ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G. Moreno, Jesus Lovon

ViQuAE, a Dataset for Knowledge-based Visual Question Answering about Named Entities

Paul Lerner
Université Paris-Saclay, CNRS, LISN
Orsay, France
paul.lerner@lisn.upsaclay.fr

Olivier Ferret
Université Paris-Saclay, CEA, List
Palaiseau, France
olivier.ferret@cea.fr

Camille Guinaudeau
Université Paris-Saclay, CNRS, LISN
Orsay, France
camille.guinaudeau@lisn.upsaclay.fr

Hervé Le Borgne
Romaric Besançon
herve.le-borgne@cea.fr
romaric.besancon@cea.fr
Université Paris-Saclay, CEA, List
Palaiseau, France

Jose G Moreno
Jesús Lovón Melgarejo
jose.moreno@irit.fr
jesus.lovon@irit.fr
IRIT, Université Paul Sabatier
Toulouse, France

ABSTRACT

Whether to retrieve, answer, translate, or reason, multimodality opens up new challenges and perspectives. In this context, we are interested in answering questions about named entities grounded in a visual context using a Knowledge Base (KB). To benchmark this task, called KVQAE (Knowledge-based Visual Question Answering about named Entities), we provide ViQuAE, a dataset of 3.7K questions paired with images. This is the first KVQAE dataset to cover a wide range of entity types (*e.g.* persons, landmarks, and products). The dataset is annotated using a semi-automatic method. We also propose a KB composed of 1.5M Wikipedia articles paired with images. To set a baseline on the benchmark, we address KVQAE as a two-stage problem: Information Retrieval and Reading Comprehension, with both zero- and few-shot learning methods. The experiments empirically demonstrate the difficulty of the task, especially when questions are not about persons. This work paves the way for better multimodal entity representations and question answering. The dataset, KB, code, and semi-automatic annotation pipeline are freely available at https://github.com/PaulLerner/ViQuAE.

CCS CONCEPTS

• **Information systems** → **Question answering**; **Test collections**; **Multimedia and multimodal retrieval**.

KEYWORDS

dataset, knowledge-based visual question answering, multimodal

ACM Reference Format:
Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, Jose G Moreno, and Jesús Lovón Melgarejo. 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), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3477495.3531753

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531753

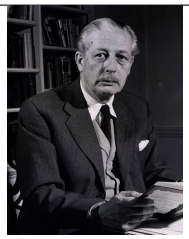
Query (input)	Relevant item in the Knowledge Base
 *"Which constituency did this man represent when he was Prime Minister?"*	 "Macmillan indeed lost Stockton in the landslide Labour victory of 1945, but returned to Parliament in the November 1945 by-election in **Bromley**."
 "*In which year did this 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: Example of questions in the ViQuAE dataset along with their grounding image and answer source (part of the Knowledge Base).

1 INTRODUCTION

Fusing multiple modalities, such as image and text, to retrieve relevant information is a long-standing problem that is nontrivial because these modalities carry semantics at different levels [46]. This is particularly true in the case of Knowledge-based Visual Question Answering about named Entities (KVQAE), the task considered

in this article, where different types of relations can stand between a question and its grounding image. In Visual Question Answering (VQA), the content of the contextual image, such as the color of an object or the number of objects, is the target of the question [2]. On the other hand, Knowledge-based VQA [30, 33, 49, 50] uses the image as a context to ask questions grounded in Knowledge Bases (KBs). However, both lines of work mostly target coarse-grained object categories, resulting in a reliance on an object detection preprocessing step (see for instance [1, 18]). For example, in Figure 1, one could ask about the kind of boat: "*Is this a fishing boat?*" Instead, our work is focused on questions that require knowledge about named entities, such as the boat *Queen Elizabeth 2*. We release the ViQuAE dataset for this purpose[1]. Our dataset was designed as a benchmark to track the progress of KVQAE systems. Indeed, we argue that KVQAE is a clear, well-defined task that can be evaluated easily, making it suitable to track the progress of multimodal entity representation's quality. Multimodal entity representation is a central issue that will allow to make human-machine interactions more natural. For example, while watching a movie, one might wonder "*Where did I already see this actress?*" or "*Did she ever win an Oscar?*"

Questions about named entities are highly challenging since current KBs contain millions of them. Therefore, using each modality independently is insufficient to retrieve relevant information with respect to users' needs. For example, in the images of Figure 1, it is fairly complex to recognize *Harold Macmillan* out of a KB of millions of *persons*. However, one can infer from the question that he was *prime minister*, filtering down the candidates to a few hundred.

Shah et al. [43] have previously worked on KVQAE but were limited to person-named entities. Instead, ViQuAE covers a wide range of entity types. This diversity is a central issue in KVQAE, notably because of the resulting heterogeneity in visual representations. Figure 2 displays a few examples of entity types targeted in our work. Obviously, smartphones and mountains do not look quite alike, but, additionally, it is worth noticing the great diversity among the same entity type or even the same entity. For example, the first row shows two ways of depicting a person (here Louis Philippe I), namely through a photograph or a painting. The heterogeneity is even greater in some sense for organizations that can be depicted through a building (*e.g.* headquarters), a known manufactured product they sell, or simply their logo. This requires a multimodal knowledge representation, which clearly distinguishes KVQAE from image retrieval. It also illustrates the need to study other entities than persons, which can be recognized from their face. Additionally, it is worth pointing out that the very same picture can be used to ask questions about different entities: for example, Figure 2b could be used to ask about Louis Philippe I, but also about the painter or even the painting itself (*e.g.* "*Who painted it?*" or "*Where can I see it?*").

As demonstrated in the next section, there are numerous tasks that mix text and image, and one cannot expect to build datasets of 100K+ samples for each. Zero- and few-shot learning underwent several breakthroughs in various fields of academic research, under the prism of Foundation Models [4]: *e.g.* GPT-3 [5] in Natural Language Processing; CLIP [37] in Computer Vision; DALL-E [39] in Text-to-Image Generation. With only 3.7K samples, ViQuAE is too small to train huge neural networks from scratch. Instead, we expect it to foster research towards transferable model architectures and zero- or few-shot learning techniques, which are essential to any KVQAE system. Note that by "zero-shot", we refer to models that are not fine-tuned using ViQuAE's training set. They are sometimes referred to as "off-the-shelf" models in the Information Retrieval (IR) literature.

Our main contributions are as follows: (i) we provide a new dataset for KVQAE, the first to cover a wide range of entity types, along with an extensible pipeline for semi-automatic annotation; (ii) we redistribute a multimodal KB of 1.5M entities based on Wikipedia; (iii) we propose and open-source strong baselines for both zero- and few-shot methods to address KVQAE, being the first to treat the task on diverse entity types and using a text-based KB.

(a) Person (Louis Philippe I)

(b) Person (Louis Philippe I) or painting (by Franz Xaver Winterhalter)

(c) Artificial Landmark (Eiffel Tower)

(d) Natural Landmark (Mount St. Helens)

(e) Organization (Apple Inc.) or building (Apple Fifth Avenue)

(f) Organization (Apple Inc.) or product (iPhone 1)

Figure 2: Some depictions of different entity types and different depictions of the same entity type considered in our work.

