Un jeu de données pour répondre à des questions visuelles à propos d'entités nommées

Paul Lerner* — **Salem Messoud*** — **Olivier Ferret**** — **Camille Guinaudeau*** — **Hervé Le Borgne**** — **Romaric Besançon**** — **Jose G. Moreno***** — **Jesús Lovón Melgarejo*****

* *Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France*
** *Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France*
*** *IRIT, UMR 5505 CNRS, Université Paul Sabatier, Toulouse, France*
paul.lerner@lisn.upsaclay.fr, olivier.ferret@cea.fr

RÉSUMÉ. *Dans le contexte des approches multimodales, nous nous intéressons à la tâche de réponse à des questions visuelles à propos d'entités nommées en utilisant des bases de connaissances (KVQAE). Nous mettons à disposition ViQuAE, un nouveau jeu de données de 3 700 questions associées à des images, annoté à l'aide d'une méthode semi-automatique. C'est le premier jeu de données de KVQAE comprenant des types d'entités variés associé à une base de connaissances composée de 1,5 million d'articles Wikipédia, incluant textes et images. Nous proposons également un modèle de référence de KVQAE en trois étapes : recherche d'information initiale, réordonnancement, puis extraction des réponses. Les résultats de nos expériences démontrent empiriquement la difficulté de la tâche et ouvrent la voie à une meilleure représentation multimodale des entités nommées.*

MOTS-CLÉS : *jeu de données, question-réponse visuelle, bases de connaissances, multimodalité.*

ABSTRACT. *In the context of multimodal processing, we focus our work on Knowledge-based Visual Question Answering about named Entities (KVQAE). We provide ViQuAE, a novel dataset of 3,700 questions paired with images, annotated using a semi-automatic method. It is the first KVQAE dataset to cover a wide range of entity types, associated with a knowledge base composed of 1.5M Wikipedia articles paired with images. To set a baseline on the benchmark, we address KVQAE as a three-stage problem: initial Information Retrieval, Re-Ranking, and Reading Comprehension. The experiments empirically demonstrate the difficulty of the task and pave the way towards better multimodal entity representations.*

KEYWORDS: *Dataset, Knowledge-based Visual Question Answering, Multimodality.*

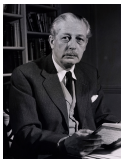
Requête (entrée)		Article pertinent dans la base de connaissances	
	*« Which constituency did this man represent when he was Prime Minister ? »*		« Macmillan indeed lost Stockton in the landslide Labour victory of 1945, but returned to Parliament in the November 1945 by-election in **Bromley**. »
	*« 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 1. *Exemple de questions du jeu de données ViQuAE avec leur image contextuelle et la source de la réponse (issue de la base de connaissances)*

1. Introduction

La fusion de modalités telles que l'image et le texte pour rechercher des informations est un problème reconnu comme difficile du fait de la différence de niveau de leurs sémantiques respectives (Srihari *et al.*, 2000). Ce constat est particulièrement vrai pour répondre à des questions visuelles à propos d'entités nommées (KVQAE[1]), où différents types de relations peuvent lier une question et l'image qui lui est associée en tant que contexte (cf. figure 1).

Dans la tâche classique de réponse à des questions visuelles (VQA), le contenu de l'image associée, par exemple la couleur d'un objet ou le nombre d'objets, est le sujet de la question (Antol *et al.*, 2015). La VQA fondée sur les connaissances (Wang *et al.*, 2017 ; Wang *et al.*, 2018 ; Marino *et al.*, 2019) utilise quant à elle l'image comme contexte pour poser des questions et trouver des réponses dans des bases de connaissances (BC), structurées ou non. Cependant, ces deux champs de recherche se focalisent principalement sur des catégories d'objets génériques en s'appuyant sur un prétraitement de détection d'objets (Anderson *et al.*, 2018 ; Gardères *et al.*, 2020). Dans cette optique, la seconde question de la figure 1 pourrait typiquement porter sur le type de bateau en prenant la forme : « Est-ce un bateau de pêche ? » Au contraire, notre travail se concentre sur des questions nécessitant des connaissances à propos des entités nommées, comme le *Queen Elizabeth 2* dans le cas présent. Nous avons conçu et publions le jeu de données ViQuAE dans ce but[2]. Notre jeu de données a été conçu comme un cadre d'évaluation pour diagnostiquer et suivre les progrès des systèmes de KVQAE. Nous pensons en effet que la KVQAE est une tâche bien définie et facilement évaluable. Elle est ainsi bien appropriée pour rendre compte des progrès de

1. Pour le sigle anglais de *Knowledge-based Visual Question Answering about named Entities*.
2. Disponible *via* `https://github.com/PaulLerner/ViQuAE`.

(a) Personne (Louis-Philippe l^{er})

(b) Personne (Louis-Philippe l^{er}) ou tableau (de Franz Xaver Winterhalter)

(c) Attraction touristique artificielle (tour Eiffel)

(d) Attraction touristique naturelle (mont Saint Helens)

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

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

FIGURE 2. *Quelques exemples d'images de différents types d'entités et différentes images du même type d'entité considérées dans notre travail*

la qualité des représentations multimodales d'entités nommées sur un plan plus général. La représentation multimodale des entités ouvre aussi la voie vers différentes applications, permettant par exemple de rendre les interactions homme-machine plus naturelles : 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 ? ». Les questions sur les entités nommées sont très difficiles car les BC actuelles en contiennent des millions. De ce point de vue, utiliser chaque modalité indépendamment n'est pas suffisamment discriminant pour répondre au besoin de l'utilisateur. À titre d'exemple, dans les images de la figure 1, il est assez complexe de reconnaître *Harold Macmillan* au sein d'une BC contenant des millions de *personnes*. Cependant, on peut déduire de la question qu'il était *Premier ministre* et réduire ainsi les candidats à quelques centaines.

Shah *et al.* (2019) ont déjà travaillé sur la KVQAE mais se sont limités aux entités nommées de type personne. Au contraire, ViQuAE comprend divers types d'entités. Cette diversité est une question centrale dans la KVQAE, notamment en raison de l'hétérogénéité des représentations visuelles qui en résulte. Entre autres entités, les entreprises peuvent être ainsi représentées par un bâtiment (par exemple leur siège), un produit manufacturé qu'elles vendent ou simplement leur logo (cf. figure 2). La KVQAE nécessite donc une représentation multimodale des connaissances, ce qui la distingue clairement de la recherche d'image par le contenu. Cette diversité implique également

la nécessité d'étudier d'autres types d'entités que les personnes, qui peuvent assez bien être reconnues visuellement à partir de leur seul visage. Par ailleurs, Shah *et al.* (2019) utilisent une BC structurée et donc des méthodes assez différentes des nôtres, qui avons opté pour une BC multimodale constituée de textes non structurés et d'images (cf. section 3.4).

Sur un autre plan, ViQuAE, avec ses 3 700 questions, s'inscrit dans le courant des travaux sur l'apprentissage sans (*zero-shot)* ou avec peu d'exemples (*few-shot*), avec une double idée : d'une part, la diversité des tâches unissant texte et image ne permet pas de développer des jeux de données d'une taille suffisante pour entraîner de gros modèles à partir de zéro ; d'autre part, les percées des travaux reposant sur les *Foundation Models* (Bommasani et *et al.*, 2021) permettent de s'affranchir d'un tel entraînement. Nous espérons ainsi que ViQuAE encouragera les études vers des modèles transférables ou vers des techniques d'apprentissage sans ou avec peu d'exemples, nécessaires pour la KVQAE.

Cet article est une version étendue de Lerner *et al.* (2022b)[3]. En plus de présenter le jeu de données ViQuAE (section 3), cette version plus détaillée contient une revue des travaux connexes actualisée (section 2), des analyses supplémentaires (sections 5 et 7) ainsi qu'une nouvelle contribution concernant le réordonnancement des résultats de la recherche d'information initiale (section 6 et par conséquent, section 7).

