Passage Retrieval for Outside-Knowledge Visual Question Answering

Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, Erik Learned-Miller

University of Massachusetts Amherst
Amherst, MA, United States
{chenqu,zamani,lyang,croft,elm}@cs.umass.edu

ABSTRACT
In this work, we address multi-modal information needs that contain text questions and images by focusing on passage retrieval for outside-knowledge visual question answering. This task requires access to outside knowledge, which in our case we define to be a large unstructured passage collection. We first conduct sparse retrieval with BM25 and study expanding the question with object names and image captions. We verify that visual clues play an important role and captions tend to be more informative than object names in sparse retrieval. We then construct a dual-encoder dense retriever, with the query encoder being LXMERT [35], a multi-modal pre-trained transformer. We further show that dense retrieval significantly outperforms sparse retrieval that uses object expansion. Moreover, dense retrieval matches the performance of sparse retrieval that leverages human-generated captions.

CCS CONCEPTS
• Information systems → Question answering; Multimedia and multimodal retrieval.

KEYWORDS
Dense Retrieval; Multi-Modal; Visual Question Answering

ACM Reference Format:

1 INTRODUCTION
Recent work on Question Answering (QA) [30, 33, 44] mostly focuses on uni-modal information needs, i.e., text- or voice-based questions (voice input can be considered as text after automatic transcription). However, many information needs, such as the one in Fig. 1, would be inconvenient or hard to explain without a picture. This motivates the study of methods that can handle multi-modal information needs containing both text questions and images [7, 22].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR ’21, July 11–15, 2021, Virtual Event, Canada
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8037-9/21/07... $15.00
https://doi.org/10.1145/3404835.3462987

Specifically, we focus on a Visual QA (VQA) task referred to as Outside-Knowledge VQA (OK-VQA). Classic VQA benchmarks [1, 12, 25, 46, 47] and models [3, 10, 17, 23, 26, 41] mainly focus on questions about counting, visual attributes, or other visual detection tasks, whose answers can be found in the given image. In contrast, images in our task help to define the information need, instead of simply being the knowledge source by which the question is answered. OK-VQA resembles open-domain QA [37] in the sense that both tasks require access to an outside and open knowledge resource, e.g., a large collection of passages, to answer the questions. Open-domain QA systems typically follow a retrieve-and-read paradigm [5, 16, 20, 30–32, 43], where the system first retrieves a number of documents (passages) from a collection and then extracts answers from them. This paradigm is less studied for multi-modal information needs, which is the focus of this paper. In this work, we focus on the retrieval phase for OK-VQA as illustrated in Fig. 1.

Unlike previous knowledge-based VQA work that retrieves knowledge from a knowledge base [11, 21, 28, 29, 38–40, 45, 48] or using a Wikipedia Search API [27], we systematically study passage retrieval for OK-VQA with generic information retrieval approaches so that our methods can be applied to a wider range of unstructured knowledge resources. In particular, we seek answers to the following research questions: (RQ1) How helpful are the visual signals in OK-VQA? (RQ2) What is the most effective way to incorporate visual signals into sparse retrieval models that are based on term matching? (RQ3) How well does dense retrieval [13, 16, 20, 24, 42] work with multi-modal information needs?

To answer these important research questions, we study passage retrieval for OK-VQA queries with a large Wikipedia passage collection. We first conduct sparse retrieval with BM25. We investigate the performance of expanding the original question with different human-annotated object names and image captions. We further study the impact of using different rank fusion methods for different expansion types. We verify that visual clues play an important role in our task. In particular, captions tend to be more informative than object names in sparse retrieval. We further reveal that it is desirable to exploit the most salient matching signal (CombMAX [9, 19]).
We are given an information need (query) denoted as $Q_i = (q_i, v_i)$. It consists of a text question $q_i$ and an image $v_i$. The task is to retrieve $k$ passages that can be used to fulfill $Q_i$, from a large passage collection. Following the work on open-domain QA [16, 20, 30], a passage is deemed as relevant if it contains the ground truth answer.

2.2 Sparse Retrieval

The backbone of our sparse retrieval approach is BM25, which works with text queries. Therefore, we expand $q_i$ with different textual descriptions of visual clues to construct the BM25 queries. Visual signals in an image are typically expressed in two forms. The first form is a set of object names $\{o_1^i, o_2^i, \ldots\}$ produced by an object detector. Each object is a Region of Interest (RoI) that reveals a meaningful component of the image. The second form is a set of captions $\{c_1^i, c_2^i, \ldots\}$ produced by an image descriptor to describe the image. We adopt human-annotated object names and captions for sparse retrieval. Although the human annotations do not necessarily give the performance upper bound, they would be strong baselines for sparse retrieval and make sure that our analysis will not be affected by the quality of automatic annotations produced by object detectors and image descriptors. We study different expansion of visual signals as follows:

