant. 2021. SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, pages 2288–2292, New York, NY, USA. Association for Computing Machinery.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4) :128–135.

Andrea Frome, Greg S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, et Tomas Mikolov. 2013. DeViSE : a deep visual-semantic embedding model. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 2121–2129, Red Hook, NY, USA. Curran Associates Inc.

Marlen Fröhlich, Christine Sievers, Simon W. Townsend, Thibaud Gruber, et Carel P. van Schaik. 2019. Multimodal communication and language origins: integrating gestures and vocalizations. *Biological Reviews*, 94(5) :1809–1829.

Xingyu Fu, Ben Zhou, Ishaan Chandratreya, Carl Vondrick, et Dan Roth. 2022. There'sa time and place for reasoning beyond the image. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1138–1149.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, et Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *Conference on Empirical Methods in Natural Language Processing*, pages 457–468. ACL.

Hengxin Fun, Sunil Gandhi, et Sujith Ravi. 2021. Efficient Retrieval Optimized Multi-task Learning. *arXiv :2104.10129 [cs]*. ArXiv : 2104.10129.

G. W. Furnas, T. K. Landauer, L. M. Gomez, et S. T. Dumais. 1987. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11) :964–971.

Etienne Gadeski, Hervé Le Borgne, et Adrian Popescu. 2017. Fast and robust duplicate image detection on the web. *Multimedia Tools and Applications*, 76(9) :11839–11858.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, et Jianfeng Gao. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Found. Trends. Comput. Graph. Vis.*, 14(3–4) :163–352.

Luyu Gao et Jamie Callan. 2021. Condenser: a Pre-training Architecture for Dense Retrieval. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luyu Gao et Jamie Callan. 2022. Unsupervised corpus aware language model pre-training for dense passage retrieval. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 2843–2853, Dublin, Ireland. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, et Jamie Callan. 2021a. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 3030–3042, Online. Association for Computational Linguistics.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, et Yu Qiao. 2021b. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. ArXiv :2110.04544 [cs].

Diego Garcia-Olano, Yasumasa Onoe, et Joydeep Ghosh. 2022. Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection. *Companion Proceedings of the Web Conference 2022*, WWW '22, page 705–715, New York, NY, USA. Association for Computing Machinery.

François Gardères et Maryam Ziaeefard. 2020. ConceptBert: Concept-Aware Representation for Visual Question Answering. *Findings of the Association for Computational Linguistics : EMNLP 2020*, page 10.

Abbas Ghaddar, Philippe Langlais, Ahmad Rashid, et Mehdi Rezagholizadeh. 2021. Context-aware Adversarial Training for Name Regularity Bias in Named Entity Recognition. *Transactions of the Association for Computational Linguistics*, 9 :586–604.

Ross Girshick. 2015. Fast r-cnn. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, et Chris Olah. 2021. Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3) :e30.

Jonathan Gordon et Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. *Proceedings of the 2013 workshop on Automated knowledge base construction*, AKBC '13, pages 25–30, New York, NY, USA. Association for Computing Machinery.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, et Devi Parikh. 2017. Making the v in vqa matter : Elevating the role of image understanding in visual question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Zenzi M Griffin. 2010. Retrieving personal names, referring expressions, and terms of address. *Psychology of learning and motivation*, volume 53, pages 345–387. Elsevier.

Paul Grimal. 2022. Image, text, knowledge : Explore integrated representation. Master's thesis, Université de Technologie de Compiègne.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O.K. Li, et Richard Socher. 2018. Non-autoregressive neural machine translation. *International Conference on Learning Representations*.

Wenzhong Guo, Jianwen Wang, et Shiping Wang. 2019. Deep Multimodal Representation Learning: A Survey. *IEEE Access*, 7 :63373–63394. Conference Name : IEEE Access.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, et Jianfeng Gao. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 87–102, Cham. Springer International Publishing.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, et Jeffrey P. Bigham. 2018. VizWiz Grand Challenge: Answering Visual Questions From Blind People. pages 3608–3617.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, et Mingwei Chang. 2020. Retrieval Augmented Language Model Pre-Training. *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938. PMLR. ISSN : 2640-3498.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D : Nonlinear Phenomena*, 42(1) :335–346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, et Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Yu-Jung Heo, Eun-Sol Kim, Woo Suk Choi, et Byoung-Tak Zhang. 2022. Hypergraph Transformer: Weakly-supervised multi-hop reasoning for knowledge-based visual question answering. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 373–390, Dublin, Ireland. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, et Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Jack Hessel et Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, et Laurent Sifre. 2022. Training compute-optimal large language models.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, et Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 113–122, New York, NY, USA. Association for Computing Machinery.

Wu Hong, Zhuosheng Zhang, Jinyuan Wang, et Hai Zhao. 2022. Sentence-aware Contrastive Learning for Open-Domain Passage Retrieval. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1062–1074, Dublin, Ireland. Association for Computational Linguistics.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, et Amos Storkey. 2022. Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9) :5149–5169.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, et Ming-Wei Chang. 2023a. Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities. ArXiv :2302.11154 [cs].

Ziniu Hu, Ahmet Iscen, Chen Sun, Kai-Wei Chang, Yizhou Sun, David A. Ross, Cordelia Schmid, et Alireza Fathi. 2023b. AVIS: Autonomous Visual Information Seeking with Large Language Models. ArXiv :2306.08129 [cs].

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, et Alireza Fathi. 2023c. Reveal : Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379.

Xiaolong Huang et QianKun Li. 2022. Runner-up solution to google universal image embedding competition 2022. *arXiv preprint arXiv :2210.08735*.