2. Travaux connexes

La KVQAE est intrinsèquement une tâche complexe mêlant à la fois des problématiques de recherche d'information (RI) et d'extraction d'information en faisant intervenir plusieurs médias tout à la fois au niveau de la requête et des documents cibles. Elle se retrouve donc à l'interface de plusieurs domaines. Le fait d'utiliser le texte comme source de réponse dans notre cas la rapproche d'abord des systèmes de question-réponse (QA) textuels se situant dans la lignée de Voorhees et Tice (2000) et traitant la tâche en deux étapes, avec une phase initiale de RI suivie d'une extraction de la réponse (*reading comprehension*). Au cours des dernières années, une attention particulière a été accordée à l'extraction de la réponse, avec des jeux de données de plus en plus grands (Rajpurkar *et al.*, 2016 ; Joshi *et al.*, 2017 ; Kwiatkowski *et al.*, 2019). Nous profitons de ces derniers pour bâtir notre propre jeu de données, comme expliqué à la section suivante, avec le même tropisme pour les questions factuelles que la plupart des travaux dans le domaine (Chen *et al.*, 2017).

De son côté, bien qu'initialement axée sur le texte, la RI s'est rapidement étendue aux documents multimodaux. Srihari *et al.* (2000) et Clough *et al.* (2004), par exemple, partageaient déjà un certain nombre de problèmes avec la KVQAE, comme la fusion d'informations multimodales. Cependant, les modalités en RI multimodale sont souvent redondantes alors qu'elles sont complémentaires avec la KVQAE.

3. Également résumé et traduit en français dans Lerner *et al.* (2022a).

L'usage de plusieurs modalités en QA prend quant à elle souvent une forme cross-modale (Kembhavi *et al.*, 2017 ; Sampat *et al.*, 2020 ; Talmor *et al.*, 2021 ; Chang *et al.*, 2022 ; Reddy *et al.*, 2021) assimilable à de l'extraction de réponse par le biais de plusieurs modalités (texte, tableaux ou images). La source de la réponse, quelle que soit la modalité, est fournie en même temps que la question contextuelle et les deux sont interdépendantes. Ainsi, Reddy *et al.* (2021) construisent leur corpus à partir d'articles de presse, où le système a accès aux métadonnées des images, telles que leur légende. Par conséquent, la tâche relève davantage du raisonnement logique que de la RI, contrairement aux questions de KVQAE, qui sont autosuffisantes.

Pour sa part, la VQA fondée sur la connaissance (Wang *et al.*, 2017 ; Wang *et al.*, 2018 ; Marino *et al.*, 2019 ; Jain *et al.*, 2021 ; Schwenk *et al.*, 2022) se concentre sur des questions de sens commun concernant des catégories d'objets génériques. En outre, les jeux de données VQA (fondés sur la connaissance ou pas) sont généralement construits à partir des images du jeu de données *Common Objects in Context* (COCO, Lin *et al.* (2014)). Pour ces deux raisons, la VQA fondée sur la connaissance a été largement traitée en s'appuyant sur des détecteurs d'objets entraînés sur COCO, ce qui facilite la RI (Gardères *et al.*, 2020).

Le premier jeu de données KVQAE a quant à lui été présenté par Shah *et al.* (2019) : il s'agit de KVQA, fondé sur Wikidata et focalisé sur les entités de type personne. Malgré sa grande taille, ce jeu de données présente plusieurs limites : (i) il est restreint aux entités de type personne, avec une RI se réduisant à la reconnaissance faciale ; (ii) les questions sont générées automatiquement à partir de patrons et de Wikidata. De ce fait, elles sont assez répétitives et limitées par le schéma de Wikidata : la plupart des questions portent sur l'identité de la personne, son lieu de naissance, sa date de naissance ou son emploi. Au contraire, nous visons à construire un jeu de données couvrant divers types d'entités avec une expression riche et des questions couvrant de nombreux sujets.

3. Jeu de données et base de connaissances ViQuAE

3.1. *Annotation automatique*

Pour limiter les efforts d'annotation manuelle, nous nous sommes appuyés sur des jeux de données de question-réponse existants, qui comprennent des questions couvrant différents sujets et entités. Nous avons ainsi décidé d'utiliser le jeu de données textuel TriviaQA en raison de sa taille et de la typologie de ses questions (Joshi *et al.*, 2017). L'idée principale de notre processus 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 (par exemple « cet homme »). De cette façon, il n'est pas possible de répondre à la question sans s'appuyer sur l'image contextuelle. Dans le premier exemple de la figure 1, la mention de l'entité nommée « *Harold Macmillan* » de la question originale est ainsi remplacée par la mention ambiguë « *this man* ».

Notre processus débute par une analyse syntaxique et une identification des entités nommées dans les questions à l'aide de `spaCy`[4], ce qui permet d'obtenir environ 0,9 mention valide par question. L'analyse des dépendances permet de ne conserver que certaines mentions d'entités, par exemple le sujet de la question. À partir de ces mentions d'entité, puisque la réponse à la question est connue, la désambiguïsation peut être effectuée en vérifiant si la réponse est présente dans l'article Wikipédia de l'entité candidate. Cette étape a en fait été réalisée par Joshi *et al.* (2017) avec TAGME (Ferragina et Scaiella, 2010) lorsqu'ils ont initialement conçu TriviaQA pour l'extraction de réponse ; nous avons simplement fait correspondre nos mentions d'entités avec leurs entités désambiguïsées. Environ 55 % des mentions d'entités (donc de questions potentielles) ont été désambiguïsées, laissant 45 % de côté. Wikidata permet de recueillir des informations sur les entités désambiguïsées : leur type, leur profession, leur genre et leur catégorie Commons. Nous avons utilisé cette dernière pour trouver une image pertinente tandis que les autres sont nécessaires pour générer une mention ambiguë. Les personnes sont référencées par leur profession (par exemple « cet écrivain ») et les autres entités par leur type (par exemple « cette attraction touristique »). De plus, si le genre était disponible, nous avons également utilisé « *this man/woman* » et « *he-him-his/she-her-hers* » selon la dépendance syntaxique de la mention originale. 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. De plus, pour se conformer à la RGPD[5], et étant donné que beaucoup de questions portent sur des personnes, nous avons conservé seulement les questions portant sur des personnes décédées. Cette étape écarte 31 % de questions supplémentaires. Les images sont récupérées à partir de la catégorie Commons de l'entité. 3 % des questions n'avaient pas d'images disponibles et ont donc été écartées. Grâce aux contributeurs de Wikimedia Commons, toutes les images du jeu de données sont soit sous licence libre[6], 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.

3.2. *Annotation manuelle*

L'annotation automatique décrite ci-dessus présente quelques inconvénients. 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[7]. Pour remédier à ce problème, une interface d'annotation

4. `https://spacy.io/`

5. `https://eur-lex.europa.eu/eli/reg/2016/679/oj`

6. `https://freedomdefined.org/Definition`

7. Par exemple, « *Which constituency did this man represent when he was Prime Minister,* ***succeeding Sir Edward Campbell****?* » contient trop d'informations. Il faudrait la reformuler pour retrouver l'exemple de la figure 1.

a été conçue à l'aide de Label Studio[8]. L'annotateur peut reformuler librement la question (des mentions complémentaires sont suggérées) tant que la réponse n'est pas modifiée. Il doit également choisir parmi huit images candidates, si celle sélectionnée n'est pas appropriée, en utilisant comme référence l'image de référence de la BC (cf. section 3.4) tout en s'assurant qu'il ne s'agit pas d'un quasi-doublon. En dernier recours, l'annotateur peut simplement rejeter la question. L'interface et les instructions d'annotation sont partagées avec le reste de notre code.