- **BM25-Orig**: taking the original $q_i$ only, i.e., $Q_i^{\text{orig}} = \{q_i\}$.
- **BM25-Obj** (object expansion): appending each one of the object names to $q_i$, i.e., $Q_i^{\text{obj}} = \{q_i, o_1^i, q_i, o_2^i, \ldots\}$.
- **BM25-Cap** (caption expansion): appending each one of the captions to $q_i$, i.e., $Q_i^{\text{cap}} = \{q_i, c_1^i, q_i, c_2^i, \ldots\}$.
- **BM25-All**: taking the union of the above queries, i.e., $Q_i^{\text{all}} = Q_i^{\text{orig}} \cup Q_i^{\text{obj}} \cup Q_i^{\text{cap}}$.

Since $Q_i^{\text{obj/cap/all}}$ contains multiple BM25 queries for the same information need $Q_i$, we need rank fusion methods to consolidate the ranked lists $R$ generated by queries within each query set. This resembles an ensemble process to combine results obtained with different visual signals. We consider **CombMAX** [9, 19] (taking the maximum score of a passage in different ranked lists), **CombSUM** [9, 19] (taking the sum of scores of a passage in different ranked lists), and **RRF** (Reciprocal Rank Fusion) [6] (the fusion score for a passage $p$ is defined as $\frac{1}{\sum_{r \in R \cap \{\text{pos}\}} r(p)}$, where $r(\cdot)$ is the rank of $p$). CombMAX could help the model be more robust to distracting visual signals while the other two fusion approaches make sure that the impacts of lower-ranked passages do not vanish.

2.3 Dense Retrieval

Following previous work [16, 20, 30, 42], we adopt a dual-encoder architecture to construct a learnable retriever. The retrieval process is “dense” in the sense that the queries and passages are encoded to low-dimensional dense vectors, as opposed to the high-dimensional sparse vectors used in sparse retrieval. As shown in Fig. 2, the retriever consists of a query encoder and a passage encoder.  

2.3.1 Query Encoder. We adopt the LXMERT model [35] as the query encoder since it can encode both the question and image components of $Q_i$. LXMERT is a pre-trained Transformer model [36] designed to learn vision and language connections. It consists of three encoders, an object relationship encoder, a language encoder, and a cross-modality encoder. The first two single-modality encoders function in a similar way to BERT [8] except that the object relationship encoder works with a set of object detections produced by a Faster R-CNN [34] model pre-trained on Visual Genome [2, 18]. Each detection representation can be considered as an “image token embedding” that consists of its RoI features (fixed) and position features (trainable). The cross-modality encoder conducts bi-directional cross attention between vision and language representations. We refer our readers to the LXMERT paper [35] for further details. We project the cross-modality output of LXMERT to an n-dimensional query representation. The dense retriever with a LXMERT query encoder is referred to as Dense-LXMERT. To adopt a deeper analytical view, we also consider BERT as the query encoder, which only works with the question component of the query, resulting in the Dense-BERT model.

2.3.2 Passage Encoder. We use BERT as the passage encoder and project the [CLS] representation to an n-dimensional passage representation. The retrieval score is defined as the dot product of the query and passage representations. After training, we encode all passages in the collection during an offline process. At inference time, we use Faiss [15] for maximum inner product search.

2.3.3 Training. We train the dual-encoder retriever with a set of training instances. Each instance is denoted as $(Q_i, p_+^i, p_-^i)$, where $p_+^i$ is a positive passage that contains the answer while $p_-^i$ is a negative passage that does not contain the answer. In our case,  

---

1https://github.com/predwh/okvqa-release
we select $p_i^+$ from the top passages retrieved by a sparse retrieval method. Thus, $p_i^+$ can be referred to as a retrieved negative. We present more details on constructing the training data in Sec. 3.1.2. In addition to the retrieved negatives, one might want to take advantage of the other passages in the batch as in-batch negatives. Although in-batch negatives resemble randomly sampled negatives that can be less effective [16], it is extremely efficient since passage representations can be reused within the batch. Karpukhin et al. [16] studied combining in-batch negatives with retrieved negatives for uni-modal queries. We further dig into this topic for multi-modal queries. We consider the following negative sampling strategies:

- **R-Neg**: using the retrieved negative passage only.
- **R-Neg+IB-Neg**: using the retrieved negative, along with all other in-batch negative passages of other instances.
- **R-Neg+IB-Pos**: using the retrieved negative, along with all other in-batch positive passages of other instances.
- **R-Neg+IB-All**: using the retrieved negative, along with all other in-batch passages, except for $p_i^+$. The same query can be paired with different positive and negative passages to augment the training data as suggested in Sec. 3.1.2. Therefore, the queries within a batch can coincide even with random batching. In this case, the misuse of a positive passage as negative may hinder the learning process. We empirically examine whether this concern holds by comparing R-Neg+IB-All/Pos to R-Neg+IB-Neg.