Jing Huo, Yang Gao, Yinghuan Shi, Wanqi Yang, et Hujun Yin. 2018. Heterogeneous Face Recognition by Margin-Based Cross-Modality Metric Learning. *IEEE Transactions on Cybernetics*, 48(6) :1814–1826. Conference Name : IEEE Transactions on Cybernetics.

Daiyaan Ijaz, Fady Nissan, Henry Dare, Jonathan Williams, Sean Radatz, William Parker, Sherrene Bogle, et Mohammed Mahmoud. 2021. The Impact of Social Media On HCI. *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1421–1431.

Gautier Izacard et Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, et Edouard Grave. 2022. Few-shot Learning with Retrieval Augmented Language Models. ArXiv :2208.03299 [cs].

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, et Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12) :248 :1–248 :38.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, et Xiang Ren. 2022. A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models. Technical Report arXiv :2110.08484, arXiv. ArXiv :2110.08484 [cs] type : article.

J. Johnson, M. Douze, et H. Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3) :535–547.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, et Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, Honolulu, HI. IEEE.

Mandar Joshi, Eunsol Choi, Daniel Weld, et Luke Zettlemoyer. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Kushal Kafle et Christopher Kanan. 2017. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163 :3–20.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, et Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, et Hannaneh Hajishirzi. 2017. Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yova Kementchedjhieva, Mareike Hartmann, et Anders Søgaard. 2019. Lost in Evaluation: Misleading Benchmarks for Bilingual Dictionary Induction. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3336–3341, Hong Kong, China. Association for Computational Linguistics.

Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, et Mubarak Shah. 2022. Transformers in vision: A survey. *ACM Comput. Surv.*, 54(10s).

Omar Khattab et Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 39–48, New York, NY, USA. Association for Computing Machinery. Https ://ai.stanford.edu/blog/retrieval-based-NLP/.

Jin-Hwa Kim, Jaehyun Jun, et Byoung-Tak Zhang. 2018. Bilinear Attention Networks. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, et R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1564–1574. Curran Associates, Inc.

Wonjae Kim, Bokyung Son, et Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Diederik P. Kingma et Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR (Poster)*.

Jana Kludas, Eric Bruno, et Stephane Marchand-Maillet. 2007. Information fusion in multimedia information retrieval. *International Workshop on Adaptive Multimedia Retrieval*, pages 147–159. Springer.

Simon Kornblith, Jonathon Shlens, et Quoc V. Le. 2019. Do Better ImageNet Models Transfer Better? pages 2661–2671.

Emiel Krahmer. 2010. Last words : What computational linguists can learn from psychologists (and vice versa). *Computational linguistics*, 36(2) :285–294.

Alex Krizhevsky, Ilya Sutskever, et Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25 :1097–1105.

Robert Krovetz. 1997. Homonymy and polysemy in information retrieval. *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 72–79, Madrid, Spain. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, et Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7 :452–466.

Carlos Lassance, Hervé Dejean, et Stéphane Clinchant. 2023. An experimental study on pretraining transformers from scratch for ir. *Advances in Information Retrieval*, pages 504–520, Cham. Springer Nature Switzerland.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 153–163.

Phuc H. Le-Khac, Graham Healy, et Alan F. Smeaton. 2020. Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8 :193907–193934. Conference Name : IEEE Access.

Yann LeCun, Yoshua Bengio, et Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553) :436–444. Number : 7553 Publisher : Nature Publishing Group.

Kenton Lee, Ming-Wei Chang, et Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et others. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2) :167–195. Publisher : IOS Press.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno, et Jesús Lovón Melgarejo. 2022. ViQuAE, a dataset for knowledge-based visual question answering about named entities. *Proceedings of The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, New York, NY, USA. Association for Computing Machinery.

Brian Lester, Rami Al-Rfou, et Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, et Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, et Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, et Daxin Jiang. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07) :11336–11344. Number : 07.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, et Qi Liu. 2023. M^3IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning. ArXiv :2306.04387 [cs].

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, et Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language.

Anne-Laure Ligozat, Julien Lefevre, Aurélie Bugeau, et Jacques Combaz. 2022. Unraveling the hidden environmental impacts of ai solutions for environment life cycle assessment of ai solutions. *Sustainability*, 14(9).

Chin-Yew Lin. 2004. Rouge : A package for automatic evaluation of summaries. *Text summarization branches out*, pages 74–81.

Jimmy Lin, Rodrigo Nogueira, et Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4) :1–325.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, et C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham. Springer International Publishing.

Hugo Liu et Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4) :211–226.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3) :225–331.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, et Ge Yu. 2023. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. *The Eleventh International Conference on Learning Representations*.

Ilya Loshchilov et Frank Hutter. 2019. Decoupled Weight Decay Regularization. *arXiv :1711.05101 [cs, math]*. ArXiv : 1711.05101.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60 :91–110.

Jiasen Lu, Dhruv Batra, Devi Parikh, et Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in Neural Information Processing Systems*, 32 :13–23.

Alexandra Sasha Luccioni, Sylvain Viguier, et Anne-Laure Ligozat. 2022. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. ArXiv :2211.02001 [cs].

Man Luo, Yankai Zeng, Pratyay Banerjee, et Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziyang Luo, Pu Zhao, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, Jing Ma, Qingwen lin, et Daxin Jiang. 2023. LexLIP: Lexicon-Bottlenecked Language-Image Pre-Training for Large-Scale Image-Text Retrieval. ArXiv :2302.02908 [cs].

Xueguang Ma, Kai Sun, Ronak Pradeep, et Jimmy Lin. 2021a. A Replication Study of Dense Passage Retriever. *arXiv :2104.05740 [cs]*.

