Cross-modal Retrieval for Knowledge-based Visual Question Answering

Paul Lerner, Olivier Ferret, Camille Guinaudeau

Cross-modal Retrieval for Knowledge-based Visual Question Answering*

Paul Lerner[1], Olivier Ferret[2], and Camille Guinaudeau[3]

[1] Sorbonne Université, CNRS, ISIR, 75005, Paris, France
lerner@isir.upmc.fr
[2] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
olivier.ferret@cea.fr
[3] Université Paris-Saclay, CNRS, LISN, 91400, Orsay, France
camille.guinaudeau@lisn.upsaclay.fr

Abstract. Knowledge-based Visual Question Answering about Named Entities is a challenging task that requires retrieving information from a multimodal Knowledge Base. Named entities have diverse visual representations and are therefore difficult to recognize. We argue that cross-modal retrieval may help bridge the semantic gap between an entity and its depictions, and is foremost complementary with mono-modal retrieval. We provide empirical evidence through experiments with a multimodal dual encoder, namely CLIP, on the recent ViQuAE, InfoSeek, and Encyclopedic-VQA datasets. Additionally, we study three different strategies to fine-tune such a model: mono-modal, cross-modal, or joint training. Our method, which combines mono- and cross-modal retrieval, is competitive with billion-parameter models on the three datasets, while being conceptually simpler and computationally cheaper.

Keywords: Visual Question Answering · Multimodal · Cross-modal Retrieval · Named Entities.

1 Introduction

The work we present in this article takes place in the context of Multimodal Information Retrieval, a field at the intersection between Information Retrieval (IR), Computer Vision, and Machine Learning. More precisely, we focus on Knowledge-based Visual Question Answering about named Entities (KVQAE), which has two specificities in regards to multimodal modeling [6,57,4,20]: (i) images represent named entities; (ii) multimodal interactions are complex and may be combined as both questions and retrieved passages are (text, image) pairs. Indeed, KVQAE consists in answering questions about named entities grounded

* We thank the anonymous reviewers for their helpful comments, as well as Antoine Chaffin for fruitful discussions about CLIP and cross-modal retrieval. Paul Lerner did this work during his PhD at LISN. 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.

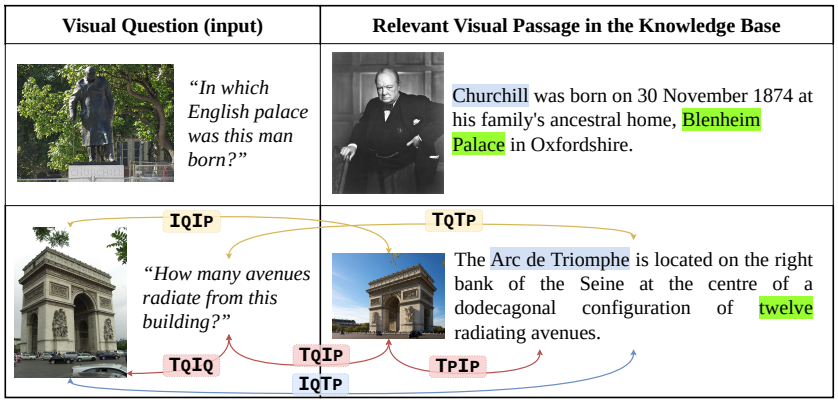

Fig. 1. Two visual questions from the ViQuAE dataset along with relevant visual passages from its Knowledge Base. The different types of mono- and cross-modal interactions studied are also shown for the second question. The acronyms of the interactions are composed of the letters T (Text), I (Image), Q (question) and P (passage).

in a visual context [43,31]. We focus here on Entity Retrieval based on this visual context, similarly to Visual Named Entity Linking [46]. Figure 1 shows two examples of visual questions along with corresponding relevant visual passages from the ViQuAE dataset [31] and its multimodal Knowledge Base (KB), i.e. the set of multimedia documents in which the answers to the questions are searched.

The first example shows how heterogeneous depictions of named entities can be: Winston Churchill is depicted through a *statue* in the visual question and by a *standard photograph* in the KB. This heterogeneity makes mono-modal image retrieval difficult. On the other hand, cross-modal retrieval may bridge the semantic gap between the two representations by using a more abstract representation of the entity, e.g. its name, here *Winston Churchill*.

We formalize these different multimodal interactions in the framework exemplified in Figure 1. The cross-modal interaction between the image of the question and the text of the passage is noted IQTP, while the mono-modal interaction between the two images is noted IQIP. This work is inspired by [30] who studied early multimodal fusion methods, also modeling the TQIQ (resp. TPIP) interaction within the visual question (resp. passage), but found that IQTP was the most important multimodal interaction.

KVQAE differs from standard Visual Question Answering (VQA [3]), which targets the content of the image (e.g., "*What color is the car?*"), and therefore does not require IR. Commonsense VQA [37,42] falls in between standard VQA and KVQAE but (i) focuses on Commonsense Knowledge; (ii) is limited to coarse-grained object categories, e.g., *person* and *building*, instead of *Winston Churchill* and *Arc de Triomphe*, which makes image retrieval straightforward using an object detector [18].

KVQAE was introduced in [43] and received an increased interest recently, with a shift towards unstructured KBs [31,30] and later Large Language Models (LLMs), which do not use an explicit KB but rather generate an answer from the knowledge implicitly stored in their parameters [8,24,38,33]. Given the results of [8,38] and the various caveats of LLMs for factual information generation (hallucinations, lack of generalization and updatability [26,56]), our work adopts a more classical Question Answering architecture, also exploited by [31,30], in which a first IR step is followed by an answer extraction step.

