Recherche cross-modale pour répondre à des questions visuelles

Paul Lerner[1] Olivier Ferret[2] Camille Guinaudeau[3]
(1) Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
(2) Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
(3) Université Paris-Saclay, CNRS, JFLI, 101-0003, Tokyo, Japon
prenom.nom@lisn.upsaclay.fr, prenom.nom@cea.fr

RÉSUMÉ

Répondre à des questions visuelles à propos d'entités nommées (KVQAE) est une tâche difficile qui demande de rechercher des informations dans une base de connaissances multimodale. Nous étudions ici comment traiter cette tâche avec une recherche cross-modale et sa combinaison avec une recherche mono-modale, en se focalisant sur le modèle CLIP, un modèle multimodal entraîné sur des images appareillées à leur légende textuelle. Nos résultats démontrent la supériorité de la recherche cross-modale, mais aussi la complémentarité des deux, qui peuvent être combinées facilement. Nous étudions également différentes manières d'ajuster CLIP et trouvons que l'optimisation cross-modale est la meilleure solution, étant en adéquation avec son pré-entraînement. Notre méthode surpasse les approches précédentes, tout en étant plus simple et moins coûteuse. Ces gains de performance sont étudiés intrinsèquement selon la pertinence des résultats de la recherche et extrinsèquement selon l'exactitude de la réponse extraite par un module externe. Nous discutons des différences entre ces métriques et de ses implications pour l'évaluation de la KVQAE.

ABSTRACT

Cross-modal retrieval for Knowledge-based Visual Question Answering

Knowledge-based Visual Question Answering about named Entities (KVQAE) is a challenging task that requires retrieving information from a multimodal Knowledge Base. To tackle this task, we study cross-modal retrieval and its combination with mono-modal retrieval. We focus on the CLIP model, a multimodal model trained on images paired with their textual caption. We show that cross-modal outperforms mono-modal retrieval but also that the two are complementary and can be easily combined. We show that cross- outperforms mono-modal retrieval but also that the two are complementary and can be easily combined. We also study different fine-tuning strategies for CLIP and find that cross-modal is again the best solution, as it matches its pre-training. Our method outperforms previous approaches, while being conceptually simpler and computationally cheaper. These performance gains are studied intrinsically according to the relevance of the retrieved documents and extrinsically according to the accuracy of the answer extracted by an external module. We discuss the differences between these metrics and their implications for KVQAE evaluation.

MOTS-CLÉS : questions visuelles, multimodalité, recherche cross-modale, entités nommées.

KEYWORDS: visual question answering, multimodality, cross-modal retrieval, named entities.

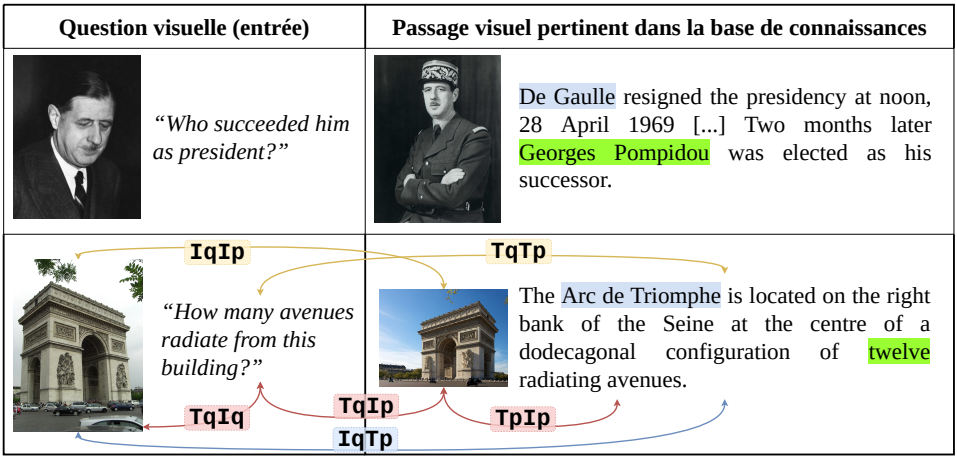

FIGURE 1 – Deux exemples de questions visuelles du jeu de données ViQuAE accompagnées de passages visuels pertinents issus de sa base de connaissances ainsi que d'une illustration des différents types d'interactions mono- et cross-modales étudiés (montrée seulement pour la deuxième question). Les sigles des interactions sont composés des lettres T (Texte), I (Image), Q (question) et P (passage).

1 Introduction

Répondre à des questions visuelles à propos d'entités nommées (KVQAE [1]) est une tâche difficile qui demande d'analyser des données multimodales (Shah *et al.*, 2019; Lerner *et al.*, 2022, 2023). Les représentations multimodales d'entités nommées y sont centrales, ce qui lie cette tâche à la désambiguïsation multimodale d'entités nommées (Adjali *et al.*, 2020). La KVQAE, comme d'autres tâches multimodales, vise à fluidifier et rendre plus naturelle l'interaction entre l'utilisateur et la 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. La Figure 1 montre deux exemples de questions visuelles ainsi que des passages visuels pertinents correspondants, tirés du jeu de données ViQuAE (Lerner *et al.*, 2022) et de sa Base de Connaissances (BC) multimodale. Une question visuelle est constituée plus précisément d'une question textuelle, accompagnée d'une image qui lui est liée, 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. Par ailleurs, à la différence de sa définition la plus courante en représentation des connaissances, le terme de *Base de Connaissances* renvoie ici à un contenu non structuré, formé de passages visuels.

Contrairement aux questions visuelles classiques (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. D'autres travaux se situent à mi-chemin mais sont limités à des catégories d'objet « gros grain », par exemple « personne » et « monument » au lieu de « Charles de Gaulle » et « Arc de Triomphe » (Marino *et al.*, 2019). Dans cette étude, 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

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

passage visuel (TPIP) ou entre les deux [2] (IQTP), comme illustré par la Figure 1. Nous nous focalisons principalement sur cette dernière interaction. La KVQAE ayant été introduite très récemment, elle reste largement à explorer. Shah *et al.* (2019) et Lerner *et al.* (2022) l'ont abordée en s'appuyant sur des représentations spécialisées pour les images, en l'occurrence en lien avec les visages, tandis que Lerner *et al.* (2023) ont proposé une méthode de pré-entraînement pour la Recherche d'Information (RI) multimodale. Leur méthode combine implicitement recherche mono- et cross-modale mais demande un pré-entraînement coûteux et n'exploite pas les modèles pré-entraînés existant pour la recherche cross-modale, tels que CLIP (Radford *et al.*, 2021).