Zhengyi Ma, Zhicheng Dou, Wei Xu, Xinyu Zhang, Hao Jiang, Zhao Cao, et Ji-Rong Wen. 2021b. Pre-training for Ad-hoc Retrieval: Hyperlink is Also You Need. *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1212–1221. Association for Computing Machinery, New York, NY, USA.

Mateusz Malinowski et Mario Fritz. 2014. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, et K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1682–1690. Curran Associates, Inc.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, et Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Aditya Menon, Sadeep Jayasumana, Ankit Singh Rawat, Seungyeon Kim, Sashank Reddi, et Sanjiv Kumar. 2022. In defense of dual-encoders for neural ranking. *Proceedings of the 39th International Conference on Machine Learning*, pages 15376–15400. PMLR. ISSN : 2640-3498.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, et Vittorio Ferrari. 2023. Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. ArXiv :2306.09224 [cs].

Salem Messoud. 2022. Reranking methods for knowledge-based visual question answering about named entities. Master's thesis, Université Paris-Saclay.

Donald Metzler, Yi Tay, Dara Bahri, et Marc Najork. 2021. Rethinking Search: Making Domain Experts out of Dilettantes. *ACM SIGIR Forum*, 55(1) :1–27. ArXiv : 2105.02274.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, et Andrew Zisserman. 2021. Thinking Fast and Slow: Efficient Text-to-Visual Retrieval With Transformers. pages 9826–9836.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, et Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, et Luke Zettlemoyer. 2020. AmbigQA: Answering ambiguous open-domain questions. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Sewon Min, Minjoon Seo, et Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, pages 510–517, Vancouver, Canada. Association for Computational Linguistics.

Margaret Mitchell, Kees Van Deemter, et Ehud Reiter. 2013. Generating expressions that refer to visible objects. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 1174–1184.

Seungwhan Moon, Leonardo Neves, et Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 2000–2008.

Roberto Navigli et Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193 :217–250.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, et Li Deng. 2016. Ms marco : A human generated machine reading comprehension dataset. *choice*, 2640 :660.

Nils J. Nilsson. 1965. Learning machines. Foundations of trainable pattern-classifying systems. McGraw-Hill Series in Systems Science. New York-St. Louis-San Francisco-Toronto-London-Sydney : McGraw-Hill Book Company. xi, 137 p. (1965).

Rodrigo Nogueira et Kyunghyun Cho. 2020. Passage Re-ranking with BERT. ArXiv :1901.04085 [cs].

Aaron van den Oord, Yazhe Li, et Oriol Vinyals. 2019. Representation Learning with Contrastive Predictive Coding. ArXiv :1807.03748 [cs, stat].

OpenAI. 2023. GPT-4 Technical Report. ArXiv :2303.08774 [cs].

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, et Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32.

Chunlei Peng, Nannan Wang, Jie Li, et Xinbo Gao. 2019. DLFace: Deep local descriptor for cross-modality face recognition. *Pattern Recognition*, 90 :161–171.

Jeffrey Pennington, Richard Socher, et Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, et Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 1756–1765, Vancouver, Canada. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, et Luke Zettlemoyer. 2018. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, et Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, et Alexander Miller. 2019. Language models as knowledge bases? *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pouya Pezeshkpour, Liyan Chen, et Sameer Singh. 2018. Embedding Multimodal Relational Data for Knowledge Base Completion. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3208–3218.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, et Andrew Zisserman. 2007. Object retrieval with large vocabularies and fast spatial matching. *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, et Andrew Zisserman. 2008. Lost in quantization : Improving particular object retrieval in large scale image databases. *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, et Matthijs Douze. 2022. A Self-Supervised Descriptor for Image Copy Detection. ArXiv :2202.10261 [cs].

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, et Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. pages 2641–2649.

Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, et Erik Learned-Miller. 2021a. Passage retrieval for outside-knowledge visual question answering. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1753–1757, New York, NY, USA. Association for Computing Machinery.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, et Haifeng Wang. 2021b. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 5835–5847, Online. Association for Computational Linguistics.

Rodrigo Quian Quiroga. 2012. Concept cells : the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 13(8) :587–597.

F. Radenović, G. Tolias, et O. Chum. 2019. Fine-Tuning CNN Image Retrieval with No Human Annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7) :1655–1668. Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence.

Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, et Ondřej Chum. 2018. Revisiting oxford and paris: Large-scale image retrieval benchmarking. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, et Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, et Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. page 24.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, et Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21 :1–67.

Thilina C. Rajapakse et Maarten de Rijke. 2023. Improving the Generalizability of the Dense Passage Retriever Using Generated Datasets. *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 94–109, Cham. Springer Nature Switzerland.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, et Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, et Omer Levy. 2021. Few-Shot Question Answering by Pretraining Span Selection. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 3066–3079, Online. Association for Computational Linguistics.

Ori Ram, Gal Shachaf, Omer Levy, Jonathan Berant, et Amir Globerson. 2022. Learning to Retrieve Passages without Supervision. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 2687–2700, Seattle, United States. Association for Computational Linguistics.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, et Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. *Proceedings of the 18th ACM international conference on Multimedia*, MM '10, pages 251–260, New York, NY, USA. Association for Computing Machinery.

Revant Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avirup Sil, Shih-Fu Chang, et al. 2022. Mumuqa : Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11200–11208.

Nils Reimers et Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Mengye Ren, Ryan Kiros, et Richard Zemel. 2015. Exploring Models and Data for Image Question Answering. *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, et Mike Gatford. 1995. Okapi at TREC-3. *Third Text REtrieval Conference (TREC-3)*, volume 500-225 of *NIST Special Publication*, pages 109–126. National Institute of Standards and Technology (NIST).