Cette annotation manuelle a été réalisée par sept annotateurs internes (les auteurs de l'article, à l'exception de Salem Messoud). L'interface a permis 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, dont 2 000 ont été écartées, principalement parce qu'elles étaient surspécifiées ou que l'image n'était pas pertinente. Finalement, le jeu de données ViQuAE est constitué de 3 700 questions, réparties aléatoirement en ensembles de taille égale pour l'entraînement, la validation et 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 cinq mots entre la version initiale et la version éditée.

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. En effet, déterminer si une question est surspécifiée ou non peut être assez subjectif[9]. De plus, la reformulation de certaines questions surspécifiées peut être subtile. 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*.

3.3. *Analyse des données*

Le jeu de données ViQuAE se compose de 3 700 questions contextualisées par 3 300 images uniques, dont deux exemples sont présentés en figure 1. Les questions comportent en moyenne 12 mots, pour un vocabulaire de 4 700 mots. Sur les 3 700 réponses, les plus fréquentes, « France » et « Turquie », n'apparaissent que 13 fois, soit 0,3 % du total, ce qui montre la quasi-absence de biais *a priori* sur les réponses pour un classifieur indépendant de la question. 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.

8. `https://labelstud.io/`

9. Par exemple « This inner planet and which other planet in our solar system has no moon ? »

	ViQuAE	KVQA
# Questions	3 700	183 000
# Questions par image	1,1	7,4
Vocabulaire	4 700	600*
Taille moyenne des questions	12,4	10,1
Réponse la plus probable *a priori*	0,3 %	15,9 %
Chevauchement entre les réponses	25,3 %	89,4 %
Chevauchement entre les entités	18,1 %	40,6 %
# Questions par entité	1,5	9,7
# Types d'entités	980	1

TABLEAU 1. *Statistiques du jeu de données par rapport à KVQA (Shah* et al.*, 2019). **En moyenne, sur 49 sous-ensembles aléatoires de la même taille que ViQuAE, le vocabulaire du jeu de données KVQA entier se compose de 8 400 tokens.*

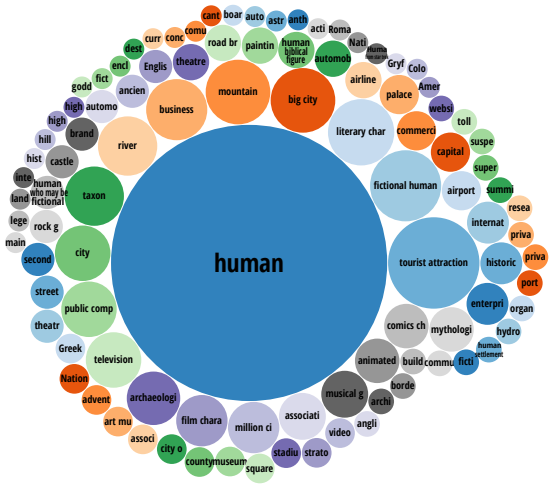


FIGURE 3. *Les 100 types d'entités (non-exclusifs) les plus fréquents dans le jeu de données*

Une contribution importante du jeu de données est sa diversité de types d'entités, qui est l'un des principaux défis pour les représentations multimodales (cf. section 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 présente un graphique à bulles des 100 types d'entités les plus

fréquents. Un résumé des statistiques comparées avec le jeu de données KVQA de Shah *et al.* (2019) est rapporté dans le tableau 1. Nous pouvons constater que, malgré sa petite taille, ViQuAE est plus diversifié sous certains aspects.

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 de l'article. Bien sûr, la question peut toujours nécessiter un raisonnement sur plusieurs phrases ou paragraphes de l'article. En revanche, Shah *et al.* (2019) comprennent 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) comprennent 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).

3.4. *La base de connaissances ViQuAE*

La BC ViQuAE 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. 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 de 1,5 million d'articles, dont 542 000 à propos de personnes, chacun associé à une image. La BC obtenue est donc cent fois plus grande que celle des expériences de Shah *et al.* (2019). 95 % des images de la base de connaissances sont uniques.

4. Approche de base pour la KVQAE et expérimentations

Nous traitons le problème de la KVQAE en trois étapes successives : recherche d'information initiale (cf. section 5), réordonnancement (cf. section 6) et extraction des réponses (*reading comprehension* ; cf. section 7), avec des métriques d'évaluation dédiées pour chacune. Nous reprenons ainsi une décomposition classiquement adoptée en QA textuelle (Chen *et al.*, 2017 ; Karpukhin *et al.*, 2020).

L'évaluation de cette approche de base a été réalisée sur notre jeu de données ViQuAE. Plus précisément, l'évaluation finale de la tâche est toujours effectuée sur l'ensemble de test de 1 257 questions tandis que les hyperparamètres sont optimisés

sur l'ensemble de validation de 1 250 questions et uniquement pour les modèles *few-shot*, l'apprentissage est effectué sur l'ensemble d'entraînement de 1 190 questions. Conformément à Joshi *et al.* (2017), les alias Wikipédia d'une réponse donnée sont considérés comme des réponses valides.

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[10]. De plus, il comprend 29 % de questions booléennes (réponse oui/non) pour lesquelles on ne peut pas évaluer la pertinence du passage de texte/de l'article Wikipédia automatiquement.

Bien que certains détails soient omis dans cette section et les suivantes en raison des contraintes d'espace, toutes les expériences peuvent être reproduites en utilisant notre code[11].

5. Recherche d'information initiale

La recherche d'information initiale a pour but de filtrer la base de connaissances afin d'obtenir des passages de texte candidats pertinents par rapport à la requête (question et image). Contrairement au réordonnancement (cf. section suivante), la RI initiale est contrainte du point de vue calculatoire par la grande taille de la BC.

Nous adoptons une approche de fusion tardive au niveau des modalités : la recherche est effectuée indépendamment avec la question et l'image puis les résultats sont fusionnés au niveau des scores. Notre implémentation s'appuie sur Elasticsearch[12] et Faiss (Johnson *et al.*, 2019), respectivement pour la recherche parcimonieuse et dense, toutes deux *via* la bibliothèque Datasets de Hugging Face (Lhoest et *et al.*, 2021).

5.1. *Recherche de texte initiale*

En amont de la recherche, nous filtrons les données semi-structurées des articles, comme les tableaux et les listes (Karpukhin *et al.*, 2020 ; Wang *et al.*, 2019). Chaque article est ensuite divisé 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. Comme modèle *zero-shot*[13], nous utilisons BM25 (Robertson *et al.*, 1995) et optimisons ses hyperparamètres sur l'ensemble de validation en utilisant une recherche par dichotomie. Pour

10. On peut estimer grossièrement 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.
11. `https://github.com/PaulLerner/ViQuAE`
12. `https://www.elastic.co/`
13. La notion de *zero-shot* renvoie ici au fait qu'il n'y a pas d'optimisation des paramètres d'un modèle mais seulement de ses hyperparamètres.

définir également une référence *few-shot*, nous utilisons DPR (Karpukhin *et al.*, 2020). DPR est un modèle de recherche dense fondé sur deux modèles BERT (Devlin *et al.*, 2019) : un pour la question et un pour le passage. DPR est entraîné à minimiser l'entropie croisée des similarités entre les questions et les passages (avec un seul passage pertinent par question). La sélection des passages négatifs utilisés lors de l'entraînement est faite à l'aide de BM25 afin de garantir sa difficulté et sa qualité. DPR est préentraîné sur TriviaQA, filtré de toutes les questions utilisées dans ViQuAE, avant d'être ajusté sur ViQuAE. Nous considérons également le modèle sans ajustement, entraîné uniquement sur TriviaQA, comme une autre référence *zero-shot*. La validation est effectuée sur les questions TriviaQA utilisées pour générer l'ensemble de validation ViQuAE. Pour l'entraînement, nous utilisons les mêmes hyperparamètres que Karpukhin *et al.* (2020).