CLIP est fondé sur un double encodeur, un pour chaque modalité, entraîné à partir d'images et de leurs légendes de façon cross-modale. Cette architecture permet d'être très efficient à la fois pour la recherche cross-modale mais aussi pour l'entraînement dans un cadre d'apprentissage contrastif, c'est-à-dire en comparant les représentations d'exemples selon leur similarité sémantique. CLIP s'est imposé comme un modèle fondateur (Bommasani *et al.*, 2021) de par ses multiples applications en vision par ordinateur, traitement automatique des langues et recherche cross-modale (Radford *et al.*, 2021; Ramesh *et al.*, 2021; Mokady *et al.*, 2021; Wolfe & Caliskan, 2022). Dans ce travail, nous montrons comment aborder la KVQAE grâce à CLIP, à la fois pour la recherche mono-modale mais aussi cross-modale, les deux pouvant être combinées aisément. Lerner *et al.* (2022) ont utilisé l'encodeur visuel de CLIP (CLIP_V) pour la recherche visuelle mais pas son encodeur textuel (CLIP_T). Lerner *et al.* (2023) ont combiné CLIP_V et BERT (Devlin *et al.*, 2019) pour la RI multimodale mais suggèrent que les bénéfices de leur modèle viennent des interactions cross-modales entre l'image de la question et le texte du passage visuel (IQTP, en bleu dans la Figure 1). Or, celles-ci sont possibles directement avec CLIP_T. De plus, la méthode de pré-entraînement de Lerner *et al.* (2023) est coûteuse. Nous montrons ici comment ajuster CLIP avec le peu de données de ViQuAE (Lerner *et al.*, 2022), sans pré-entraînement supplémentaire.

Nos résultats suggèrent que la recherche cross-modale est supérieure à la mono-modale, ce qui est intéressant car les modèles cross-modaux tels que CLIP peuvent être entraînés de façon faiblement supervisée. De plus, nous montrons que les deux sortes de recherche peuvent être combinées très facilement, sans entraînement supplémentaire, et que cette recherche hybride apporte des gains de performance conséquents. Enfin, nos expérimentations suggèrent que l'ajustement (*fine-tuning*) de CLIP est plus efficace en condition cross-modale, comme pour son pré-entraînement. Les gains de performance sont étudiés intrinsèquement selon la pertinence des résultats de la RI et extrinsèquement selon l'exactitude de la réponse extraite par un module externe. Nous discutons des différences entre ces métriques.

2 Travaux connexes

Notre travail se situe à l'intersection de plusieurs domaines de la RI : textuelle, visuelle et cross-modale, appliquée à des questions ou plus généralement à des requêtes. L'apprentissage de représentations denses pour la RI est transverse à ces trois domaines. Avec une représentation adéquate, la RI se ramène alors à un problème de recherche des plus proches voisins.

2. Nous ne nous intéressons pas à l'interaction TQIP entre l'image de la question et le texte du passage comme expliqué à la Section 3.

Recherche cross-modale Couairon *et al.* (2022) explorent les possibilités d'analogie multimodale avec les représentations de CLIP. Par exemple, la reine est au roi ce qu'une photo de femme est à une photo d'homme. Ils montrent que les représentations de CLIP ne sont pas à l'origine adaptées pour cet usage mais peuvent être ajustées facilement avec une simple projection linéaire en gardant le même objectif d'entraînement. Puisque ces analogies prennent la forme d'opérations arithmétiques sur des vecteurs, elles combinent recherche mono- et cross-modale de manière assez semblable à notre travail. Sun *et al.* (2022) 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 que CLIP surpasse un modèle de reconnaissance faciale, même en *zero-shot*. Nos résultats suggèrent le contraire mais Sun *et al.* (2022) utilisent un modèle de reconnaissance faciale différent du nôtre et ne précisent pas 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 fixes et donc, en ajoutant un perceptron multi-couches. Il reprend en cela le modèle CLIP-Adapter de Gao *et al.* (2021), qui supposent que l'ajustement de l'ensemble des paramètres de CLIP conduirait inévitablement au sur-apprentissage. Toutefois, Gao *et al.* (2021) 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. Wang *et al.* (2022) utilisent quant à eux un objectif d'apprentissage contrastif à la fois mono- et cross-modal ayant pour fin la recherche cross-modale. Contrairement à notre cadre où l'on vise à rapprocher une image du nom de l'entité représentée et de son image de référence, les auteurs regroupent les représentations mono-modales par catégorie « gros grain » (par exemple les 20 catégories du jeu de données Pascal ; Rashtchian *et al.*, 2010).

Questions cross-modales Liu *et al.* (2023) travaillent sur le jeu de questions cross-modales WebQA (Chang *et al.*, 2022). Bien que WebQA ait à l'origine été proposé comme une évaluation de *reading comprehension*, Liu *et al.* (2023) traitent ses questions cross-modales sorties de leur contexte et travaillent donc sur la partie RI, visant justement à retrouver le contexte pertinent. On pourrait ainsi qualifier WebQA de « questions visuelles sans images » car les questions sont purement textuelles mais les réponses se trouvent dans une image. Par exemple, on peut répondre « *helmet* » à la question « *What is the sculpted bust at the Baroque library, Prague wearing on its head ?* » à condition de trouver une image pertinente. Liu *et al.* (2023) utilisent CLIP pour chercher des images à partir de la question. Ils exploitent également la légende des images et combinent ainsi les informations de manière similaire à la nôtre (cf. Section 3), sauf que leur requête est textuelle et non pas visuelle.

Questions visuelles de sens commun Gui *et al.* (2022) intègrent CLIP dans un encodeur-décodeur T5 (Raffel *et al.*, 2020) entraîné à générer la réponse à la manière de Lewis *et al.* (2020). CLIP, qui reste fixé, sert alors à la recherche cross-modale entre l'image de la question et le nom d'une entité accompagné de sa description dans un sous-graphe de Wikidata. Les auteurs expérimentent avec le jeu de données OK-VQA (Marino *et al.*, 2019), qui se focalise sur des questions de sens commun (*commonsense*) et vise ainsi des catégories d'objets « gros grain » plutôt que des entités nommées.