More precisely, we focus on Entity Retrieval and propose to use a multimodal dual encoder [16], namely CLIP [41], for both mono- and cross-modal retrieval, i.e. modeling IqIp and IqTp, respectively. Multimodal dual encoders like CLIP are used as foundation models for a set of diverse tasks such as multimodal analogy [10], Visual Named Entity Linking [46], Cross-modal Question Answering [35], and Commonsense VQA [19]. We show that both mono- and cross-modal retrieval are complementary and can be simply yet effectively combined. We provide empirical evidence through experiments on the ViQuAE, InfoSeek, and Encyclopedic-VQA datasets, being as such the first comparative study of these recently introduced datasets. Furthermore, we study three different strategies to fine-tune such a model, which has been pre-trained in a cross-modal fashion, in this context: mono-modal, cross-modal, or joint training.

2 Related Work

In this section, we present a review of datasets and methods for KVQAE.

Datasets KVQA was the first KVQAE dataset proposed in [43]. Despite its apparent large size, it has several limitations as pointed out by [31]: (i) only one entity type is considered, namely *person*; (ii) it is generated automatically, and thus, has a limited diversity of topics, lexicon, and syntax. Another key difference with the other datasets is that KVQA was designed for structured KBs, in particular Wikidata, from which it was generated, and not an unstructured KB like the following works. To address the limitations of KVQA, ViQuAE was introduced in [31]. It has fewer visual questions but they are manually annotated and it covers a broad range of topics, lexicon, and syntax, as showed in Table 1. Above all, ViQuAE comprises a large number of different entity types, including for example landmarks and organizations in addition to persons. Recently, two other datasets were proposed, aiming at larger size than ViQuAE and with fewer textual bias: InfoSeek [8] and Encyclopedic-VQA (EVQA [38]). InfoSeek is split into two subsets according to the annotation method: manual (ISM) or automatic (ISA). Unfortunately, since neither ISM nor the test set of ISA is available at the time of writing, we can evaluate our model only on the validation set of ISA. As its annotation is automatic, it shares part of the caveats of KVQA but covers more diverse entity types. EVQA alleviates these by using more sophisticated question generation techniques than templates. However, it is sometimes biased towards text, with questions such as "*Which republic celebrated the vendémiaire*

Table 1. Key features of different KVQAE datasets: ViQuAE [31], InfoSeek [8], Encyclopedic-VQA (EVQA [38]), and KVQA [43]. InfoSeek is split into two subsets according to the annotation method: manual (ISM) or automatic (ISA). *Computed on a subset of 500 questions by [8].

	ViQuAE	**ISM**	**ISA**	**EVQA**	**KVQA**
# Visual questions	3,700	8,900	1,356,000	1,036,000	183,000
# Unique questions (text-only)	3,562	2,022	1,498	175,000	8,310
# Unique POS sequences	2,759	1,056	267	91,945	376
# Questions per image	1.1	1.0	1.4	2.0	7.4
Vocabulary (# words)	4,700	1,307	725	40,787	8,400
Average question length (# words)	12.4	7.8	8.9	11.6	10.1
Answer prior	0.3%	–	0.6%	0.4%	15.9%
Answer overlap	25.3%	–	48.1%	59.6%	89.4%
Entity overlap	18.1%	–	20.1%	82.0%	40.6%
# Questions per entity	1.5	11.0	117.6	62.5	9.7
# Entity types	980	527	2,739	–	1
Requires knowledge*	95.2%	95.6%	–	–	–

in the month that the growing season ends for this tree?", a type of overspecified questions that were typically filtered by the manual annotation in ViQuAE [31]. Some key features of these datasets are summarized in Table 1. Question length is expressed in number of words provided by spaCy's English tokenizer. Answer prior is computed as the most likely answer in the training set, independently of the question. All datasets are limited to the English language.

Methods Because the KVQA dataset is limited to *person*-named entities, it was addressed through face recognition in [43]: a Wikidata subgraph is constructed from the recognized entities and processed by a *memory network* to output an answer [50]. A few other studies were carried out on KVQA but the comparison with the rest of the state of the art is made difficult as their systems take the image *caption* as input, making the image *itself* redundant [48,17,22].

Our work is closer to [31], which uses an unstructured KB, a collection of visual passages (as in Figure 1). The authors tackle the task in two steps, where Reading Comprehension follows IR. Their retrieval is a combination of two mono-modal retrievals: textual with DPR [28] and visual with a combination of CLIP, ArcFace [12], and a ResNet model trained on ImageNet [21,11]. We aim at simplifying this system by (i) removing the dependency on ArcFace and ImageNet, two supervised models that provide *a priori* less generic representations than CLIP; (ii) taking full advantage of CLIP by combining mono-modal and cross-modal retrieval. After the IR step, answers are extracted using Multi-passage BERT [49]. This work was then extended in [1], by combining the text retrieval of DPR with Wikidata embeddings, but in doing so, it sets aside multimodal interactions and the image of the visual question. For their part, [30] have, like us, focused on IR. In order to model cross-modal interactions, they jointly represent text and

image using a multimodal Transformer [29,16]. However, this model requires an expensive pre-training and the authors ultimately suggest that it mostly leverages the IqTp interaction. Our conclusions converge because our model outperforms theirs — without additional pre-training — by explicitly modeling IqTp via CLIP, as described in the next section.

Very recently, following the overall trend in our domains, there has been a handful of works aiming to tackle KVQAE with (multimodal) LLMs, directly generating an answer from the visual question, without IR [8,24,38,33]. The same conclusions are reached in [8] and [38]: multimodal LLMs suffer from the same caveats as text-only LLMs and underperform compared to retrieval-augmented models. As a consequence, a sophisticated planning method using a tool-augmented LLM as agent was proposed in [24]. However, [24] and [38] share the same experimental protocol problem: they query the whole Web for image or text retrieval through Google APIs, although the images of the visual questions are public and indexed by Google, which leads to overoptimistic and non-reproducible results. On the contrary, we follow the methodology of [31,30,8], using a controlled, publicly available KB. As for [33], they tackle KVQAE with a multimodal LLM, which is only able to generate long explanatory answers. Therefore, [33] evaluate it on ViQuAE using ROUGE-L [34], after paraphrasing the ground-truth answers with ChatGPT. For that reason, their results are unfortunately not comparable with the rest of the state of the art.