5.2. *Recherche d'image initiale*

Pour la recherche d'images, nous utilisons deux représentations différentes de manière alternative : ArcFace (Deng *et al.*, 2019) pour les visages, si au moins un visage est détecté ; ImageNet-ResNet (He *et al.*, 2016) et CLIP (Radford *et al.*, 2021) pour l'image complète. Par conséquent, la BC est divisée en deux parties : les personnes avec un visage détecté et les non-personnes, en faisant l'hypothèse que les visages ne sont pertinents que pour les personnes. Comme Deng *et al.* (2019), nous utilisons MTCNN (Zhang *et al.*, 2016) pour la détection des visages. Les cinq points de repère du visage (les yeux, le nez et les coins de la bouche) sont adoptés pour effectuer une transformation de similarité afin qu'ils soient toujours à la même position dans l'image, quelle que soit la pose originale de la personne. Si plusieurs visages sont détectés, seul celui associé à la plus forte probabilité est conservé. 6,6 % des personnes de la BC n'ont pas de visage détecté et ont donc été écartées.

ArcFace est une méthode d'apprentissage de représentation pour la reconnaissance et la vérification des visages très efficace. Il est préentraîné sur MS-Celeb (Guo *et al.*, 2016), composé de photos de célébrités. Ses entités ont un certain chevauchement avec ViQuAE, qui est analysé dans la section suivante. Cette approche est assez comparable à celle utilisée par Shah *et al.* (2019), bien qu'ils aient opté pour FaceNet (Schroff *et al.*, 2015) et ne donnent pas de détails sur le jeu de données d'entraînement[14].

Le modèle ResNet, dont les connexions résiduelles permettent de construire des réseaux très profonds, est très utilisé pour l'apprentissage de représentations visuelles, par exemple dans ArcFace. Nous désignons par « ImageNet-ResNet » le modèle entraîné sur mille catégories d'objets d'ImageNet (Deng *et al.*, 2009), le jeu de données de préentraînement le plus populaire pour la classification d'images. Les caractéristiques extraites de la dernière couche convolutive d'ImageNet-ResNet se sont en effet avérées être efficaces pour la recherche d'images (Sharif Razavian *et al.*, 2014 ; Ra-

14. C'est ce qui a motivé notre choix pour ArcFace car FaceNet a originellement été proposé dans une version entraînée avec un jeu de données propriétaire de Google.

denović *et al.*, 2018). Nous utilisons le *max-pooling* pour en réduire la carte de caractéristiques (*feature map*), compte tenu des résultats rapportés dans Radenović *et al.* (2018).

CLIP (Radford *et al.*, 2021) est une architecture permettant d'apprendre des représentations visuelles à partir d'une faible supervision textuelle. L'objectif d'apprentissage est similaire à celui de DPR, bien que CLIP associe des images à des légendes pertinentes au lieu de requêtes à des documents pertinents. CLIP a été entraîné sur un jeu de données de 400 millions de paires image-légende. Nous ne nous intéressons qu'à l'encodeur visuel de CLIP et laissons de côté son encodeur textuel.

Tous ces modèles sont gelés dans nos expériences, c'est-à-dire qu'ils ne sont pas ajustés. Dans un souci de comparaison équitable, nous utilisons systématiquement une architecture ResNet-50 pour toutes les représentations visuelles. La recherche dense est effectuée au moyen du produit scalaire, équivalent à la similarité cosinus car les représentations sont normalisées au préalable (sauf pour DPR).

5.3. *Fusion multimodale*

Les résultats de la recherche par l'image sont ensuite mis en correspondance avec les passages pour la fusion avec la recherche textuelle. Les scores des résultats de ces modèles ayant des distributions très différentes, ils sont centrés-réduits avant de les fusionner. La fusion est faite *via* une combinaison linéaire (Karpukhin *et al.*, 2020 ; Ma *et al.*, 2021) : $P = \alpha_b B + \alpha_d D + \mathbf{F}\alpha_a A + (1 - \mathbf{F})(\alpha_i I + \alpha_c C)$. On note $B,\ D,\ A,\ I,\ C$, les scores respectifs de BM25, DPR, ArcFace, ImageNet-ResNet et CLIP, chacun étant pondéré par l'hyperparamètre α_j. $\mathbf{F} \in \{0, 1\}$ dénote la détection d'un visage. Seuls les 100 premiers passages sont considérés. Par conséquent, si, compte tenu d'une requête, un passage n'est pas retrouvé par un système donné, il lui est attribué le score minimal des autres passages retrouvés par ce système (Ma *et al.*, 2021). Les passages sont ensuite réordonnés par rapport au score P. Les hyperparamètres d'interpolation α_j sont réglés sur l'ensemble de validation en utilisant une recherche par dichotomie pour maximiser le rang réciproque moyen. Pour limiter l'espace de recherche et permettre une comparaison directe entre BM25 et DPR, nous contraignons $\sum_j \alpha_j = 1$ et n'utilisons qu'un seul modèle pour la recherche texte : nous avons donc $\alpha_b = 0$ ou $\alpha_d = 0$.

5.4. *Résultats*

Puisqu'il est fondé sur TriviaQA (Joshi *et al.*, 2017), ViQuAE n'est supervisé que de façon distante, c'est-à-dire qu'un document est jugé pertinent s'il contient la réponse. Nous évaluons la RI avec la précision à K (P@K) et le rang réciproque moyen (MRR) ainsi que Hits@K. Hits@K représente la proportion de questions pour lesquelles la RI récupère *au moins un* document pertinent parmi les K premiers. Les résultats sont présentés dans les tableaux 2 et 3. Les tests de significativité statistique

#	Modèle	MRR	P@1	P@20	Hits@20
-	$\mathbf{F}A$ (ArcFace, visage détecté)	54,3	50,2	5,5	65,3
a	$(1-\mathbf{F})I$ (ImageNet, *pas* de visage détecté)	17,5	11,9	4,9	36,1
b	$(1-\mathbf{F})C$ (CLIP, *pas* de visage détecté)	**27,5**^{a}	**20,5**^{a}	**9,5**^{a}	**53,1**^{a}
a	B (BM25, texte seulement)*	23,2	16,5	7,1	45,3
b	D_0 (DPR *zero-shot*, texte seulement)*	35,5^{a}	24,9^{a}	**17,6**^{ac}	66,5^{ac}
c	$\mathbf{F}0,3A+(1-\mathbf{F})(0,1I+0,3C)$	**41,4**^{ab}	**35,6**^{abd}	7,4	59,5^{a}
d	D_f (DPR *few-shot*, texte seulement)*	38,2^{ab}	27,8^{ab}	17,5^{ac}	**66,7**^{ac}

TABLEAU 2. *Résultats de la RI initiale évaluée au niveau de l'article : sur deux sous-ensembles (visage détecté ou pas) et sur le test complet. *Chaque article se voit assigner le score maximal de ses passages. Les exposants dénotent des différences significatives selon le test de randomisation de Fisher avec $p \leq 0{,}01$. Hits@1 est omis car il est équivalent à P@1.*

sont effectués à l'aide du test de randomisation de Fisher (Fisher, 1937 ; Smucker *et al.*, 2007). Nous présentons également comme référence les performances de BM25 et de DPR utilisant seulement le texte.

Pour étudier l'apport de chaque modalité séparément, nous évaluons les résultats au niveau de l'article. Cette comparaison suppose deux subtilités :

– les différentes représentations visuelles (section 5.2) sont d'abord évaluées séparément sur des sous-ensembles du jeu de données selon la détection des visages, puis en combinaison sur l'ensemble du jeu de test ;

– DPR fonctionne au niveau du passage car il est fondé sur BERT, qui ne peut pas traiter directement de longs articles. De plus, le passage sert comme unité par la suite pour l'extraction des réponses, elle aussi fondée sur BERT. Dans le tableau 2 nous avons donc simplement assigné à l'article le score maximal de ses passages pour avoir un point de comparaison, mais la performance au niveau du passage est étudiée dans le tableau 3.