2 RELATED WORK

Since our approach to KVQAE relies on a text-based KB, it is strongly linked to text Question Answering (QA). Text QA gained popularity with the TREC QA evaluations [48]. It has largely been addressed as a two-stage problem, with an IR stage followed by a Reading Comprehension (RC) stage, and a global focus on factoid questions (*e.g.* [7]). Our work is no exception. In the last few

[1] Available at https://github.com/PaulLerner/ViQuAE.

Table 1: Summary of common points and differences between KVQAE and related tasks. All share two modalities: vision and language. Named entities are often opposed to coarse-grained object categories. *Unclear.

Task	Question Answering	Common-sense	Information Retrieval	Named Entities
KVQAE [43]	✔	✘	✔	✔
Multimodal IR [46]	✘	✘	✔	✔
Cross-modal VQA [40]	✔	✘	✘	*
Knowledge-based VQA [33]	✔	✔	✔	✘
VQA [2]	✔	*	✘	✘

years, increased attention has been paid to RC, spawning ever-larger datasets [22, 27, 38, 53]. We take advantage of the latter to build our own dataset, as explained in the next section.

While initially focused on text, IR was rapidly extended to multimodal documents. Srihari et al. [46] and Clough et al. [9] for instance already shared a number of issues with KVQAE, such as multimodal information fusion. However, modalities in multimodal IR are often redundant, while they are complementary in KVQAE.

On the contrary, cross-modal QA [6, 24, 40, 42, 47] can be seen as RC across multiple modalities (*e.g.* text, tables, images...). The answer source, whatever the modality, is provided along with the contextual question and both are interdependent. For example, Reddy et al. [40] build their corpus upon news articles, where the system has access to image metadata, such as its caption. Hence, the task is more about logical reasoning than IR, unlike KVQAE, which is factoid.

Knowledge-based VQA [30, 33, 49, 50] focuses on commonsense questions about coarse-grained object categories. Furthermore, (Knowledge-based) VQA datasets are based on the images of the *Common Objects in Context* (COCO) dataset [31]. For these two reasons, Knowledge-based VQA has largely been addressed with an object detection preprocessing step, the object detector being trained on the images of COCO, which facilitates IR (*e.g.* [18]). Common points and differences between KVQAE and related tasks are summarized in Table 1.

Shah et al. [43] introduce the first KVQAE dataset: KVQA, based on Wikidata and restricted to person entities. An important difference with our work is their use of a KB based on a knowledge graph instead of unstructured text. Despite its large size, their dataset has several limitations: (i) it is restricted to person entities and in this case, person recognition boils down to face recognition; (ii) questions are automatically generated from templates and Wikidata schema. Thus, they are quite repetitive and limited by the schema: most questions are about the person's identity, place of birth, date

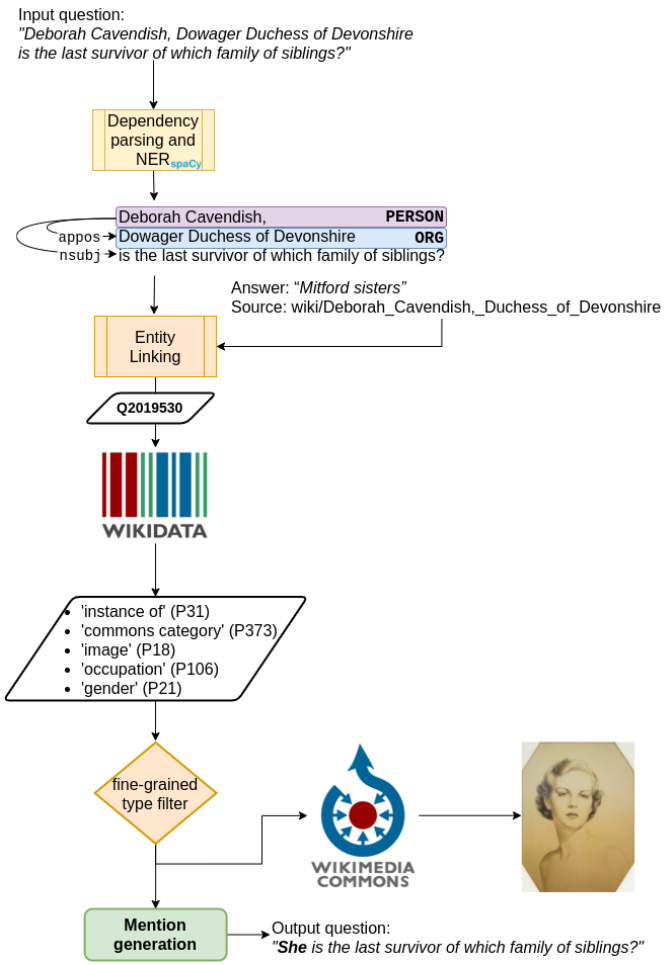


Figure 3: Overview of the automatic annotation pipeline. Note that not only the entity mention ("Deborah Cavendish") but also its syntactic children ("Dowager Duchess of Devonshire") are replaced by the ambiguous mention.

of birth, or job. Instead, we aim at building a dataset covering various entity types with a rich language and questions spanning over many topics.

3 THE VIQUAE DATASET

3.1 Automatic annotation

We build upon existing QA datasets, which provide a wide range of questions spanning over various topics and entities. Additionally, this limits manual annotation efforts. The main idea of the process is to replace the entity mention in the question with a depiction of the entity (see Figure 3). The entity is then referenced by an ambiguous mention (*e.g. "she"*). In this way, one cannot answer the question without relying on the grounding image.

To implement this process, we first need to recognize and disambiguate named entities in the question. We must also find relevant

depictions of the entities. Finally, entities need to be referenced by an ambiguous mention. Referring expression generation has been extensively studied [26], but our approach is quite different since we are looking for images that depict a single entity. In this case, the referring expression does not need to include any distinctive property of the entity, which can be simply referred to by a pronoun or hypernym [10].

To address these challenges, we use Wikipedia[2], Wikidata[3], and Wikimedia Commons[4], where entities are uniquely identified.

Among the various QA datasets mentioned in the previous section, we decided to use TriviaQA because of its large scale and question typology [22]. More precisely, we use the KILT version of TriviaQA [35]. KILT is a benchmark for knowledge-intensive Natural Language Processing tasks, such as QA and Entity Linking. Our automatic annotation pipeline could be applied effortlessly to other QA datasets in KILT.

3.2 Application on TriviaQA

First, dependency parsing and named entity recognition are applied using `spaCy`[5], yielding around 0.9 valid mentions per question. Dependency parsing enables to keep only some entity mentions, *e.g.* the subject of the question. These entity mentions are then matched with the entities disambiguated by Joshi et al. [22], who used TAGME [15]. Note that this entity disambiguation was very precise because candidate entities were discarded if their Wikipedia page did not contain the answer to the question.

Wikidata allows to gather information about the disambiguated entities: their type, occupation, gender, and Wikimedia Commons category. The latter is used to find a relevant depiction, while the others are needed to generate an ambiguous mention. Humans are mentioned by their occupation (*e.g.* "*this writer*") and other entities by their type (*e.g.* "*this tourist attraction*"). Furthermore, if the gender is available, we also use "*this man/woman*" and "*he-him-his/she-her-hers*" according to the syntactic dependency of the original mention.

Because some abstract entities such as countries or nationalities are often mentioned in questions but are not relevant for KVQAE, the entity type is restricted to be part, or a subclass, of a handcrafted list of types, available along with the dataset. Moreover, to comply with GDPR [14], and since the number of questions about humans is quite large, only questions about deceased persons are kept. This step discards another 31% of questions.