Pedro Rodriguez et Jordan Boyd-Graber. 2021. Evaluation Paradigms in Question Answering. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9630–9642, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anna Rogers, Matt Gardner, et Isabelle Augenstein. 2021. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *arXiv :2107.12708 [cs]*. ArXiv : 2107.12708 version : 1.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, et Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3) :211–252.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, et Bryan Catanzaro. 2021. End-to-End Training of Neural Retrievers for Open-Domain Question Answering. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 6648–6662, Online. Association for Computational Linguistics.

Toshiyuki Sakai, Makoto Nagao, et Takeo Kanade. 1972. *Computer analysis and classification of photographs of human faces*. Kyoto University.

Shailaja Keyur Sampat, Yezhou Yang, et Chitta Baral. 2020. Visuo-Linguistic Question Answering (VLQA) Challenge. *Findings of the Association for Computational Linguistics : EMNLP 2020*, pages 4606–4616, Online. Association for Computational Linguistics.

Manisha M Sawant et Kishor M Bhurchandi. 2019. Age invariant face recognition : a survey on facial aging databases, techniques and effect of aging. *Artificial Intelligence Review*, 52 :981–1008.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, et Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

Florian Schroff, Dmitry Kalenichenko, et James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. pages 815–823.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, et Jenia Jitsev. 2022. LAION-5b: An open large-scale dataset for training next generation image-text models. *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, et Danqi Chen. 2021. Simple Entity-Centric Questions Challenge Dense Retrievers. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sanket Shah, Anand Mishra, Naganand Yadati, et Partha Pratim Talukdar. 2019. KVQA: Knowledge-Aware Visual Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8876–8884.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, et Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Amanda J. C. Sharkey. 1999. *Linear and Order Statistics Combiners for Pattern Classification*, pages 127–161. Springer London, London.

Tom Sherborne et Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Fatemeh Shiri, Fatih Porikli, Richard Hartley, et Piotr Koniusz. 2018. Identity-Preserving Face Recovery from Portraits. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 102–111.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, et Jason Weston. 2021. Retrieval Augmentation Reduces Hallucination in Conversation. *Findings of the Association for Computational Linguistics : EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

T. Sim, S. Baker, et M. Bsat. 2002. The CMU Pose, Illumination, and Expression (PIE) database. *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 53–58.

Oriane Siméoni, Yannis Avrithis, et Ondrej Chum. 2019. Local features and visual words emerge in activations. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11651–11660.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, et Douwe Kiela. 2022. FLAVA: A Foundational Language and Vision Alignment Model. pages 15638–15650.

Sivic et Zisserman. 2003. Video Google: a text retrieval approach to object matching in videos. *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2.

A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, et R. Jain. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern*

Analysis and Machine Intelligence, 22(12) :1349–1380. Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence.

Mark D. Smucker, James Allan, et Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, pages 623–632, New York, NY, USA. Association for Computing Machinery.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, et Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1) :11–21.

Lucia Specia, Stella Frank, Khalil Sima'an, et Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. *Proceedings of the First Conference on Machine Translation : Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Rohini K. Srihari, Zhongfei Zhang, et Aibing Rao. 2000. Intelligent Indexing and Semantic Retrieval of Multimodal Documents. *Information Retrieval*, 2(2) :245–275.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, et Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, et Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. *International Conference on Learning Representations*.

Fabian M. Suchanek, Gjergji Kasneci, et Gerhard Weikum. 2007. Yago: a core of semantic knowledge. *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. Association for Computing Machinery.

Wen Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, et Xueqi Cheng. 2022. Visual named entity linking: A new dataset and a baseline. *Findings of the Association*

for Computational Linguistics : EMNLP 2022, pages 2403–2415, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Deni Sutisna, Arif Widodo, Nursaptini Nursaptini, Umar Umar, Muhammad Sobri, et Dyah Indraswati. 2020. An analysis of the use of smartphone in students' interaction at senior high school. *Proceedings of the 1st Annual Conference on Education and Social Sciences (ACCESS 2019)*, pages 221–224. Atlantis Press.

Alon Talmor et Jonathan Berant. 2018. The Web as a Knowledge-Base for Answering Complex Questions. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, et Jonathan Berant. 2021. MultiModalQA: Complex Question Answering over Text, Tables and Images. *ICLR 2021*.

Manveer Singh Tamber, Ronak Pradeep, et Jimmy Lin. 2023. Pre-processing Matters! Improved Wikipedia Corpora for Open-Domain Question Answering. *Advances in Information Retrieval*, Lecture Notes in Computer Science, pages 163–176, Cham. Springer Nature Switzerland.

Fuwen Tan, Jiangbo Yuan, et Vicente Ordonez. 2021. Instance-level image retrieval using reranking transformers. *proceedings of the IEEE/CVF international conference on computer vision*, pages 12105–12115.

Hao Tan et Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W Cohen, et Donald Metzler. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, volume 35, pages 21831–21843. Curran Associates, Inc.

Wilson L Taylor. 1953. "cloze procedure" : A new tool for measuring readability. *Journalism quarterly*, 30(4) :415–433.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, et Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv :2104.08663 [cs]*. ArXiv : 2104.08663.

Kushal Tirumala, Aram Markosyan, Luke Zettlemoyer, et Armen Aghajanyan. 2022. Memorization without overfitting : Analyzing the training dynamics of large language models. *Advances in Neural Information Processing Systems*, 35 :38274–38290.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, et Felix Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. *Advances in Neural Information Processing Systems*, volume 34, pages 200–212. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, et Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, pages 5998–6008.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, et Raquel Urtasun. 2016. Order-Embeddings of Images and Language. *ICLR 2016*.

Peter Vickers, Nikolaos Aletras, Emilio Monti, et Loïc Barrault. 2021. In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2 : Short Papers)*, pages 468–475, Online. Association for Computational Linguistics.