Dans le tableau 2, nous constatons que la recherche *via* l'image obtient d'assez bons résultats, notamment avec ArcFace quand un visage est détecté. Comparativement, Shah *et al.* (2019) obtiennent une P@1 de 73,5 avec FaceNet mais leur BC est cent fois plus petite. Par ailleurs, CLIP surpasse largement ImageNet 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. Enfin, nous remarquons une dynamique très différente entre les modèles visuels et textuels, notamment entre ArcFace et DPR : ArcFace est très précis mais avec un rappel relativement mauvais tandis que c'est l'inverse pour DPR. Ceci s'explique par l'objet des questions, qui permet de deviner la réponse avec un nombre suffisamment grand d'essais sans regarder l'image. Pour l'exemple de la figure 1, DPR pourrait ainsi retourner toutes les circonscriptions (*constituency*) du Royaume-Uni, ordonnées aléatoirement.

#	Modèle	MRR	P@1	P@20	Hits@20
a	B (BM25, texte seulement)	19,0	13,1	5,9	39,5
b	D_0 (DPR *zero-shot*, texte seulement)	30,5^{a}	21,2^{a}	16,2^{ac}	60,5^{ac}
c	$0,3(B + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	27,9^{a}	20,4^{a}	10,1^{a}	50,5^{a}
d	$0,3(D_0 + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	36,0^{abce}	26,7^{abce}	17,1^{ac}	65,2^{abce}
e	D_f (DPR *few-shot*, texte seulement)	32,8^{abc}	22,8^{a}	16,4^{ac}	61,2^{ac}
f	$0,3(D_f + \mathbf{F}A) + 0,2(1 - \mathbf{F})(I + C)$	**37,9**^{abcde}	**27,8**^{abce}	**17,5**^{ac}	**65,7**^{abce}

TABLEAU 3. *Résultats de la RI évaluée au niveau du passage avec les* baseline *textuelles et la fusion de la recherche multimodale, dans les deux configurations d'apprentissage : sans ou avec peu d'exemples*

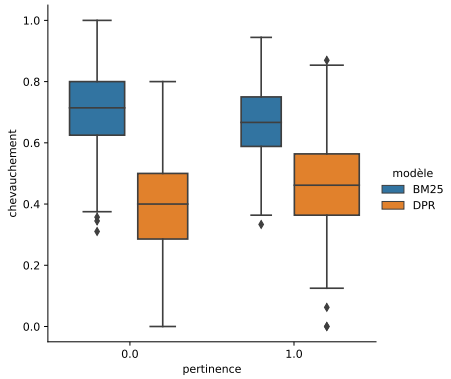


FIGURE 4. *Chevauchement entre les lemmes de la question et du premier passage retourné par BM25 et DPR en fonction de la pertinence du passage. Chaque boîte montre les quartiles tandis que ses moustaches s'étendent pour montrer le reste de la distribution, à l'exception des valeurs extrêmes.*

Le gain de performance de DPR par rapport à BM25 est important, y compris dans sa version *zero-shot* où il surpasse significativement BM25 et même la recherche multimodale fondée sur BM25, pour P@20 et Hits@20. Contrairement à BM25, DPR est capable de trouver des passages pertinents, même avec très peu de chevauchement lexical, grâce à ses représentations sémantiques distribuées, comme on peut le voir sur la figure 4. Toutefois, ses passages pertinents tendent tout de même à avoir un chevauchement supérieur avec la question par rapport à ses passages non pertinents. DPR pose par ailleurs le problème d'être plus sensible aux biais des jeux de données que BM25. Par exemple, pour une question telle que « Dans quel pays est-ce que cette personne est née ? », le modèle peut être biaisé par la distribution selon le jeu d'entraînement si une nationalité est surreprésentée.

Par ailleurs, il faut noter que la fusion multimodale apporte des gains de performance significatifs. Ce gain diffère selon le type de l'entité-sujet de la question. Pour

#	Modèle	MRR	P@1	P@20	Hits@20
a	B (BM25, texte seulement)	19,8	14,4	6,1	37,6
b	D_0 (DPR *zero-shot*, texte seulement)	28,0^{a}	19,2^{a}	14,4^{a}	57,9^{a}
c	$0,3(B + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	32,4^{a}	24,4^{a}	11,9^{a}	56,0^{a}
d	$0,3(D_0 + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	37,9^{abce}	28,9^{abe}	17,4^{abce}	67,4^{abce}
e	D_f (DPR *few-shot*, texte seulement)	31,1^{ab}	21,7^{a}	15,2^{ac}	57,5^{a}
f	$0,3(D_f + \mathbf{F}A) + 0,2(1 - \mathbf{F})(I + C)$	**40,4**^{abce}	**29,8**^{abce}	**18,4**^{abcde}	**67,8**^{abce}
a	B (BM25, texte seulement)	18,3	12,1	5,8	41,0
b	D_0 (DPR *zero-shot*, texte seulement)	32,7^{ac}	22,9^{ac}	**17,7**^{ac}	62,6^{ac}
c	$0,3(B + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	24,1^{a}	17,1^{a}	8,5^{a}	45,9^{a}
d	$0,3(D_0 + \mathbf{F}A) + (1 - \mathbf{F})(0,1I + 0,3C)$	34,3^{ac}	24,7^{ac}	16,9^{ac}	63,4^{ac}
e	D_f (DPR *few-shot*, texte seulement)	34,1^{ac}	23,8^{ac}	17,4^{ac}	**64,3**^{ac}
f	$0,3(D_f + \mathbf{F}A) + 0,2(1 - \mathbf{F})(I + C)$	**35,7**^{abc}	**26,0**^{ac}	16,8^{ac}	64,0^{ac}

TABLEAU 4. *Résultats de la RI évaluée au niveau du passage pour les questions à propos de personnes (partie supérieure) et de non-personnes (partie inférieure)*

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 d'images, 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 (cf. tableau 4). En outre, sur le sous-ensemble d'entités qui se chevauchent avec MS-Celeb (le jeu de données de préentraînement d'ArcFace), la valeur de P@1 monte jusqu'à 25,7, ce qui représente une amélioration de 5 % par rapport au score mesuré sur toutes les personnes. De manière similaire, la fusion multimodale apporte un gain significatif avec DPR, même en tenant compte du fait que sa *baseline* textuelle est meilleure.

Plus globalement, ces premiers résultats montrent des tendances intéressantes mais également une marge d'amélioration importante, laissant la place à de futurs travaux sur la fusion multimodale. En attendant, nous présentons une *baseline* pour le réordonnancement des résultats de cette étape à la section suivante.

6. Réordonnancement

La relative faiblesse de la précision de la RI initiale est une conséquence, en particulier pour les approches denses, d'une modélisation limitée par la taille de la BC à considérer. Par conséquent, il est intéressant de réordonnancer les passages issus de cette première phase, beaucoup plus restreints en termes de volume, afin d'améliorer la précision. Nous adoptons une approche de fusion tardive similaire à celle de la RI initiale. Le réordonnancement est effectué indépendamment avec le texte et l'image puis les résultats sont fusionnés.

#	Modèle	MRR	P@1	P@20	Hits@20
a	RRT (visage détecté)	37,4	25,2	13,2^{b}	73,7
b	$\mathbf{F}A$ (ArcFace, visage détecté)	49,6^{a}	42,9^{a}	12,2	76,3
c	$\mathbf{F}A + \mathrm{RRT}$ (visage détecté)	**52,4**^{ab}	**43,0**^{a}	**13,3**^{b}	**77,4**^{a}
a	RRT	39,2	27,5	14,1	73,7
b	$\mathbf{F}A + \mathrm{RRT}$	**47,0**^{a}	**36,7**^{a}	**14,2**	**75,6**^{a}

TABLEAU 5. *Évaluation au niveau de l'article du réordonnancement de la RI initiale, selon la détection d'un visage. Les exposants dénotent des différences significatives dans le test de randomisation de Fisher avec* $p \leq 0,01$.