To find relevant depictions of the entity in its Commons category, several heuristics are designed to sort the images: first, the image should be tagged as *depicting* the entity in Commons structured data; then, the entity label should be included in: (i) the image's title; (ii) the image's description; (iii) *all* of the image's Commons categories. If several images are available, a unique one is used for each question about a given entity. Of course, the reference image of the entity (see Section 5) is excluded. Thanks to the Wikimedia Commons contributors, all images of the dataset are either freely licensed[6] or in the public domain, allowing us to redistribute them to ensure reproducibility. Around 3% of questions lacked available images and were discarded.

We describe how to refine the automatic pipeline in the next section.

3.3 Manual refinement

The automatic annotation described above has some caveats. Two major sources of errors are the selected image, which might be irrelevant, and the specificity of the question: *e.g.* "***Bonar Law** is the only Prime Minister not born in the UK. In which country was he born?*" is processed into "***He** is the only Prime Minister not born in the UK. In which country was he born?*" which can be answered without looking at the image. To tackle this, an annotation interface has been designed using Label Studio[7]. The annotator is allowed to rephrase the question freely (some alternative mentions are suggested) as long as it *does not change the answer*. They should also choose among eight candidate images if the selected one is not appropriate (based on the reference image of the entity; see Section 5). As a last resort, the annotator may also plainly discard the question. A screenshot of the interface is available in Appendix C, and annotators' instructions are part of our codebase.

Given the subtleties of the annotation process and the staggering reports of Marino et al. [33] who had to discard 73K out of 87K questions in their dataset, we decided to rely on seven in-house annotators (the authors of the paper). Once familiar with the interface, the annotators were able to process ≈ 120 questions per hour. The proportion of questions about humans was balanced to ensure the diversity of the dataset. We annotated 5.7K generated questions, *i.e.* spent around 48 hours of manual annotation in total. Among those 5.7K questions, 2K were discarded, mostly because they were over-specified or lacked a relevant image. Hence, the ViQuAE dataset consists of 3.7K questions, randomly split in training, validation, and test equally-sized sets such that images do not overlap. The majority (55%) of the valid questions were edited by the annotators. Edited questions had an average Levenshtein distance of 5 words from their generated question.

To measure inter-annotator agreement, a subset of 103 questions was annotated by at least 3 different annotators. The agreement on whether to discard the question was computed with Fleiss' Kappa [17]. The annotators showed a fair agreement, with $\kappa = 0.33$. Indeed, whether a question is over-specified or not can be quite subjective. Moreover, some over-specified questions' reformulation can be subtle and not obvious to all annotators. However, one should bear in mind that, in our case, inter-annotator disagreement does not concern *answering* the question but only filtering the automatically generated dataset, as both questions and answers are defined in TriviaQA and the annotator *cannot change the answer*.

We analyze the resulting dataset in the following section.

4 DATA ANALYSIS

The ViQuAE dataset consists of 3.7K questions grounded in 3.3K unique images. Two examples are shown in Figure 1. Questions are 12.4 words long on average, with a vocabulary of 4.7K words. In contrast, left-out questions of TriviaQA are 16.4 words long on average. There are close to no answer priors; among 3.7K answers,

[2] https://www.wikipedia.org/
[3] https://www.wikidata.org/
[4] https://commons.wikimedia.org/
[5] https://spacy.io/
[6] https://freedomdefined.org/Definition
[7] https://labelstud.io/

Table 2: Dataset statistics compared to KVQA [43].

	ViQuAE	KVQA
# Questions	3.7K	183K
# Questions per image	1.1	7.4
Vocabulary	4.7K	0.6K[8]
Average q. length	12.4	10.1
Answer prior	0.3%	15.9%
Answer overlap	25.3%	89.4%
Entity overlap	18.1%	40.6%
# Questions per entity	1.5	9.7
# Entity types	980	1

the most common, "*France*" and "*Turkey*", only occur 13 times, that is 0.3%. Moreover, there is only a 25% answer overlap between the train and test sets, which is very low compared to the reports of Lewis et al. [28], who found 72% answer overlap between the train and test sets of TriviaQA, among other datasets. Because nearly all images in our dataset are unique, and there is no overlap between the subsets' images, there is only an 18% overlap between the entities in the test and train sets. Those three points further highlight the difference between KVQAE and (Knowledge-based) VQA and demonstrate that treating KVQAE as a classification task would be inefficient.

A significant contribution of the dataset is its entity diversity. As discussed in Section 1, it is one of the key challenges for multimodal representations. ViQuAE comprises nearly a thousand different entity types (in the Wikidata ontology; as defined by the property P31 of the entities) among its 2.4K unique entities. Note, however, that those types are not exclusive: each entity has 1.6 types on average. There are 43% humans in the dataset, but other entities are human-like, such as fictional or mythological characters, or groups of humans, *e.g.* a music band. A bubble chart of the top-100 most frequent entity types is shown in Figure 4. A summary of the statistics compared with the KVQA dataset of Shah et al. [43] is reported in Table 2. We can see that, despite its small size, ViQuAE is more diverse in several aspects.

The ViQuAE dataset also has some limitations. A downside of our annotation process, more precisely the named entity disambiguation, is that answers are guaranteed to be in the Wikipedia page of the entity, *i.e.* the questions are one-hop at the document level. However, the question might require reasoning over multiple sentences or paragraphs in the document. In contrast, Shah et al. [43] include several multi-hop questions that, while they might not sound very natural, efficiently probe the reasoning capacities of the model.

In the following section, we describe the Knowledge Base (KB) used to answer the questions of ViQuAE.

5 THE VIQUAE KNOWLEDGE BASE

The KB is built upon Wikipedia, more precisely, the 2019/08/01 dump available in KILT [35], which consists of 5.9M articles. Each one is mapped to a Wikidata entity. Hence, we use both terms interchangeably. To get a visual representation of the entity, a single image is retrieved from Wikidata, in the following order of preference of its image properties: (i) P18 "image" (it is roughly equivalent to the *infobox* image in Wikipedia articles); (ii) P154 "logo image"; (iii) P41 "flag image"; (iv) P94 "coat of arms image"; (v) P2425 "service ribbon image". Articles without any image are discarded, leaving us with a KB of 1.5M articles – including 542K humans – each paired with an image. This is more than two orders of magnitude greater than the "open-world" experiments of Shah et al. [43]. 95% of the images in the KB are unique.

In the next section, we describe how to index this KB to address KVQAE along with the rest of our baseline.

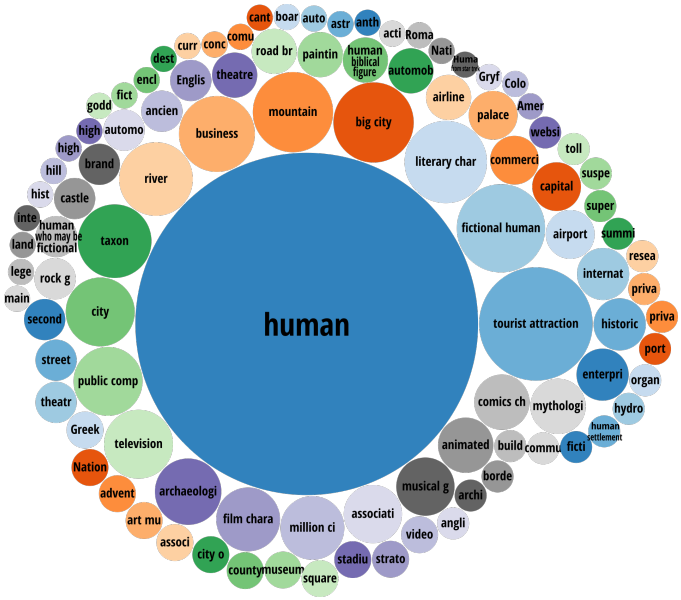


Figure 4: Bubble chart of the top-100 most frequent, non-exclusive, entity types in the dataset.

