A Symmetric Dual Encoding Dense Retrieval Framework for Knowledge-Intensive Visual Question Answering

Alireza Salemi
University of Massachusetts Amherst
United States
asalemi@cs.umass.edu

Juan Altmayer Pizzorno
University of Massachusetts Amherst
United States
jpizzorno@cs.umass.edu

Hamed Zamani
University of Massachusetts Amherst
United States
zamani@cs.umass.edu

ABSTRACT

Knowledge-Intensive Visual Question Answering (KI-VQA) refers to answering a question about an image whose answer does not lie in the image. This paper presents a new pipeline for KI-VQA tasks, consisting of a retriever and a reader. First, we introduce DEDR, a symmetric dual encoding dense retrieval framework in which documents and queries are encoded into a shared embedding space using uni-modal (textual) and multi-modal encoders. We introduce an iterative knowledge distillation approach that bridges the gap between the representation spaces in these two encoders. Extensive evaluation on two well-established KI-VQA datasets, i.e., OK-VQA and FVQA, suggests that DEDR outperforms state-of-the-art baselines by 11.6% and 30.9% on OK-VQA and FVQA, respectively.

Utilizing the passages retrieved by DEDR, we further introduce MM-FiD, an encoder-decoder multi-modal fusion-in-decoder model, for generating a textual answer for KI-VQA tasks. MM-FiD encodes the question, the image, and each retrieved passage separately and uses all passages jointly in its decoder. Compared to competitive baselines in the literature, this approach leads to 5.5% and 8.5% improvements in terms of question answering accuracy on OK-VQA and FVQA, respectively.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Question answering**; **Multimedia and multimodal retrieval**; • **Computing methodologies** → **Computer vision**.

KEYWORDS

Dense Retrieval; Knowledge Distillation; Visual Question Answering; Multi-Modal Retrieval

ACM Reference Format:
Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. 2023. A Symmetric Dual Encoding Dense Retrieval Framework for Knowledge-Intensive Visual Question Answering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591629

SIGIR '23, July 23–27, 2023, Taipei, Taiwan.

ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3591629

Figure 1: An example KI-VQA question. Answering these question requires external knowledge.
Image © zrim, https://www.flickr.com/photos/zrimshots/2788695458

1 INTRODUCTION

Knowledge-intensive visual question answering[1] (KI-VQA) is a variant of visual question answering tasks whose questions cannot be answered using the image alone. Therefore, accessing external knowledge sources is necessary to answer these questions. KI-VQA has a large number of real-world applications. Imagine customers of e-commerce websites taking a photo of a product or a part of a product and asking a question about it. In the context of education, students can ask a question about an image in their textbook. Users can take a photo of a visual sign or a piece of art and ask questions about its meaning or history. These are just a few examples of KI-VQA applications. Figure 1 shows an example of KI-VQA tasks: the image is sufficient to identify the "animals" as giraffes, and likely also the the subspecies, but not to answer how tall they get.

The majority of prior work on KI-VQA, such as [12, 14, 37, 50, 57], assumes that the external knowledge can be obtained from a structured knowledge base. However, a high-quality and complete knowledge base may not be available for some domains [1]. Besides, maintaining knowledge bases with up-to-date information is challenging [54]. To prevent these issues, following Qu et al. [40], we take an alternative approach to KI-VQA: using a large text corpus as the external knowledge source. In this setting, a two-stage pipeline for KI-VQA systems is to first retrieve a list of passages for a given question-image pair and then process the retrieved passages to generate an answer.[2] This pipeline is depicted in Figure 2.

The effective performance of dense retrieval models in various information retrieval tasks [23, 62, 63] and their extension flexibility to multi-modal[3] input have motivated us to focus on dense retrieval for implementing the first stage of the KI-VQA pipeline (i.e., passage

[1]This task is also referred to as outside-knowledge visual question answering (OK-VQA) in the literature [38]. OK-VQA is also the name of a dataset used in this paper. To avoid confusion, we use "KI-VQA" to refer to the task.
[2]This is similar to the retriever and reader stages in open-domain question answering tasks [4].
[3]In this paper, multi-modality refers to the combination of text and image.

retrieval). A property of this retrieval task is that it deals with asymmetric input modalities: the user information need is multi-modal (question-image pair) while the information items (passages) are uni-modal. As a result of this property, Qu et al. [40] recently showed that a KI-VQA dense retrieval model that uses a multi-modal encoder for representing the question-image pair and a text encoder for representing the passages in the collection leads to state-of-the-art passage retrieval performance. We argue that using such an asymmetric bi-encoder architecture is sub-optimal, since the encoders produce outputs in different semantic spaces and fine-tuning the encoders cannot always close this gap. We first study two alternatives for developing symmetric dense retrieval models:[4] (1) producing a textual representation of the image and using a symmetric uni-modal bi-encoder architecture for dense retrieval, and (2) converting passages to a multi-modal input format and using a symmetric multi-modal bi-encoder architecture. We observe that both alternatives suffer from information loss, but also that they produce complementary representations. This observation motivates us to not only combine these two encodings, but also transfer knowledge between them. In more detail, we propose an iterative knowledge distillation approach to transfer knowledge between these two alternative symmetric dense retrieval models. The proposed symmetric dual encoding approach leads to 11.6% and 30.9% MRR improvements compared to the state-of-the-art baseline on OK-VQA [38] and FVQA [56] test sets, respectively.

For the second stage of the pipeline, unlike much prior work on answer span detection for KI-VQA [12, 14, 37, 50, 57] (i.e., answer extraction from the retrieved passages), we focus on retrieval-augmented autoregressive answer generation. We propose MM-FiD, a simple yet effective extension of the Fusion-in-Decoder (FiD) [20] architecture to multi-modal input. FiD is a retrieval-augmented text generation model that has recently shown effective performance in question answering tasks [20]. MM-FiD uses a multi-modal encoder to represent the question, the image, and the retrieved passages and uses a uni-modal decoder that generates an answer and is trained using the maximum likelihood objective. Extensive experiments on both OK-VQA and FVQA datasets demonstrate that MM-FiD significantly outperforms alternative approaches for answer generation. It also performs better than answer span detection baselines. In more detail, our end-to-end pipeline achieves 5.5% and 8.5% improvement compared to the baselines on OK-VQA and FVQA question answering tasks, respectively. We open-source our code and release our learned model parameters for research purposes[5].

2 RELATED WORK

(Multi-Modal) Dense Retrieval. Using dense vectors for retrieving textual documents related to a textual query has been studied since the emergence of Latent Semantic Analysis [7]. However, dense retrievers' performance remained inferior to that of sparse retrievers like BM25 until Karpukhin et al. [23]'s Dense Passage Retriever (DPR), which uses the `[CLS]` token output by BERT [8], a pre-trained language model. While many dense retrievers only use a single vector to represent the query and the document [23, 41, 59], using multiple vectors per document and query has been also studied [11, 19, 24, 36, 49].

Multi-modal dense retrieval has recently been investigated in different forms: (1) uni-modal query and multi-modal documents [15, 34, 52], (2) multi-modal query and uni-modal documents [40], (3) multi-modal query and multi-modal documents [51], and (4) uni-modal query and uni-modal documents with queries and documents from different modalities, i.e., cross-modal retrieval [21, 42].