3 Entity Retrieval from Visual Context

3.1 Method

Before being able to extract the answer to the question from a visual passage, or even retrieve such a passage, we focus here on Entity Retrieval, given the image of the question $\mathbf{i_q}$ and a collection of entities $(\mathbf{t_p}, \mathbf{i_p})$, where $\mathbf{t_p}$ denotes the name of the entity and $\mathbf{i_p}$ its reference image. To do so, we define the following similarity function, which combines mono- and cross-modal similarities:

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

where the parameters $\alpha_{\{I,C\}}$ weigh each similarity. We focus on CLIP, a multimodal dual encoder, to implement $s_I(\mathbf{i_q}, \mathbf{i_p})$ and $s_C(\mathbf{i_q}, \mathbf{t_p})$, which models the IqIp and IqTp interactions, respectively (see Figure 1). The objective is thus to bring the image of the question closer to the image of this entity in the KB (*mono-modal training*), or to its name (*cross-modal training*), or both jointly.

More formally, the objective underlying our IR model is to maximize $s(\mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p})$ if the two images $\mathbf{i_q}$ and $\mathbf{i_p}^{(+)}$ depict the same entity, named with the textual form $\mathbf{t_p}^{(+)}$, and to minimize it otherwise. In such a contrastive approach, the other entities of a batch, for which the textual and visual representations are respectively noted $\mathbf{t_p}^{(j)}$ and $\mathbf{i_p}^{(j)}$, are used as negatives. To implement this approach, we jointly train $s_I(\mathbf{i_q}, \mathbf{i_p})$ and $s_C(\mathbf{i_q}, \mathbf{t_p})$ for each $\mathbf{i_q}$ image of the batch

by minimizing the following objective, given the temperature τ:

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

Since we implement $s_C(\mathbf{i_q}, \mathbf{t_p})$ with CLIP, we have:

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

If $\alpha_I = 0$ and $\alpha_C = 1$ (cross-modal training only), the objective is equivalent to the one used during the pre-training of CLIP, except that it is asymmetric (the softmax function expresses the probabilities according to $\mathbf{i_q}$ and not according to $\mathbf{t_p}$). Since $\mathbf{i_q}$, $\mathbf{t_p}$, and $\mathbf{i_p}$ are encoded independently, this objective leverages all the other images and texts of the batch in a highly efficient way (we only need a matrix product to compute the denominator of Equation 2). We implement $s_I(\mathbf{i_q}, \mathbf{i_p})$ in a similar way: $s_I(\mathbf{i_q}\mathbf{i_p}) = \cos\left(\mathrm{CLIP}_V(\mathbf{i_q}), \mathrm{CLIP}_V(\mathbf{i_p})\right)$. The same method could be applied to any multimodal dual encoder [16].

3.2 Data

As mentioned in the introduction, our evaluations are performed on the ViQuAE, ISA, and EVQA datasets. For ViQuAE and ISA, we use the KB proposed in [31], which consists of 1.5 million Wikipedia articles and images of corresponding Wikidata entities. Unfortunately, the KB proposed by [8] has yet to be made available; so our results on ISA will not be directly comparable to theirs. Indeed, 11.5% of ISA entities are missing from our KB, which filters down the training set by 28%. On the contrary, only a few entities from ViQuAE are missing from the KB. For EVQA, we use the corresponding KB of [38], which consists of 2 million Wikipedia articles and corresponding images in WIT [45].

ViQuAE contains 3,700 visual questions about 2,400 different entities, randomly divided into equal-sized sets for training, validation, and testing, with no overlap between images. As a result, the overlap between entities in the training and test sets is quite small, only 18%. Likewise, the entity overlap in ISA is of 20%. Our models must therefore learn to generalize not only to new images but also to new entities. On the contrary, the entity overlap of EVQA is of 82%.

3.3 Hyperparameters

We use the `ViT-B/32` version of CLIP unless otherwise mentioned. To take full advantage of the entities associated with the other images in the batch $\mathbf{t_p}^{(j)}$ and $\mathbf{i_p}^{(j)}$, we use a batch of the largest possible size, here 3,072 $(\mathbf{i_q}, \mathbf{t_p}^{(+)}, \mathbf{i_p}^{(+)})$ triples, i.e., more than the whole training set of ViQuAE. We use a single NVIDIA V100 GPU with 32 GB of RAM. The large batch size is partly enabled by gradient checkpointing.

Because the training set of ViQuAE is so small, training is very cheap: our best model converges, i.e., starts to overfit, after 11 steps/epochs, in less than

15 minutes, which is negligible compared to the pre-training of 8,000 steps in three days of [30] with the same hardware ([30] reports a carbon footprint of 1.7 kgCO2e for 3 days of GPU power consumption). On the larger ISA and EVQA datasets, our models converge roughly after 500 steps in 5 hours.

We use a very small learning rate of 2×10^{-6}, increasing linearly for 4 steps and then decreasing for 50 steps on ViQuAE (or 1,000 on ISA and EVQA) if training is not interrupted before. We use the AdamW optimizer [36], with a weight decay of 0.1. For joint training, we initialize $\alpha_I = \alpha_C = 0.5$ and assign them a learning rate of 0.02, much larger than the rest of the model. Like [41], the temperature τ remains trainable but, given the small learning rate, it remains close to its initial value, 4.6.[4] These hyperparameters were set manually through experiments on the validation set of ViQuAE.

Early stopping is done according to the *in-batch* mean reciprocal rank on the validation set, i.e., by reranking the images or texts of the batch according to the similarity score s, to avoid computing the representations of the whole KB at each epoch.