6 EXPERIMENTAL SETTINGS

We divide the KVQAE problem into two steps: Information Retrieval and Reading Comprehension, with dedicated evaluation metrics for each. We emphasize on IR as we argue that it is the most challenging step. Following Joshi et al. [22] and Petroni et al. [35], Wikipedia aliases of a given answer are considered valid answers. Final evaluation is always carried out on the test set of 1,257 questions, while hyperparameters were tuned on the validation set of 1,250 questions and, for few-shot baselines, models were trained on the training set of 1,190 questions. Additional details are given in Appendix B and all the experiments can be reproduced using our codebase[9].

7 INFORMATION RETRIEVAL

IR aims at retrieving relevant sources of knowledge from the KB with respect to the query (question and image).

[8] Averaged over 49 random subsets of the same size as ViQuAE, the vocabulary over the whole KVQA dataset consists of 8.4K tokens.

[9] https://github.com/PaulLerner/ViQuAE

7.1 Methods

We follow a late fusion approach: search is done independently with the question and the image. Results are then fused at the score level. Our implementation is based on Elasticsearch[10] and Faiss [21] for sparse and dense retrieval respectively, both via Hugging Face's `datasets` library [29].

7.1.1 Text Retrieval. Following Wang et al. [51] and Karpukhin et al. [23], articles are stripped of their semi-structured data, such as tables and lists. Each article is then split into disjoint passages of 100 words for text retrieval while preserving sentence boundaries, which leads to 12M passages ($\approx$ 8 passages per article). The title of the article is appended to the beginning of each passage. As a zero-shot baseline, we use BM25 [41] and optimize its hyperparameters on the validation set using grid search. To also set a few-shot baseline, we rely on DPR [23]. DPR is a neural retrieval model built upon two BERT [13] models: one for the question and one for the passage. DPR is trained to minimize the cross-entropy of the similarities between questions and passages (with a single relevant passage per question). Crucially, hard negatives are mined using BM25. Because of the small size of ViQuAE, the model is first pre-trained on TriviaQA, filtered out of all questions used in ViQuAE, even those that were discarded. We also consider the model only trained on TriviaQA as another zero-shot baseline[11]. The validation is done on the TriviaQA questions used to generate the ViQuAE validation set. For training, the hyperparameters are set as in [23].

7.1.2 Image Retrieval. For image retrieval, two different representations are used in an exclusive manner. ArcFace [12] for faces, if at least one is detected, and, if not, ImageNet-ResNet [20] and CLIP [37] for the full image. Therefore, the KB is split into two parts: humans with a detected face and non-humans, as we (naively) assume that faces are only relevant for human entities. Following Deng et al. [12], we use MTCNN [54] for face detection. The 5 face landmarks (two eyes, nose, and two mouth corners) are adopted to perform similarity transformations so that they are always at the same position in the image, regardless of the original pose of the person. If several faces are detected, only the one associated with the highest probability is kept. 6.6% of the humans in the KB lacked a detected face and were hence discarded.

ArcFace is a state-of-the-art representation learning method for face recognition and verification. We use the model pre-trained on MS-Celeb [19], consisting of celebrities' pictures. Its entities have some overlap with ViQuAE, which is analyzed in the next section.

ResNet is a milestone in the history of deep neural networks as its "skip-connections" allow it to train hundred layer deep Convolutional Neural Networks (CNNs). It is widely used as a backbone in representation learning, *e.g.* in ArcFace. We denote "ImageNet-ResNet" the model trained on ImageNet [11], the most popular pre-training dataset for image classification over 1,000 object categories. Indeed, the features extracted from the last convolutional layer of ImageNet-ResNet have been shown to be a competitive baseline for image retrieval [36, 44]. We rely on max-pooling to reduce the feature map, given the results reported in [36].

CLIP [37] is a dual-encoder framework to learn visual representations from language supervision. The training objective is akin to DPR, although CLIP matches images with relevant captions instead of questions with relevant answers. CLIP has been trained on a dataset of 400M image-caption pairs. We are only interested in the visual encoder of CLIP and discard its text encoder. Indeed, CLIP was trained on image captions, so we do not expect its text encoder to be suited for QA.

For the sake of fair comparison, we systematically use a ResNet-50 backbone for all visual representations. Note that all these models are used off-the-shelf and are *not* fine-tuned.

7.1.3 Multimodal fusion. Dense search is carried out with maximum inner product search, equivalent to cosine similarity, as features are normalized beforehand (except for DPR). The image results are then mapped to their associated passages to enable fusion with the text search.

The result scores of these models have very different distributions. Therefore, before fusing them, they are normalized to have a zero mean and unit variance. Following Karpukhin et al. [23] and Ma et al. [32], the scores are fused through a linear interpolation:

$$P = \alpha_b B + \alpha_d D + \mathbf{F}\alpha_a A + (1 - \mathbf{F})(\alpha_i I + \alpha_c C) \quad (1)$$

where B, D, A, I, C stands for BM25, DPR, ArcFace, ImageNet-ResNet, and CLIP, respectively, and each has an interpolation hyperparameter α_j (with $\sum_j \alpha_j = 1$). $\mathbf{F} \in \{0, 1\}$ denotes the detection of a face. Only the top-100 passages are considered. Therefore, if, given a query, a passage is not retrieved by a given system, then it is assigned to the minimum score of the other passages retrieved by that system. Passages are then re-ordered with respect to the score P. Interpolation hyperparameters α_j are tuned on the validation set using grid search to maximize Mean Reciprocal Rank. To limit the search space and facilitate direct comparison between BM25 and DPR, we use a single model for text search, *i.e.* $\alpha_b = 0$ or $\alpha_d = 0$.

7.2 Results

Since it is based on TriviaQA [22], ViQuAE is only distantly supervised, *i.e.* a passage is deemed relevant if it contains the answer. We evaluate IR with Precision@K (P@K) and Mean Reciprocal Rank (MRR) along with Hits@K. Hits@K represents the proportion of questions for which IR retrieves *at least one* relevant passage in top-K. Metrics are computed with `ranx` [3].

Results are reported in Table 3. Statistical significance tests are carried out using Fisher's randomization test [16, 45]. We also report the text-only performance of BM25 and DPR as baselines.

7.2.1 DPR vs. BM25. DPR's performance gain over BM25 is impressive, even in the zero-shot setting where it significantly outperforms BM25, and even the BM25-based multimodal search in P@K and Hits@K when $K \geq 5$. Unlike BM25, DPR is able to find relevant passages even with very few lexical overlap thanks to its abstract semantic representations. For example, in the question "*This art museum*[12] *is in which Russian city?*", zero-shot DPR is able to guess

[10] https://www.elastic.co/

[11] Even if the ViQuAE questions are built from a subset of the TriviaQA questions, we consider that the rephrasing of the TriviaQA questions is important enough for considering this setting as a kind of zero-shot baseline. TriviaQA and ViQuAE have very different questions lengths (see Section 4).