In this work, we focus on the second case, where the query is multi-modal while the documents only contain text. Qu et al. [40] utilized an asymmetric bi-encoder with LXMERT [53], a pre-trained vision-language model based on BERT [8] for encoding queries, and BERT itself for encoding documents. As we show, such an asymmetric architecture is sub-optimal; utilizing different encoders creates a semantic "gap" in the embedding space and fine-tuning cannot easily overcome the issue. We instead propose a new symmetric dual encoding framework that addresses this issue.

Knowledge Distillation for Dense Passage Retrieval. Due to the vast number of learnable parameters in dense passage retrievers, sometimes available datasets are insufficient to train them [62]. Consequently, knowledge distillation, in which a teacher model provides labels for a student model, has become a standard approach for training dense retrieval models and has shown compelling outcomes [17, 30]. Existing work in this area often uses cross-encoder rerankers that input both query and document as teacher models for dense retrieval models with a bi-encoder architecture [44]. Another approach is to distill knowledge from multi-vector dense retrieval models, such as ColBERT [24] to single-vector dense retrievers [63].

In our experiments, we did not find knowledge distillation from cross-encoder rerankers helpful for KI-VQA datasets. Thus, we introduce a novel approach: iterative knowledge distillation for dual encoding, in which knowledge distillation happens iteratively between two bi-encoders that use different modalities in their inputs. Accordingly, each model learns the perspective of the other, adjusting their representation spaces for more effective dense retrieval.

Knowledge-Intensive Visual Question Answering. *Knowledge-intensive* refers to a category of retrieval-enhanced machine learning problems [61] whose inputs are not sufficient to produce the output and external information should be provided. KILT [39] is a benchmark for natural language knowledge-intensive tasks such as open-domain question answering, fact-checking, entity-linking, slot-filling, and knowledge-based dialogue. All the mentioned tasks are text-only knowledge-intensive tasks. To the best of our knowledge, there is no unified benchmark on multi-modal knowledge-intensive tasks, which is relatively less explored. Therefore, this paper focuses on knowledge-intensive visual question answering.

Visual question answering (VQA) is a multi-modal question answering task whose goal is to answer a natural language question about an image [3]. VQA is primarily designed to measure the ability of models in representing and comprehending multi-modal data. Therefore, questions in VQA are often related to visual features (e.g., color or shape of the objects) and sometimes require commonsense knowledge. In other words, a human can answer VQA questions by just looking at the image, without accessing external information. Given this formulation, VQA has limited real-world use cases. In contrast to VQA, knowledge-intensive visual question answering

[4]Symmetric dense retrieval refers to a bi-encoder architecture with shared parameters.
[5]https://github.com/alirezasalemi7/DEDR-MM-FiD

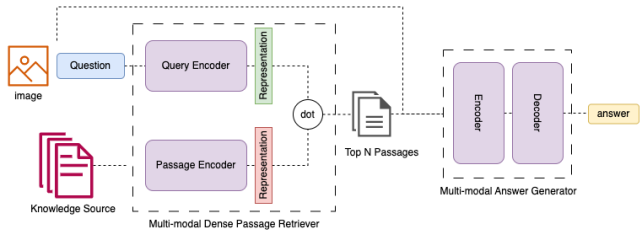

Figure 2: A pipeline for KI-VQA tasks that retrieves unstructured text in response to the input question-image pair and uses the retrieved passages as supporting documents to generate textual answer.

is the task of answering a question about an image that needs an external piece of information not available in the image to answer the questions. Fact-based Visual Question Answering (FVQA) [56] is a visual question-answering dataset in which answering the questions about an image needs the model to consider a relevant fact. Alternatively, outside-knowledge visual question answering (OK-VQA) [38] is a dataset similar to FVQA, but the required knowledge is not limited by facts.

Current approaches for OK-VQA utilize different strategies to solve the problem: some rely on implicit knowledge stored in language or vision-language models to answer the questions without using any external source of knowledge [48, 60], while others use external knowledge sources for this purpose in addition to implicit knowledge [12–14, 37, 57]. Using OCR and dense object labels can also be effective for this task [10, 32], but is beyond the scope of this paper. In order to use an explicit knowledge source, it is necessary to design a *retriever* that retrieves a small set of relevant passages to the image and question [40] and a *reader* that selects or generates the response from the retrieved passages [13, 14, 35]. This paper proposes effective solutions for both of these steps.

3 PROBLEM STATEMENT

In knowledge-intensive visual question answering, a user asks a natural language question about an image, which requires access to external information. In other words, the answer to the question does not exist in the image, which necessitates the utilization of external resources. See Figure 1 for an example of KI-VQA tasks. These resources can be in many different forms, from structured and semi-structured knowledge bases to unstructured text retrieved from the web. In this paper, we consider a scenario where the answer should be retrieved and extracted from a collection of unstructured natural language passages. In the following, we specify more formally the KI-VQA task studied in this paper.

Let $T = \{(Q_1, I_1, A_1, R_1), (Q_2, I_2, A_2, R_2), \cdots, (Q_n, I_n, A_n, R_n)\}$ denote the training set for a KI-VQA task. Each training instance consists of a natural language question Q_i, an image I_i, a set of short textual answers A_i, and a set of relevant passages R_i. That means each question *may* have multiple answers in the training set (i.e., $|A_i| \geq 1$), which are often semantically the same but syntactically different. Similarly, there may exist multiple passages that are relevant to the question (i.e., $|R_i| \geq 1$). All relevant passages are selected and annotated from a large-scale collection C. Therefore, $R_i \subseteq C : \forall 1 \leq i \leq n$. We study the following two related tasks:

Passage Retrieval for KI-VQA: the retrieval task is to use the training set T to train a retriever that retrieves relevant passages from the collection C for a given question-image pair (Q, I).

Retrieval-Augmented Answer Generation for KI-VQA: the retrieval-augmented answer generation task is to generate a short textual answer for any unseen question-image pair (Q, I) by having access to the collection C. Therefore, models in this task naturally retrieve passages from C and utilize them for generating an answer.

We first propose a symmetric dual encoding architecture for dense retrieval in KI-VQA and then introduce a multi-modal fusion-in-decoder model as a retrieval-enhanced answer generation approach.

4 DEDR: DUAL ENCODING DENSE RETRIEVER FRAMEWORK

Figure 2 depicts a pipeline for knowledge-intensive visual question answering tasks. As shown in the pipeline, the input to the dense retrieval model is asymmetric – query encoder takes multi-modal input (i.e., a question and an image), while the passage encoder takes a uni-modal text input (i.e., a passage from C). This asymmetric property in the input modalities makes it challenging to design an effective symmetric dense retrieval model. This is why the current state-of-the-art dense retrieval model proposed by Qu et al. [40] uses an asymmetric architecture, where a pre-trained multi-modal language model (i.e., LXMERT [53]) is used for query encoding and a pre-trained uni-modal language model (i.e., BERT [8]) is used for document encoding. Since such asymmetric architectures start from fundamentally different embedding spaces, they suffer from slow convergence speed and sub-optimal dense retrieval performance. Conversely, extensive research on dense retrieval for uni-modal data (textual queries and documents) suggests that symmetric architectures lead to significantly better performance. State-of-the-art dense passage retrieval models, such as TAS-B [17], ColBERT [24, 49], RocketQA [41, 44], and CLDRD [62], use symmetric architectures. Motivated by this observation, our goal is to learn a *symmetric dense retrieval model* for KI-VQA tasks.