Our implementation is based on Lightning,[5] PyTorch [39], and Transformers [53] for training, and Datasets [32], Faiss [27], and Ranx [5] for IR, based on the codebase of [31]. Our code is freely available at `https://github.com/PaulLerner/ViQuAE` to ensure the reproducibility of our results.

3.4 Results

We evaluate Entity Retrieval according to the relevance of the Wikipedia article associated with the target entity, which is determined automatically according to the presence of the answer after standard preprocessing (lowercasing, stripping articles, and punctuation). Additionally, because ISA contains a large portion of numerical answers, we follow the same soft matching method as [8] for ISA (years can be off by one and there is a 10% tolerance for measures and various numerical answers). We focus on the single-hop questions of EVQA, following [38]. The metrics used are Precision at 1 (P@1) and Mean Reciprocal Rank (MRR).[6]

We first explore in Table 2 three training strategies and three ways of using a multimodal dual encoder through experiments conducted on the validation set. These three strategies can be defined from Equation 1:

- Mono-modal (image-image) retrieval/training, i.e., $\alpha_I = 1$, $\alpha_C = 0$;
- Cross-modal (image-text) retrieval/training, i.e., $\alpha_I = 0, \alpha_C = 1$;
- Hybrid retrieval or joint training, i.e., $\alpha_I > 0, \alpha_C > 0$.

For hybrid retrieval, the weights $\alpha_{\{I,C\}}$ are set through a grid search over the validation set to maximize the mean reciprocal rank while constraining their

[4] We kept the formulation of [41] but the temperature is usually expressed as $\frac{1}{\tau'}$ and not e^{τ}, which would be equivalent to $\tau' = \frac{1}{100}$ here.

[5] `https://www.pytorchlightning.ai/`

[6] The results are consistent with precision and recall at higher cutoffs, which we omit for the sake of space.

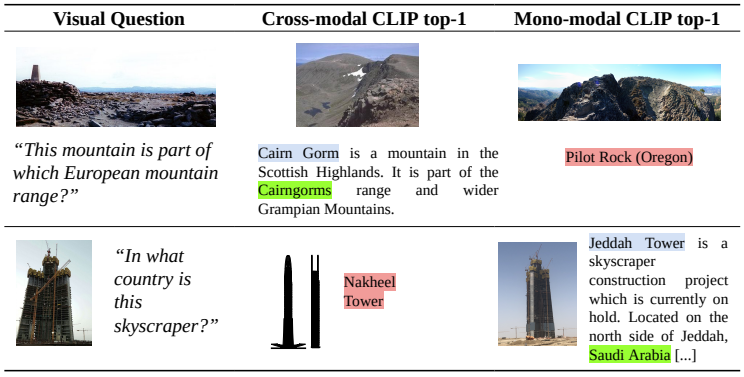

Fig. 2. Strengths and weaknesses of mono- and cross-modal retrieval exemplified through CLIP results (not fine-tuned) on ViQuAE's validation set.

sum to 1, so as to fairly compare joint training with mono- and cross-modal training. Note that the retrieval is independent from the training strategy, as shown in Table 2. Recall that CLIP's pre-training is only cross-modal [41], as most multimodal dual encoders [16].

Mono- or Cross-modal Retrieval? Before comparing the different training methods, we can first notice that cross-modal IR outperforms[7] mono-modal IR on both ViQuAE and ISA,[8] especially without fine-tuning (first lines of each block in Table 2), which may seem curious since proper nouns are not *a priori* very meaningful. Therefore, it is surprising that CLIP generalizes[9] to new entity names. Nevertheless, some names carry meaning. For example, a name can indicate the gender of a person or suggest their nationality.[10] Moreover, we are working here with titles of Wikipedia articles, which are also likely to contain the nature of the entity (e.g., the profession of a person or the type of a monument). These features can thus be mapped to visual attributes.

Foremost, we mainly attribute the success of cross-modal IR to its adequacy with the pre-training of CLIP: the representation space of CLIP is organized to bring together similar texts and images, which the mono-modal proximity of images is only an indirect consequence of. We show examples of successes and failures in Figure 2. In line with the results of [30], we observe that mono-modal retrieval may be more sensitive to superficial image details (color vs. black-and-

[7] Significantly according to Fisher's randomization test [15,44] with $p \leq 0.01$.

[8] An exception is EVQA, for which mono-modal retrieval outperforms cross-modal retrieval. This is surprising as both EVQA and ISA stem from the iNaturalist [47] and Google Landmarks [51] datasets. Further investigations are required.

[9] Unless its pre-training dataset contains enough entities from ViQuAE and ISA so that it circumvents generalization. We develop this discussion in Section 5.

[10] An interactive visualization is provided at `https://paullerner.github.io/ViQuAE/#text-embedding-cross-modal`.

Table 2. Entity Retrieval with a multimodal dual encoder, CLIP, on the validation subsets of ViQuAE, InfoSeek-Automatic (ISA), and EVQA (single-hop). Mono- and cross-modal retrieval model the IqIp and IqTp interactions, respectively. The best results are marked in bold for each type of retrieval. Hybrid retrieval of disjoint training combines *mono-modal trained* mono-modal retrieval and *cross-modal trained* cross-modal retrieval.

Retrieval	Training	ViQuAE		ISA		EVQA	
		MRR	P@1	MRR	P@1	MRR	P@1
Mono-modal	–	29.4	21.8	28.3	18.1	26.1	15.2
	Mono-modal	30.0	21.8	**31.4**	**20.5**	**32.6**	**21.7**
	Cross-modal	29.8	21.4	29.0	18.3	29.7	18.9
	Joint	**30.4**	**22.0**	30.5	19.9	30.7	19.6
Cross-modal	–	32.7	23.1	32.8	22.4	20.9	12.2
	Mono-modal	31.6	21.9	33.0	22.0	20.5	12.0
	Cross-modal	**37.1**	**26.9**	**34.7**	**23.8**	**23.2**	**13.8**
	Joint	30.8	21.3	31.1	20.3	22.4	12.9
Hybrid	–	39.6	30.6	36.2	25.8	28.7	18.7
	Mono-modal	40.1	31.8	38.2	27.4	33.8	22.9
	Cross-modal	**44.1**	**34.9**	38.5	27.8	33.8	23.3
	Joint	41.0	32.6	37.6	26.9	34.0	23.3
	Disjoint	43.7	34.5	**40.0**	**29.6**	**37.4**	**27.8**

white photography, subject pose...). Here, the two photographs at the top of two mountains, showing the horizon, are judged to be similar even though they are different mountains. In contrast, the mono-modal retrieval is more effective in the second example, where the two photographs of the Jeddah Tower are taken from similar vantage points. These qualitative results support our hypothesis that cross-modal retrieval might help addressing the heterogeneity of visual representations of named entities.