[12] Referring to the Hermitage Museum.

Table 3: IR results with the text-only baselines and the fusion of text and image searches, in both zero- and few-shot settings. Superscripts denote significant differences in Fisher's randomization test with $p \leq 0.01$. Hits@1 is omitted as it is equivalent to P@1.

#	Model	MRR@100	P@1	P@5	P@20	Hits@5	Hits@20	Hits@100
a	B (BM25, text-only)	19.0	13.1	8.7	5.9	23.9	39.5	62.1
b	D_0 (DPR zero-shot, text-only)	30.5^{a}	21.2^{a}	19.1^{ac}	16.2^{ac}	40.3^{ac}	60.5^{ac}	76.9^{ac}
c	$0.3(B + \mathbf{F}A) + (1 - \mathbf{F})(0.1I + 0.3C)$	27.9^{a}	20.4^{a}	13.8^{a}	10.1^{a}	35.2^{a}	50.5^{a}	69.8^{a}
d	$0.3(D_0 + \mathbf{F}A) + (1 - \mathbf{F})(0.1I + 0.3C)$	36.0^{abce}	26.7^{abce}	21.4^{abc}	17.1^{ac}	46.0^{abc}	65.2^{abce}	81.3^{abc}
e	D_f (DPR few-shot, text-only)	32.8^{abc}	22.8^{a}	20.0^{ac}	16.4^{ac}	43.6^{abc}	61.2^{ac}	79.1^{ac}
f	$0.3(D_f + \mathbf{F}A) + 0.2(1 - \mathbf{F})(I + C)$	$\mathbf{37.9}^{abcde}$	$\mathbf{27.8}^{abce}$	$\mathbf{22.5}^{abce}$	$\mathbf{17.5}^{ac}$	$\mathbf{49.5}^{abcde}$	$\mathbf{65.7}^{abce}$	$\mathbf{82.3}^{abce}$

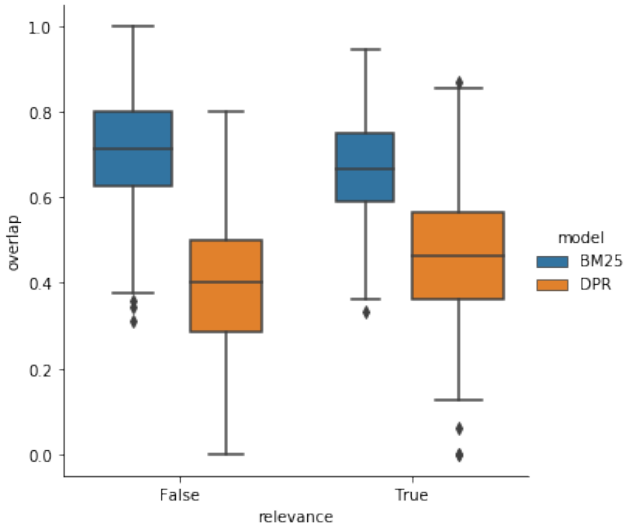


Figure 5: Overlap between the lemmas of the question and the top-1 passage retrieved by BM25 and DPR zero-shot against the passage's relevance. The box shows the quartiles while the whiskers extend to show the rest of the distribution, except for outliers.

the answer ("*Saint Petersburg*"), while BM25 is fooled by the following passage that includes the "art" and "museum" terms several times: "*Ramat Gan [SEP] Man and the Living World* ***Museum*** *is a natural history* ***museum*** *and the Maccabi* ***Museum*** *focuses on the history of Jewish sports since 1898. The Ramat Gan Safari, a zoo housing 1,600 animals, is the largest animal collection in the Middle East. Other* ***museums*** *in the city include the* ***Museum*** *of Israeli* ***Art****, Kiryat Omanut which houses sculpture galleries and a ceramics studio, the* ***Museum*** *of* ***Russian Art****, the* ***Museum*** *of Jewish* ***Art****, and the Yehiel Nahari* ***Museum*** *of Far Eastern* ***Art***." In Figure 5, we can see that DPR has very little lexical overlap compared to BM25, while being more precise. However, its relevant passages tend to overlap more with the question.

7.2.2 Mono- vs. Multi-modal. Fusing BM25 with image search provides a tremendous gain: +56% in P@1. Fusing DPR with image search also results in significant performance gains, both in the zero- and few-shot settings. It is worth pointing out that, in the few-shot setting, the optimal α hyperparameters are $\alpha_d = \alpha_a = 0.3$ and $\alpha_i = \alpha_c = 0.2$, *i.e.* the three modalities (text, face, and full image) are near-equally represented and ImageNet-ResNet and CLIP equally share the full-image modality. The performance gain brought by the multimodal fusion can be analyzed according to the type of entity the question is about. On questions about humans, P@1 jumps from 14.4 with BM25-only to 24.4 when fusing BM25 with image search, which is a 70% improvement. In comparison, the 41% improvement in P@1 on questions about non-humans is weaker. Furthermore, on the subset of entities that overlap with MS-Celeb (ArcFace's pre-training dataset), P@1 further boosts to 25.7, which is a 5% improvement compared to all humans. The trend is similar with DPR, although it starts higher with its text-only performance.

7.2.3 Conclusion. Despite the improvement brought by DPR and the multimodal fusion, there is still a lot of room for improvement, which highlights the difficulty of the task, especially for questions about non-human entities. This can be explained by the specialized image representation of ArcFace, whereas ImageNet-ResNet and CLIP are more general. It also highlights the need to study visual representation of non-human entities, as exemplified in Section 1.

8 READING COMPREHENSION

Given a selected list of passages (*e.g.* from IR), RC aims at extracting a concise answer to the question.

8.1 Methods

To set a baseline on our dataset, we rely on a text-only reader as we argue that, *once the relevant passage has been retrieved* (and only then), the question can be answered without looking at the image (see *e.g.* Figure 1). RC is done with Multi-passage BERT [51]. This model takes as input the concatenation of the question and passage and encodes them with BERT [13]. The representations are then fed into two different fully-connected layers, trained independently to predict the start and end positions of the answer span, respectively. At inference, the answer span probability is the product of the start and end probabilities. In order to make answer scores comparable across passages, Multi-passage BERT leverages the global normalization technique of Clark and Gardner [8] so that all passages share the same softmax normalization. For irrelevant passages, the model is trained to predict the first position, *i.e.* that of the special token `[CLS]`. Furthermore, following Karpukhin et al. [23], since

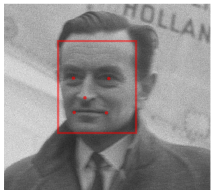
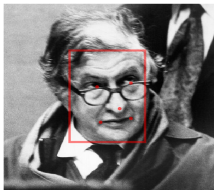
Query	1st result	2nd result	3rd result
			
"*This arch bridge spans what river?*"	"Marlow Bridge [SEP] [...] The current suspension bridge was designed by William Tierney Clark and was built between 1829 and 1832, replacing a wooden bridge further downstream which collapsed in 1828. The Széchenyi Chain Bridge, spanning the River *Danube* in Budapest, was also designed by William Clark and it is a larger scale version of Marlow bridge."	"Hudson River [SEP] The width of the Lower *Hudson River* required major feats of engineering to cross; the results are today visible in the George Washington Bridge and the 1955 Tappan Zee Bridge (replaced by the New Tappan Zee Bridge) as well as the Lincoln and Holland Tunnels and the PATH and Pennsylvania Railroad tubes. [...]"	"Pont de la Tournelle [SEP] The location of the is the site of successive structures. The first, a wooden bridge, was built in 1620. This bridge connected the Eastern bank of the **Seine** (le quai Saint-Bernard) to l'ile Saint-Louis. It was subsequently washed away by ice in 1637, and again on 21 January 1651. [...]"
			