Jette Viethen et Robert Dale. 2008. The use of spatial relations in referring expression generation. *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67, Salt Fork, Ohio, USA. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, et Dumitru Erhan. 2016. Show and tell : Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4) :652–663.

Ellen M. Voorhees. 2001. The TREC question answering track. *Natural Language Engineering*, 7(4) :361–378. Publisher : Cambridge University Press.

Ellen M Voorhees. 2002. The philosophy of information retrieval evaluation. *Evaluation of Cross-Language Information Retrieval Systems : Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001 Darmstadt, Germany, September 3–4, 2001 Revised Papers 2*, pages 355–370. Springer.

Ellen M. Voorhees. 2019. *The Evolution of Cranfield*, pages 45–69. Springer International Publishing, Cham.

Ellen M. Voorhees et Dawn M. Tice. 2000. Building a question answering test collection. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00)*, pages 200–207, Athens, Greece. ACM Press.

Denny Vrandečić et Markus Krötzsch. 2014. Wikidata : a free collaborative knowledgebase. *Communications of the ACM*, 57(10) :78–85.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, et Anton van den Hengel. 2018. FVQA: Fact-Based Visual Question Answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10) :2413–2427.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, et Anton Van Den Henge. 2017. Explicit knowledge-based reasoning for visual question answering. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1290–1296.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, et Hongxia Yang. 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, et Bing Xiang. 2019. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5878–5882, Hong Kong, China. Association for Computational Linguistics.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23.

Jason Weston, Sumit Chopra, et Antoine Bordes. 2014. Memory networks.

Tobias Weyand, Andre Araujo, Bingyi Cao, et Jack Sim. 2020. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2575–2584.

W. X. Wilcke, P. Bloem, V. de Boer, R. H. van t Veer, et F. A. H. van Harmelen. 2020. End-to-End Entity Classification on Multimodal Knowledge Graphs. *arXiv :2003.12383 [cs]*. ArXiv : 2003.12383.

Adina Williams, Nikita Nangia, et Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, et Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv :1910.03771 [cs]*.

Robert Wolfe et Aylin Caliskan. 2022. Contrastive Visual Semantic Pretraining Magnifies the Semantics of Natural Language Representations. *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers), pages 3050–3061.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, et Anton van den Hengel. 2017. Visual question answering : A survey of methods and datasets. *Computer Vision and Image Understanding*, 163 :21–40.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, et Maosong Sun. 2017. Image-embodied knowledge representation learning. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 3140–3146, Melbourne, Australia. AAAI Press.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, et Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval.

Jinxi Xu et W. Bruce Croft. 1996. Query expansion using local and global document analysis. *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, page 4–11, New York, NY, USA. Association for Computing Machinery.

Lanling Xu, Jianxun Lian, Wayne Xin Zhao, Ming Gong, Linjun Shou, Daxin Jiang, Xing Xie, et Ji-Rong Wen. 2022. Negative sampling for contrastive representation learning: A review.

Jheng-Hong Yang, Carlos Lassance, Rafael Sampaio de Rezende, Krishna Srinivasan, Miriam Redi, Stéphane Clinchant, et Jimmy Lin. 2023. AToMiC: An Image/Text Retrieval Test Collection to Support Multimedia Content Creation. ArXiv :2304.01961 [cs].

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, et Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3) :3081–3089.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, et Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5123–5132, Hong Kong, China. Association for Computational Linguistics.

Desen Yuan. 2021. Language bias in Visual Question Answering: A Survey and Taxonomy. ArXiv :2111.08531 [cs].

Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, et Michael Bendersky. 2022. Retrieval-Enhanced Machine Learning. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 2875–2886, New York, NY, USA. Association for Computing Machinery.

Jingtao Zhan, Xiaohui Xie, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, et Shaoping Ma. 2022. Evaluating interpolation and extrapolation performance of neural retrieval models. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 2486–2496.

Dongxiang Zhang, Rui Cao, et Sai Wu. 2019. Information fusion in visual question answering: A survey. *Information Fusion*, 52 :268–280.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, et Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10) :1499–1503. Conference Name : IEEE Signal Processing Letters.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, et Nan Duan. 2022. Multi-View Document Representation Learning for Open-Domain Dense Retrieval. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 5990–6000, Dublin, Ireland. Association for Computational Linguistics.

Liang Zheng, Yi Yang, et Qi Tian. 2018. SIFT Meets CNN: A Decade Survey of Instance Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5) :1224–1244. Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence.

Jiawei Zhou, Xiaoguang Li, Lifeng Shang, Lan Luo, Ke Zhan, Enrui Hu, Xinyu Zhang, Hao Jiang, Zhao Cao, Fan Yu, et al. 2022. Hyperlink-induced pre-training for passage retrieval in open-domain question answering. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, pages 7135–7146.

Xuan Zou, Josef Kittler, et Kieron Messer. 2007. Illumination invariant face recognition: A survey. *2007 First IEEE International Conference on Biometrics : Theory, Applications, and Systems*, pages 1–8.

Annexe A

Bilan carbone partiel

Nous prenons dans cette annexe du recul par rapport aux problématiques scientifiques liées à la KVQAE, en replaçant notre travail dans le contexte beaucoup plus global du changement climatique. Les chercheurs de nos domaines sont vecteurs de progrès, technique et scientifique, mais qui est parfois en opposition avec la sobriété énergétique nécessaire pour limiter, notamment, le réchauffement climatique. Citons, par exemple en TAL, BLOOM, un gros modèle de langue de la même taille que GPT-3, dont l'entraînement a émis 50 500 kgCO2e (ou seulement 24 700 kgCO2e si on se limite à la consommation électrique des GPUs [1]) et dont l'inférence émet environ 19 kgCO2e par jour (Luccioni et al., 2022). Estimer les émissions en carbone d'un modèle d'apprentissage automatique est un problème difficile à multiples facteurs (Ligozat et al., 2022). Nous nous limitons ici aux émissions liées à la consommation électrique des GPUs utilisées pendant l'entraînement ou l'inférence du modèle, sans prendre en compte, par exemple, les émissions liées à la fabrication de ces GPUs. Un autre facteur d'émission, qui concerne de façon beaucoup plus générale tous les chercheurs, est lié aux voyages en avion. Nous trouvons que, malgré les nombreuses expériences menées au cours de cette thèse, l'avion est de loin le facteur d'émission le plus important, bien que nous nous soyons limités à des trajets en Europe.