To this aim, we study two alternative solutions. First, we convert all model inputs to a uni-modal textual form and then use uni-modal language models for both query and document encoding (Section 4.1). Second, we convert all inputs to the same multi-modal (text and image) form and then use multi-modal language models for both encoders (Section 4.2). We hypothesize that these two models learn complementary representations for the following reasons: (1) they take different input formats, and (2) the pre-training process and data in uni-modal and multi-modal language models are different. Our experimental results also validate this hypothesis (see Section 6.3). Following this observation, we propose an iterative knowledge distillation approach that alternates between these two encoding approaches as teacher and student models. Finally, by combining these two encoding approaches DEDR learns a symmetric dual uni-modal/multi-modal encoder for both queries and documents.

4.1 Unified Uni-Modal Encoding

In order to use a shared uni-modal (textual) encoder for representing both queries and passages, we need to convert the image in the

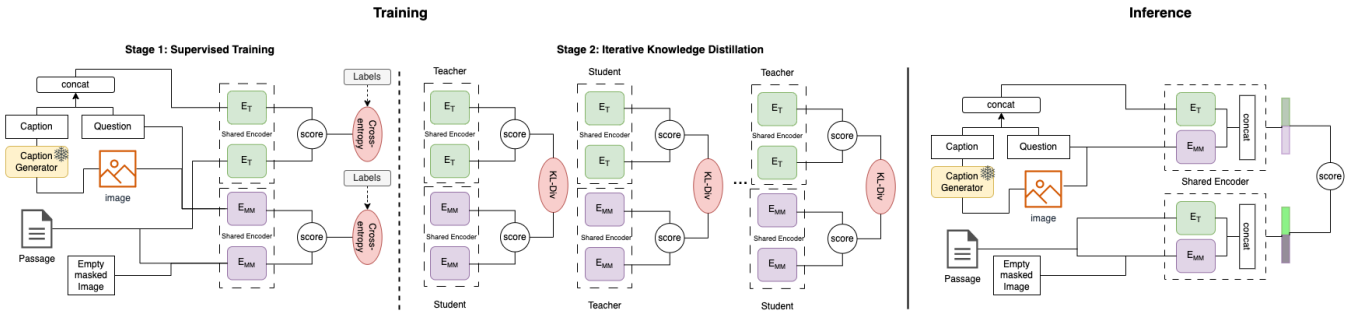

Figure 3: The training and inference procedure in the DEDR framework. DEDR first trains uni-modal and multi-modal encoders in isolation (left), then uses iterative knowledge distillation to adjust both representation spaces (middle). At inference, DEDR uses the aggregation of both encodings to construct a symmetric dual encoding dense retriever (right).

query to text. This model can be formulated as follows:

$$S_T((Q, I), P) = E_T(\text{concat}(Q, \phi_{I \to T}(I))) \cdot E_T(P) \tag{1}$$

where $\cdot$ denotes dot product between two vectors, E_T is a uni-modal text encoder, and $\phi_{I \to T}$ is a modality converting module that takes an image and produces a textual description for it.

There are several approaches to implement the modality converter $\phi_{I \to T}$. One approach is to generate the name of objects that are in the image using object detection approaches. We take an alternative approach by using image captioning objectives to train the modality converter model $\phi_{I \to T}$. The reason is that image captioning approaches produce open-ended descriptions of images, as opposed to predefined categories in object detection models. In addition, collecting training data for image captioning models is cheap, given the availability of large-scale images with captions on the web. In more detail, we use the ExpansionNet v2 [18] architecture to implement $\phi_{I \to T}$. Expansion V2 is an encoder-decoder architecture designed on top of the Swin-Transformer [33] extended by Block Dynamic and Static Expansion [18] and multi-head attention [55]. The model is first pre-trained using images and captions from the Microsoft COCO dataset [31]. Then, the self-critical optimization [46] is performed to complete the model's training. Once the model is trained for producing textual descriptions of images, we freeze the ExpansionNet v2's parameters and only optimize the text encoder E_T. This substantially reduces the number of parameters that need to be learned using the KI-VQA training set.

For implementing the text encoder E_T, there are numerous language models available. In our experiments, we use BERT-base [8] and the representation associated with the `[CLS]` token is considered as the output of the encoder E_T. Note that in Equation (1), query and passage encoders are the same and they use shared parameters, guaranteeing a symmetric architecture for dense retrieval.

4.2 Unified Multi-Modal Encoding

Even though using a multi-modal encoder for both query and passage encoding seems straightforward, most multi-modal language models do not accept text-only inputs. Therefore, we need to develop a technique to fill this modality gap. Our unified multi-modal encoding approach can be formulated as follows:

$$S_{MM}((Q, I), P) = E_{MM}(Q, I) \cdot E_{MM}(P, I_{\texttt{[MASKED]}}) \tag{2}$$

where E_{MM} is a pre-trained multi-modal encoder that represents a pair of text and image as input. To address the modality gap between the query and passage sides, we use a multi-modal language model that has used the masking technique in the visual side during pre-training. In more detail, we use LXMERT [53] that uses Faster R-CNN [45] to recognize 36 objects in the given image and generate their representations and bounding boxes. Then, this information about objects in addition to the text tokens are fed to a dual-encoder transformer network with a cross-modality encoder on top. Finally, the `[CLS]` token is used to represent the whole input. Since LXMERT has been pre-trained with different pre-training objectives, including Masked Object Prediction, it is a perfect fit for our unified multi-modal representation learning.

In order to overcome the aforementioned problem with encoding textual only data for passage representation, we propose a simple yet effective technique that we call Passage Expansion using Masked Image Representation (PEMIR). In this technique, we feed a passage to LXMERT as the textual input and zero (masked) as the visual input with bounding boxes of $[0.0, 0.0, 1.0, 1.0]$. This visual input means 36 masked objects with bounding boxes of the whole image. Intuitively, we ask the model to generate a representation for the input passage while trying to generate the best visual object representations based on the textual input data. Thus, this is roughly equivalent to expanding the passage with image representation based on the passage content.

Using this approach, we can generate the representation for queries and documents using the same multi-modal encoder with shared parameters. This is beneficial because it helps the dual encoder architecture start from the same shared embedding space. Additionally, we can use only a single encoder as both query and document encoder, which results in decreasing the number of parameters of the model.

4.3 Dual Encoding Optimization via Iterative Knowledge Distillation

E_T and E_{MM} represent the KI-VQA inputs from different perspectives; E_T only considers the textual representation of the inputs, while E_{MM} considers their multi-modal representations. These models also use unique pre-training data and objectives. Hence, we hypothesize that these two representation learning models provide complementary information. Our empirical analysis validates this

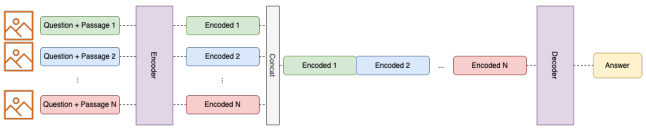

Figure 4: The architecture of MM-FiD. It uses multi-modal encoder to encode each question-image-passage triplet separately and then concatenates their encodings as input to the decoder for knowledge aggregation and answer generation.

hypothesis, see Figure 5. Given this observation, we propose to use a knowledge distillation approach to improve both of these models.

Isolated Training of Encoders. We first optimize the parameters of these two models separately using the retrieval training set for each KI-VQA task. Following DPR [23], we use a contrastive loss function based on cross-entropy to train the models:

$$L_{\text{isolated}} = -\log \frac{e^{S_X((Q,I),P_{pos})}}{e^{S_X((Q,I),P_{pos})} + \sum_{P' \in \mathbf{P_{neg}}} e^{S_X((Q,I),P')}} \quad (3)$$

where $X \in \{T, MM\}$ is the encoding approach used to generate scores, and P_{pos} is a positive (relevant) passage and $\mathbf{P_{neg}}$ is a set of negative passages for the question-image pair (Q, I). To construct $\mathbf{P_{neg}}$, we use a hard negative sampling approach used in [40] in addition to in-batch negatives – in which all the positive and negative documents of other queries in the training batch are considered as negative documents to the query.

Iterative Knowledge Distillation among Encoders. In order to adjust the representations space in E_T and E_{MM} and improve their generalization, we design an iterative knowledge distillation approach. In this method, we first use the more effective encoder, based on the performance of the validation set, as the teacher and the other encoder as the student. Then, we train the student using the scores provided by the teacher. We use the following listwise KL-divergence loss function:

$$L_{\text{IKD}} = -\sum_{P' \in \mathbf{P_{neg}} \cup \{P_{pos}\}} S'_Y((Q,I),P') \log \frac{S'_X((Q,I),P')}{S'_Y((Q,I),P')} \quad (4)$$

where X and Y respectively denote the student and teacher model. $S'((Q,I),P')$ is the normalized score of $S((Q,I),P')$ for $P' \in \mathbf{P_{neg}} \cup \{P_{pos}\}$ generated by the student or teacher model using the softmax function. Similar to Equation (3), P_{pos} is a positive passage, and the same method is used for negative sampling. We continue the training of the student using the teacher's scores until a stopping criterion is met: either the computing budget finishes (i.e., it reaches the maximum number of epochs set in the experiment) or early stopping based on validation performance.

In the next round of distillation, we swap the teacher and the student. In other words, the student of the previous round acts as the teacher in this round to provide scores for the previous teacher's training in this round. This iterative approach to knowledge distillation is helpful because, in each round, the student model learns the perspective of the teacher model in scoring documents, especially when these two models rely on two different embedding spaces to generate scores. We continue this iterative distillation process and use early stopping based on validation performance to terminate.

4.4 Retrieval Inference using Dual Encoding

As we mentioned earlier, E_T and E_{MM}, introduced in previous sections, retrieve passages in response to a multi-modal query using separate embedding spaces. The former focuses on textual embedding space, while the latter encodes queries and passages into a multi-modal embedding space. An idea to combine these two models in a single embedding space is to concatenate each model's representation for its inputs together. If each of these encoders produce a d-dimensional encoding for each input, the final embedding space consists of $2d$ dimensions.

There are two methods to use this combined embedding space: (1) we can use the concatenated representation of the models to train a new ranker from scratch using the loss function in Equation 3, and (2) we can use the best rankers of each type after knowledge distillation and combine their representations without further training to generate representations for queries and passages. The latter does not need training because each model has been trained previously. We just combine their representations to index the embeddings using Faiss [22] and search the index to retrieve passages.

5 MULTI-MODAL FUSION-IN-DECODER

Fusion-in-Decoder (FiD) [20] is a state-of-the-art generative encoder-decoder model based on T5 [43]. It is intended to aggregate knowledge across multiple passages and generate a single response based on them. It has produced strong performance on a wide range of knowledge-intensive language tasks (KILTs). The current state-of-the-art model on six benchmarks from the KILT leaderboard[6] is based on a light variant of FiD [16]. We extend FiD architecture to multi-modal data and propose MM-FiD.

To formalize the task, for a query consisting of an image I and a question Q, we assume a set $P = \{p_1, ..., p_n\}$ including n supporting passages is provided (e.g., by a retrieval model). The goal of the multi-modal fusion-in-decoder (MM-FiD) is to generate a textual answer to Q about the I by considering all passages in the P.

We use VL-T5 [6] architecture, a multi-modal text generative visual-language model pre-trained with different text generation tasks, as a start point to design multi-modal fusion-in-decoder architecture. VL-T5 takes a piece of text and image objects' features detected by Faster R-CNN [45] as input and generates a piece of text as output. The simplest solution to the mentioned problem formulation in this section is to feed VL-T5 with the concatenation of the question, image, and each passage to generate an answer based on each supporting document; however, the aggregation and reduction of the generated answers for each document are challenging because the model might generate a different answer for each document. Another approach to solving the problem is to concatenate the question, image, and all documents and feed it to VL-T5 to generate a single answer based on them; this approach suffers from two shortcomings: (1) concatenating all passages results in a long sequence, which decreases the speed of the model and increases the memory consumption and may also reach the maximum token limit, and (2) concatenation of different passages together makes it hard for the model to consider the context of each passage because the passage set P can be about different unrelated subjects.

[6] https://eval.ai/web/challenges/challenge-page/689/leaderboard/

MM-FiD uses an alternative approach, shown in Figure 4. In this architecture, the question and image are concatenated with each document in the supporting set independently and fed to the encoder of VL-T5 to be encoded. Then, the encoded representation of each question, image, and each passage is concatenated together to be fed to the decoder of VL-T5. In this approach, the image, question, and each document are encoded separately, which helps the model decrease memory use and consider the context better than the two other alternatives mentioned above. Additionally, concatenating the encoded representations of all passages at the decoder helps the model consider the information in all documents for generating a single answer to the question about the image.

In order to train the MM-FiD for text generation, we use the cross-entropy loss function using the following formulation:

$$L_{\text{mm-fid}} = -\sum_{k} \log P(y_k | y_{i<k}, \{I, Q, p_1\}, \{I, Q, p_2\}, ..., \{I, Q, p_n\})$$

where y_k is k^{th} output token, I is the image, Q is the question, and p_i is the i^{th} supporting (e.g., retrieved) passage.

6 EXPERIMENTS

6.1 Datasets

Outside-Knowledge Visual Question Answering (OK-VQA) [38]: This dataset consists of triplets, including an image, a question about the image, and an answer to the mentioned question. Answering most of the questions in this dataset needs a piece of information that is not provided in the image. Therefore, accessing an external source of information is required for this task. A retrieval dataset based on a Wikipedia dump[7] with 11 million passages was later constructed by Qu et al. [40], which we use to train and evaluate our retrievers. This dataset contains 9009 questions for training, 2523 questions for validation, and 2523 for testing [41]. We also use original OK-VQA dataset to evaluate the performance of our end-to-end retrieval and answer generation pipeline.

Fact-based Visual Question Answering (FVQA) [56]: Each data point in this dataset consists of an image, a question about the image, the answer, and a fact supporting the answer. This dataset has also provided an unstructured knowledge source containing all the facts (i.e., sentences) we need to answer the question about the image. We use 70% of samples (4077) in this dataset for train set and augment them similar to Qu et al. [40] with five hard negatives retrieved by BM25, 15% for validation (874), and 15% for test set (874). We use the original FVQA dataset to evaluate the performance of our end-to-end retrieval and answer generation pipeline.

6.2 Experimental Setup

Retriever Training Setup. In our experiments, we use the Adam optimizer [26] with a batch size of 16 and a learning rate of 10^{-5}. We use a linear learning rate scheduler with 10% of total steps as warm-up steps. We also use gradient clipping with the value of 1.0. The maximum input length of each encoder is set to 400 tokens. We train each model for 2 epochs on OK-VQA and 4 epochs on FVQA. All the experiments are conducted on a machine with a Nvidia RTX8000 GPU with 49GB of memory and 256GB of RAM. We use Faiss [22] to index the learned embeddings with a flat index for efficient dense retrieval. For BM25, Pyserini is used.

Multi-Modal Fusion-in-Decoder Training Setup. We use the AdamW optimizer with a batch size of 1 with 32 gradient accumulation steps, which results in an effective batch size of 32. We utilize a learning rate of 5×10^{-5} and weight decay of 0.1 for training the MM-FiD model. Given the training dataset sizes, we use a linear learning rate scheduler with 800 and 200 warm-up steps for the OK-VQA and FVQA datasets, respectively. We train the model for 5000 (OK-VQA) and 2000 (FVQA) gradient update steps. We create a checkpoint of the model every 500 training steps and select the best checkpoint based on its performance on the validation set. We also use gradient clipping at 1.0 for training. The maximum encoder's input length of each question and passage pair is set to 420 (OK-VQA) and 64 (FVQA) tokens. Since the answers in both datasets are short, we set the MM-FiD's output length to 16. MM-FiD's decoder uses beam search [58] with a beam size of 2 for answer generation. We train MM-FiD using 32 (OK-VQA) and 5 (FVQA) supporting passages for each question and image pair, where the supporting passages are retrieved using the proposed DEDR model. The MM-FiD experiments are conducted on a machine with a single Nvidia RTX8000 GPU with 49GB of memory and 128GB of RAM. Following [56], we use five-fold cross-validation for the FVQA dataset. Note that we train an individual DEDR for each fold to avoid data leaks.

Evaluation Metrics. To be consistent with the literature, we use the common metrics suggested for each dataset. Following Qu et al. [40], we use mean reciprocal rank (MRR) and precision of the top five retrieved documents (MRR@5 and P@5) for evaluating the retrieval models on the OK-VQA dataset. We use MRR@5 and P@1 for retrieval evaluation on the FVQA dataset. Since the FVQA dataset provides only one ground truth passage per question, we use Precision@1 as the evaluation metric. We use a two-tailed paired t-test with Bonferroni correction to identify statistically significant improvements (p-value < 0.05).

For the evaluation of the answer generation model for the OK-VQA dataset, we follow the official evaluation script provided by Marino et al. [38]. It uses the evaluation metric for VQA [3] task, which relies on human annotations. For evaluation on the FVQA dataset, we follow Wang et al. [56] and use Top-1 Accuracy or Exact Match (EM) as the evaluation metric, in which we lowercase answers and remove articles (e.g. a, an, the) and punctuation.

6.3 Passage Retrieval Results for KI-VQA Tasks

Baselines. We compare the proposed dense retrieval framework with the following baselines:

- **Sparse (Term Matching) Retrieval Models:** We use two sparse retrieval baselines: (1) **BM25:** this baseline uses the BM25 formulation [47] with questions as queries, ignoring the images, and passages as documents. (2) **BM25-Obj (CombMax):** this approach extracts 36 objects from the image (objects are generated by a Faster R-CNN [45] model pre-trained on Visual Genome [2, 27]) and concatenates each object's name to the question as the query and uses the BM25 formulation to retrieve passages. Then it uses CombMax [9, 28] to aggregate these 36 ranked lists.

[7] This Wikipedia collection is available at https://ciir.cs.umass.edu/downloads/ORConvQA/all_blocks.txt.gz.

Table 1: Passage retrieval performance for KI-VQA tasks on OK-VQA and FVQA datasets. The superscript * denotes statistically significant improvement compared to all the baselines based on two-tailed paired t-test with Bonferroni correction ($p < 0.05$).

Dataset	OK-VQA				FVQA			
Model	Validation		Test		Validation		Test	
	MRR@5	P@5	MRR@5	P@5	MRR@5	P@1	MRR@5	P@1
Sparse Retrievers								
BM25	0.2565	0.1772	0.2637	0.1755	0.3368	0.2700	0.3509	0.2848
BM25-Obj (CombMax)	0.3772	0.2667	0.3686	0.2541	0.3903	0.3272	0.4057	0.3421
Symmetric Single-Encoding Dense Retrievers								
Dense-BERT (question, passage)	0.4555	0.3155	0.4325	0.3058	0.3860	0.3089	0.3836	0.3020
Dense-BERT (question + caption, passage) → Our E_T	0.5843	0.4445	0.5797	0.4420	0.4409	0.3558	0.4292	0.3409
Dense-LXMERT → Our E_{MM}	0.5722	0.4276	0.5465	0.4066	0.4293	0.3478	0.4269	0.3409
Asymmetric Dual-Encoding Dense Retrievers								
BERT-LXMERT	0.4704	0.3364	0.4526	0.3329	0.1455	0.1006	0.1477	0.1029
Symmetric Dual-Encoding Dense Retrievers								
DEDR	**0.6260***	**0.4890***	**0.6469***	**0.5059***	**0.5833***	**0.4931***	**0.5618***	**0.4713***
% relative improvement w.r.t. the best baseline	7.1% ↑	10.0% ↑	11.6% ↑	14.5% ↑	32.3% ↑	38.6% ↑	30.9% ↑	37.8% ↑

Qu et al. [40] explored other rank aggregation approaches as well and CombMax was found to be the best solution for this task.

- **Symmetric Single-Encoding Dense Retrieval Models:** We use three baselines in this category: (3, 4) **Dense-BERT:** a BERT-based dense retrieval model (similar to DPR [23]) with the same training objective as ours. We provide the results for two variations of this model, an image-independent approach whose query encoder only encodes the question, and an image-dependent approach whose query encoder takes the concatenation of the question and the image caption (captions are generated by ExpansionNet v2 [18]). The latter is the same as our E_T encoder. (5) Dense-LXMERT is a model that uses a multi-modal encoder, i.e., LXMERT, to encode queries and passages. It uses masked image tokens on the passage side. This approach is our E_{MM} encoder.
- **Asymmetric Dual-Encoding Dense Retrieval Models:** In this category, we use BERT-LXMERT, proposed in [40], that uses BERT for passage encoding and LXMERT for query encoding.[8]

For fair comparison, we use the same training and evaluation process for all (our and baseline) models. To the best of our knowledge, our baseline results are the highest reported in the literature.

Comparison Against Retrieval Baselines. The passage retrieval results are reported in Table 1. We observe that dense retrieval models generally outperform sparse retrieval baselines, confirming our design choice to focus on dense retrieval for KI-VQA tasks. Interestingly, sparse retrieval models achieve higher MRR on FVQA than on OK-VQA, while dense retrieval models perform significantly better on the OK-VQA dataset. This observation suggests that relevant documents in FVQA perhaps have higher term overlap with the questions and image objects. The results show that image-independent models (i.e., BM25 and Dense-BERT that only encodes the question) underperform their own variant with image information (i.e., generated objects or captions). This highlights the importance of representing images in KI-VQA tasks. Furthermore, Table 1 shows that the asymmetric dense retrieval baseline (BERT-LXMERT) does not perform as well as symmetric dense retrieval baselines. In particular, BERT-LXMERT shows a poor performance

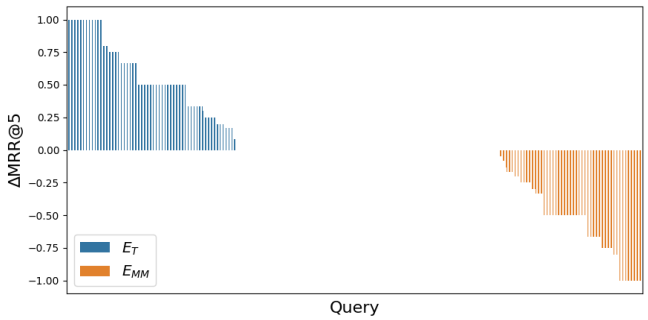

Figure 5: Difference between reciprocal rank (RR) obtained by E_T and E_{MM} for each query on the OK-VQA test set. The blue / orange color denotes the queries where E_T / E_{MM} wins.

on the FVQA dataset. That can be due to the smaller training set in FVQA,[9] as asymmetric models often require more training data. Another observation from these results is that performance on validation and test sets are relatively close, therefore it is safe to argue that performances on the validation sets are generalizable to the test sets and no aggressive overfitting is observed. The results show that our own encoders E_T and E_{MM} are the best performing baselines. We conducted an experiment to see if these two encoding approaches provide complementary information. To this aim, we compute the reciprocal rank (RR) obtained by E_T and E_{MM} for each query in OK-VQA and plot their differences in Figure 5. For better visualization, this figure sorts the queries with respect to their ΔMRR in descending order. Figure 5 shows that E_T and E_{MM} perform similarly for about half of the queries, while E_T performs better for about 25% of queries and E_{MM} performs better for the other ~ 25%. **This shows that these two encoders contain complementary information and their aggregation can improve the results – confirming the motivation of designing DEDR.**

Results obtained by DEDR suggest statistically significant improvements compared to all the baselines. We achieve 11.6% higher MRR and 14.5% higher P@5 than the best baseline (including our own encoders E_T and E_{MM}) on the OK-VQA test set and 30.9%

[8]The original paper [40] refers to this baseline as Dense-LXMERT. We rename it to BERT-LXMERT to avoid confusion.

[9]FVQA not only has fewer training questions, but only has a single relevant passage per question.

Table 2: Ablation study for iterative knowledge distillation in DEDR. The superscript * denotes statistically significant improvement compared to all the ablation cases based on two-tailed paired t-test with Bonferroni correction ($p < 0.05$).

Dataset	OK-VQA				FVQA			
Model	Validation		Test		Validation		Test	
	MRR@5	P@5	MRR@5	P@5	MRR@5	P@1	MRR@5	P@1
DEDR (joint encoders, no KD)	0.5930	0.4507	0.5405	0.4263	0.4669	0.3764	0.4498	0.3684
DEDR (isolated encoders, no KD)	0.6236	0.4771	0.6330	0.4972	0.5551	0.4668	0.5313	0.4439
DEDR	**0.6260**	**0.4890***	**0.6469***	**0.5059***	**0.5833***	**0.4931***	**0.5618***	**0.4713***

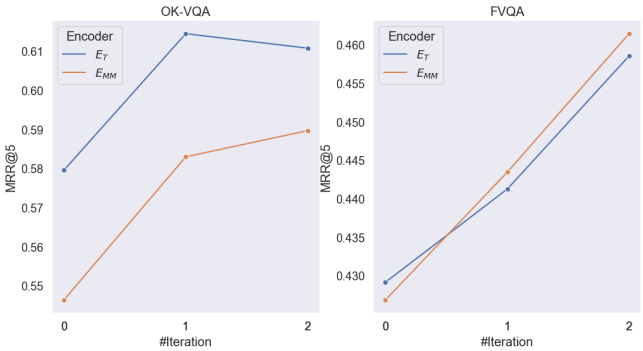

Figure 6: DEDR performance at different iterations of the proposed iterative knowledge distillation approach on the test set of both OK-VQA and FVQA datasets.

higher MRR and 37.8% higher P@1 on the FVQA test set. We believe that our symmetric dual encoding approach works well without requiring a large scale training set, which justifies substantially larger gain on the FVQA dataset. Note that DEDR also uses BERT and LXMERT and the obtained improvements are not due to larger model parameters or different pretraining. However, compared to Dense-BERT and Dense-LXMERT, DEDR has the ability to take advantage of knowledge from both uni- and multi-modal language models. BERT-LXMERT, however, could not take advantage of such extra implicit "knowledge" effectively due to its asymmetric design.

DEDR Ablations. To empirically study the impact of each decision we made in designing DEDR, we report ablation results in Table 2. The first row of the table is associated with a model that jointly used both encoders (E_T and E_{MM}) at both training and evaluation. The second row, on the other hand, train each encoder separately until convergence and only concatenates them at inference. There is no knowledge distillation in either of these two models. The results show that DEDR outperforms both of these models, demonstrating the effectiveness of the proposed iterative knowledge distillation approach for dual encoding. The improvements are statistically significant in nearly all cases, except for MRR@5 on the OK-VQA validation set. We further plot the retrieval performance at each knowledge distillation training step in Figure 6. We observe that in the first three iterations the performance of both encoders generally increases on both datasets. Models show different behavior on the different datasets in Figure 6. This shows that the number of iterations in the knowledge distillation approach is dataset-dependent and should be tuned for best results.

6.4 Question Answering Results for KI-VQA

In this section, we report and discuss the end-to-end retrieval and answer generation results. A wide range of question answering methods has been applied to KI-VQA tasks. Not all of these methods are publicly available and not all of them use the same knowledge source. We compare our methods to the best performing models in the literature with relatively similar model size. Note that MM-FiD contains 220 million parameters. In our first set of experiments, we also exclude the models that use GPT-3 (175 billion parameters) as an implicit "knowledge" source. The results are reported in Table 3. As mentioned in the table, different approaches on OK-VQA use different knowledge sources, such as ConceptNet, Wikidata, Wikipedia, Google Search, and Google Images. The FVQA dataset released a fact corpus which is used by several models. We observe that MM-FiD that uses passages retrieved by DEDR for answer generation outperforms all the baselines listed in Table 3. We observe 8.5% improvements on FVQA compared to the best performing baseline. This table also includes the results for MM-FiD without any supporting document (i.e., without retrieval). We observe that the MM-FiD is able to produce competitive performance even without utilizing retrieval results. The reason is that these large language models contain a lot of information in their parameters from pretraining phase and they can answer many questions based on their internal implicit "knowledge". MM-FiD without retrieval even outperforms all the baselines on OK-VQA. Note that the number of parameters in MM-FiD (220M) is comparable to the baselines. The results suggest that employing retrieval results leads to larger gain on FVQA than on OK-VQA, possibly due to the nature of the questions in fact-based visual question answering.

Note that some models use the outputs produced by GPT-3 as a knowledge source and often outperform those models above that use explicit knowledge. However, it is difficult to draw conclusions, as the GPT-3 training set is unknown and it has 175B parameters, making it extremely expensive to run and not truly comparable to any other model mentioned above in terms of capacity. That being said, we do compare our model against state-of-the-art baselines with comparable model size that use GPT-3's output as supporting evidence. The results on the OK-VQA dataset are reported in Table 4.[10] When our model uses GPT-3's output in addition to passages, it still outperforms its alternatives, but with a smaller margin, highlighting the impact of document quality in KI-VQA tasks.

Question Answering Ablations and Analysis. For a deeper understanding of the proposed answer generation solution, we conduct careful ablation studies whose results are reported in Table 5. The results on both datasets suggest that when using the same retriever (i.e., DEDR), MM-FiD outperforms its uni-modal variation, FiD [20], that has been also used for KI-VQA tasks, for example in KAT [13]. Moreover, Table 5 demonstrates the impact of different retrieval models on the final answer generation. For example,

[10]We do not have access to the GPT-3's output for FVQA questions.

Table 3: Question answering performance for models with explicit knowledge source on both OK-VQA and FVQA datasets. All these models are relatively comparable in terms of model size and contain less than 1 billion parameters.

OK-VQA			FVQA		
Model	**Knowledge Source**	**Acc.**	**Model**	**Knowledge Source**	**Top-1 Acc.**
KAT-base [13]	WikiData	40.93	BAN [25]	None	35.69
RVL [50]	ConceptNet	39.04	RVL [50]	ConceptNet	54.27
MAVEx [57]	Wikipedia, ConceptNet, Google Images	40.28	BAN [25] + KG-AUG [29]	None	38.58
UnifER+ViLT [14]	ConceptNet	42.13	UnifER + ViLT [14]	FVQA Corpus	55.04
VRR [35]	Google Search	39.20	Top1-QQmaping [56]	FVQA Corpus	52.56
LaKo-base [5]	ConceptNet, DBPedia, WebChild	42.21	Top3-QQmaping [56]	FVQA Corpus	56.91
MM-FiD	None	42.82	MM-FiD	None	52.78
DEDR +MM-FiD	Wikipedia	**44.57**	DEDR +MM-FiD	FVQA Corpus	**61.80**
% relative improvement w.r.t. the best baseline		5.5% ↑	% relative improvement w.r.t. the best baseline		8.5% ↑

Table 4: QA performance of SOTA models that rely on the output of GPT-3 as a "knowledge source" on OK-VQA.

Model	Knowledge source	Accuracy
PICa-full [60]	Frozen GPT-3	48.00
KAT-base [13]	Frozen GPT-3, Wikidata	50.58
DEDR + MM-FiD	Frozen GPT-3, Wikipedia	**51.02**

Table 5: Ablation study of the the proposed KI-VQA pipeline. The superscript * denotes statistically significant improvement compared to all the ablation cases based on two-tailed paired t-test with Bonferroni correction ($p < 0.05$).

Retriever	Answer Generator	OK-VQA Acc	FVQA Top-1 Acc
DEDR	FiD	39.48	60.85
BM25-Obj (CombMax) [40]	MM-FiD	41.97	54.65
Best retrieval baseline from Table 1	MM-FiD	41.82	52.78
DEDR	MM-FiD	**44.57***	**61.80***

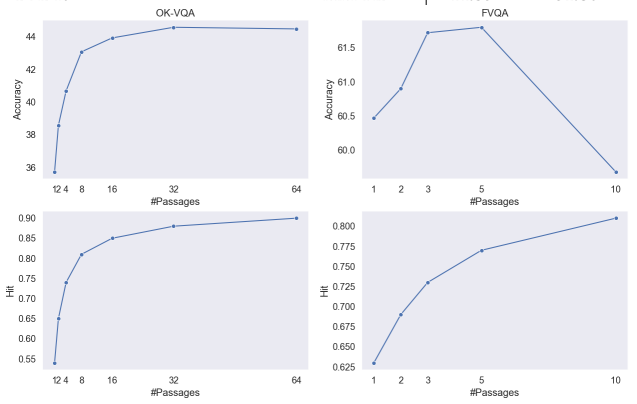

Figure 7: MM-FiD accuracy and the hit ratio in the supporting passages at different ranking cut-off levels.

using DEDR for retrieval instead of the best performing retrieval baseline from Table 1 would lead to 13% higher accuracy in FVQA, highlighting the importance of retrieval in the KI-VQA pipeline.

Figure 7 plots the sensitivity of MM-FiD performance to the number of passages retrieved by DEDR. It also plots the hit ratio, i.e., the ratio of success retrieving at least one relevant passage to be presented to MM-FiD. Generally speaking, the more documents we feed to MM-FiD on OK-VQA, the higher the question answering accuracy. Its accuracy becomes relatively stable after retrieving 16 documents. The accuracy curve follows the same behavior as the hit curve on OK-VQA. However, FVQA demonstrates a substantially different behavior. The highest accuracy is achieved when only five supporting passages are retrieved. That is due to the nature of the dataset, where there is only one relevant fact for each question and retrieving more (potentially inaccurate) facts may confuse the answer generation model. Note that DEDR reaches a hit ratio of 70% by only retrieving two passages on FVQA, while the same model needs to retrieve 4 passages to reach the same level of hit ratio on OK-VQA. This finding suggest that it is worth studying automatic prediction of ranking cut-off for KI-VQA tasks in the future.

7 CONCLUSIONS AND FUTURE WORK

This paper presented DEDR, a novel symmetric dense retrieval framework based on dual uni-modal and multi-modal encoding. We propose an iterative knowledge distillation approach for updating these two encoding representation spaces and aggregating them at inference. It also proposed MM-FiD, an extension to the fusion-in-decoder architecture [20] for multi-modal data. Extensive experiments on two well-established datasets, OK-VQA and FVQA, suggested that retrieving passages using DEDR and using them to generate answers via MM-FiD substantially outperforms state-of-the-art baselines with comparable capacity. For instance, this approach led to 37.8% retrieval improvement in terms of P@1 and 8.5% exact match accuracy improvement on FVQA test set compared to the best performing baselines. We demonstrated the impact of every design decision we made in both DEDR and MM-FiD through extensive ablation studies and highlight open areas for future explorations. For example, our results suggest accurate prediction of 'when to retrieve' is an impactful area for KI-VQA tasks. Hence, exploring retrieval performance prediction and ranking cut-off truncation in KI-VQA tasks can potentially be a fruitful future direction. We also intend to explore universal knowledge-intensive models for both textual and multi-modal inputs. We further plan to expand the applications of knowledge-intensive multi-modal tasks beyond question answering.

ACKNOWLEDGMENT

This work was supported in part by the Center for Intelligent Information Retrieval, in part by NSF grant #2106282, and in part by Lowe's. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Bilal Abu-Salih. 2021. Domain-specific knowledge graphs: A survey. *Journal of Network and Computer Applications* 185 (2021), 103076. https://doi.org/10.1016/j.jnca.2021.103076

[2] Peter Anderson, X. He, C. Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 6077–6086.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

[4] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. https://doi.org/10.48550/ARXIV.1704.00051

[5] Zhuo Chen, Yufeng Huang, Jiaoyan Chen, Yuxia Geng, Yin Fang, Jeff Z. Pan, Ningyu Zhang, and Wen Zhang. 2022. LaKo: Knowledge-driven Visual Question Answering via Late Knowledge-to-Text Injection. *CoRR* abs/2207.12888 (2022).

[6] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 1931–1942.

[7] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* 41 (1990), 391–407.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[9] Edward A. Fox and Joseph A. Shaw. 1993. Combination of Multiple Searches. In *Text Retrieval Conference*.

[10] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-Retrieve-Generate: Natural Language-Centric Outside-Knowledge Visual Question Answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5057–5067. https://doi.org/10.1109/CVPR52688.2022.00501

[11] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2020. Modularized Transfomer-based Ranking Framework. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 4180–4190. https://doi.org/10.18653/v1/2020.emnlp-main.342

[12] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. ConceptBert: Concept-Aware Representation for Visual Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 489–498. https://doi.org/10.18653/v1/2020.findings-emnlp.44

[13] Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. 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*. Association for Computational Linguistics, Seattle, United States, 956–968. https://doi.org/10.18653/v1/2022.naacl-main.70

[14] Yangyang Guo, Liqiang Nie, Yongkang Wong, Yibing Liu, Zhiyong Cheng, and Mohan Kankanhalli. 2022. A Unified End-to-End Retriever-Reader Framework for Knowledge-Based VQA. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 2061–2069. https://doi.org/10.1145/3503161.3547870

[15] Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. ManyModalQA: Modality Disambiguation and QA over Diverse Inputs. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 7879–7886. https://doi.org/10.1609/aaai.v34i05.6294

[16] Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2022. FiD-Light: Efficient and Effective Retrieval-Augmented Text Generation. *ArXiv* abs/2209.14290 (2022).

[17] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 113–122. https://doi.org/10.1145/3404835.3462891

[18] Jia Cheng Hu, Roberto Cavicchioli, and Alessandro Capotondi. 2022. ExpansionNet v2: Block Static Expansion in fast end to end training for Image Captioning. https://doi.org/10.48550/ARXIV.2208.06551

[19] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkxgnnNFvH

[20] Gautier Izacard and Edouard Grave. 2021. 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*. Association for Computational Linguistics, Online, 874–880. https://doi.org/10.18653/v1/2021.eacl-main.74

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *International Conference on Machine Learning*.

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

[23] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550

[24] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 39–48. https://doi.org/10.1145/3397271.3401075

[25] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/96ea64f3a1aa2fd00c72faacf0cb8ac9-Paper.pdf

[26] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. https://doi.org/10.48550/ARXIV.1412.6980

[27] R. Krishna, Yuke Zhu, O. Groth, J. Johnson, Kenji Hata, J. Kravitz, Stephanie Chen, Yannis Kalantidis, L. Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123 (2016), 32–73.

[28] Joon Ho Lee. 1997. Analyses of Multiple Evidence Combination. *SIGIR Forum* 31, SI (jul 1997), 267–276. https://doi.org/10.1145/278459.258587

[29] Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting Visual Question Answering with Context-aware Knowledge Aggregation. *Proceedings of the 28th ACM International Conference on Multimedia* (2020).

[30] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. In-Batch Negatives for Knowledge Distillation with Tightly-Coupled Teachers for Dense Retrieval. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*. Association for Computational Linguistics, Online, 163–173. https://doi.org/10.18653/v1/2021.repl4nlp-1.17

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

[32] Weizhe Lin and Bill Byrne. 2022. Retrieval Augmented Visual Question Answering with Outside Knowledge. https://doi.org/10.48550/ARXIV.2210.03809

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 10012–10022.

[34] Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2022. Universal Multi-Modality Retrieval with One Unified Embedding Space. https://doi.org/10.48550/ARXIV.2209.00179

[35] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-Supervised Visual-Retriever-Reader for Knowledge-based Question Answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6417–6431. https://doi.org/10.18653/v1/2021.emnlp-main.517

[36] Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Efficient Document Re-Ranking for Transformers by Precomputing Term Representations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, China) *(SIGIR '20)*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/3397271.3401093

[37] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Kumar Gupta, and Marcus Rohrbach. 2020. KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 14106–14116.

[38] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT:

a Benchmark for Knowledge Intensive Language Tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 2523–2544. https://doi.org/10.18653/v1/2021.naacl-main.200

[40] Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik Learned-Miller. 2021. Passage Retrieval for Outside-Knowledge Visual Question Answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 1753–1757. https://doi.org/10.1145/3404835.3462987

[41] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5835–5847. https://doi.org/10.18653/v1/2021.naacl-main.466

[42] 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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[43] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21, 140 (2020), 1–67. http://jmlr.org/papers/v21/20-074.html

[44] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A Joint Training Method for Dense Passage Retrieval and Passage Re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2825–2835. https://doi.org/10.18653/v1/2021.emnlp-main.224

[45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'15)*. MIT Press, Cambridge, MA, USA, 91–99.

[46] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2016. Self-Critical Sequence Training for Image Captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 1179–1195.

[47] Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1995. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC-3)*. Gaithersburg, MD: NIST, 109–126.

[48] Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. 2023. Image captioning for effective use of language models in knowledge-based visual question answering. *Expert Systems with Applications* 212 (2023), 118669. https://doi.org/10.1016/j.eswa.2022.118669

[49] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 3715–3734. https://doi.org/10.18653/v1/2022.naacl-main.272

[50] Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. 2021. Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge. In *Proceedings of the Third Workshop on Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*. Association for Computational Linguistics, Kyiv, Ukraine, 1–18. https://aclanthology.org/2021.lantern-1.1

[51] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. 2021. MIMOQA: Multimodal Input Multimodal Output Question Answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, 5317–5332. https://doi.org/10.18653/v1/2021.naacl-main.418

[52] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultiModal{QA}: complex question answering over text, tables and images. In *International Conference on Learning Representations*. https://openreview.net/forum?id=ee6W5UgQLa

[53] Hao Tan and Mohit Bansal. 2019. 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)*. Association for Computational Linguistics, Hong Kong, China, 5100–5111. https://doi.org/10.18653/v1/D19-1514

[54] Jizhi Tang, Yansong Feng, and Dongyan Zhao. 2019. Learning to Update Knowledge Graphs by Reading News. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 2632–2641. https://doi.org/10.18653/v1/D19-1265

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[56] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. FVQA: Fact-Based Visual Question Answering. *IEEE Trans. Pattern Anal. Mach. Intell.* 40, 10 (oct 2018), 2413–2427. https://doi.org/10.1109/TPAMI.2017.2754246

[57] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-Modal Answer Validation for Knowledge-Based VQA. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 3 (Jun. 2022), 2712–2721. https://doi.org/10.1609/aaai.v36i3.20174

[58] Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv* abs/1609.08144 (2016).

[59] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. In *International Conference on Learning Representations*. https://openreview.net/forum?id=zeFrfgyZln

[60] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2021. An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA. In *AAAI Conference on Artificial Intelligence*.

[61] Hamed Zamani, Fernando Diaz, Mostafa Dehghani, Donald Metzler, and Michael Bendersky. 2022. Retrieval-Enhanced Machine Learning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 2875–2886. https://doi.org/10.1145/3477495.3531722

[62] Hansi Zeng, Hamed Zamani, and Vishwa Vinay. 2022. Curriculum Learning for Dense Retrieval Distillation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 1979–1983. https://doi.org/10.1145/3477495.3531791

[63] Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing Dense Retrieval Model Training with Hard Negatives. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021).