"*What was the last film directed by this film producer?*"	"David Lean [SEP] Sir David Lean (25 March 1908-16 April 1991) was an English film director, producer, screenwriter and editor, responsible for large-scale epics such as "The Bridge on the River Kwai" (1957), "Lawrence of Arabia" (1962), "Doctor Zhivago" (1965) and "**A Passage To India**" (1984).	Bernard Herrmann [SEP] An Academy Award-winner (for "*The Devil and Daniel Webster*", 1941; later renamed "*All That Money Can Buy*"), Herrmann is particularly known for his collaborations with director Alfred Hitchcock, most famously "*Psycho*", "*North by Northwest*", "*The Man Who Knew Too Much*", and "*Vertigo*"."	"David Lean [SEP] [...] Lean recruited long-time collaborators for the cast and crew, including Maurice Jarre (who won another Academy-Award for the score), Alec Guinness in his sixth and final role for Lean, as an eccentric Hindu Brahmin, and John Box, the production designer for "*Dr. Zhivago*"."

Figure 6: Queries along with the top-3 results of multimodal IR. The answer (in the relevant passage) is printed in bold font and plausible answers in irrelevant passages are printed in italic. Face landmarks and bounding boxes, if detected, are shown in red. The passage of text has been shortened due to space constraints.

the answer may appear several times in the same passage, the training objective is to maximize the marginal log-likelihood of all the answer positions in the passage. We do not use re-ranking, as we expect that re-ranking based on text-only will only worsen the original IR order. We leave multimodal re-ranking for future work. Instead, following Wang et al. [51], we experiment with weighting the answer score a with the IR score of the passage P s.t. $a \leftarrow a \cdot P$.

The model is implemented and trained using Hugging Face's `transformers` library [52], itself based on PyTorch [34]. The hyperparameters are set as in [23], except for the ratio of relevant and irrelevant passages per question, which is set to 8:16. During inference, RC is carried out on the top-24 IR results.

As in the previous section, the model is first pre-trained on our custom subset of TriviaQA, with IR carried out using BM25 on the full 5.9M articles of KILT's Wikipedia instead of our multimodal KB. The model is then fine-tuned on ViQuAE, using the same hyperparameters, with IR done using the best model on our multimodal KB. Although the model is pre-trained, given the small size of ViQuAE, training was run 5 times with different seeds to account for the variability caused by questions' order and the random choice of relevant and irrelevant passages among the pool. Since the IR scores $\mathbf{P}$ have a zero mean and unit variance, before weighting the answer, they are updated s.t. $\forall P \in \mathbf{P},\ P \leftarrow 1 - \min(\mathbf{P})$ to ensure they are greater than 1.

Table 4: Downstream RC results on ViQuAE's test set, averaged over 5 runs for the few-shot model. Both zero- and few-shot models share the same IR results at inference (top-24 passages).

# Shots	Setting	F1	Exact Match
Zero	Reader only	20.96	18.06
Zero	+ IR weighting	21.19	18.22
Few	Reader only	25.43 ± 0.42	22.07 ± 0.54
Few	+ IR weighting	25.50 ± 0.38	22.10 ± 0.54
Few	Semi-oracle	44.10 ± 0.39	40.32 ± 0.43
Few	Full-oracle	63.17 ± 1.18	57.55 ± 1.10

8.2 Results

Following Joshi et al. [22] and Petroni et al. [35], we use Exact Match (EM) and F1-score to evaluate the downstream QA, after standard answer preprocessing (lowercasing, stripping articles, and punctuation). Results are reported in Table 4. As expected, fine-tuning the model on the training set provides a solid boost in performance: +22% in EM. Weighing the answers with the IR score brings a slight improvement but well within the standard deviation range of the few-shot runs. Results are overall quite low compared to text QA benchmarks.

To better understand these numbers, we studied two additional settings. In the *semi-oracle* setting, the top-24 IR results are filtered to contain only relevant passages (if any). This brings an impressive 83% improvement in EM compared to the baseline. This shows that the reader is unable to disambiguate between a relevant and an irrelevant passage. For instance, in both examples of Figure 6, two out of three passages are irrelevant but provide a plausible answer to the question. Compared to this setting, the improvements of the IR weighting are insignificant. This motivates future research towards better integration of the image in RC. In the *full-oracle* setting, the reader is only fed relevant passages. The performance gap keeps widening: +43% compared to the semi-oracle EM. It corroborates the results of Section 7: KVQAE is very challenging for current image representations, and future work should focus on a better multimodal information fusion. Moreover, those fairly high numbers support our hypothesis, while nuancing it: *once the relevant passage has been retrieved*, the question *may* be answered without looking at the image. These oracle results could therefore serve as a topline for future work.

9 CONCLUSION AND PERSPECTIVES

We introduce a new dataset, ViQuAE, designed as a benchmark to track the progress of KVQAE systems. The dataset has been annotated with a semi-automatic pipeline that we also provide. Questions in the dataset may be answered using a freely available KB of 1.5M Wikipedia articles paired with images. We propose a baseline along with the benchmark that addresses KVQAE as a two-stage problem: IR and RC, with both zero- and few-shot learning methods for the two stages. First, IR is carried out with well-established technologies: term-based text retrieval, CNN-based image retrieval, and face recognition, as well as recent BERT-based retrieval techniques. Then, RC also takes advantage of the ubiquitous BERT model. While both stages could be improved, the experiments highlight the need for a better IR. Indeed, our late fusion scheme neglects interaction between the modalities. Future work should focus on a better multimodal representation, ideally embedding text and image in the same space, on both the query and KB sides. Special attention should be paid to the representation of non-human entities. As exemplified in Section 1 and demonstrated in Section 7.2, humans can be clearly represented with their face, while other entities have more heterogeneous depictions. We believe that multimodal representations will also benefit the RC stage, as our experiments show that using a text-only reader is insufficient if the IR stage is noisy.

We expect that this work will foster research towards a better multimodal entity representation and question answering and, more generally, a better understanding of the links between language, vision, and knowledge.

ETHICAL CONSIDERATIONS

This paper describes the collection of a dataset to address the task of KVQAE. We made sure that we had the right to redistribute the dataset and KB, thus ensuring the reproducibility of our experiments. Questions of our dataset are released under a CC BY 4.0 License[13]. Thanks to the Wikimedia Commons contributors, all images of the dataset and the KB are either freely licensed or in the public domain. The text in the KB comes from Wikipedia and is therefore available under the CC BY-SA 3.0 License[14]. Moreover, in order to comply with the GDPR, we do not use images of persons unless they are famous and deceased.

During the automatic annotation process, some referring expressions rely on the gender of the entity, if applicable. Note, however, that the gender is not binary in Wikidata; transgender and cisgender people get the same mentions, and intersex and non-binary people[15] are mentioned using other properties (see Section 3.2).