L'immense majorité des expériences de cette thèse ont été menées sur le supercalculateur Jean Zay [2], avec une part négligeable sur Lab-IA [3]. 6 400 heures GPU V100 ont été consommées au total, en majorité sur les expériences de pré-entraînement du chapitre 5. Cela reste modeste par rapport aux pré-entraînements de modèles fondateurs tels que BLOOM ou CLIP (ViT-L/14), qui ont demandé environ 1 083 000 et 74 000 heures GPU A100 et V100, respectivement (Luccioni et al., 2022; Radford et al., 2021). Selon l'IDRIS, leurs GPUs V100 consomment 0,482 kW, ou 0,259 kW après récupération de la chaleur. Nous obtenons donc, pour nos expériences, une consommation électrique totale de 3 100 kWh, ou 1 700 kWh après récupération de la chaleur. En tenant compte du facteur d'émission moyen en France, de 0,0569

1. BLOOM est assez comparable à nos expériences car il a été entraîné sur Jean Zay, qui a la chance de bénéficier d'une électricité française peu carbonée. Luccioni et al. (2022) estiment que la consommation électrique des GPUs utilisées pour l'entraînement de GPT-3 a émis 502 000 kgCO2e, principalement à cause d'une énergie presque 7,5 fois plus carbonée.
2. http://www.idris.fr/jean-zay/
3. http://hebergement.universite-paris-saclay.fr/lab-ia/

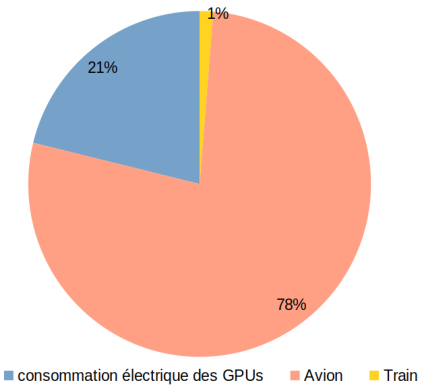


FIGURE A.1 – Bilan carbone partiel : 828 kgCO2e. La consommation électrique des GPUs montrée ici correspond à 175 kgCO2e, avant récupération de la chaleur, plus pessimiste que 94 kgCO2e après récupération.

kgCO2e/kWh selon l'Ademe [4], nous obtenons finalement des émissions carbone de 175 kgCO2e, ou 94 kgCO2e après récupération de la chaleur.

Par ailleurs, afin de participer aux conférences SIGIR 2022 et ECIR 2023 et à l'école d'été LxMLS 2023, cette thèse a donné lieu à plusieurs allers-retours en avion, depuis Paris vers Madrid, Dublin et Lisbonne. Nous ne tenons pas compte des trajets en train dont les facteurs d'émission sont négligeables par rapport à l'avion. Selon leur compagnie aérienne respective, ces trois trajets en avion ont au total produit 642 kgCO2e par passager, émissions bien supérieures donc à celles liées à la consommation électrique des GPUs. Ce bilan carbone partiel est résumé à la figure A.1. Nous avons discuté de méthodes alternatives au pré-entraînement de gros modèles au chapitre 7. Par ailleurs, le LISN fait partie de Labos 1point5 [5] et vise en cela à réduire ses émissions de gaz à effet de serre.

4. https://bilans-ges.ademe.fr/fr
5. https://labos1point5.org/

Annexe B

Guide d'annotation de ViQuAE

Nous restituons ici le guide d'annotation du jeu de données ViQuAE (cf. chapitre 3), également disponible à l'adresse suivante : `https://github.com/PaulLerner/ViQuAE/blob/main/ANNOTATION.md`. Le guide est en anglais. La section 1 se réfère à la figure 3.4. En plus des deux sections suivantes, le guide fournissait 10 exemples pour former l'annotateur (`eg_generations`), avec l'annotation de référence (`eg_annotations`).

1 Interface

— Skip button : **avoid its use** as for some reason, it will also mark this "task" as completed (although you will be able to sort tasks by number of skips then delete the fake completions)

— Submit button : hit it only when you're done

— Image is the image selected by the model that means to illustrate the Generated question

— Original question is the source from which we removed the explicit entity mention to replace it with an ambiguous one (Generated question)

— Generated question is **the** question that will end up in the dataset. This field is **editable** (so be careful) so that you can correct any potential errors (see below) **as long as it doesn't change the answer**

— Answer : the answer to both questions. Note there are actually multiple valid answers based on Wikipedia aliases. If the match between the original question and the answer seem strange, don't worry too much.

— Disambiguated entity : entity QID, description and illustrative image on Wikidata. This should serve as a reference for spotting errors.

— Other available mentions : mentions extracted from Wikidata that might suit the entity. Can be used if the one selected is not appropriate (see below).

— Alternative images : 8 images along with their caption (left-to-right). You should tick the corresponding caption of the image you want to replace the question-image with (only if the latter is inappropriate obviously)

— Discard question : different reasons why you choose to discard the (image, question, answer) triple. **don't tick any of the boxes otherwise** (although you can tick and untick if you've made a mistake).