Why choose? We show that mono- and cross-modal retrievals are complementary: their results can be simply combined at the score level (as in Equation 1). Thus, without fine-tuning (first lines of each block in Table 2), fusing the two retrievals brings a relative improvement of 32% in P@1 for ViQuAE (and 15% for ISA, 23% for EVQA) compared to the best single retrieval (significant with Fisher's $p \leq 0.01$). It would be interesting to study whether these results generalize to other tasks. For example, this method could benefit Content-based Image Retrieval in a Web browsing context. Overall, hybrid retrieval gives the best performance, on all three datasets.

How to fine-tune multimodal dual encoders? We see that fine-tuning with a given strategy (e.g. mono-modal) always enhances the performance of retrieval with the same strategy. However, it also sometimes decreases retrieval with

another strategy (e.g. cross-modal). Therefore, we find it best to combine models trained disjointly: a *mono-modal trained* mono-modal retrieval and a *cross-modal trained* cross-modal retrieval.

4 Retrieving Passages and Extracting Answers

4.1 Methods

While we have focused on Entity Retrieval through cross-modal retrieval, we are ultimately interested in answering questions about these entities. To do so, we follow the same framework as [31], where Entity Retrieval results are mapped to the corresponding passages to enable fusion with a text passage retrieval method, such as DPR.

This implies redefining s as follows:

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

where $s_T(\mathbf{t_q}, \mathbf{t_p})$ models the TQTP interaction between the text of the question and of the passage and is implemented with DPR. We note this model DPR_{V+T} as it combines DPR, CLIP_V, and CLIP_T, or $\text{DPR}_{\mathbf{V+T}}$ (in bold font) when CLIP is fine-tuned.[11] The weights $\alpha_{\{T,I,C\}}$ are set through a grid search on the validation set like in the previous section (see Figure 3 for an illustration of the impact of these hyperparameters on MRR). DPR is a dual encoder model that combines two BERT encoders, one for the question and one for the passage [28].

Answers are then extracted from these passages using Multi-passage BERT [49], which also models the TQTP interaction.

4.2 Data and implementation

The 1.5 (resp. 2) million articles of the KB of ViQuAE [31] (resp. EVQA [38]) are divided into 12 (resp. 27) million 100-word passages, while preserving sentence boundaries, as in [31].

Both DPR and Multi-passage BERT are pre-trained on TriviaQA, filtered out of all questions used in [31] to generate ViQuAE,[12] before being fine-tuned on the downstream KVQAE dataset, following [31]. Both models are built upon the uncased version of BERT-base [13]. We refer the reader to [31] for further implementation details.

4.3 Baselines

We compare our approach to the DPR_{V+R+A} model of [31], which combines DPR, CLIP_V, ArcFace, and an ImageNet-trained ResNet model. The results of the four models are combined in the same way as in Equation 4, where DPR

[11] DPR is always fine-tuned as described in the next section.

[12] `https://huggingface.co/datasets/PaulLerner/triviaqa_for_viquae`

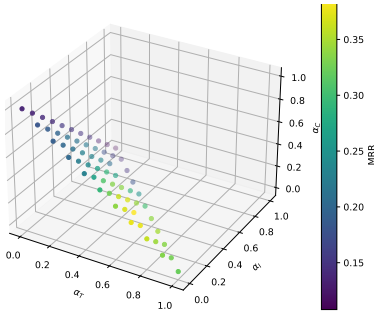

Fig. 3. Passage-level MRR on the validation set of ViQuAE depending on the $\alpha_{\{T,I,C\}}$ hyperparameters.

implements $s_T(\mathbf{t_q}, \mathbf{t_p})$, CLIP_V, ArcFace, and ImageNet compose $s_I(\mathbf{i_q}, \mathbf{i_p})$, and there is no cross-modal similarity, i.e., $s_C(\mathbf{i_q}, \mathbf{t_p}) = 0$.

We also compare our methods to the ECA_V and ILF_V models of [30]. ECA (Early Cross-Attention) early-fuses modalities through an attention mechanism. The similarity is computed as $s(\mathbf{t_q}, \mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = \text{ECA}(\mathbf{t_q}, \mathbf{i_q}) \cdot \text{ECA}(\mathbf{t_p}, \mathbf{i_p})$ and thus combines all the multimodal interactions shown in Figure 1. ILF (Intermediate Linear Fusion) fuses modalities with a simple linear projection and thus has, like our method, neither TQIQ nor TPIP interactions since the similarity can be reduced to:

$$s(\mathbf{t_q}, \mathbf{i_q}, \mathbf{t_p}, \mathbf{i_p}) = s_T(\mathbf{t_q}, \mathbf{t_p}) + s_{C'}(\mathbf{t_q}, \mathbf{i_p}) + s_I(\mathbf{i_q}, \mathbf{i_p}) + s_C(\mathbf{i_q}, \mathbf{t_p}) \quad (5)$$

Note that [31,30] use CLIP_V with the ResNet architecture while we use ViT [14] in most of our experiments (but compare the two in the next section and find no significant difference).