ACKNOWLEDGMENTS

We thank the SIGIR reviewers for their useful comments. This work was supported by the ANR-19-CE23-0028 MEERQAT project. This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012846 made by GENCI.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Santiago, Chile, 2425–2433. https://doi.org/10.1109/ICCV.2015.279

[3] Elias Bassani. 2022. ranx: A Blazing-Fast Python Library for Ranking Evaluation and Comparison. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty (Eds.). Springer International Publishing, Cham, 259–264. https://doi.org/10.1007/978-3-030-99739-7_30

[4] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]* (Aug. 2021). http://arxiv.org/abs/2108.07258 arXiv: 2108.07258.

[5] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[6] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2021. WebQA: Multihop and Multimodal QA. *arXiv:2109.00590 [cs]* (Sept. 2021). http://arxiv.org/abs/2109.00590

[7] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1870–1879. https://aclanthology.org/P17-1171/

[13] http://creativecommons.org/licenses/by/4.0/

[14] https://creativecommons.org/licenses/by-sa/3.0/

[15] In practice, there were none in TriviaQA.

[8] Christopher Clark and Matt Gardner. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 845–855. https://doi.org/10.18653/v1/P18-1078

[9] Paul Clough, Mark Sanderson, and Henning Müller. 2004. The CLEF Cross Language Image Retrieval Track (ImageCLEF) 2004. In *Image and Video Retrieval (Lecture Notes in Computer Science)*, Peter Enser, Yiannis Kompatsiaris, Noel E. O'Connor, Alan F. Smeaton, and Arnold W. M. Smeulders (Eds.). Springer, Berlin, Heidelberg, 243–251. https://doi.org/10.1007/978-3-540-27814-6_31

[10] Robert Dale and Nicholas Haddock. 1991. Generating Referring Expressions Involving Relations. In *Fifth Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Berlin, Germany. https://aclanthology.org/E91-1028

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848 ISSN: 1063-6919.

[12] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] The European Parliament. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). http://data.europa.eu/eli/reg/2016/679/oj/eng Legislative Body: EP, CONSIL.

[15] Paolo Ferragina and Ugo Scaiella. 2010. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM'10)*. Association for Computing Machinery, New York, NY, USA, 1625–1628. https://doi.org/10.1145/1871437.1871689

[16] R. A. Fisher. 1937. The design of experiments. *The design of experiments*. 2nd Ed (1937). https://www.cabdirect.org/cabdirect/abstract/19371601600 Publisher: Oliver & Boyd, Edinburgh & London..

[17] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76, 5 (1971), 378–382. https://doi.org/10.1037/h0031619

[18] François Gardères and Maryam Ziaeefard. 2020. ConceptBert: Concept-Aware Representation for Visual Question Answering. *Findings of the Association for Computational Linguistics: EMNLP 2020* (2020), 10. https://aclanthology.org/2020.findings-emnlp.44/

[19] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *Computer Vision – ECCV 2016 (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 87–102. https://doi.org/10.1007/978-3-319-46487-9_6

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778. https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

[21] J. Johnson, M. Douze, and H. Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547. https://doi.org/10.1109/TBDATA.2019.2921572

[22] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 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)*. Association for Computational Linguistics, Vancouver, Canada, 1601–1611. https://doi.org/10.18653/v1/P17-1147

[23] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://www.aclweb.org/anthology/2020.emnlp-main.550

[24] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2017/html/Kembhavi_Are_You_Smarter_CVPR_2017_paper.html

[25] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*. http://arxiv.org/abs/1412.6980

[26] Emiel Krahmer and Kees van Deemter. 2019. Computational Generation of Referring Expressions: An Updated Survey. In *The Oxford Handbook of Reference*, B. Abbott and J. Gundel (Eds.). Oxford University Press. https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199687305.001.0001/oxfordhb-9780199687305-e-19

[27] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (March 2019), 452–466. https://doi.org/10.1162/tacl_a_00276

[28] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021. Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1000–1008. https://aclanthology.org/2021.eacl-main.86

[29] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 175–184. https://aclanthology.org/2021.emnlp-demo.21

[30] Guohao Li, Hang Su, and Wenwu Zhu. 2017. Incorporating External Knowledge to Answer Open-Domain Visual Questions with Dynamic Memory Networks. *arXiv:1712.00733 [cs]* (Dec. 2017). http://arxiv.org/abs/1712.00733

[31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014 (Lecture Notes in Computer Science)*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[32] Xueguang Ma, Kai Sun, Ronak Pradeep, and Jimmy Lin. 2021. A Replication Study of Dense Passage Retriever. *arXiv:2104.05740 [cs]* (April 2021). http://arxiv.org/abs/2104.05740

[33] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3195–3204. https://ieeexplore.ieee.org/document/8953725/

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems* 32 (2019). https://papers.nips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[35] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2523–2544. https://doi.org/10.18653/v1/2021.naacl-main.200

[36] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. 2018. Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://openaccess.thecvf.com/content_cvpr_2018/html/Radenovic_Revisiting_Oxford_and_CVPR_2018_paper.html

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[38] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2383–2392. https://doi.org/10.18653/v1/D16-1264

[39] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation.

In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

[40] Revanth Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avirup Sil, Shih-Fu Chang, Alexander Schwing, and Heng Ji. 2021. MuMuQA: Multimedia Multi-Hop News Question Answering via Cross-Media Knowledge Extraction and Grounding. (Dec. 2021). https://arxiv.org/abs/2112.10728v1

[41] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline M. Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. In *Third Text REtrieval Conference (TREC-3) (NIST Special Publication, Vol. 500-225)*, Donna K. Harman (Ed.). National Institute of Standards and Technology (NIST), 109–126. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.32.9922&rep=rep1&type=pdf

[42] Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. 2020. Visuo-Linguistic Question Answering (VLQA) Challenge. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 4606–4616. https://doi.org/10.18653/v1/2020.findings-emnlp.413

[43] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: Knowledge-Aware Visual Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8876–8884. https://144.208.67.177/ojs/index.php/AAAI/article/view/4915

[44] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. https://www.cv-foundation.org/openaccess/content_cvpr_workshops_2014/W15/html/Razavian_CNN_Features_Off-the-Shelf_2014_CVPR_paper.html

[45] Mark D. Smucker, James Allan, and Ben Carterette. 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)*. Association for Computing Machinery, New York, NY, USA, 623–632. https://doi.org/10.1145/1321440.1321528

[46] Rohini K. Srihari, Zhongfei Zhang, and Aibing Rao. 2000. Intelligent Indexing and Semantic Retrieval of Multimodal Documents. *Information Retrieval* 2, 2 (May 2000), 245–275. https://doi.org/10.1023/A:1009962928226

[47] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultiModalQA: Complex Question Answering over Text, Tables and Images. In *ICLR 2021*. https://openreview.net/forum?id=ee6W5UgQLa

[48] Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'00)*. ACM Press, Athens, Greece, 200–207. https://doi.org/10.1145/345508.345577

[49] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. 2017. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. 1290–1296. https://dl.acm.org/doi/abs/10.5555/3171642.3171825

[50] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. FVQA: Fact-Based Visual Question Answering. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2018), 2413–2427. https://doi.org/10.1109/TPAMI.2017.2754246

[51] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 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)*. Association for Computational Linguistics, Hong Kong, China, 5878–5882. https://doi.org/10.18653/v1/D19-1599

[52] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771 [cs]* (July 2020). http://arxiv.org/abs/1910.03771

[53] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2369–2380. https://doi.org/10.18653/v1/D18-1259

[54] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters* 23, 10 (Oct. 2016), 1499–1503. https://doi.org/10.1109/LSP.2016.2603342 Conference Name: IEEE Signal Processing Letters.