Avoid using the "update" feature or the "previous"/"next" arrows when you want to correct an error (it seems to save the initial "generated question" along your modifications). Instead delete the completion and hit "label" again.

2 Common errors and how to fix them

See Table B.1.

Last source of error : *entity linking*, e.g. Washington, D.C. has been disambiguated as George Washington.

Please have a look at `eg_generations` to get familiar with the interface and try to reproduce the annotations in `eg_annotations` !

original question	**image**	**generated question**	**error category**	**explanation**
Which country is bordered by **Cambodia** and Laos to the west and China to the north ?	-	Which country is bordered by **this sovereign state** to the west and China to the north ?	*entity type*	Cambodia is a country, hard to illustrate thus it doesn't really fit our needs. These should have been filtered automatically but your annotation might allow to refine this filter.
Where did **Richard III** 'imprison' his two young nephews [...]		Where did **he** 'imprison' his two young nephews [...]	*image*	Image is related to the entity but doesn't depict it directly. You should try to select an alternative image if possible.
When was **the president Barack Obama** born ?		When was **he** born ?	*image*	Related issue : the entity is depicted but is not prominent. Although you can select an alternative image, you may get creative with the mention, e.g. "the man on the right". Avoid near-duplicates of the reference image. Also, if possible, avoid images where the entity name is written (e.g. plaque on a statue, photograph caption) although this might not be a problem or might be post-processed.
To within five years either way, in which year was the Boy Scout movement founded by **Robert Baden-Powell** ?	-	-	*overspecified*	The original question is not really about Robert Baden-Powell so it doesn't make sense to illustrate it with him.
Bonar Law is the only Prime Minister not born in the UK. In which country was he born ?	-	**He** is the only Prime Minister not born in the UK. In which country was he born ?	*overspecified*	Related issue, however here you are able to modify the question to make it more "visual" **as long as it doesn't change the answer**, e.g. simply keeping the last part "In which country was he born ?"
Who plays **Han Solo** in Star Wars ?		-	*GDPR*	In order to avoid having trouble with GDPR, we only use pictures of deceased celebrities thus questions about those have been filtered. However in this corner case we have a common folk who is *prominently* depicted in the image. Please select an alternative image if possible. (Note all images that have "cosplay" in their category have been filtered out but your annotation might provide useful new keywords).
Who founded **Stanford University** ?		Who founded **this open-access publisher** ?	-	Although Stanford University *is* an open-access publisher, it is not its prime function, please edit the question by using another mention (you may use one available in "Other available mentions")
In the game of **Bingo**, 'Get up and run' represents which number ?"	-	In the game of **this game of chance**, 'Get up and run' represents which number ?	-	The whole "the game of Bingo" should have been replaced in the original question. Modify the question to make it grammatical.
If you landed at **'Santa Catarina Airport** on which Island would you be in ?	-	If you landed at **this airport** ?	-	Related issue : here, the parser removed *too much* of the original question, nothing a little copy-paste can't fix ;)

TABLEAU B.1 – Common errors and how to fix them

Annexe C

Inverse Cloze Task multimodale : résultats sur le jeu de validation

Comme pour les résultats en amont sur WIT, nous présentons ici le MRR et la P@1 *au sein du lot*. Le lot en question comprend 625 questions visuelles appariées à 1 250 passages visuels (un pertinent et un négatif difficile par question). Puisque le jeu de validation sert à interrompre l'entraînement (*early stopping*), il s'agit des meilleurs résultats au cours de l'entraînement du modèle, avec un nombre d'itérations variable selon le modèle.

Les résultats sont présentés dans les tableaux C.1 et C.2. On peut voir que les différences entre les modèles, et plus particulièrement entre les versions d'ECA, observées au chapitre 5 se retrouvent ici, et donc que ces résultats sont robustes et non pas excessivement optimistes.

Modèle	**ICT**	**MRR**	**P@1**
$\mathrm{ECA}_V(l = \mathrm{NA})$	✗	58,7	45,3
$\mathrm{ECA}_V(l = 6)$	✓	63,9	49,8
$\mathrm{ECA}_V(l = 0)$	✓	63,8	50,1
$\mathrm{ILF}_V(l = 12)$	✓	65,0	50,3
$\mathrm{ECA}_{V+R+A}(l = 6)$	✓	**65,1**	**52,4**

TABLEAU C.1 – Évaluation sur le jeu de validation de ViQuAE des modèles pré-entraînés au question-réponse sur TriviaQA et ajustés sur ViQuAE. l correspond au nombre de dernières couches figées dans le modèle pendant l'ICT multimodale, si applicable.

Modèle	**ICT**	**MRR**	**P@1**
$\mathrm{ECA}_V(l = 0)$	✓	58,7	45,3
$\mathrm{ILF}_V(l = 12)$	✓	**60,4**	**46,5**
$\mathrm{ECA}_{V+R+A}(l = 0)$	✓	59,2	45,4

TABLEAU C.2 – Évaluation en aval sur le jeu de validation de ViQuAE, des modèles sans pré-entraînement au question-réponse sur TriviaQA mais *avec* ajustement sur ViQuAE.

Annexe D

Droits d'auteurs des images utilisées dans les figures

Si la figure contient plusieurs images, les auteurs sont crédités de gauche à droite et de haut en bas. Toutes les images utilisées sont sous licence libre (par exemple Creative Commons).