6.1. *Réordonnancement pour l'image*

Pour le réordonnancement d'image, nous utilisons deux représentations d'image différentes : ArcFace pour les visages, si au moins un visage est détecté par MTCNN, comme pour la RI initiale ; Re-Ranking Transformers (RRT) (Tan *et al.*, 2021) pour l'image complète. Si aucun visage n'est détecté, nous utilisons uniquement RRT pour réordonnancer les images ; sinon, nous combinons ArcFace et RRT en utilisant encore une fois la technique du « minimum par défaut » de Ma *et al.* (2021).

RRT utilise l'architecture Transformer (Vaswani *et al.*, 2017) et son mécanisme d'auto-attention pour combiner les caractéristiques globales et locales obtenues à partir de DELG (Cao *et al.*, 2020), un extracteur de caractéristiques, les deux méthodes étant remarquablement efficaces. Le modèle RRT prend en entrée une séquence de représentations globales et locales obtenues à partir d'une paire d'images (associées à la question et au passage) et utilise le plongement du token spécial `[CLS]` pour apprendre une métrique de similarité. RRT et DELG sont entraînés sur Google Landmarks v2 (GLDv2) (Weyand *et al.*, 2020), composé de photos de monuments, et ne sont pas ajustés dans nos expériences. Les URLs de ces photos ont un chevauchement de 9 % avec les images des questions de ViQuAE, qui est analysé à la section 6.4.

Les scores d'ArcFace et RRT n'étant pas comparables, nous les normalisons d'abord en fonction du rang puis nous utilisons PosFuse (Lillis *et al.*, 2010) pour fusionner les scores normalisés. La normalisation par le rang est très utile pour minimiser l'effet des valeurs extrêmes. Elle consiste à attribuer à chaque passage le score $1 - \frac{r-1}{K}$, où r est le rang du passage et K, le nombre total de passages à ordonner (100 dans nos expériences). PosFuse est quant à elle une méthode supervisée qui apprend la probabilité qu'un passage apparaissant à une position donnée soit pertinent. Elle est optimisée sur le jeu de validation en utilisant une recherche par dichotomie.

6.2. *Réordonnancement pour le texte*

Comme Wang *et al.* (2019), notre réordonnanceur de texte prend en entrée la concaténation d'une paire question-passage et l'encode au moyen de BERT. De façon

#	Modèle	MRR	P@1	P@20	Hits@20
a	RI initiale	37,9	27,8	17,5	65,7
b	Texte	47,4^{a}	37,8^{a}	23,7^{a}	73,1^{a}
c	Texte + $\mathbf{F}A$ + RRT	**52,7**^{ab}	**43,7**^{ab}	**25,1**^{ab}	**75,3**^{ab}

TABLEAU 6. *Évaluation au niveau du passage du réordonnancement de la RI initiale, fondé seulement sur le texte ou en fusionnant texte et image*

comparable à RRT, la représentation associée au token [CLS] est introduite dans un perceptron pour prédire un score de pertinence unique pour chaque passage. Le modèle est entraîné sur 24 passages (avec un seul passage pertinent) échantillonnés parmi les 100 meilleurs passages retournés par la RI. Comme à la section précédente, le modèle est d'abord préentraîné sur notre sous-ensemble de TriviaQA, avec une RI effectuée avec BM25 sur les 5,9 millions d'articles de la Wikipédia de KILT au lieu de notre BC multimodale. Le modèle est ensuite ajusté sur ViQuAE en utilisant les mêmes hyperparamètres, la RI étant alors effectuée avec le modèle multimodal fondé sur DPR.

6.3. *Fusion multimodale*

Les scores du réordonnanceur textuel sont ensuite fusionnés avec ceux du réordonnanceur d'image *via* une simple combinaison linéaire, comme pour la RI initiale, après une normalisation min-max. Les méthodes de normalisation, de fusion et les hyperparamètres de ces dernières ont été choisis sur le jeu de validation, avec une recherche par dichotomie.

6.4. *Résultats*

Les performances des réordonnanceurs sont présentées dans les tableaux 5 et 6 de façon comparable à la section précédente. Le modèle pour le texte améliore l'ordonnancement initial de 10 points de MRR et P@1, ce qui représente des améliorations de 25 % et 36 % respectivement. La fusion avec le modèle d'image augmente encore les performances de 5,9 points de P@1, soit une amélioration totale de 57 %. Le chevauchement de 9 % avec GLDv2 (le jeu de données de préentraînement de RRT) est probablement trop faible pour en tirer des conclusions significatives : l'amélioration relative entre les réordonnanceurs texte et multimodal est similaire avec ou sans chevauchement.

L'amélioration de l'ordonnancement de la RI initiale est sans doute due pour l'essentiel au mécanisme d'auto-attention des Transformers appliqué aux paires question-passage, avec BERT dans le cas du texte et avec RRT dans le cas de l'image. L'auto-attention est un mécanisme coûteux mais qui permet de capturer des interactions plus riches qu'un simple produit scalaire (Karpukhin *et al.*, 2020).

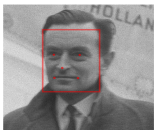
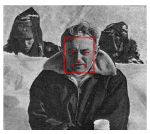
Requête	1er résultat	2ème résultat	3ème résultat
			
« *This arch bridge spans what river ?* »	« Marlow Bridge [SEP] [...] The Széchenyi Chain Bridge, spanning the River *Danube* in Budapest [...] »	« Hudson River [SEP] The width of the Lower *Hudson River* required major feats of engineering to cross [...]	« 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] [...] responsible for large-scale epics such as "The Bridge on the River Kwai" (1957), [...] and "**A Passage To India**" (1984). »	« Bernard Herrmann [SEP] [...] is particularly known for [...] "*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 5. *Requêtes accompagnées des trois premiers résultats de la RI multimodale initiale. 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.*

7. Extraction des réponses

7.1. *Méthodes*

Le but de cette étape est d'extraire une réponse concise à partir d'un passage de texte candidat (provenant par exemple des étapes précédentes de RI). Pour établir notre référence sur ViQuAE, nous nous limitons à un modèle textuel car nous faisons l'hypothèse qu'*une fois le passage pertinent retrouvé en associant texte et image*, il est *possible* de répondre à la question sans utiliser l'image (cf. exemple de la figure 1). L'extraction des réponses est réalisée avec le modèle BERT multipassage de Wang *et al.* (2019). Ce modèle prend en entrée la concaténation de la question et du passage et les encode avec BERT, comme le réordonnanceur. Les représentations sont ensuite données à deux perceptrons différents, entraînés indépendamment à prédire les positions de début et de fin de la réponse. Lors de l'inférence, la probabilité de la position de la réponse est le produit des probabilités de début et de fin. Afin de rendre les scores de réponse comparables d'un passage à l'autre, BERT multipassage exploite la technique de la normalisation globale de Clark et Gardner (2018) afin que

# Exemples	Entrée	F1	Appariement exact (EM)
Aucun	Top 5 RI initiale	22,1	18,5
Aucun	Top 5 réordonnancement	27,7	24,3
Aucun	+ pondération	29,4	26,1
Peu	Top 5 RI initiale	25,5 $\pm$ 0,7	21,4 $\pm$ 0,8
Peu	Top 5 réordonnancement	32,7 $\pm$ 0,4	28,9 $\pm$ 0,4
Peu	+ pondération	33,8 $\pm$ 0,5	30,1 $\pm$ 0,6
Peu	Mi-oracle	43,1 $\pm$ 0,2	39,1 $\pm$ 0,4
Peu	Oracle complet	66,5 $\pm$ 0,7	60,7 $\pm$ 0,9

TABLEAU 7. *Résultats de l'extraction de réponses sur l'ensemble de test de ViQuAE. Pour le modèle few-shot, moyennes sur 5 entraînements avec des graines aléatoires différentes. En inférence, les modèles zero et few-shot prennent les 5 premiers passages en entrée.*

tous les passages partagent la même normalisation softmax. Pour les passages non pertinents, le modèle est entraîné à prédire la première position, c'est-à-dire celle du token spécial [CLS]. De plus, puisque la réponse peut apparaître plusieurs fois dans le même passage, l'objectif d'entraînement, à l'instar de Karpukhin *et al.* (2020), est de maximiser la log-vraisemblance marginale de toutes les positions de réponse dans le passage. Pour prendre en compte les scores P associés aux passages par l'étape précédente de réordonnancement, nous pondérons le score de réponse a tel que $a \leftarrow a \cdot P$ (Wang *et al.*, 2019).