Moving away from the Retrieval+Extraction framework of [31], we compare our results to [8,38], who both use the PaLM LLM [9], either as is or augmented with the image caption and in-context learning examples (denoted PromptCap [23]). [8] also experiment with FiD [25], augmented with CLIP retrieval results.

4.4 Results

Metrics Extracted answers are evaluated using Exact Match (EM) and token-level F1 score on ViQuAE following [31], using the soft matching score defined by [8] on ISA (see Section 3.4), and using both F1 and BEM [7] on EVQA. The results for these three benchmarks are reported in Table 3.

Hybrid retrieval effectiveness When comparing the DPR_V and $\text{DPR}_{\mathbf{V+T}}$ models, we see that the effectiveness of combining mono- and cross-modal retrieval observed earlier indeed translates to more accurate answers, on all three datasets. Therefore, our model also outperforms the previously proposed models of [31,30]

Table 3. Reading Comprehension results on the test set of ViQuAE, the validation set of ISA, and the test single-hop questions of EVQA. As in [28], the reader takes as input the top-24 of different IR systems listed in the "Method" column (except for the methods of [8,38]). The results of [8], in gray, are provided as reference but use a different, yet unavailable, smaller KB, which perfectly covers ISA. *CLIP is based on ViT's architecture instead of ResNet. †Our re-implementation of the reader, which fixes the loss function.

Method	**# Param. (M)**	**ViQuAE**		**ISA**	**EVQA**	
		EM	**F1**	**Soft Match**	**BEM**	**F1**
PaLM few-shot (text-only) [8]	540,000	31.5	–	4.8	–	–
CLIP + FiD [8]	1,170	–	–	20.9	–	–
PaLM zero-shot (text-only) [38]	540,000	–	–	–	19.7	–
PromptCap + PaLM [38]	540,870	–	–	–	29.7	–
DPR (text-only) [31]	330	16.9	20.1	–	–	–
DPR_V [30]	432	19.0	22.3	–	–	–
DPR_V* (baseline)	481	19.7	23.3	–	–	–
DPR_V*† (baseline)	481	26.4	29.1	7.7	27.4	25.4
DPR_{V+R+A} [31]	500	22.1	25.4	–	–	–
ECA_V [30]	432	20.6	24.4	–	–	–
ILF_V [30]	432	21.3	25.4	–	–	–
$\text{DPR}_{\mathbf{V+T}}$* (this work)	481	24.7	28.7	–	–	–
$\text{DPR}_{\mathbf{V+T}}$*† (this work)	481	**30.9**	**34.3**	**12.4**	**29.1**	**26.6**
Oracle retrieval + FiD [8]	Oracle + 770	–	–	52.5	–	–
Oracle retrieval† (this work)	Oracle + 110	68.3	72.7	46.8	65.3	59.7

on ViQuAE, while being conceptually simpler and computationally cheaper (emitting hundred times less CO2 than [30]). Furthermore, we found a bug in the implementation of the reader's loss provided by [31]. Fixing it consistently improved results, for both DPR_V and $\text{DPR}_{\mathbf{V+T}}$. Our model is also competitive with the method of [38] on EVQA, while using 1,000 times less parameters.[13]

Knowledge base incompleteness The results of [8] are provided as reference but are hardly comparable to the others. Apart from PaLM being three order of magnitude greater than the other models and partly trained on ViQuAE's test set,[14] they use a different KB. This KB, yet unavailable, is fifteen times smaller than ours, so contains less distractors, and covers 100% of the entities and questions of ISA. In contrast, our KB lacks 11.5% of ISA entities and is not guaranteed to contain the answers for the 88.5% remaining, because of differences between the Wikipedia versions.

[13] We focus on the single-hop subset of EVQA following [38]. On the two-hop questions, the model using $\text{DPR}_{\mathbf{V+T}}$ achieves 31.1 BEM and 25.6 F1, and 9.8 BEM/3.8 F1 on the multi-answer questions.

[14] According to [9], around 20% of TriviaQA is contained in PaLM's pre-training dataset. ViQuAE was derived from TriviaQA [31].

Oracle retrieval We conduct additional experiments in an "oracle retrieval" setting, where the reader only takes relevant passages as input, similarly to the oracle experiments of [31,8]. In agreement with their results, we find a large gap between our best retrieval model and the oracle, showing that IR is still the main bottleneck of KVQAE. Compared to the FiD model of [8], with 770M parameters, we approach its performance, although Multi-passage BERT is seven times smaller and our KB does not fully cover ISA.

5 Conclusion

This paper studies cross-modal retrieval and its combination with mono-modal retrieval for Knowledge-based Visual Question Answering about named Entities (KVQAE). Retrieval is carried out with a multimodal dual encoder, namely CLIP. Our results demonstrate the superiority of cross-modal retrieval over mono-modal retrieval, but also the complementarity of the two, which can be easily combined.

We argue that cross-modal retrieval may help addressing the heterogeneity of visual representations of named entities, consistently with prior work. It would be interesting to study whether these results generalize to other tasks. For example, this method could benefit Content-based Image Retrieval, in a Web browsing context.

Although it was the abundance of cross-modal data that enabled CLIP's training in the first place, which would have been difficult with a mono-modal annotation, this limits our results because it is difficult to control such a large amount of data and thus to estimate CLIP's generalization capabilities. We hypothesize that mono-modal retrieval is better suited to generalize to new entities.

We show that the effectiveness of cross-modal retrieval leads to more accurate answers, on all three studied datasets. Therefore, our method outperforms our baseline (mono-modal retrieval) but also the methods of [31,30], while being conceptually simpler and computationally cheaper. Furthermore, it is competitive with billion-scale parameters models on ISA and EVQA. As such, this is the first comparative study of the recently introduced ViQuAE, ISA, and EVQA datasets. We find that ISA is more challenging as it is less biased towards text, but advocate for further studies on all three datasets — which all have their pros and cons — with diverse methods.

Consistently with [31,8], we find a large gap between our best retrieval model and oracle retrieval, showing that entity retrieval is the main bottleneck of KVQAE. For future work, we plan to combine our unstructured KB with a structured one, such as Wikidata, to enable the modeling of links between the entities [54,40,52,2], which would further address the heterogeneity of their visual representations. A more IR perspective on the matter could cast KVQAE as a query expansion problem, with an initial ambiguous textual query which would benefit from pseudo-relevant feedback [55].

References

1. Adjali, O., Grimal, P., Ferret, O., Ghannay, S., Le Borgne, H.: Explicit knowledge integration for knowledge-aware visual question answering about named entities. In: Proceedings of the 2023 ACM International Conference on Multimedia Retrieval. p. 29–38. ICMR '23, Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3591106.3592227, https://doi.org/10.1145/3591106.3592227
2. Alberts, H., Huang, N., Deshpande, Y., Liu, Y., Cho, K., Vania, C., Calixto, I.: VisualSem: a high-quality knowledge graph for vision and language. In: Proceedings of the 1st Workshop on Multilingual Representation Learning. pp. 138–152. Association for Computational Linguistics, Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.mrl-1.13, https://aclanthology.org/2021.mrl-1.13
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: 2015 IEEE International Conference on Computer Vision (ICCV). pp. 2425–2433. IEEE, Santiago, Chile (Dec 2015). https://doi.org/10.1109/ICCV.2015.279, http://ieeexplore.ieee.org/document/7410636/
4. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (Feb 2019). https://doi.org/10.1109/TPAMI.2018.2798607, conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence
5. Bassani, E.: ranx: A Blazing-Fast Python Library for Ranking Evaluation and Comparison. In: Hagen, M., Verberne, S., Macdonald, C., Seifert, C., Balog, K., Nørvåg, K., Setty, V. (eds.) Advances in Information Retrieval. pp. 259–264. Lecture Notes in Computer Science, Springer International Publishing, Cham (2022). https://doi.org/10.1007/978-3-030-99739-7_30
6. Bokhari, M.U., Hasan, F.: Multimodal information retrieval: Challenges and future trends. International Journal of Computer Applications **74**(14) (2013), publisher: Foundation of Computer Science
7. Bulian, J., Buck, C., Gajewski, W., Börschinger, B., Schuster, T.: Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 291–305. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.emnlp-main.20, https://aclanthology.org/2022.emnlp-main.20
8. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can Pre-trained Vision and Language Models Answer Visual Information-Seeking Questions? (Feb 2023). https://doi.org/10.48550/arXiv.2302.11713, http://arxiv.org/abs/2302.11713, arXiv:2302.11713 [cs]
9. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
10. Couairon, G., Douze, M., Cord, M., Schwenk, H.: Embedding arithmetic of multimodal queries for image retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4950–4958 (June 2022)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and

Pattern Recognition. pp. 248–255 (Jun 2009). https://doi.org/10.1109/CVPR.2009.5206848, iSSN: 1063-6919
12. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019), https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://aclanthology.org/N19-1423
14. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
15. Fisher, R.A.: The design of experiments. The design of experiments. (2nd Ed) (1937), https://www.cabdirect.org/cabdirect/abstract/19371601600, publisher: Oliver & Boyd, Edinburgh & London.
16. Gan, Z., Li, L., Li, C., Wang, L., Liu, Z., Gao, J.: Vision-language pre-training: Basics, recent advances, and future trends. Found. Trends. Comput. Graph. Vis. **14**(3–4), 163–352 (dec 2022). https://doi.org/10.1561/0600000105, https://doi.org/10.1561/0600000105
17. Garcia-Olano, D., Onoe, Y., Ghosh, J.: Improving and diagnosing knowledge-based visual question answering via entity enhanced knowledge injection. In: Companion Proceedings of the Web Conference 2022. p. 705–715. WWW '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3487553.3524648, https://doi.org/10.1145/3487553.3524648
18. Gardères, F., Ziaeefard, M.: ConceptBert: Concept-Aware Representation for Visual Question Answering. Findings of the Association for Computational Linguistics: EMNLP 2020 p. 10 (2020), https://aclanthology.org/2020.findings-emnlp.44/
19. Gui, L., Wang, B., Huang, Q., Hauptmann, A., Bisk, Y., Gao, J.: KAT: A Knowledge Augmented Transformer for Vision-and-Language. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 956–968. Association for Computational Linguistics, Seattle, United States (Jul 2022), https://aclanthology.org/2022.naacl-main.70
20. Guo, W., Wang, J., Wang, S.: Deep Multimodal Representation Learning: A Survey. IEEE Access **7**, 63373–63394 (2019). https://doi.org/10.1109/ACCESS.2019.2916887, conference Name: IEEE Access
21. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016), https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf
22. Heo, Y.J., Kim, E.S., Choi, W.S., Zhang, B.T.: Hypergraph Transformer: Weakly-supervised multi-hop reasoning for knowledge-based visual question answering.

In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 373–390. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.acl-long.29, https://aclanthology.org/2022.acl-long.29
23. Hu, Y., Hua, H., Yang, Z., Shi, W., Smith, N.A., Luo, J.: Promptcap: Prompt-guided task-aware image captioning (2023)
24. Hu, Z., Iscen, A., Sun, C., Chang, K.W., Sun, Y., Ross, D.A., Schmid, C., Fathi, A.: AVIS: Autonomous Visual Information Seeking with Large Language Models (Jun 2023), http://arxiv.org/abs/2306.08129, arXiv:2306.08129 [cs]
25. Izacard, G., Grave, E.: Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 874–880. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.74, https://aclanthology.org/2021.eacl-main.74
26. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of Hallucination in Natural Language Generation. ACM Computing Surveys **55**(12), 248:1–248:38 (Mar 2023). https://doi.org/10.1145/3571730, https://dl.acm.org/doi/10.1145/3571730
27. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019). https://doi.org/10.1109/TBDATA.2019.2921572
28. 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. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6769–6781. Association for Computational Linguistics, Online (Nov 2020), https://www.aclweb.org/anthology/2020.emnlp-main.550
29. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM Comput. Surv. **54**(10s) (sep 2022). https://doi.org/10.1145/3505244, https://doi.org/10.1145/3505244
30. Lerner, P., Ferret, O., Guinaudeau, C.: Multimodal inverse cloze task for knowledge-based visual question answering. In: Advances in Information Retrieval (ECIR 2023). pp. 569–587. Springer Nature Switzerland, Cham (2023). https://doi.org/10.1007/978-3-031-28244-7_36
31. 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. In: 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 (2022). https://doi.org/10.1145/3477495.3531753, https://hal.archives-ouvertes.fr/hal-03650618
32. Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., Wolf, T.: Datasets: A Community Library for Natural Language Processing. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 175–184. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021), https://aclanthology.org/2021.emnlp-demo.21