A KNOWLEDGE BASE DETAILS

As explained in Section 5, our KB is built upon Wikipedia, more precisely, the version available along with KILT. While KILT provides a near 1-1 mapping between the 5.9M articles of Wikipedia and their corresponding Wikidata entities, 11K entities (that is 0.2%) are mapped to more than one article. Therefore, to build the KB, we pruned some articles to obtain a 1-1 mapping using the following heuristics: (i) keep the article that provides an answer for the TriviaQA dataset; (ii) discard articles with "disambiguation" in the title to remove disambiguation pages; (iii) keep the longest article for the same purpose.

Questions in ViQuAE are grounded in an image, as are the articles in the KB. A question about a given entity always uses a different image than the one in the KB. However, other entities in the KB might use the same image as a question in ViQuAE. For example, a question about Odin uses the same image as Hugin and Munin in the KB, or, a question about the Severn Bridge the same as the M48 motorway in the KB. Out of the 3.3K images in ViQuAE and the 1.4M in the KB, there is an intersection of 98 images that correspond to 108 questions, that is 3% of ViQuAE. However, this is not necessarily a bias that will lead to over-optimistic results. Indeed, only 54 of these 108 questions have an answer in the article of the KB that uses the same image.

B EXPERIMENT DETAILS

While our codebase allows to reproduce our experiments, we discuss a few details here, left out of sections 7 and 8 to facilitate reading.

All experiments were carried out with NVIDIA V100 GPUs with 32GB of RAM.

B.1 Information Retrieval

For training DPR, we use the same hyperparameters as Karpukhin et al. [23]. We train DPR using 4 V100 GPUs of 32GB, allowing a total batch size of 256 (32 questions $\times$ 2 passages each $\times$ 4 GPUs). This is crucial because each question uses all passages paired with other questions in the batch as negative examples. Each question is paired with 1 relevant passage and 1 irrelevant passage mined with BM25. Both the question and passage encoder are initialized from "bert-base-uncased". We use the Adam optimizer [25] with a learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$. The learning rate is scheduled linearly with 1,237 warm-up steps. Gradients' norms are clipped at 2.

B.2 Reading Comprehension

As explained in Section 3, Joshi et al. [22] use entity linking to find relevant passages of text for the questions of TriviaQA (upon which our dataset is built). They also retrieve additional passages using Bing Search web API. The reader is trained in priority on the passages retrieved by the IR system, but, if the IR returns only irrelevant passages, the pool of Joshi et al. [22] is used.

For training the reader, we use the same hyperparameters as Karpukhin et al. [23], except for the ratio of relevant and irrelevant passages per question, which is set to 8:16. We use a single V100 GPU with a batch size of 72 (3 questions $\times$ 24 passages each). We use

the Adam optimizer with a constant learning rate of 10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$. Gradients' norms are clipped at 1.

C ANNOTATION INTERFACE

The manual annotation process is described in Section 3.3. The user interface is depicted in Figure 7. The annotator is allowed to rephrase the question freely (some ambiguous mentions are suggested) as long as it does not change the answer. They should also choose among the available images if the one selected (on the top-left) is not appropriate (based on the reference image of the entity, shown on the right). As a last resort, the annotator may also plainly discard the question.

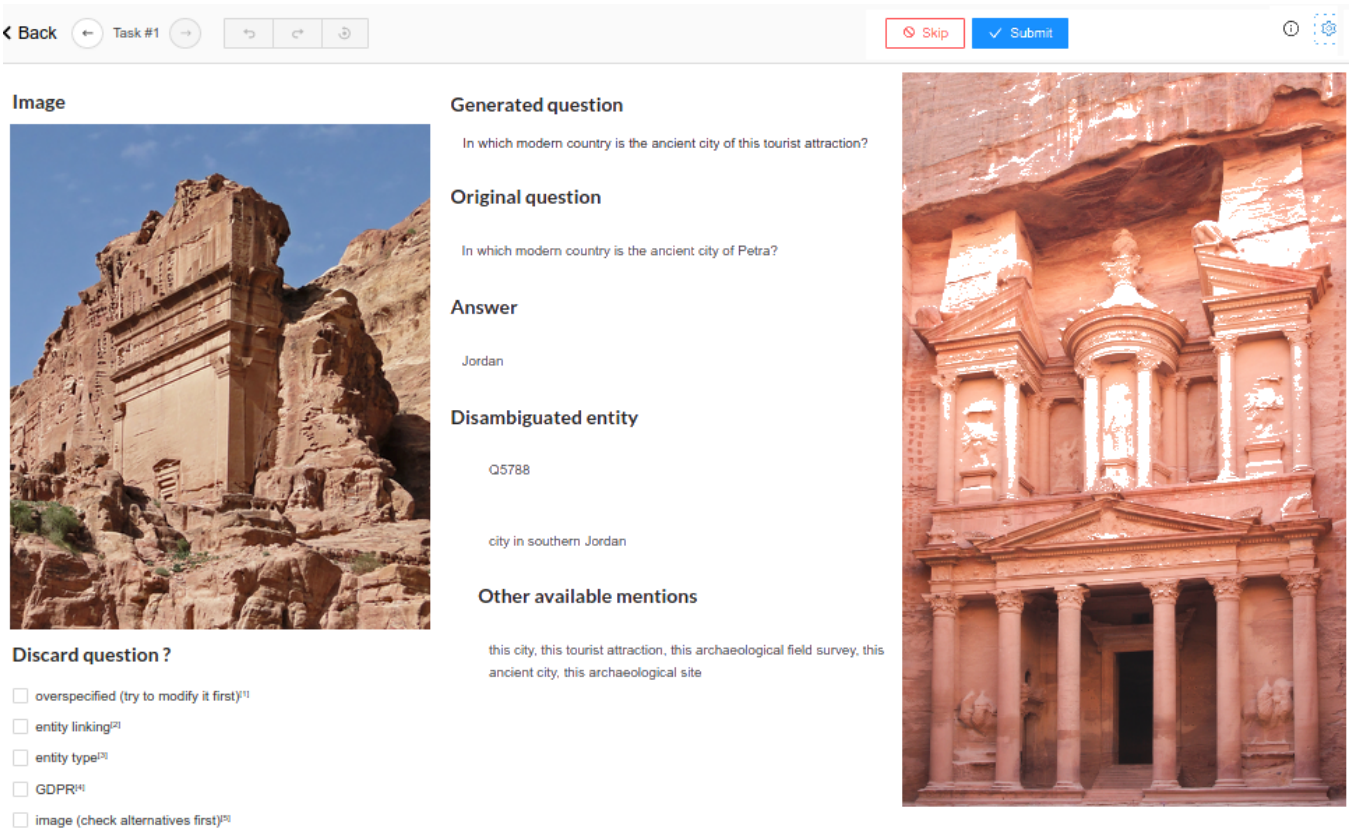


Figure 7: User interface to refine the automatic annotation pipeline.