- Figure 1.1 :
  - *domaine public*
  - The National Archives UK
  - Novalis31
  - Jiuguang Wang
- Figure 1.2 :
  - *domaine public*
  - *domaine public*
  - Rüdiger Wölk, Münster
  - *domaine public*
  - Ed Uthman, MD
  - Rafael Fernandez
- Figure 2.1 :
  - Novalis31
  - Institut national de l'information géographique et forestière
- Figure 2.3 : issue de Kim et al. (2018)
- Figure 2.4 : issue de Bugliarello et al. (2021)
- Figure 2.5 : issue de Bugliarello et al. (2021)
- Figure 3.1 :
  - *domaine public*
  - *domaine public*
  - Indigodelta

    - Trondheim Havn
- Figure 3.2 : William Acton
- Figure 3.3 : *domaine public*
- Figure 3.4 :
    - Bernard Gagnon
    - Diego Delso
    - Bernard Gagnon
    - Hobe / Holger Behr
    - Odilia
    - Dror Feitelson
    - Dror Feitelson
    - Dror Feitelson
    - Dror Feitelson
    - Dror Feitelson
- Figure 4.4 :
    - jfgornet
    - Chris j wood
    - Dan DeLuca
    - Mathieu.clabaut
    - Harry Pot / Anefo
    - Kalervo Manninen
    - *domaine public*
- Figure 5.1 :
    - Ed Uthman, MD
    - Jorge Láscar
    - Arne Müseler
- Figure 5.2 :
    - Cecil W. Stoughton
- Figure 5.3 :
    - Official White House Photo by Pete Souza
    - United States Senate
- Figure 5.4 :
    - Studio Harcourt
    - Georges Biard
    - L.E.rewi-sor
    - Hullbr3ach
- Figure 5.5 :

  - Michael
  - Jvhertum
  - Allan Ramsay, National Portrait Gallery
  - Yousuf Karsh. Library and Archives Canada, e010751643
  - Diego Delso
  - Vangeliste, engraved after a painting by Richard Brompton
  - Herbert Art Gallery and Museum
  - Mark Fosh
- Figure 6.2 :
  - Michael Graham
  - Paul Kennedy
  - Little Mountain 5
  - Ammar shaker
  - Rama
  - S.Nitzold
- Figure 6.3 :
  - Theo Crazzolara
  - Thomas Wolf
  - Official Administration photograph
  - Andrew Bossi
  - Anna Belial
  - Margoz
- Tableau B.1 :
  - Poliphilo
  - Staff Sergeant Teddy Wade, United States Army
  - Roger Murmann
  - King of Hearts

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : Répondre aux questions visuelles à propos d'entités nommées

Mots clés : questions visuelles, recherche d'information multimodale, apprentissage de représentation, entités nommées, pré-entraînement, système de question-réponse

Résumé : Cette thèse se positionne à l'intersection de plusieurs domaines de recherche, le traitement automatique des langues, la Recherche d'Information (RI) et la vision par ordinateur, qui se sont unifiés autour des méthodes d'apprentissage de représentation et de pré-entraînement. Dans ce contexte, nous avons défini et étudié une nouvelle tâche multimodale : répondre aux questions visuelles à propos d'entités nommées (KVQAE). Dans ce cadre, nous nous sommes particulièrement intéressé aux interactions cross-modales et aux différentes façons de représenter les entités nommées. Nous avons également été attentifs aux données utilisées pour entraîner mais surtout évaluer les systèmes de question-réponse à travers différentes métriques. Plus précisément, nous avons proposé à cet effet un jeu de données, le premier de KVQAE comprenant divers types d'entités. Nous avons également défini un cadre expérimental pour traiter la KVQAE en deux étapes grâce à une base de connaissances non-structurée et avons identifié la RI comme principal verrou de la KVQAE, en particulier pour les questions à propos d'entités non-personnes. Afin d'améliorer l'étape de RI, nous avons étudié différentes méthodes de fusion multimodale, lesquelles sont pré-entraînées à travers une tâche originale : l'*Inverse Cloze Task* multimodale. Nous avons trouvé que ces modèles exploitaient une interaction cross-modale que nous n'avions pas considéré à l'origine, et qui permettrait de traiter l'hétérogénéité des représentations visuelles des entités nommées. Ces résultats ont été renforcés par une étude du modèle CLIP qui permet de modéliser cette interaction cross-modale directement. Ces expériences ont été menées tout en restant attentif aux biais présents dans le jeu de données ou les métriques d'évaluation, notamment les biais textuels qui affectent toute tâche multimodale.

Title : Knowledge-based Visual Question Answering about Named Entities

Keywords : Visual Question Answering, Multimodal Information Retrieval, Representation Learning, Named Entities, Pre-training, Question Answering System

Abstract : This thesis is positioned at the intersection of several research fields, Natural Language Processing, Information Retrieval (IR) and Computer Vision, which have unified around representation learning and pre-training methods. In this context, we have defined and studied a new multimodal task : Knowledge-based Visual Question Answering about Named Entities (KVQAE). In this context, we were particularly interested in cross-modal interactions and different ways of representing named entities. We also focused on data used to train and, more importantly, evaluate Question Answering systems through different metrics. More specifically, we proposed a dataset for this purpose, the first in KVQAE comprising various types of entities. We also defined an experimental framework for dealing with KVQAE in two stages through an unstructured knowledge base and identified IR as the main bottleneck of KVQAE, especially for questions about non-person entities. To improve the IR stage, we studied different multimodal fusion methods, which are pre-trained through an original task : the Multimodal Inverse Cloze Task. We found that these models leveraged a cross-modal interaction that we had not originally considered, and which may address the heterogeneity of visual representations of named entities. These results were strengthened by a study of the CLIP model, which allows this cross-modal interaction to be modeled directly. These experiments were carried out while staying aware of biases present in the dataset or evaluation metrics, especially of textual biases, which affect any multimodal task.