33. Li, L., Yin, Y., Li, S., Chen, L., Wang, P., Ren, S., Li, M., Yang, Y., Xu, J., Sun, X., Kong, L., Liu, Q.: M^{3}IT: A Large-Scale Dataset towards Multi-Modal Multilingual Instruction Tuning (Jun 2023). https://doi.org/10.48550/arXiv.2306.04387, http://arxiv.org/abs/2306.04387, arXiv:2306.04387 [cs]
34. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
35. Liu, Z., Xiong, C., Lv, Y., Liu, Z., Yu, G.: Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PQOlkgsBsik
36. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=Bkg6RiCqY7
37. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: OK-VQA: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3195–3204 (2019), https://ieeexplore.ieee.org/document/8953725/
38. Mensink, T., Uijlings, J., Castrejon, L., Goel, A., Cadar, F., Zhou, H., Sha, F., Araujo, A., Ferrari, V.: Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3113–3124 (October 2023)
39. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: 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
40. Pezeshkpour, P., Chen, L., Singh, S.: Embedding Multimodal Relational Data for Knowledge Base Completion. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 3208–3218 (2018)
41. 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. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
42. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European Conference on Computer Vision. pp. 146–162. Springer (2022)
43. Shah, S., Mishra, A., Yadati, N., Talukdar, P.P.: KVQA: Knowledge-Aware Visual Question Answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8876–8884 (2019), https://144.208.67.177/ojs/index.php/AAAI/article/view/4915
44. Smucker, M.D., Allan, J., Carterette, B.: A comparison of statistical significance tests for information retrieval evaluation. In: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management. pp. 623–632. CIKM '07, Association for Computing Machinery, New York, NY, USA (Nov 2007). https://doi.org/10.1145/1321440.1321528, https://doi.org/10.1145/1321440.1321528
45. Srinivasan, K., Raman, K., Chen, J., Bendersky, M., Najork, M.: Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development

in Information Retrieval. p. 2443–2449. SIGIR '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3404835.3463257, https://doi.org/10.1145/3404835.3463257
46. Sun, W., Fan, Y., Guo, J., Zhang, R., Cheng, X.: Visual named entity linking: A new dataset and a baseline. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 2403–2415. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022). https://doi.org/10.18653/v1/2022.findings-emnlp.178, https://aclanthology.org/2022.findings-emnlp.178
47. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The iNaturalist Species Classification and Detection Dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8769–8778 (2018), https://openaccess.thecvf.com/content_cvpr_2018/html/Van_Horn_The_INaturalist_Species_CVPR_2018_paper.html
48. Vickers, P., Aletras, N., Monti, E., Barrault, L.: In Factuality: Efficient Integration of Relevant Facts for Visual Question Answering. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 468–475. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-short.60, https://aclanthology.org/2021.acl-short.60
49. Wang, Z., Ng, P., Ma, X., Nallapati, R., Xiang, B.: 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). pp. 5878–5882. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1599, https://www.aclweb.org/anthology/D19-1599
50. Weston, J., Chopra, S., Bordes, A.: Memory networks (2014). https://doi.org/10.48550/ARXIV.1410.3916, https://arxiv.org/abs/1410.3916
51. Weyand, T., Araujo, A., Cao, B., Sim, J.: Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2575–2584 (2020), https://openaccess.thecvf.com/content_CVPR_2020/html/Weyand_Google_Landmarks_Dataset_v2_-_A_Large-Scale_Benchmark_for_Instance-Level_CVPR_2020_paper.html
52. Wilcke, W.X., Bloem, P., de Boer, V., Veer, R.H.v.t., van Harmelen, F.A.H.: End-to-End Entity Classification on Multimodal Knowledge Graphs. arXiv:2003.12383 [cs] (Mar 2020), http://arxiv.org/abs/2003.12383, arXiv: 2003.12383
53. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs] (Jul 2020), http://arxiv.org/abs/1910.03771
54. Xie, R., Liu, Z., Luan, H., Sun, M.: Image-embodied knowledge representation learning. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. pp. 3140–3146. IJCAI'17, AAAI Press, Melbourne, Australia (Aug 2017)
55. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: Proceedings of the 19th Annual International ACM SIGIR Conference on Research

and Development in Information Retrieval. p. 4–11. SIGIR '96, Association for Computing Machinery, New York, NY, USA (1996). https://doi.org/10.1145/243199.243202, https://doi.org/10.1145/243199.243202

56. Zamani, H., Diaz, F., Dehghani, M., Metzler, D., Bendersky, M.: Retrieval-Enhanced Machine Learning. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 2875–2886. SIGIR '22, Association for Computing Machinery, New York, NY, USA (Jul 2022). https://doi.org/10.1145/3477495.3531722, https://doi.org/10.1145/3477495.3531722
57. Zhang, D., Cao, R., Wu, S.: Information fusion in visual question answering: A survey. Information Fusion **52**, 268–280 (2019), https://www.sciencedirect.com/science/article/pii/S1566253518308893