Passage Reranking and Document Expansion with Image Representation for Outside-Knowledge Visual Question Answering Passage Retrieval

Alireza Salemi

Manning College of Information & Computer Sciences

University of Massachusetts, Amherst

asalemi@cs.umass.edu

Tai Dang

Manning College of Information & Computer Sciences

University of Massachusetts, Amherst

ttdang@umass.edu

Abstract

Outside-knowledge visual question-answering (OK-VQA) passage retrieval is retrieving documents from a knowledge source like Wikipedia that answers a question about an image. Standard methods in this subject use dual-encoder architecture with a text encoder to encode passages and a multi-modal encoder to encode questions and images. This approach suffers because textual encoders, such as BERT, and multi-modal encoders, such as LXMERT, do not share the same embedding space, which decreases the speed of convergence of the model. This paper uses only one multi-modal encoder as both passage and query encoders. We expand the passage representation by feeding a masked image to the encoder and letting the model generate a representation based on the passage. Additionally, we use a multi-modal re-ranker on top of the dual-encoder ranker to increase the model's performance on OK-VQA passage retrieval. Our experiments show that using a single multi-modal encoder with document expansion with image representation increases the performance of multi-modal dense passage retrieval by roughly 9% in MRR and Precision. Additionally, we show that re-ranking on top of a ranker is an effective method for increasing the performance of rankers in multi-modal scenarios.

1. Introduction

Visual question answering is the task of answering a question written in natural language related to an image [2]. Question-answering tasks are mostly targeted at single-modal information needs such as text- or voice-based questions. But there are many information needs containing multiple modalities of information. One instance of such multi-modal information-seeking tasks could be Visual Question Answering (VQA), which combines text questions and images. VQA is aimed at answering visual questions that target different areas of an image, such as background details or underlying context. The classic VQA benchmarks and models mainly revolve around questions that are about counting, visual attributes, or other visual detection tasks. The answer to all these questions can be found given only the image without any supplementary information. However, another class of VQA tasks referred to as Outside-Knowledge VQA (OK-VQA) task [18], is a specific setting of VQA in which the images, rather than simply acting as the knowledge source of the question, are a part of the question itself, helping to define the exact information need. Thus, OK-VQA questions require an outside knowledge source, e.g., a large collection of passages, to be answered. Therefore, designing retrieval systems that retrieve proper information related to a question and image pair plays an essential role in achieving a good performance in this task [20].

Performing the VQA task requires models that pro-

duce joint representation for image and text pairs. Recently, some multi-modal models have been trained with self-supervised tasks to learn a joint representation for text and image pairs [12, 13, 15, 23, 25]. Another idea for combining language and vision is to learn a similar representation for an image and text pair by using text representation as a supervision method for learning image representation, which has been shown to be effective when trained on a large scale [8, 21]. However, these models are trained using the captions of images, which are usually short texts and can decrease the model's performance when it deals with long passages.

Similar to the retrieve-and-read paradigm in open-domain QA systems, where the system first retrieves several documents (passages) from a collection and then extracts answers from them, the external information need of the OK-VQA task could be addressed by retrieving a document related to the image-question pair and then extracting the answer to the question from the document. This approach is studied by using dense passage retrieval with LXMERT [25] as the image-question pair encoder, and BERT [4] as the passage encoder [20]. However, this approach suffers from the fact that query and passage encoders start from different embedding spaces, which decreases the speed of the convergence of the model and its performance. Additionally, dual-encoder dense passage retrievers are known to be weaker than cross-encoders because they only observe queries and passages independently.

To address the mentioned issues, this paper uses a single shared multi-modal encoder as both a passage and query encoder. To use a multi-modal encoder for encoding passages, we feed the model with the passage and masked images and let the model generate a representation. Indeed, we expand the document representation with an image representation by masking the image and letting the model generate the representation based on the passage. Additionally, we train a re-ranker cross-encoder model to further improve the model's performance. This re-ranker is used on top of the ranker to re-rank the retrieved documents by the main ranker and select the top K.

Our results show that using a shared multi-modal encoder as both a passage and query encoder significantly increases the performance on the OK-VQA passage retrieval task. Our experiments show that this approach increases the model's performance by 9% in MRR@5 and Precision@5 metrics. Additionally, using a re-ranker on top of a ranker effectively improves performance for this task. We found that using a re-ranker is generally effective but more favorable for weaker rankers. We conducted several experiments to investigate the effects of different architectures for rankers and re-rankers in advance.

This paper tries to answer the following questions:

- **Q1:** Which multi-modal model works better as a ranker for dual-encoder architecture?
- **Q2:** Which multi-modal model works better as a re-ranker for cross-encoder architecture?
- **Q3:** Does a shared encoder for encoding queries and passages help the model's performance?
- **Q4:** Is starting from the same checkpoint that has the same embedding space help the model's performance?
- **Q5:** Is document expansion with masked image representation effective for generating a better representation of passages?
- **Q6:** Does a re-ranker improve the results in multi-modal scenarios?

2. Related Work

Apart from single-modal settings, many real-world cases involve multi-modal information needs that contain both textual questions and images. The MultiModalQA (MMQA) dataset [24] is specifically designed for the multi-modal question-answering task to address QA that requires joint reasoning over text, tables, and images. One instance of information needs beyond a single mode of text would be in a cross-modal information retrieval application for image-to-image recipe retrieval [14] using images of ingredients as input queries and retrieving cuisine images. Another instance is the task of Visual Question Answering (VQA) [2], aiming to answer visual questions about different aspects of an image. Other benchmark datasets for the task of visual QA have also been established [7, 29, 30]. One end-to-end approach to this task is Neural-Image-QA [17], jointly trained on a combination of CNN and LSTM. Another work [28] studies the dynamic memory network architecture, which combines a memory component and the attention mechanism, applied to the setting and modality of this task. Also explored is the co-attention mechanism for VQA that jointly performs question-guided visual attention and image-guided question attention [16].

In contrast to these types of visual question-answering tasks, where the questions are mostly concerning visual attributes of the image and the answers can be found given only the image, the focus of this project is on Outside-Knowledge VQA (OK-VQA) [18]. In OK-VQA, rather than acting as the knowledge source for the question, images are a crucial part of the question itself, helping to define the information needed. Hence, OK-VQA requires access to an outside and open knowledge resource, e.g., a large collection of passages, to answer the questions. Most of the work in this domain retrieves knowledge from a structured knowledge base, such as a knowledge graph. ConceptBert [6] jointly learns from visual, language, and knowledge graph embeddings and captures image-question-knowledge

specific interactions used to predict the answer. In another work, the entities and relations of the knowledge graph are embedded into a continuous feature space and attended to by dynamic memory networks to implement complex reasoning over several facts [11]. The Ahab approach [27] detects relevant content in the image and relates it to the information in a knowledge base. The natural question is then converted to a knowledge base query using the combined image and knowledge-base information to retrieve the answer from the query's response.

LEXMERT [25] is a transformer-based [26] visual language model that was trained with masked cross-modality language modeling and masked object prediction. For each image and text pair, 36 objects in the image are detected by the Faster R-CNN object detection model [22]. Then, the image representation of detected objects using a Resnet model fine-tuned on this task in previous works [1] and word embeddings of text are fed to a transformer model to learn a joint representation. Next, the model is asked to predict the masked words in the sentence by observing other tokens, which include both textual and visual tokens. Similarly, the model is asked to predict the class of masked objects using other textual and visual tokens. In this way, the model must consider textual and visual features to predict correct answers, which forces the model to learn a joint representation for image and text input.

CLIP [21] is another multi-modal model ties to learning visual and textual representation for image and text pairs using contrastive learning [3]. Using contrastive loss, this model tries to maximize the similarity between an image and text pair that are related but also minimizes the similarity between each image and other texts that are not paired with that image in the current batch. The same happens for each text and other images that are not related to that in a batch. The formula for contrastive loss is shown in equation 3:

$$L_{image2text} = -\frac{1}{N}\sum_{i} Nlog\frac{exp(x_i.y_i^t/\delta)}{\sum_j^N exp(x_i.y_j^t/\delta)} \quad (1)$$

$$L_{text2image} = -\frac{1}{N}\sum_{i} Nlog\frac{exp(y_i.x_i^t/\delta)}{\sum_j^N exp(y_i.x_j^t/\delta)} \quad (2)$$

$$L_{CLIP} = L_{text2image} + L_{image2text} \quad (3)$$

x_i and y_j are the normalized representations of the ith and jth image and text in the batch. A transformer model produces the representation for texts, and the representation for images is produced by Resnet, or ViT [5] model. Additionally, they gathered an extensive dataset of image and text pairs to stabilize learning from noisy labels and used several unsupervised filters to filter out unrelated texts and images. ALIGN [8] also uses the same technique and loss function but on a larger scale, which results in even better performance.

Flava [23] is another transformer-based visual-language model. This model was trained with image-only objectives, text-only objectives, and multi-modal pre-training objectives, which makes it more robust to longer texts. This model was also trained with a similar objective as CLIP and ALIGN, making it among the best-performed visual-language models.

While cross-encoder methods for textual similarity comparison work better than bi-encoder methods, they are inefficient when there are many passages in a knowledge source because indexing is impossible for cross-encoder methods. On the other hand, cross-encoders can be used as re-ranking models, which re-rank the retrieved passages of a weaker ranking approach [19]. Re-ranking using BERT [4] has shown improvement in open-domain question-answering tasks [19]. In this method, first, we use a ranking model to retrieve k documents with the highest scores based on a query. Then, we score each retrieved document using the BERT re-ranker model and obtain a more robust and realistic similarity score between documents and the query [19].

One of the first works on outside knowledge visual question-answering passage retrieval uses LEXMERT to encode a query, which consists of a question and image, and BERT to encode passages [20]. Utilizing contrastive loss, the model tries to decrease the similarity between the representation for question and image pair and the representation for passages. However, this approach has several shortcomings. First, BERT and LEXMERT are two different models trained with different objectives, which results in dissimilarity between their representations. Thus, this approach needs many training steps to fulfill its mission. Additionally, using a bi-encoder method decreases performance, which a re-ranker model can solve. This paper tries to solve the mentioned issues using the same encoder to encode both passages and queries. Additionally, we use a re-ranker on top of the ranker to improve the results.

3. Task Definition & Dataset

Outside-knowledge visual question answering is the task of answering a question that needs some extra piece of knowledge about an image that is not available inside the image. Figure 1 shows some examples of the OK-VQA dataset. It can be seen that this dataset contains samples from a various range of subjects.

In this paper, we experiment with the outside-knowledge visual question answering (OK-VQA) dataset [18]. However, we use a version that is designed for outside-knowledge visual question-answering passage retrieval [20], in which we must retrieve a document that contains the answer to a question about an image. Each sample in

Figure 1. Some examples of the OK-VQA dataset. It can be seen that the answer to the questions is not available in images. However, a deep understanding of the image and the question is helpful in finding the answer to the question by utilizing a knowledge source like Wikipedia.

this dataset consists of the following items:

1. Image: This is an image that a question about it will be asked.
2. Question: This is a question about the image, but the answer to this image is not available in the image itself.
3. Answer: This is the answer to the aforementioned question about the image.
4. Positive Passage: This passage contains the answer to the question.
5. Negative Passage: This is a passage retrieved by BM25 using the query, but it does not contain the answer to the question.

Therefore, the main purpose of this task is to retrieve a document that contains the answer to the question about the image.

4. Methodology

The primary purpose of this work is to retrieve passages from Wikipedia, a knowledge source with 11 million passages, that can answer a question about one image. To do this efficiently, we use a two-stage neural pipeline depicted in Figure 2.

In the first stage, we use a bi-encoder architecture to independently encode passages and queries (question-image pairs), then use the dot product between their representations as the similarity score. Using this approach, we can efficiently retrieve N ($N << 11M$) documents corresponding to the query with the highest scores. However, we know that bi-encoder architecture works weaker than cross-encoder retrievers because, in bi-encoders, the interactions between query and documents are only limited to dot product. On the other hand, cross-encoders can model interactions of queries and documents using an attention mechanism. At the same time, applying them to a large set of documents is impossible. Therefore, we use a cross-encoder architecture as a re-ranker, which selects the top K documents among the top N documents retrieved by the bi-encoder retriever ($K < N$). This section explains the architecture and training procedure of each stage.

4.1. Dual-encoder Multi-modal Dense Passage Retriever

Suppose that E_q is a transformer model which encodes queries, consisting of images and questions, into a d-dimensional space. Similarly, E_p is a transformer model which encodes passages into a d-dimensional space. We define the similarity score between a passage and a query as follows:

$$sim(p, q, im) = E_p(p)_{CLS}.E_q(q, im)^T_{CLS} \quad (4)$$

where $E_p(p)_{CLS}$ is the *CLS* token derived from feeding the passage p to encoder E_p and $E_q(q, im)_{CLS}$ is the *CLS*

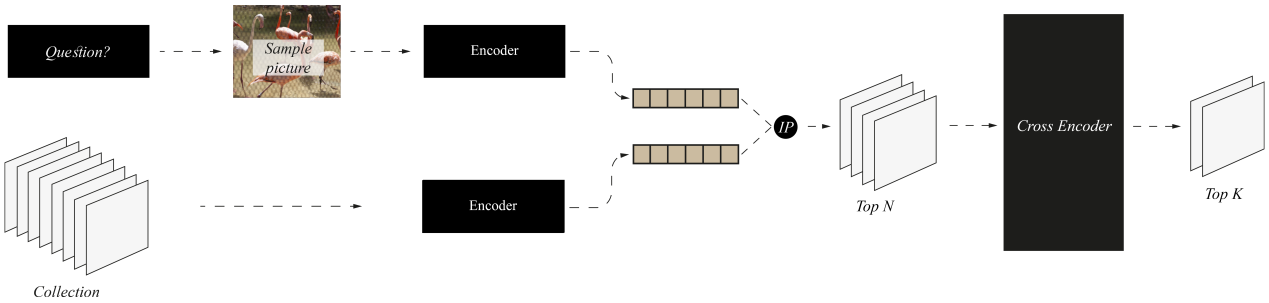

Figure 2. An overview of the retrieval pipeline for outside-knowledge visual question-answering passage retrieval. The pipeline consists of two models: 1) a bi-encoder ranking model to efficiently select top N documents from a large collection of documents, and 2) a cross-encoder re-ranking model that selects top K documents from top N documents retrieved by the first stage ranker.

token derived from feeding the image and question q, im to encoder E_q.

The main purpose of the dual encoder architecture is to generate a separate representation for passages and queries that are similar to each other. To force the encoders to produce a similar representation for related queries and passages and a dissimilar representation for a query and unrelated passages, we use the following loss function to train the model:

$$L = -log \frac{e^{sim(p_{pos},q,im)}}{e^{sim(p_{pos},q,im)} + \sum_{p \in P_{neg}} e^{sim(p,q,im)}} \quad (5)$$

where p_{pos} is the related document to the question and image, and P_{neg} is the set of unrelated documents to the question and image.

To provide negative documents for each query, we use in batch negative technique with a hard negative sample. In this technique, all the documents in a batch except those related to a query are considered negative samples. Additionally, we use the hard negative sample provided by the OK-VQA passage retrieval dataset as the hard negative sample.

The choice of E_q and E_p that provide representation for passages and documents is important. In previous work [20], E_q was LXMERT [25], which encodes images and questions into a single space, and BERT [4] was used to generate passages' representations. Considering that BERT and LXMERT are two different models that produce representations in two different vector spaces, we believe that using a single model to encode both queries and passages can increase the speed of convergence and the model's performance. To address this, we use Flava [23] visual-language model as both the passage and the query encoder.

To use Flava for generating a representation for passages, which do not have any images, we use document expansion. In this method, we only feed the model with text only and use the model's ability learned from its pre-training to generate a representation for the input image. Indeed, we mask all the image tokens for passages and let the model generate the corresponding representation for them by itself. Since Flava was pre-trained to predict images based on text, it can expand text-only documents with a representation of images by itself. We do the same for LXMERT too. For LXMERT, we mask all the tokens in the image for passages and use $[0, 0, 1, 1]$ as the boundaries of bounding boxes of the images. Similarly, LXMERT can generate a representation for the masked image because it has learned this ability in its pre-training.

Moreover, it has been shown that Dense Passage Retriever (DPR) [10] with a shared E_p and E_q outperforms DPR with two encoders. Similarly, we use a single encoder as E_p and E_q. This is possible because we expand passages with masked image representations, which let us use a single model as both encoders.

4.2. Cross-encoder Multi-modal Re-ranker

A cross-encoder model is a single encoder that considers the passage and query in a single encoder. Since the interaction between the passage and document happens in the same encoder, it can better evaluate the similarity between a query and a document. However, since we need to feed the query and each document to the model to find their score, it is impossible to use them on a large corpus. Hence, cross-encoders are used as re-rankers.

Suppose E_rr is a transformer-based vision-language model, such as LXMERT or Flava, which encoders an image and a text into a d-dimensional vector space. Then, we calculate the similarity score as follows:

$$Sim(p, q, im) = sigmoid(E_{rr}(p, q, im)^{1\times d}_{CLS}.W^{d\times 1}) \quad (6)$$

where p is a passage, im, q are question and image, and $W^{d\times 1}$ is a learnable linear layer that maps the d-dimensional vector to a single similarity score. In order to train the re-ranker model to learn to score the documents, we use the following objective function:

$$L = -log(sim(p_{pos}, q, im)) - log(1 - sim(p_{neg}, q, im)) \quad (7)$$

where p_{pos} is the positive passage for the query and p_{neg} is the negative passage for the query, both provided by the training dataset. Due to memory constraints and efficiency, we do not use in batch negative sampling for training the re-ranker. We use both Flava and LXMERT as E_{rr} for the re-ranker and compare their performance.

It should be noted that re-rankers are usually much slower than the dual-encoder rankers because they have to run the encoder for each document separately. On the other hand, in dual-encoder architecture, it is enough to generate the representation for each document separately in an offline session and use the computed representation in inference time. This is not possible for cross-encoder models because the representation the model generates is dependent on the query. Therefore, it is not possible to compute anything in an offline session.

4.3. End-to-End Inference with Ranker and Re-ranker

Figure 2 shows the end-to-end pipeline of our retrieval system. To retrieve documents efficiently, we first index all passages in the collection using the trained dual-encoder model. Then, we store the generated representations using the faiss [9] library for efficient vector search. Same as training, we use dot product as the similarity metric between vectors. It should be noted that indexing corpus happens in an offline session, which decreases retrieval latency when we want to retrieve documents for a query.

Next, for each given query, we generate the representation of that query using the dual-encoder model and find the top N documents using faiss library. Then, we feed the top N documents to the re-ranker model trained using the aforementioned objective function to generate a similarity score for each retrieved document. Finally, we select the top K documents with the highest score generated by the re-ranker as the final retrieved documents.

5. Experiments

In this section, we explain our experiments and the results we have obtained in detail. Additionally, we explain the setup in which we conduct our experiments and the metrics we used to evaluate the results.

5.1. Setup

All experiments are conducted on a single Nvidia RTX8000 GPU with 49GB memory. Following the previous work [20], we use a batch size of 16 for all experiments. We use Adam optimizer with a learning rate of 10^{-5} for LXMERT and 5×10^{-5} for Flava ranker and 10^{-5} for Flava re-ranker. A linear learning rate scheduler with 10% of total training steps as warmup steps is used to train the models. The representation of each model is in a 768 dimensions vector space in this architecture. Additionally, gradient clipping with a clipping value of 1 is used in the training procedure. The maximum length of passages and queries for each encoder is 400 tokens. Additionally, we use faiss library on the cpu only version. More speedups can be achieved if the GPU version of this library is used. We train each model for two epochs on the training data and evaluate the model for each 1K steps. We use early stopping to report the best checkpoint results.

5.2. Dynamic Evaluation Strategy and Evaluation Metrics

Following the previous work [20], we dynamically evaluate the models. In this method, we consider each retrieved document that contains the answer to the question as a positive document. Therefore, using recall-based metrics for the evaluation of the system is meaningless since they rely on the total number of relevant documents to each query, which is not available in our dynamic evaluation strategy.

After retrieving and finding positive documents to a query, we use Mean Reciprocal Rank (MRR@N) and Mean Precision (P@N) as evaluation metrics. Suppose we have a ranked list of documents $l = \{d_1, d_2, ..., d_N\}$. Then:

$$MRR@N = \frac{1}{rank\ of\ the\ first\ positive\ document} \quad (8)$$

$$P@N = \frac{number\ of\ positive\ documents}{N} \quad (9)$$

We average the aforementioned metrics on the whole test set to find MRR@N and P@N. In these experiments, we use $N = 5$ for re-rankers and $N = 25$ for rankers unless we explicitly mention other numbers.

5.3. Results

In this section, we explain the results of our experiments. We evaluate the models from different aspects and report the results in the following sections.

5.3.1 Re-ranking Accuracy

We use accuracy as the metric to evaluate the re-ranker models separately from the ranker. Indeed, we feed the model with each sample in the OK-VQA passage retrieval dataset with a positive and negative passage. Then, we check how often the model assigns a higher probability to the positive passage, which is the definition of the accuracy in this case.

Model	Accuracy
LXMERT Re-ranker	0.762
Flava Re-ranker	0.860

Table 1. The results of the evaluation of re-rankers separately from the whole pipeline. The results are reported using the accuracy metric, which shows how often a re-ranker assigns a higher probability to the positive document than the negative document for a sample.

We use LXMERT and Flava as re-rankers in the architecture introduced in section 4.2. The results of these models are reported in Table 1. The results in Table 1 show that Flava performs better than LXMERT as a re-ranker model (answer to **Q1**). Based on these results, we use Flava re-ranker as the main ranker in our final evaluation pipeline. Additionally, we evaluate the whole pipeline in an end-to-end mode to investigate the results effect of all components together on the performance of the whole system.

5.3.2 Dual-encoder Ranker results

We trained different versions of the dual-encoder multi-modal dense passage retriever based on the architecture introduced in section 4.1. We trained the models based on the following scenarios:

1. LXMERT as the query embedding and BERT as passage embedding, without any document expansion (baseline).
2. A shared LXMERT as both the query and passage embedding with document expansion. We expand the document by masking the image part of the input and letting the model predict a representation of the image based on the text.
3. Flava as the query embedding and Flava's text encoder as passage embedding, without any document expansion.
4. Flava as the query embedding and BERT as passage embedding, without any document expansion.
5. A shared Flava as both the query and passage embedding with document expansion. We expand the document by masking the image part of the input and letting the model predict a representation of the image based on the text.
6. Two Flava as both the query encoder and passage encoder without parameter sharing and with document expansion. We expand the document by masking the image part of the input and letting the model predict a representation of the image based on the text.

The results of these experiments in passage retrieval are reported in Table 2. It can be seen that, generally, LXMERT performs better than Flava as a backbone for dual-encoder models (answer to **Q2**). Indeed, whenever LXMERT is used, the model's performance is higher than those with Flava as the encoder. This might be because LXMERT was trained on visual question answering as one of its pre-training objectives, which resulted in generating a better representation for the task of outside-knowledge visual question answering.

Additionally, it can be seen that using a shared Flava model as both query and passage encoder resulted in a better performance (0.327 for shared Flava in MRR@5 vs. 0.251 for not shared Flava in MRR@5). Therefore, sharing parameters between encoders can be helpful for the final retrieval task (answer to **Q3**).

Finally, the results Table 2 shows that using the same LXMERT model as both encoders with the help of document expansion using the image representation can significantly improve the results of the dual-encoder architecture (0.56 vs. 0.47 in MRR@5) (answer to **Q4**). Since multi-modal models always need an image as input to generate a multi-modal representation, it is not possible to investigate the effect of starting from a shared space independent of document expansion. Indeed, document expansion using masked image representation and starting from a shared embedding space are two methods that always happen together in our experiments, which makes it impossible to separate them (answer to **Q5**).

Therefore, our experiments show that our assumptions about dual-encoder models were correct. To shed light on this, using a shared encoder for encoding queries and passages improves the results. Additionally, using the masked image representation and starting from the same embedding space and document expansion helps the model generate a better representation for documents, which results in a better performance.

5.3.3 End-to-end Evaluation of the Ranker and Re-ranker Pipeline

Among the trained re-ranker in previous sections, we use Flava re-ranker as the re-ranker for the full pipeline. Among the trained dual-encoder rankers, we use the model with the best results as the main ranker of the pipeline, which is the model with shared LXMERT for both encoders with document expansion. We also use the best Flava-based ranker with a re-ranker to investigate the effect of re-ranking.

In order to evaluate the model, we retrieve 25 documents using the ranker model from the documents collection. Then, we use the re-ranker to score each document independently and sort them based on the re-ranker's score. Finally, we keep the top 5 documents with the highest score.

Model	MRR@5	P@5
Dual-Encoder Multi-modal Dense Passage Retriever(LXMERT, BERT) [20]	0.470	0.336
Dual-Encoder Multi-modal Dense Passage Retriever(shared LXMERT) with Document Expansion	**0.560**	**0.420**
Dual-Encoder Multi-modal Dense Passage Retriever(Flava, Flava text encoder)	0.218	0.139
Dual-Encoder Multi-modal Dense Passage Retriever(Flava, BERT)	0.421	0.296
Dual-Encoder Multi-modal Dense Passage Retriever(shared Flava) with Document Expansion	0.327	0.212
Dual-Encoder Multi-modal Dense Passage Retriever(Flava, Flava) with Document Expansion	0.251	0.168

Table 2. The results of the dual-encoder multi-modal dense passage retriever on outside-knowledge visual question answering passage retrieval task. The results are reported using MRR@5 and Precision@5. It can be seen that using a shared LXMERT model as both query and passage encoders with document expansion using masked image representation has achieved the best results.

Model	MRR@5	P@5
Dual-Encoder multi-modal DPR (Shared LXMERT) with document expansion + Flava Re-ranker	**0.578**	**0.446**
Dual-Encoder multi-modal DPR (LXMERT, BERT) + Flava Re-ranker	0.559	0.408
Dual-Encoder multi-modal DPR (Shared Flava) with document expansion + Flava Re-ranker	0.471	0.327

Table 3. The results of the full pipeline consist of a ranker and a re-ranker on outside-knowledge visual question answering passage retrieval task. The results are reported using MRR@5 and Precision@5 for the re-ranker and MRR@25, and Precision@25 for the ranker. In this experiment, we first retrieve 25 documents using the ranker. Then, we re-rank the retrieved documents using the re-ranker and keep the top 5 documents.

The system's end-to-end evaluation results are reported in Table 3. The results in Tables 3 and 2 show that using a re-ranker on top of a ranker can always improve MRR@5 and Precision@5 (answer to **Q6**). However, the amount of improvement is different for different models. Indeed, weaker rankers usually benefit more from using a re-ranker model than stronger rankers. It can be seen that the results of the model with shared LXMERT and document expansion, which was our best ranker in the previous section, only improved by roughly 2% while other models achieved 9% and 11% improvements. Therefore, we conclude that, generally, re-ranking is an effective technique to improve the base ranker's results.

6. Conclusion

Common approaches for outside-knowledge visual question-answering passage retrieval usually use different text and multi-modal encoders to encode passages, images, and questions. However, using two different encoders increases the convergence time since these encoders encode passages, images, and questions in different embedding spaces. Additionally, it is known that while dual-encoder architecture for dense passage retrieval is efficient, it is weaker than cross-encoder methods. In this paper, we define a dual-encoder multi-modal dense passage architecture that uses a shared vision-language model as an encoder for both modalities. Additionally, we expand the documents by feeding a blank mask as an image to the encoder and letting the model generate a representation of the image based on the passage. Moreover, we train a cross-encoder re-ranker to re-rank the outputs of the first-level ranker. Our results show that using a shared multi-modal encoder to encode both passages and queries (images and questions) significantly improves ranking results on the OK-VQA passage retrieval dataset. Additionally, we show that re-ranking is an effective technique to increase the ranker's performance even in multi-modal scenarios.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 1, 2

[3] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, 2005. 3

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. 2, 3, 5

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[6] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online, Nov. 2020. Association for Computational Linguistics. 2

[7] Yash Goyal, Tejas Khot, Douglas Summers Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. pages 6325–6334, 07 2017. 2

[8] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2, 3

[9] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 6

[10] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, Nov. 2020. Association for Computational Linguistics. 5

[11] Guohao Li, Hang Su, and Wenwu Zhu. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks. 12 2017. 3

[12] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language, 2019. 2

[13] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, August 2020. 2

[14] Yen-Chieh Lien, Hamed Zamani, and W. Bruce Croft. Recipe retrieval with visual query of ingredients. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1565–1568, New York, NY, USA, 2020. Association for Computing Machinery. 2

[15] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[16] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 289–297, Red Hook, NY, USA, 2016. Curran Associates Inc. 2

[17] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. 12 2015. 2

[18] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[19] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *ArXiv*, abs/1901.04085, 2019. 3

[20] Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik Learned-Miller. Passage retrieval for outside-knowledge visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1753–1757, New York, NY, USA, 2021. Association for Computing Machinery. 1, 2, 3, 5, 6, 8

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2, 3

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region

proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 3

[23] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, June 2022. 2, 3, 5

[24] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodal{qa}: complex question answering over text, tables and images. In *International Conference on Learning Representations*, 2021. 2

[25] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. 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)*, pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 2, 3, 5

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 3

[27] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Henge. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 1290–1296. AAAI Press, 2017. 3

[28] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 2397–2406. JMLR.org, 2016. 2

[29] Licheng Yu, Eunbyung Park, Alexander Berg, and Tamara Berg. Visual madlibs: Fill in the blank image generation and question answering. 05 2015. 2

[30] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. pages 4995–5004, 06 2016. 2