Questions visuelles à propos d'entités nommées (KVQAE) Garcia-Olano *et al.* (2022) et Heo *et al.* (2022) travaillent sur la KVQAE avec le jeu de données de Shah *et al.* (2019). Cependant, il est difficile de comparer leurs approches à la nôtre car leurs systèmes prennent en entrée la légende de

l'image, ce qui rend l'image elle-même redondante. Shah *et al.* (2019) ont proposé KVQA, le premier jeu de données pour la KVQAE. Ils 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). Notre travail est plus proche de Lerner *et al.* (2022), qui utilisent une BC fondée sur Wikipédia, faite donc d'images et de textes non-structurés (comme dans la Figure 1). Ils traitent la tâche en deux étapes, où l'extraction des réponses suit la RI. La RI est une combinaison de deux recherches mono-modales : textuelle avec DPR (Karpukhin *et al.*, 2020) et visuelle avec une combinaison de CLIP_V, ArcFace (Deng *et al.*, 2019) et un modèle ResNet entraîné sur ImageNet (He *et al.*, 2016; Deng *et al.*, 2009). Nous cherchons d'une part à simplifier ce système en supprimant la dépendance vis-à-vis d'ArcFace et ImageNet, deux modèles supervisés qui fournissent a priori des représentations moins génériques que CLIP, et d'autre part à exploiter pleinement CLIP en combinant recherche mono-modale et cross-modale. Après la RI, Lerner *et al.* (2022) extraient les réponses des passages de texte grâce à BERT multi-passage (Wang *et al.*, 2019). Pour leur part, Lerner *et al.* (2023) se sont, comme nous, focalisés sur la RI. Afin de modéliser les interactions cross-modales TQIQ et TPIP, ils représentent de façon jointe le texte et l'image, inspirés par les BERT multimodaux qui dominent les travaux sur les questions visuelles classiques ces dernières années (Khan *et al.*, 2022; Gan *et al.*, 2022). Néanmoins, ces architectures demandent un pré-entraînement coûteux et Lerner *et al.* (2023) suggèrent finalement que leur modèle exploite surtout l'interaction IQTP. Nos conclusions se rejoignent car notre modèle surpasse le leur — sans pré-entraînement supplémentaire — en modélisant explicitement IQTP via CLIP, comme décrit dans la section suivante.

3 Méthodes

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) (cf. Figure 1). 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 (cf. Figure 1) :

$$s(\mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = \alpha_I s_I(\mathbf{i_q}, \mathbf{i_p}) + \alpha_C s_C(\mathbf{i_q}, \mathbf{t_p}) \quad (1)$$

où les paramètres $\alpha_{\{I,C\}}$ pondèrent chaque similarité. Cette décomposition nous permet d'exploiter directement des modèles pré-entraînés. Plus précisément, dans cette étude, 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 (*optimisation mono-modale*) ou bien de son nom (*optimisation cross-modale*) ou les deux de façon jointe.

3.1 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 batch, 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 batch. 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 batch en minimisant l'objectif suivant, étant donné τ la température :

$$-\log \frac{\exp\left(s(\mathbf{i_q}, \mathbf{t_p}^{(+)}, \mathbf{i_p}^{(+)})e^{\tau}\right)}{\exp\left(s(\mathbf{i_q}, \mathbf{t_p}^{(+)}, \mathbf{i_p}^{(+)})e^{\tau}\right) + \sum_j \exp\left(s(\mathbf{i_q}, \mathbf{t_p}^{(j)}, \mathbf{i_p}^{(j)})e^{\tau}\right)} \tag{2}$$

Puisque nous implémentons $s_C(\mathbf{i_q}, \mathbf{t_p})$ avec CLIP :

$$s_C(\mathbf{i_q}, \mathbf{t_p}) = \cos\left(\mathrm{CLIP}_V(\mathbf{i_q}), \mathrm{CLIP}_T(\mathbf{t_p})\right) \tag{3}$$

Cet objectif correspond à celui utilisé pendant le pré-entraînement de CLIP si $\alpha_I = 0$ et $\alpha_C = 1$ (optimisation cross-modale 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 batch de manière très efficace (il suffit d'un produit matriciel pour calculer le dénominateur de l'équation 2). Nous implémentons $s_I(\mathbf{i_q}, \mathbf{i_p})$ de manière similaire : $\cos\left(\mathrm{CLIP}_V(\mathbf{i_q}), \mathrm{CLIP}_V(\mathbf{i_p})\right)$.

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 :

$$s(\mathbf{t_q}, \mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = \alpha_T s_T(\mathbf{t_q}, \mathbf{t_p}) + \alpha_I s_I(\mathbf{i_q}, \mathbf{i_p}) + \alpha_C s_C(\mathbf{i_q}, \mathbf{t_p}) \tag{4}$$

Nous discutons des difficultés à optimiser ces trois similarités de façon jointe dans la Section 4. De ce fait, $s_T(\mathbf{t_q}, \mathbf{t_p})$ est implémenté par un modèle entraîné séparément et les poids $\alpha_{\{T,I,C\}}$ sont déterminés par dichotomie sur le jeu de validation pour maximiser le rang réciproque moyen en contraignant leur somme à 1.

3.2 Baselines

Nous comparons notre approche au modèle de Lerner *et al.* (2022), qui combine DPR, CLIP_V, ArcFace et un modèle ResNet entraîné sur ImageNet. DPR est fondé sur deux encodeurs BERT (Devlin *et al.*, 2019) [3] : un pour la question et un pour le passage. Dans notre cas, il implémente $s_T(\mathbf{t_q}, \mathbf{t_p}) = \mathrm{DPR}(\mathbf{t_q}, \mathbf{t_p}) = \mathrm{BERT}_q(\mathbf{t_q})_{[\mathrm{CLS}]} \cdot \mathrm{BERT}_p(\mathbf{t_p})_{[\mathrm{CLS}]}$. Il est d'abord pré-entraîné sur TriviaQA (Joshi *et al.*, 2017) avant d'être ajusté sur ViQuAE. Les autres modèles sont disponibles publiquement et ne sont pas ajustés [4]. Les résultats des quatre modèles sont combinés de la même façon que dans l'équation 4, où DPR implémente $s_T(\mathbf{t_q}, \mathbf{t_p})$; CLIP_V, ArcFace, et ImageNet composent

3. `bert-base-uncased` disponible dans la bibliothèque Transformers.

4. ArcFace est disponible à https://github.com/deepinsight/insightface et ImageNet dans `torchvision`. Les deux utilisent une architecture ResNet-50.