Le modèle est implémenté et entraîné en utilisant la bibliothèque Transformers de Hugging Face (Wolf *et al.*, 2020), elle-même fondée sur PyTorch (Paszke *et al.*, 2019). 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 $8{:}16$. Nous avons également étudié sur le jeu de validation l'effet de la variation du nombre de passages sur l'extraction de réponses lors de l'inférence. Avec les 5 premiers passages, les modèles sont nettement meilleurs. Dans la suite, l'extraction est ainsi appliquée sur les 5 premiers résultats de la RI, initiale ou réordonnancée.

Comme pour le réordonnanceur de texte, le modèle est préentraîné sur TriviaQA et ensuite ajusté sur ViQuAE. Bien que le modèle soit préentraîné, étant donné la petite taille de ViQuAE, l'entraînement a été effectué 5 fois avec des graines aléatoires différentes pour tenir compte de la variabilité causée par l'ordre des questions et le choix aléatoire des passages pertinents et non pertinents parmi leurs ensembles respectifs.

7.2. *Résultats*

Conformément à Joshi *et al.* (2017) ainsi qu'à Petroni *et al.* (2021), nous utilisons l'appariement exact (EM) et le score F1 pour évaluer l'extraction de la réponse après un prétraitement standard (normalisation en minuscule, suppression des articles et de la ponctuation). Les résultats sont présentés dans le tableau 7. Sans surprise,

l'ajustement du modèle sur l'ensemble d'entraînement améliore les performances : + 16 % d'EM. Dans les deux cas, le réordonnancement apporte une amélioration notable de plus de 32 % et la pondération réalisée avec son score est également bénéfique.

Toutefois, les résultats sont globalement assez faibles par rapport à l'état de l'art en QA textuelle. Nous pouvons les comparer aux performances du modèle sur le sous-ensemble de TriviaQA qui a servi à générer le test de ViQuAE : 62,9 de F1 et 59,2 d'EM en prenant en entrée le top 24 de BM25[15], ce qui est du même ordre de grandeur que les résultats obtenus par Wang *et al.* (2019) et par Karpukhin *et al.* (2020) sur les sous-ensembles officiels de validation et de test respectivement. On observe ainsi une amélioration relative de 147 % (F1) et 177 % (EM) en passant de ViQuAE à TriviaQA pour le même ensemble initial de questions.

Pour mieux comprendre ces chiffres, nous avons étudié deux configurations différentes. Premièrement, *mi-oracle*, où les 5 premiers résultats du réordonnancement 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 35 % 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, même si le réordonnancement permet de réduire l'écart[16]. Par exemple, dans la figure 5, deux passages sur trois ne sont pas pertinents mais fournissent une réponse plausible à la question. De futurs travaux pourraient se focaliser sur une meilleure intégration de l'image dans l'extraction de la réponse. Enfin, nous avons considéré la configuration *oracle complet*, où le modèle ne reçoit que des passages pertinents[17]. L'écart de performance continue de se creuser : + 55 % en EM par rapport à *mi-oracle*, qui souffre des résultats modestes de la RI. Ce constat corrobore les résultats de la section 5 : la KVQAE est très difficile pour les représentations d'images actuelles et de futurs travaux devraient porter sur une meilleure fusion des informations multimodales. 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 *oracle* pourraient servir de référence haute aux futures études.

8. Conclusion et perspectives

Nous présentons un nouveau jeu de données, ViQuAE, conçu comme un cadre d'évaluation pour suivre les progrès des systèmes de KVQAE. ViQuAE a été annoté selon une procédure semi-automatique que nous fournissons également. Ses questions ont pour cible une base de connaissances librement disponible de 1,5 million d'articles

15. 63,3 de F1 et 59,7 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.

16. Les résultats *mi-oracle* sont similaires, qu'ils proviennent de la RI initiale ou réordonnancée.

17. Pour *oracle complet*, les résultats de la RI sont filtrés de la même manière que pour *mi-oracle* mais s'il n'y en a aucun, on utilise ceux liés à l'article Wikipédia de l'entité-sujet.

Wikipédia associés à des images. 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 de l'article et ses images représentent une seule et unique entité, sauf dans certains cas exceptionnels. Nos résultats suggèrent 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.

Nous proposons aussi une approche de la KVQAE en trois étapes, distinguant recherche d'information initiale, réordonnancement et extraction des réponses, avec des méthodes d'apprentissage sans ou avec peu d'exemples. Un résultat notable de cette première référence est l'apport positif de l'association du texte et de l'image dans ces différentes configurations. Sans négliger l'extraction des réponses, les évaluations soulignent par ailleurs la nécessité d'une meilleure RI. En effet, notre stratégie de fusion tardive néglige l'interaction entre les modalités. Les travaux futurs devront se concentrer sur une meilleure représentation multimodale, idéalement en intégrant le texte et l'image dans le même espace, tant du côté de la requête que du côté de la BC. Une attention particulière devra être accordée à la représentation des entités non-personnes. Ces représentations multimodales pourront aussi bénéficier à l'étape d'extraction des réponses car nos expériences montrent que l'utilisation d'un modèle textuel seul est insuffisante si la RI est bruitée, bien que le réordonnancement permette en partie de pallier ce cas de figure. D'autre part, bien que nous ayons démontré l'efficacité de notre BC, un système de KVQAE pourrait tirer bénéfice d'une BC plus riche visuellement, avec plusieurs images par entité, afin de prendre en compte la diversité des représentations. Nous espérons plus globalement que ce travail encouragera la recherche vers une meilleure représentation multimodale des entités nommées.

Remerciements

Nous remercions les relecteurs anonymes pour leurs retours constructifs. 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 2021-AD011012846 attribuée par GENCI.

9. Bibliographie

Anderson P., He X., Buehler C., Teney D., Johnson M., Gould S., Zhang L., « 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)*, June, 2018.

Antol S., Agrawal A., Lu J., Mitchell M., Batra D., Zitnick C. L., Parikh D., « VQA : Visual Question Answering », *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Santiago, Chile, December, 2015.

Bommasani R., *et al.*, « On the Opportunities and Risks of Foundation Models », *arXiv :2108.07258 [cs]*, August, 2021. arXiv : 2108.07258.

Cao B., Araujo A., Sim J., « Unifying deep local and global features for image search », *European Conference on Computer Vision*, Springer, 2020.

Chang Y., Narang M., Suzuki H., Cao G., Gao J., Bisk Y., « WebQA : Multihop and Multimodal QA », *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 16495-16504, June, 2022.

Chen D., Fisch A., Weston J., Bordes A., « Reading Wikipedia to Answer Open-Domain Questions », *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 2017.

Clark C., Gardner M., « Simple and Effective Multi-Paragraph Reading Comprehension », *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, July, 2018.

Clough P., Sanderson M., Müller H., « The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004 », *in* P. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, A. W. M. Smeulders (eds), *Image and Video Retrieval*, Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2004.

Deng J., Dong W., Socher R., Li L.-J., Li K., Fei-Fei L., « ImageNet : A large-scale hierarchical image database », *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June, 2009. ISSN : 1063-6919.

Deng J., Guo J., Xue N., Zafeiriou S., « ArcFace : Additive Angular Margin Loss for Deep Face Recognition », *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2019.

Devlin J., Chang M.-W., Lee K., Toutanova K., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », *Proceedings of the 2019 NAACL, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, June, 2019.

Ferragina P., Scaiella U., « 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)*, Association for Computing Machinery, New York, NY, USA, October, 2010.

Fisher R. A., *The Design of Experiments*, 2ème edn, Oliver & Boyd, Edinburgh & London., 1937.

Fleiss J. L., « Measuring nominal scale agreement among many raters », *Psychological Bulletin*, 1971.

Gardères F., Ziaeefard M., Abeloos B., Lecue F., « ConceptBert : Concept-Aware Representation for Visual Question Answering », *Findings of the Association for Computational Linguistics : EMNLP 2020*, Association for Computational Linguistics, Online, November, 2020.

Guo Y., Zhang L., Hu Y., He X., Gao J., « MS-Celeb-1M : A Dataset and Benchmark for Large-Scale Face Recognition », *in* B. Leibe, J. Matas, N. Sebe, M. Welling (eds), *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2016.

He K., Zhang X., Ren S., Sun J., « Deep residual learning for image recognition », *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Jain A., Kothyari M., Kumar V., Jyothi P., Ramakrishnan G., Chakrabarti S., « Select, Substitute, Search : A New Benchmark for Knowledge-Augmented Visual Question Answering », *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021.

Johnson J., Douze M., Jégou H., « Billion-scale similarity search with GPUs », *IEEE Transactions on Big Data*, 2019.

Joshi M., Choi E., Weld D., Zettlemoyer L., « 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)*, Association for Computational Linguistics, Vancouver, Canada, July, 2017.

Karpukhin V., Oguz B., Min S., Lewis P., Wu L., Edunov S., Chen D., Yih W.-t., « Dense Passage Retrieval for Open-Domain Question Answering », *Proceedings of the 2020 EMNLP (EMNLP)*, Association for Computational Linguistics, Online, November, 2020.

Kembhavi A., Seo M., Schwenk D., Choi J., Farhadi A., Hajishirzi H., « 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)*, July, 2017.

Kwiatkowski T., Palomaki J., Redfield O., Collins M., Parikh A., Alberti C., Epstein D., Polosukhin I., Devlin J., Lee K., Toutanova K., Jones L., Kelcey M., Chang M.-W., Dai A. M., Uszkoreit J., Le Q., Petrov S., « Natural Questions : A Benchmark for Question Answering Research », *Transactions of the Association for Computational Linguistics*, March, 2019.

Lerner P., Ferret O., Guinaudeau C., Le Borgne H., Besançon R., Moreno J. G., Lovón Melgarejo J., « 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) 2022.*, ATALA, Avignon, France, 2022a.

Lerner P., Ferret O., Guinaudeau C., Le Borgne H., Besançon R., Moreno J. G., Lovón Melgarejo J., « 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, Association for Computing Machinery, New York, NY, USA, 2022b.

Lhoest Q., *et al.*, « Datasets : A Community Library for Natural Language Processing », *Proceedings of the 2021 EMNLP : System Demonstrations*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, November, 2021.

Lillis D., Zhang L., Toolan F., Collier R. W., Leonard D., Dunnion J., « Estimating probabilities for effective data fusion », *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, 2010.

Lin T.-Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C. L., « Microsoft COCO : Common Objects in Context », *in* D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (eds), *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2014.

Ma X., Sun K., Pradeep R., Lin J., « A Replication Study of Dense Passage Retriever », *arXiv :2104.05740 [cs]*, April, 2021.

Marino K., Rastegari M., Farhadi A., Mottaghi R., « OK-VQA : A visual question answering benchmark requiring external knowledge », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Paszke A., *et*, *al.*, « PyTorch : An Imperative Style, High-Performance Deep Learning Library », *Advances in Neural Information Processing Systems*, 2019.

Petroni F., Piktus A., Fan A., Lewis P., Yazdani M., De Cao N., Thorne J., Jernite Y., Karpukhin V., Maillard J., Plachouras V., Rocktäschel T., Riedel S., « 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*, Association for Computational Linguistics, Online, June, 2021.

Radenović F., Iscen A., Tolias G., Avrithis Y., Chum O., « Revisiting Oxford and Paris : Large-Scale Image Retrieval Benchmarking », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2018.

Radford A., Kim J. W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J. *et al.*, « Learning transferable visual models from natural language supervision », *International Conference on Machine Learning*, PMLR, 2021.

Rajpurkar P., Zhang J., Lopyrev K., Liang P., « SQuAD : 100,000+ Questions for Machine Comprehension of Text », *Proceedings of the 2016 EMNLP*, Association for Computational Linguistics, Austin, Texas, November, 2016.

Reddy R. G., Rui X., Li M., Lin X., Wen H., Cho J., Huang L., Bansal M., Sil A., Chang S.-F., Schwing A., Ji H., « MuMuQA : Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding », December, 2021.

Robertson S. E., Walker S., Jones S., Hancock-Beaulieu M. M., Gatford M., « Okapi at TREC-3 », *in* D. K. Harman (ed.), *Third Text REtrieval Conference (TREC-3)*, vol. 500-225 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 1995.

Sampat S. K., Yang Y., Baral C., « Visuo-Linguistic Question Answering (VLQA) Challenge », *Findings of the Association for Computational Linguistics : EMNLP 2020*, Association for Computational Linguistics, Online, November, 2020.

Schroff F., Kalenichenko D., Philbin J., « FaceNet : A Unified Embedding for Face Recognition and Clustering », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2015.

Schwenk D., Khandelwal A., Clark C., Marino K., Mottaghi R., « A-OKVQA : A Benchmark For Visual Question Answering Using World Knowledge », *Computer Vision – ECCV 2022 : 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*, Springer-Verlag, Berlin, Heidelberg, p. 146–162, 2022.

Shah S., Mishra A., Yadati N., Talukdar P. P., « KVQA : Knowledge-Aware Visual Question Answering », *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019.

Sharif Razavian A., Azizpour H., Sullivan J., Carlsson S., « CNN Features Off-the-Shelf : An Astounding Baseline for Recognition », *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June, 2014.

Smucker M. D., Allan J., Carterette B., « 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, Association for Computing Machinery, New York, NY, USA, November, 2007.

Srihari R. K., Zhang Z., Rao A., « Intelligent Indexing and Semantic Retrieval of Multimodal Documents », *Information Retrieval*, May, 2000.

Talmor A., Yoran O., Catav A., Lahav D., Wang Y., Asai A., Ilharco G., Hajishirzi H., Berant J., « MultiModalQA : Complex Question Answering over Text, Tables and Images », *ICLR 2021*, 2021.

Tan F., Yuan J., Ordonez V., « Instance-Level Image Retrieval Using Reranking Transformers », *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, p. 12105-12115, October, 2021.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. u., Polosukhin I., « Attention is All you Need », *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds), *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

Voorhees E. M., Tice D. M., « Building a question answering test collection », *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00)*, ACM Press, Athens, Greece, 2000.

Wang P., Wu Q., Shen C., Dick A., Van Den Henge A., « Explicit knowledge-based reasoning for visual question answering », *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017.

Wang P., Wu Q., Shen C., Dick A., van den Hengel A., « FVQA : Fact-Based Visual Question Answering », *IEEE transactions on pattern analysis and machine intelligence*, 2018.

Wang Z., Ng P., Ma X., Nallapati R., Xiang B., « Multi-passage BERT : A Globally Normalized BERT Model for Open-domain Question Answering », *Proceedings of the 2019 EMNLP and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, November, 2019.

Weyand T., Araujo A., Cao B., Sim J., « 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)*, 2020.

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., Le Scao T., Gugger S., Drame M., Lhoest Q., Rush A., « Transformers : State-of-the-Art Natural Language Processing », *Proceedings of the 2020 EMNLP : System Demonstrations*, Association for Computational Linguistics, Online, p. 38-45, October, 2020.

Zhang K., Zhang Z., Li Z., Qiao Y., « Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks », *IEEE Signal Processing Letters*, October, 2016.