Following previous work [16, 30], we use cross entropy loss to maximize the probability of the positive passage given the negatives identified above. We then average the losses for queries in the batch.

### 3 EXPERIMENTS

#### 3.1 Experimental Setup

**Dataset**. Our retrieval dataset is based on the OK-VQA dataset [27], where all questions require outside knowledge. The images in the OK-VQA dataset are from the COCO dataset [12]. We take the original training queries as our training queries and $p_i^+$ can be referred to as a retrieved negative. We present more details on constructing the training data in Sec. 3.1.2.

In addition to the retrieved negatives, one might want to take advantage of the other passages in the batch as in-batch negatives. Although in-batch negatives resemble randomly sampled negatives that can be less effective [16], it is extremely efficient since passage representations can be reused within the batch. Karpukhin et al. [16] studied combining in-batch negatives with retrieved negatives for uni-modal queries. We further dig into this topic for multi-modal queries. We consider the following negative sampling strategies:

- **R-Neg**: using the retrieved negative passage only.
- **R-Neg+IB-Neg**: using the retrieved negative, along with all other in-batch negative passages of other instances.
- **R-Neg+IB-Pos**: using the retrieved negative, along with all other in-batch positive passages of other instances.
- **R-Neg+IB-All**: using the retrieved negative, along with all other in-batch passages, except for $p_i^+$. The same query can be paired with different positive and negative passages to augment the training data as suggested in Sec. 3.1.2. Therefore, the queries within a batch can coincide even with random batching. In this case, the misuse of a positive passage as negative may hinder the learning process. We empirically examine whether this concern holds by comparing R-Neg+IB-All/Pos to R-Neg+IB-Neg.

Following previous work [16, 30], we use cross entropy loss to maximize the probability of the positive passage given the negatives identified above. We then average the losses for queries in the batch.

#### 3.1.3 Evaluation Metrics

We focus on passage retrieval for multi-modal information needs as the first step in the OK-VQA pipeline. The output of the retrieval process will be used by a reader model to extract the answer. Therefore, following previous work [5, 16],

we use precision-oriented metrics to evaluate the performance of our models. Precisely, we use Mean Reciprocal Rank and Precision with the ranking cut-off of 5 (MRR@5 and P@5) as our metrics.

### 3.2 Main Results

#### 3.2.1 Sparse Retrieval

We present the results for sparse retrieval in Tab. 2 to answer RQ1 and RQ2 raised in Sec. 1. First, we observe that approaches that consider visual signals outperform BM25-Orig by a large margin, verifying that visual clues are helpful in our task. We then compare BM25 with different forms of visual clues. Methods with captions (BM25-Cap/All) outperform object expansion, indicating that captions are more informative than object names. This makes sense since captions typically cover important objects descriptively. BM25-All does not benefit from incorporating both objects and captions. On the contrary, objects can be distracting and hurt the performance gain from captions. Finally, we compare different rank fusion methods. When objects are being considered (BM25-Obj/All), CombMAX yields the best performance since it is robust to potentially misleading objects by only considering objects with the best matching score. On the other hand, CombSUM and RRF work well with caption expansion. Their ability to consider the impact of all captions is desirable since captions are closely connected to the image and can be diverse and complementary. The best performing approach is BM25-Cap with CombSUM.

#### 3.2.2 Dense Retrieval

We present the dense retrieval results, along with the best sparse retrieval results in Tab. 3 to answer RQ3. We first compare retrieval without visual clues: we observe that Dense-BERT outperforms BM25-Orig by a large margin, verifying

### Table 1: Data statistics.

<table>
<thead>
<tr>
<th></th>
<th>Split</th>
<th># questions</th>
<th># BM25 queries</th>
<th># training instances</th>
<th># passages in collection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>9,009</td>
<td>81,100</td>
<td>211,200</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Val</td>
<td>2,523</td>
<td>22,352</td>
<td>N/A</td>
<td>34,059</td>
<td>11,000,000</td>
</tr>
<tr>
<td>Test</td>
<td>2,523</td>
<td>22,573</td>
<td>N/A</td>
<td>11,000,000</td>
<td></td>
</tr>
</tbody>
</table>

---

2. https://github.com/castorini/anserini
3. https://github.com/huggingface/transformers
4. https://github.com/castorini/anserini
5. https://github.com/huggingface/transformers
Table 2: Sparse retrieval results. Boldface denotes the best performance within each group and underscores denote the best overall results. ▲ i denotes that the gain with respect to the best method in group i has statistically significant with $p < 0.05$ tested by the Student’s paired t-test.