Modèle	Mono-modal		Cross-modal		
	TQTP	IQIP	TQIQ	TPIP	IQTP
DPR	✓				
DPR + CLIP mono-modal *zero-shot*	✓	✓			
DPR et reconnaissance faciale (Lerner *et al.*, 2022)	✓	✓			
ECA (Lerner *et al.*, 2023)	✓	✓	✓	✓	✓
ILF (Lerner *et al.*, 2023)	✓	✓			✓
DPR + CLIP mono- et cross-modal ajusté	✓	✓			✓

TABLE 1 – Récapitulatif des différentes interactions mono- et cross-modales utilisées par les modèles étudiés.

$s_I(\mathbf{i_q}, \mathbf{i_p})$ 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 seulement sur les entités nommées de type personne dans 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 [5] :

$$s_I(\mathbf{i_q}, \mathbf{i_p}) = FH\alpha_A s_A(\mathbf{i_q}, \mathbf{i_p}) + (1 - F)(1 - H)\left(\alpha_V s_V(\mathbf{i_q}, \mathbf{i_p}) + \alpha_R s_R(\mathbf{i_q}, \mathbf{i_p})\right) \tag{5}$$

Pour rendre les scores de ces différents modèles comparables, ils sont centrés-réduits. De plus, quand un document 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, selon la technique du « minimum par défaut » de Ma *et al.* (2021).

Nous nous comparons également aux modèles ECA et ILF de Lerner *et al.* (2023). ECA (*Early Cross-Attention*) fusionne les modalités de manière précoce à l'aide d'un mécanisme d'attention, comme son nom l'indique. La similarité est donc calculée suivant $s(\mathbf{t_q}, \mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = \text{ECA}(\mathbf{t_q}, \mathbf{i_q}) \cdot \text{ECA}(\mathbf{t_p}, \mathbf{i_p})$ et combine ainsi toutes les interactions multimodales présentées à la Figure 1. ILF (*Intermediate Linear Fusion*) 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 :

$$s(\mathbf{t_q}, \mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = s_T(\mathbf{t_q}, \mathbf{t_p}) + s_{C'}(\mathbf{t_q}, \mathbf{i_p}) + s_I(\mathbf{i_q}, \mathbf{i_p}) + s_C(\mathbf{i_q}, \mathbf{t_p}) \tag{6}$$

Les différentes interactions mono- et cross-modales utilisées par les modèles étudiés sont résumées dans le Tableau 1.

Il est à noter que Lerner *et al.* (2022) et Lerner *et al.* (2023) emploient CLIP_V avec l'architecture ResNet tandis que nous utilisons ViT (Dosovitskiy *et al.*, 2021) dans la plupart de nos expériences (mais comparons les deux dans la Section 5 sans trouver de différence significative) [6].

5. On connaît seulement le type d'entité des images $\mathbf{i_p}$ de la BC, *pas* celles des questions $\mathbf{i_q}$.
6. Plus précisément, il s'agit de `RN50` et `ViT-B/32` disponibles à `https://github.com/openai/CLIP`

4 Implémentation

4.1 Données

Nous utilisons la BC proposée par Lerner *et al.* (2022), qui consiste en 1,5 millions d'articles Wikipédia et images des entités Wikidata correspondantes. Les articles sont divisés en 12 millions de passages de 100 mots. Par conséquent, tous les passages d'un même article partagent la même image. Deux exemples de passages visuels sont montrés à la Figure 1. Par la suite, nous évaluons les méthodes à deux niveaux de RI : article et passage.

Notre étude se focalise sur ViQuAE, un des deux seuls jeux de données pour le KVQAE. Nous n'expérimentons pas avec l'autre, KVQA (Shah *et al.*, 2019), pour les mêmes raisons que Lerner *et al.* (2023) : KVQA ayant été généré automatiquement à partir de Wikidata, rien ne garantit que les réponses se trouvent dans la BC. De plus, il comprend 29 % de questions booléennes (réponse oui/non) pour lesquelles la pertinence du passage ne peut pas être évaluée automatiquement sur la base de la présence de la réponse à la question.

ViQuAE contient 3 700 questions visuelles à propos de 2 400 entités différentes, réparties aléatoirement en ensembles de taille égale pour l'entraînement, la validation et le test, sans recouvrement entre les images. Par conséquent, le recouvrement entre les entités du jeu d'entraînement et de test 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. Notons que tout le texte, des questions comme de la BC, est en anglais.

4.2 Problème de l'annotation de référence

Comme nous l'avons indiqué à la Section 3.1, l'optimisation de la similarité textuelle entre question et passage $s_T(\mathbf{t_q}, \mathbf{t_p})$, implémentée par DPR (cf. Section 3.2), avec les deux autres similarités ne s'est pas accompagnée dans nos expériences d'améliorations significatives au niveau des résultats. Nous interprétons ce constat comme une conséquence des incohérences induites par l'annotation de référence des passages concernant la modalité visuelle, ce qui nuit à l'entraînement d'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 Lerner *et al.* (2023) ont réussi à entraîner leurs modèles, ECA et ILF (cf. Section 3.2), avec l'annotation au niveau du passage. Ce succès pourrait être expliqué par la représentation jointe d'ECA (qui modélise TQIQ et TPIP) ou par l'expressivité d'ILF. Une autre explication, pas forcément 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.

À la place de cette annotation au niveau du passage, nous utilisons l'annotation au niveau de l'entité fournie par Lerner *et al.* (2022) 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 [7].

4.3 Hyperparamètres

Pour profiter au mieux des entités associées aux autres images du batch $\mathbf{t_p}^{(j)}$ et $\mathbf{i_p}^{(j)}$, nous utilisons un batch 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 batch 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 [8] 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 de Lerner *et al.* (2023) avec le même matériel [9].

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 & Hutter, 2019), avec $\lambda = 0{,}1$. Pour l'optimisation jointe, 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 $4{,}6$ [10]. 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 batch sur le jeu de validation, c'est-à-dire en réordonnant les images ou textes du batch selon le score de similarité s, pour éviter de calculer les représentations de toute la BC à chaque époque.

Notre implémentation est fondée sur Lightning [11], 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. Notre code est disponible librement à `https://github.com/PaulLerner/ViQuAE`.

5 Résultats

Nous évaluons la RI à deux niveaux :

— article (qui contient plusieurs passages ; cf. Section 4.1) ;

— passage visuel, afin de pouvoir nous comparer aux autres méthodes (Lerner *et al.*, 2022, 2023).

Dans les deux cas, un document (passage ou article entier) est jugé pertinent s'il contient la réponse après prétraitement standard (insensibilité à la casse, aux déterminants et à la ponctuation). Les métriques utilisées sont la précision à K (P@K) et le rang réciproque moyen (MRR) ainsi que Hits@K

7. Parce qu'elles n'ont pas d'images libres de droit.
8. C'est-à-dire commence à sur-apprendre.
9. Lerner *et al.* (2023) rapportent un bilan carbone de 1,7 kgCO2e pour trois jours de consommation électrique des GPUs.
10. 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.
11. `https://www.pytorchlightning.ai/`

Recherche	Optimisation	MRR	P@1	P@20	Hits@20
Mono-modale (IqIp)	Non (*zero-shot*)	29,4	21,8	9,1	53,4
	Mono-modale	30,0	21,8	9,2	55,7
	Cross-modale	29,8	21,4	**9,5**	54,7
	Jointe	**30,4**	**22,0**	**9,5**	**55,8**
Cross-modale (IqTp)	Non (*zero-shot*)	32,7	23,1	10,9	60,6
	Mono-modale	31,6	21,9	10,9	59,6
	Cross-modale	**37,1**	**26,9**	**11,9**	**67,8**
	Jointe	30,8	21,3	10,4	59,5
Fusion	Non (*zero-shot*)	39,6	30,6	11,8	63,9
	Mono-modale	40,1	31,8	11,6	63,6
	Cross-modale	***44,1***	***34,9***	***12,7***	***69,9***
	Jointe	41,0	32,6	11,6	64,9
	Disjointe	43,7	34,5	***12,7***	***69,9***

TABLE 2 – Validation des différentes méthodes d'ajustement de CLIP (ainsi que la version *zero-shot* pour référence) 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 sur le sous-ensemble de validation de ViQuAE. 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 et sont marqués en gras italique. Fusion de l'optimisation disjointe : recherche mono-modale optimisée de manière mono-modale et idem pour cross-modale.

(équivalent au rappel en considérant qu'il n'y a qu'un seul document pertinent par question visuelle). P@1 et Hits@1 sont équivalents.

Une fois la RI effectuée au niveau du passage visuel, les réponses sont extraites à l'aide du modèle BERT multi-passage entraîné par Lerner *et al.* (2022). Deux métriques sont utilisées pour évaluer les réponses : l'appariement exact et le score F1 (au niveau des sacs de mots) entre la réponse extraite et la vérité terrain [12].

5.1 Recherche d'information au niveau de l'article

Nous explorons dans un premier temps trois modes d'optimisation 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 1 :

— recherche/optimisation mono-modale entre les deux images, soit $\alpha_I = 1, \alpha_C = 0$;
— recherche/optimisation cross-modale entre l'image et le nom de l'entité, soit $\alpha_I = 0, \alpha_C = 1$;
— fusion des deux recherches ou optimisation jointe, soit $\alpha_I > 0, \alpha_C > 0$.

À noter que le mode d'optimisation n'influence pas le mode de recherche, comme en témoigne le Tableau 2. Pour mémoire, CLIP est pré-entraîné de manière cross-modale uniquement (Radford *et al.*, 2021).

12. Ou plutôt les vérités terrains car les alias Wikipédia d'une entité constituent une réponse valide.

Recherche mono- ou cross-modale ? Avant de comparer les différentes méthodes d'optimisation, nous pouvons d'ores et déjà remarquer que la RI cross-modale l'emporte [13] systématiquement sur la RI mono-modale, notamment en *zero-shot* [14], ce qui peut paraître surprenant puisque les noms propres portent a priori peu de sémantique. On s'étonne donc que CLIP parvienne à généraliser [15] 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 et suggérer sa nationalité. De plus, nous travaillons ici avec 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.

Pourquoi choisir ? Nous montrons que les recherches mono- et cross-modales sont complémentaires : leurs résultats peuvent être simplement combinés au niveau du score (comme dans l'équation 1). Pour l'optimisation jointe, nous pourrions utiliser directement les poids α optimisés par descente de gradient avec le reste du modèle sur le jeu d'entraînement, mais cela détériore légèrement les résultats. Ainsi, en *zero-shot*, 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 sur le Web.

Quelle optimisation ? On peut voir que l'optimisation cross-modale améliore légèrement la recherche mono-modale mais pas l'inverse. Dans les trois cas, l'optimisation dans un mode améliore au moins la recherche dans le même mode. Il est intéressant de noter que l'optimisation jointe détériore la RI cross-modale mais améliore la fusion (toujours par rapport au *zero-shot*). Mais au total, l'optimisation cross-modale semble être la meilleure option, surpassant significativement l'optimisation mono-modale. 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 mono-modale pénalise l'optimisation jointe. On voit également que les différences entre les modes d'optimisation de la recherche mono-modale sont très faibles : il n'est pas bénéfique de combiner la recherche mono-modale optimisée de manière mono-modale et la recherche cross-modale optimisée de manière cross-modale (« optimisation disjointe » dans le Tableau 2). Par conséquent, dans la suite de l'article nous utilisons le modèle entraîné de manière cross-modale et présentons les résultats sur le jeu de test.

5.2 Recherche d'information au niveau du passage visuel

Les résultats sont présentés dans le Tableau 3. Nous utilisons comme *baseline* DPR (recherche avec la question seulement) ainsi que sa fusion avec la recherche mono-modale de CLIP *zero-shot* (résultats rapportés par Lerner *et al.*, 2023). Puisque ces résultats, ainsi que ceux des autres modèles de Lerner

13. Significativement selon le test de randomisation de Fisher avec $p \leq 0{,}01$ (Fisher, 1937; Smucker *et al.*, 2007).
14. C'est-à-dire sans ajustement sur ViQuAE.
15. À moins que son jeu de pré-entraînement ne contienne suffisamment d'entités de ViQuAE pour que ce ne soit pas nécessaire. Nous développons cette discussion dans la Section 7.

Modèle	MRR	P@1	P@20	Hits@20
DPR	32,8	22,8	16,4	61,2
DPR + CLIP mono-modal *zero-shot*	34,5	24,8	15,8	61,8
DPR + CLIP* mono-modal *zero-shot*	34,7	24,3	16,0	62,8
DPR et reconnaissance faciale (Lerner *et al.*, 2022)	**37,9**	27,8	17,5	65,7
ECA (Lerner *et al.*, 2023)	37,8	26,7	**19,5**	**67,6**
ILF (Lerner *et al.*, 2023)	37,3	26,8	19,1	66,9
DPR + CLIP* mono- et cross-modal ajusté	37,6	**28,6**	16,3	63,6

TABLE 3 – Résultats de la RI évaluée au niveau du passage visuel sur le jeu de test de ViQuAE. *CLIP fondé sur l'architecture ViT au lieu de ResNet.

Recherche d'Information	Appariement exact	F1
DPR	16,9 ± 0,4	20,1 ± 0,5
DPR + CLIP mono-modal *zero-shot*	19,0 ± 0,4	22,3 ± 0,4
DPR + CLIP* mono-modal *zero-shot*	19,7 ± 0,9	23,3 ± 0,8
DPR et reconnaissance faciale (Lerner *et al.*, 2022)	22,1 ± 0,5	25,4 ± 0,4
ECA (Lerner *et al.*, 2023)	20,6 ± 0,3	24,4 ± 0,2
ILF (Lerner *et al.*, 2023)	21,3 ± 0,6	25,4 ± 0,3
DPR + CLIP* mono- et cross-modal ajusté	**24,7** ± 0,5	**28,7** ± 0,4

TABLE 4 – Résultats de l'extraction des réponses sur le jeu de test de ViQuAE. Moyennes sur 5 entraînements du modèle d'extraction avec des graines aléatoires différentes. Ce modèle prend en entrée le top-24 des différents systèmes de RI. *CLIP fondé sur l'architecture ViT au lieu de ResNet.

et al. (2022) et Lerner *et al.* (2023), 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 précision@1 de 3 % relativement au modèle de Lerner *et al.* (2022), 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 de Lerner *et al.* (2023), sans pré-entraînement supplémentaire. Les différences de MRR avec ces modèles sont très faibles, mais ECA et ILF surpassent notre méthode en P@20 et Hits@20, ce qui suggérerait un avantage de la représentation jointe d'ECA et de l'expressivité d'ILF. Nous discutons davantage de ces métriques dans la Section 6.

On peut voir que les améliorations par rapport à la baseline *DPR + CLIP mono-modal zero-shot* sont assez modestes, beaucoup plus faibles que dans la section précédente où nous étudions les résultats au niveau de l'article et avant la fusion avec DPR. Nous verrons dans la section suivante que l'impact sur l'extraction des réponses est, lui, plus important, et démontre la supériorité de notre approche.

5.3 Extraction des réponses

Nous suivons le même protocole que Lerner *et al.* (2023) pour extraire les réponses, c'est-à-dire que nous utilisons le modèle fourni par Lerner *et al.* (2022), pré-entraîné sur TriviaQA puis ajusté sur

Question visuelle (entrée)	DPR + CLIP mono- et cross-modal ajusté	ECA
 *"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 2 – Exemples qualitatifs où BERT multi-passage parvient à extraire la réponse du passage pertinent fourni par notre méthode de RI tandis qu'il est distrait par le passage fourni par ECA (Lerner *et al.*, 2023), qui contient beaucoup de réponses plausibles mais n'est pas vraiment pertinent (tout en étant considéré comme tel car il contient la réponse).

ViQuAE, qui prend 24 passages en entrée, lesquels varient selon les systèmes de RI.

Les résultats sont présentés dans le Tableau 4. On voit que la recherche cross-modale et l'ajustement de CLIP apportent 23 % à 25 % d'amélioration relativement à la baseline *DPR + CLIP mono-modal zero-shot* selon les métriques, ce qui est plus cohérent avec les résultats de la RI au niveau de l'article (cf. Section 5.1) où nous avions alors de 31 % à 60 % d'amélioration relative selon les métriques (avant la fusion avec DPR) [16]. Ainsi, notre méthode apporte des améliorations appréciables, de 12 % à 20 % selon les métriques et modèles : par rapport au modèle de Lerner *et al.* (2022), sans utiliser ArcFace, ni ImageNet, ni l'heuristique de la division de la BC entre personnes et non-personnes, et par rapport aux modèles de Lerner *et al.* (2023), sans pré-entraînement supplémentaire.

Nous discutons davantage de ces résultats et des différences entre les métriques dans la section suivante.

6 Discussion

La section précédente 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 est une évaluation extrinsèque. 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

16. La Section 5.1 présente les résultats sur le jeu de validation mais ils sont similaires sur le jeu de test.

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 deuxième exemple de la Figure 1, « *Combien y a-t-il d'avenues autour de ce bâtiment ?* », on peut énumérer les chiffres de 1 à 20, 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 dans Lerner *et al.* (2022). Au contraire, avec une RI purement visuelle, donc exempte de biais textuel, on récupère les passages de l'article Wikipédia de l'Arc de Triomphe ou d'autres monuments ressemblants. 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.

On peut observer ce phénomène quantitativement : entre la recherche purement textuelle (DPR) et la *baseline* multimodale (DPR + CLIP mono-modal *zero-shot*), il n'y a presque pas de différence pour les métriques de RI au niveau du passage (cf. Tableau 3), voire une détérioration de la précision@20, alors que les métriques d'extraction de réponse (cf. Tableau 4) montrent 16 % à 17 % d'amélioration relative pour la RI multimodale *baseline*. De la même manière, les modèles de fusion précoce de Lerner *et al.* (2023) 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 2.

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 et reconnaissance faciale*)
— le top-K (ici 24)

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

7 Conclusion

Dans cet article, nous étudions la recherche cross-modale et sa combinaison avec la recherche mono-modale pour répondre à des questions visuelles à propos d'entités nommées (KVQAE), en nous focalisant sur le modèle CLIP. Nos résultats démontrent la supériorité de la recherche cross-modale, mais aussi la complémentarité des deux, qui peuvent être combinées facilement. 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. Bien que ce soit l'abondance de données cross-modales qui ait permis d'entraîner un modèle avec la capacité de CLIP, ce qui aurait été difficile avec une annotation mono-modale, cela limite nos résultats car il est difficile de contrôler une telle masse de données et donc d'estimer les capacités de généralisation de CLIP. Nous étudions également différentes manières d'ajuster CLIP et trouvons que l'optimisation cross-modale 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 représentations.

Notre méthode surpasse la *baseline* (recherche mono-modale) mais aussi les méthodes de Lerner *et al.* (2022) et Lerner *et al.* (2023), 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 dans la Section 5.1, ou bien d'évaluer extrinsèquement les résultats via un modèle d'extraction de réponse (Section 5.3). 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'optimisation jointe 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.*, 2022).

Remerciements

Les auteurs remercient chaleureusement les relecteurs anonymes pour leur retour constructif ainsi qu'Antoine Chaffin pour les discussions à propos de CLIP et de la recherche cross-modale. Ce travail a été financé par le projet ANR-19-CE23-0028 MEERQAT. Il a en outre bénéficié d'un accès aux moyens de calcul de l'IDRIS au travers de l'allocation de ressources 2022-AD011012846R1 attribuée par GENCI.

Références

ADJALI O., BESANÇON R., FERRET O., LE BORGNE H. & GRAU B. (2020). Multimodal Entity Linking for Tweets. In J. M. JOSE, E. YILMAZ, J. MAGALHÃES, P. CASTELLS, N. FERRO, M. J. SILVA & F. MARTINS, Éds., *Advances in Information Retrieval*, Lecture Notes in Computer Science, p. 463–478, Cham : Springer International Publishing. DOI : 10.1007/978-3-030-45439-5_31.

ANTOL S., AGRAWAL A., LU J., MITCHELL M., BATRA D., ZITNICK C. L. & PARIKH D. (2015). VQA : Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, p. 2425–2433, Santiago, Chile : IEEE. DOI : 10.1109/ICCV.2015.279.

BASSANI E. (2022). ranx : A Blazing-Fast Python Library for Ranking Evaluation and Comparison. In M. HAGEN, S. VERBERNE, C. MACDONALD, C. SEIFERT, K. BALOG, K. NØRVÅG & V. SETTY, Éds., *Advances in Information Retrieval*, Lecture Notes in Computer Science, p. 259–264, Cham : Springer International Publishing. DOI : 10.1007/978-3-030-99739-7_30.

BOMMASANI R., HUDSON D. A., ADELI E., ALTMAN R., ARORA S., VON ARX S., BERNSTEIN M. S., BOHG J., BOSSELUT A., BRUNSKILL E., BRYNJOLFSSON E., BUCH S., CARD D., CASTELLON R., CHATTERJI N., CHEN A., CREEL K., DAVIS J. Q., DEMSZKY D., DONAHUE C., DOUMBOUYA M., DURMUS E., ERMON S., ETCHEMENDY J., ETHAYARAJH K., FEI-FEI L., FINN C., GALE T., GILLESPIE L., GOEL K., GOODMAN N., GROSSMAN S., GUHA N., HASHIMOTO T., HENDERSON P., HEWITT J., HO D. E., HONG J., HSU K., HUANG J., ICARD T., JAIN S., JURAFSKY D., KALLURI P., KARAMCHETI S., KEELING G., KHANI F., KHATTAB O., KOH P. W., KRASS M., KRISHNA R., KUDITIPUDI R., KUMAR A., LADHAK F., LEE M., LEE T.,

LESKOVEC J., LEVENT I., LI X. L., LI X., MA T., MALIK A., MANNING C. D., MIRCHANDANI S., MITCHELL E., MUNYIKWA Z., NAIR S., NARAYAN A., NARAYANAN D., NEWMAN B., NIE A., NIEBLES J. C., NILFOROSHAN H., NYARKO J., OGUT G., ORR L., PAPADIMITRIOU I., PARK J. S., PIECH C., PORTELANCE E., POTTS C., RAGHUNATHAN A., REICH R., REN H., RONG F., ROOHANI Y., RUIZ C., RYAN J., RÉ C., SADIGH D., SAGAWA S., SANTHANAM K., SHIH A., SRINIVASAN K., TAMKIN A., TAORI R., THOMAS A. W., TRAMÈR F., WANG R. E., WANG W., WU B., WU J., WU Y., XIE S. M., YASUNAGA M., YOU J., ZAHARIA M., ZHANG M., ZHANG T., ZHANG X., ZHANG Y., ZHENG L., ZHOU K. & LIANG P. (2021). On the Opportunities and Risks of Foundation Models. *arXiv :2108.07258 [cs]*. arXiv : 2108.07258.

CHANG Y., NARANG M., SUZUKI H., CAO G., GAO J. & BISK Y. (2022). Webqa : Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 16495–16504.

CHEN W., HU H., CHEN X., VERGA P. & COHEN W. (2022). MuRAG : Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 5558–5570, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics.

COUAIRON G., DOUZE M., CORD M. & SCHWENK H. (2022). Embedding arithmetic of multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, p. 4950–4958.

DENG J., DONG W., SOCHER R., LI L.-J., LI K. & FEI-FEI L. (2009). ImageNet : A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, p. 248–255. ISSN : 1063-6919, DOI : 10.1109/CVPR.2009.5206848.

DENG J., GUO J., XUE N. & ZAFEIRIOU S. (2019). Arcface : Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J. & HOULSBY N. (2021). An Image is Worth 16x16 Words : Transformers for Image Recognition at Scale. In *Proceedings of ICLR 2021*.

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

GAN Z., LI L., LI C., WANG L., LIU Z. & GAO J. (2022). Vision-language pre-training : Basics, recent advances, and future trends. *Found. Trends. Comput. Graph. Vis.*, **14**(3–4), 163–352. DOI : 10.1561/0600000105.

GAO P., GENG S., ZHANG R., MA T., FANG R., ZHANG Y., LI H. & QIAO Y. (2021). CLIP-Adapter : Better Vision-Language Models with Feature Adapters. arXiv :2110.04544 [cs].

GARCIA-OLANO D., ONOE Y. & GHOSH J. (2022). Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection. In *Companion Proceedings of the Web Conference 2022*, WWW '22, p. 705–715, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3487553.3524648.

GUI L., WANG B., HUANG Q., HAUPTMANN A., BISK Y. & GAO J. (2022). KAT : A Knowledge Augmented Transformer for Vision-and-Language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 956–968, Seattle, United States : Association for Computational Linguistics.

HE K., ZHANG X., REN S. & SUN J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 770–778.

HEO Y.-J., KIM E.-S., CHOI W. S. & ZHANG B.-T. (2022). Hypergraph Transformer : Weakly-supervised multi-hop reasoning for knowledge-based visual question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 373–390, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.29.

HU Z., ISCEN A., SUN C., WANG Z., CHANG K.-W., SUN Y., SCHMID C., ROSS D. A. & FATHI A. (2022). REVEAL : Retrieval-Augmented Visual-Language Pre-Training with Multi-Source Multimodal Knowledge Memory. arXiv :2212.05221 [cs], DOI : 10.48550/arXiv.2212.05221.

JOHNSON J., DOUZE M. & JÉGOU H. (2019). Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, **7**(3), 535–547. DOI : 10.1109/TBDATA.2019.2921572.

JOSHI M., CHOI E., WELD D. & ZETTLEMOYER L. (2017). TriviaQA : A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1601–1611, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1147.

KARPUKHIN V., OGUZ B., MIN S., LEWIS P., WU L., EDUNOV S., CHEN D. & YIH W.-T. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6769–6781, Online : Association for Computational Linguistics.

KHAN S., NASEER M., HAYAT M., ZAMIR S. W., KHAN F. S. & SHAH M. (2022). Transformers in vision : A survey. *ACM Comput. Surv.*, **54**(10s). DOI : 10.1145/3505244.

LERNER P., FERRET O. & GUINAUDEAU C. (2023). Multimodal inverse cloze task for knowledge-based visual question answering. In J. KAMPS, L. GOEURIOT, F. CRESTANI, M. MAISTRO, H. JOHO, B. DAVIS, C. GURRIN, U. KRUSCHWITZ & A. CAPUTO, Éds., *Advances in Information Retrieval*, p. 569–587, Cham : Springer Nature Switzerland. DOI : 10.1007/978-3-031-28244-7_36.

LERNER P., FERRET O., GUINAUDEAU C., LE BORGNE H., BESANÇON R., MORENO J. G. & LOVÓN MELGAREJO J. (2022). ViQuAE, a dataset for knowledge-based visual question answering about named entities. In *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. DOI : 10.1145/3477495.3531753.

LEWIS P., PEREZ E., PIKTUS A., PETRONI F., KARPUKHIN V., GOYAL N., KÜTTLER H., LEWIS M., YIH W.-T., ROCKTÄSCHEL T., RIEDEL S. & KIELA D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 9459–9474 : Curran Associates, Inc.

LHOEST Q., VILLANOVA DEL MORAL A., JERNITE Y., THAKUR A., VON PLATEN P., PATIL S., CHAUMOND J., DRAME M., PLU J., TUNSTALL L., DAVISON J., ŠAŠKO M., CHHABLANI G., MALIK B., BRANDEIS S., LE SCAO T., SANH V., XU C., PATRY N., MCMILLAN-MAJOR A., SCHMID P., GUGGER S., DELANGUE C., MATUSSIÈRE T., DEBUT L., BEKMAN S., CISTAC

P., GOEHRINGER T., MUSTAR V., LAGUNAS F., RUSH A. & WOLF T. (2021). Datasets : A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 175–184, Online and Punta Cana, Dominican Republic : Association for Computational Linguistics.

LIU Z., XIONG C., LV Y., LIU Z. & YU G. (2023). Universal vision-language dense retrieval : Learning a unified representation space for multi-modal retrieval. In *The Eleventh International Conference on Learning Representations*.

LOSHCHILOV I. & HUTTER F. (2019). Decoupled Weight Decay Regularization. *arXiv :1711.05101 [cs, math]*. arXiv : 1711.05101.

MA X., SUN K., PRADEEP R. & LIN J. (2021). A Replication Study of Dense Passage Retriever. *arXiv :2104.05740 [cs]*.

MARINO K., RASTEGARI M., FARHADI A. & MOTTAGHI R. (2019). OK-VQA : A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 3195–3204.

MOKADY R., HERTZ A. & BERMANO A. H. (2021). Clipcap : Clip prefix for image captioning. DOI : 10.48550/ARXIV.2111.09734.

PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., DESMAISON A., KOPF A., YANG E., DEVITO Z., RAISON M., TEJANI A., CHILAMKURTHY S., STEINER B., FANG L., BAI J. & CHINTALA S. (2019). PyTorch : An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, **32**.

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

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, **21**, 1–67.

RAMESH A., PAVLOV M., GOH G., GRAY S., VOSS C., RADFORD A., CHEN M. & SUTSKEVER I. (2021). Zero-shot text-to-image generation. In M. MEILA & T. ZHANG, Éds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 de *Proceedings of Machine Learning Research*, p. 8821–8831 : PMLR.

RASHTCHIAN C., YOUNG P., HODOSH M. & HOCKENMAIER J. (2010). Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, p. 139–147, USA : Association for Computational Linguistics.

SHAH S., MISHRA A., YADATI N. & TALUKDAR P. P. (2019). KVQA : Knowledge-Aware Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, p. 8876–8884.

SMUCKER M. D., ALLAN J. & CARTERETTE B. (2007). A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, CIKM '07, p. 623–632, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/1321440.1321528.

SUN W., FAN Y., GUO J., ZHANG R. & CHENG X. (2022). Visual Named Entity Linking : A New Dataset and A Baseline. arXiv :2211.04872 [cs].

WANG J., GONG T., ZENG Z., SUN C. & YAN Y. (2022). C3CMR : Cross-Modality Cross-Instance Contrastive Learning for Cross-Media Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, p. 4300–4308, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3503161.3548263.

WANG Z., NG P., MA X., NALLAPATI R. & XIANG B. (2019). Multi-passage BERT : A Globally Normalized BERT Model for Open-domain Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 5878–5882, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1599.

WESTON J., CHOPRA S. & BORDES A. (2014). Memory networks. DOI : 10.48550/ARXIV.1410.3916.

WOLF T., DEBUT L., SANH V., CHAUMOND J., DELANGUE C., MOI A., CISTAC P., RAULT T., LOUF R., FUNTOWICZ M., DAVISON J., SHLEIFER S., VON PLATEN P., MA C., JERNITE Y., PLU J., XU C., SCAO T. L., GUGGER S., DRAME M., LHOEST Q. & RUSH A. M. (2020). HuggingFace's Transformers : State-of-the-art Natural Language Processing. *arXiv :1910.03771 [cs]*.

WOLFE R. & CALISKAN A. (2022). Contrastive Visual Semantic Pretraining Magnifies the Semantics of Natural Language Representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3050–3061.

ZAMANI H., DIAZ F., DEHGHANI M., METZLER D. & BENDERSKY M. (2022). Retrieval-Enhanced Machine Learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, p. 2875–2886, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3477495.3531722.