<table>
<thead>
<tr>
<th>Projection size</th>
<th>Methods</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>BM25-Orig</td>
<td>0.2565</td>
<td>0.1772</td>
</tr>
<tr>
<td>384</td>
<td>BM25-Obj</td>
<td>0.3493</td>
<td>0.2395</td>
</tr>
<tr>
<td>512</td>
<td>BM25-Cap</td>
<td>0.4550</td>
<td>0.3293</td>
</tr>
<tr>
<td>768</td>
<td>BM25-All</td>
<td>0.4704</td>
<td>0.3366</td>
</tr>
</tbody>
</table>

Table 3: Dense retrieval results. Refer to Tab. 2 for notations. ▲ i denotes the statistical significance is obtained with $0.05 < p < 0.1$. Note that BM25-Obj/Cap/All has access to the ground truth object names and captions.

<table>
<thead>
<tr>
<th>Negative sampling</th>
<th>Methods</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>BM25-Orig</td>
<td>0.2565</td>
<td>0.1772</td>
</tr>
<tr>
<td>R-Neg</td>
<td>BM25-Obj</td>
<td>0.3493</td>
<td>0.2395</td>
</tr>
<tr>
<td>IB-Neg</td>
<td>BM25-Cap</td>
<td>0.4550</td>
<td>0.3293</td>
</tr>
<tr>
<td>IB-Pos</td>
<td>BM25-All</td>
<td>0.4704</td>
<td>0.3366</td>
</tr>
</tbody>
</table>

3.3 Additional Results

We provide additional analysis to study the impact of the projection size and negative sampling strategies with validation performance. The capability of dense retrieval. We further explain that this capability is contingent upon the negative sampling strategy used during training in Sec. 3.3.2. Moreover, Dense-BERT even surpasses BM25-Obj that considers visual signals. This could be due to the tendency of Dense-BERT to retrieve passages containing frequent answers. We speculate this kind of overfitting is caused by the lack of visual signals. Further analysis can be found in Sec. 3.3.

We further observe that Dense-LXMERT significantly outperforms BM25-Obj. Dense-LXMERT leverages both the RoI features and position features in object detection to learn object relations, which can be more effective than using ground truth object names in sparse retrieval. We then compare Dense-LXMERT with BM25-Cap/All. These sparse retrieval methods consider human-annotated image captions that are highly informative and descriptive. On the contrary, Dense-LXMERT has to learn the importance of the objects and the relation among them with object-level features. In this unfavorable situation, Dense-LXMERT still manages to match the performance of BM25-Cap/All. Although BM25-Cap is slightly better, the margins are statistically insignificant. Finally, we observe that Dense-LXMERT significantly outperforms Dense-BERT, further validating the use of a multi-modal query encoder.

3.3.1 Impact of projection size. We present the impact of the dimensionality $n$ of query/passage representations in Fig. 3a. We observe that a larger projection size always leads to better performance, although the performance gain seems to be insignificant for LXMERT after $n = 256$. We set $n = 768$ as reported in Sec.3.1.4 since it gives the best performance for both Dense-LXMERT and Dense-BERT. When working with a much larger collection than ours (ours has 11 million passages), one might want to use $n = 256$ since it offers similar performance with less memory consumption.

3.3.2 Impact of negative sampling. The desirable performance of dense retrieval is contingent upon the negative sampling strategy. We present the impact of different sampling methods described in Sec. 2.3.3 in Fig. 3b. We observe that combining retrieved negatives with in-batch negatives dramatically improves the model performance, verifying the observations in Karpukhin et al. [16] for multi-modal queries. Also, different choices of in-batch negatives (R-Neg+IB-Neg/Pos/All) give a similar performance for LXMERT, indicating that coinciding questions in the same batch should not be a concern for our batch size and data size reported in Sec. 3.1.

Both analyses show that Dense-BERT is more demanding on larger model capacity (larger projection size) and more negative samples. We speculate that BERT is overfitting the patterns in the training data since it lacks important visual clues for matching. In comparison, Dense-LXMERT is less sensitive to reasonably-chosen projection sizes and negative sampling strategies because it can learn matching signals from both language and vision clues.

4 CONCLUSIONS AND FUTURE WORK

We study passage retrieval for OK-VQA with sparse and dense retrieval and verify visual clues play an important role. We discover that captions are more informative than object names in sparse retrieval and CombMax works well with object expansion whereas CombSum and RFF are better for caption expansion. We further show a dense retriever with a multi-modal query encoder can significantly outperform sparse retrieval with object expansion and even matches the performance of that with human-generated captions. In the future, we will consider using automatic captions for sparse retrieval and study answer extraction to complete the QA pipeline.

ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors.