Transform-Retrieve-Generate:
Natural Language-Centric Outside-Knowledge Visual Question Answering

Feng Gao[1]   Qing Ping[2]   Govind Thattai[2]   Aishwarya Reganti [2]   Ying Nian Wu [1]   Prem Natarajan[2]

[1]University of California, Los Angeles, [2]Amazon

f.gao@ucla.edu, ywu@stat.ucla.edu, {pingqing,thattg,areganti,premknat}@amazon.com

Abstract

Outside-knowledge visual question answering (OK-VQA) requires the agent to comprehend the image, make use of relevant knowledge from the entire web, and digest all the information to answer the question. Most previous works address the problem by first fusing the image and question in the multi-modal space, which is inflexible for further fusion with a vast amount of external knowledge. In this paper, we call for an alternative paradigm for the OK-VQA task, which transforms the image into plain text, so that we can enable knowledge passage retrieval, and generative question-answering in the natural language space. This paradigm takes advantage of the sheer volume of gigantic knowledge bases and the richness of pre-trained language models. A Transform-Retrieve-Generate framework (TRiG) framework is proposed[1], which can be plug-and-played with alternative image-to-text models and textual knowledge bases. Experimental results show that our TRiG framework outperforms all state-of-the-art supervised methods by at least 11.1% absolute margin.

1. Introduction

The visual question answering (VQA) task is to provide a natural language answer to a natural language question given an image [2]. This task has been well studied in the research communities, and numerous cross-modal methods have achieved state-of-the-art performance [6, 14, 21, 30, 31, 34, 50, 63, 67, 69]. The knowledge-based visual question answering (KB-VQA) task requires more extensive learning since the questions can be answered only by referring to external general knowledge [35, 48, 48, 57, 58]. Most KB-VQA datasets come with pre-defined knowledge bases, and each question is annotated with at least one supporting knowledge fact. Moreover, the recently proposed outside-knowledge visual question answering (OK-VQA) task is the most open in the sense that any external knowledge can be

---
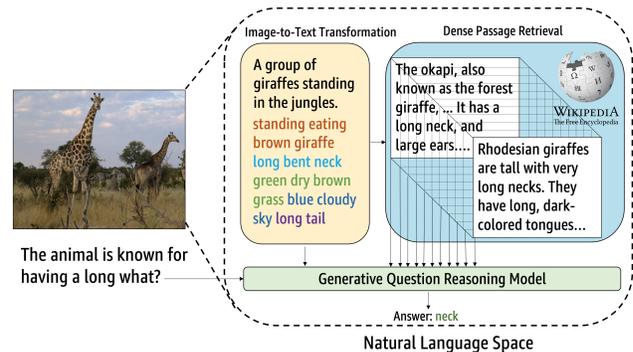
[1]The code of this work will be made public.

Figure 1. An intuitive example of our TRiG framework on the OK-VQA problem. Our Framework transform all information into language space and performs retrieved-based question answering through generative language models.

used to answer the questions.

Consider the example in Figure 1. As a human, one needs to first identify objects like *giraffes* and *trees* in the image, and associate the *giraffes* to the word *animal* in the question. Second, the human needs to apply his/her acquired commonsense knowledge about *giraffe*'s characteristics and answer the question that *giraffe* is known for having a *long neck*. For machine learning models to solve the same problem, there are several unique challenges. First, in order to answer such a question, one has to align the image, the question, and the vast amount of knowledge passages into one common space. One solution is to first fuse the image and question information in the multi-modal space with pre-trained vision-language models, and then inject knowledge into the multi-modal space. Most previous work on OK-VQA follow this paradigm, including directly injecting the knowledge embeddings [12, 49] and fusing the output of a vision-language model with the knowledge graph through graph convolutional network [39]. However, this paradigm is at the cost of squeezing the rich representation of the textual knowledge, in the magnitude of hundreds of millions, into a much smaller multi-modal space. Comparing to knowledge corpus such as BookCorpus (800M words)

and English Wikipedia (2,500M words), multi-modal pre-training datasets are much smaller such as Visual Genome with 0.01 million images and less than 2 million question-answer pairs [27], which leads to less knowledge. Therefore, we argue that it is possible to transform everything into the language space first, and then take advantage of the tremendous amount of textual knowledge for question answering. Although this seems counter-intuitive, our work proves its advantage. In this paradigm, the challenge is to be able to transform the image into language with minimum information loss. In order to tackle this, we propose three-level image-to-text transformations which significantly outperform baselines that use only captions or object labels.

The second challenge of the OK-VQA task is how to effectively retrieve the most relevant knowledge passages from gigantic knowledge bases. Previous work has explored various retrieval methods such as term-based BM25 [37], and network-based ranking [37, 61]. In the OK-VQA dataset, this task is problematic in that there is no ground-truth knowledge annotation for each question. The retrieval has to rely on either transfer learning from similar knowledge-retrieval tasks or weak supervision from pseudo signals such as whether the passage contains the answer tokens [41]. Our preliminary study finds that there is no guarantee that a passage containing the ground-truth answer will essentially relate to the question or help the answer prediction. Such signals are very weak and may introduce more noise than useful information into the retrieval model. Instead, we adopt the state-of-the-art dense passage retrieval model (DPR) [24] that is pre-trained on large question-answering dataset *Natural Questions* (NQ) [28] as our knowledge retriever, which is shown to outperform the BM25 method in terms of retrieval coverage rate.

The third challenge of the OK-VQA task is to consolidate all the multi-source input, namely the question, visual context, and the retrieved knowledge passages, to predict answers. Since now everything is in the language space, the problem can be formulated as a multi-passage question answering problem. More specifically, the model needs to not only rank the retrieved passages but also predict an answer according to the ranked passages. Most existing work utilizes extractive methods to predict the answer span in the passage [5, 7, 29, 44, 45, 60, 62]. This is not applicable in the OK-VQA dataset because there is neither annotation of ground-truth passage nor answer span in any passage. Instead, we use the generative question answering model [19] to avoid the defect in span prediction. Furthermore, we use beam-search for robust answer generation. Lastly, since the question-answering model is the last stage in the entire framework, any information distortion or loss in the image-to-text transformation and knowledge retrieval would propagate to the final question answering model. Therefore, it is important for the final question answering model to be more

transparent and interpretable to diagnose the root cause of errors. We use cross-attention scores from the decoder of the generative model to rank and highlight the top supporting knowledge passages, which helps to interpret the results of the model.

To bridge the above-mentioned research gaps, we propose the **T**ransform-**R**etrieve-**G**enerate (TRiG) framework for the OK-VQA task. At the high level, the framework aligns all the information (image, question, and knowledge) into the language space in order to take advantage of the rich semantics of textual knowledge. The framework starts with three-level image-to-text transformations, followed by dense passage retrieval to retrieve the most relevant knowledge passages. Further, the TRiG aggregates the information from all passages and generates an answer that is relatively easy to interpret. Our contributions are as follows:

- We propose a new paradigm shift for the OK-VQA task, from aligning all the information in the multi-modal space, to first transforming an image into plain text and performing knowledge retrieval and question answering all in language space.

- We propose a robust framework **T**ransform-**R**etrieve-**G**enerate (TRiG), that achieves new state-of-the-art performance on the OK-VQA dataset and leading other supervised methods by 11.1%.

2. Related Work

Visual Question Answering (VQA) The conventional visual question answering (VQA) task aims to answer questions pertaining to a given image. Multiple VQA datasets have been proposed, such as Visual Genome QA [26] VQA [2], GQA [17], CLEVR [23], MovieQA [54] and so on. Many works have shown state-of-the-art performance on VQA tasks, including task-specific VQA models with various cross-modality fusion mechanisms [14,21,25,50,63,67, 68] and joint vision-language models that are pretrained on large-scale vision-language corpus and finetuned on VQA tasks [6, 11, 30, 31, 34, 53, 69]. Please note that the conventional VQA task does not require external knowledge by definition, although studies show some VQA questions may require commonsense knowledge to answer correctly [2].

Outside Knowledge-Based VQA (OK-VQA) Beyond the above paradigm, knowledge-based visual question answering (KB-VQA) is proposed where a visual question cannot be answered without external knowledge. Several knowledge-based VQA datasets are proposed, each providing its own knowledge bases and ground-truth supporting fact [35, 48, 57]. More recently, the dataset outside-knowledge visual question answering (OK-VQA) [40] is proposed where the usage of outside knowledge is open to
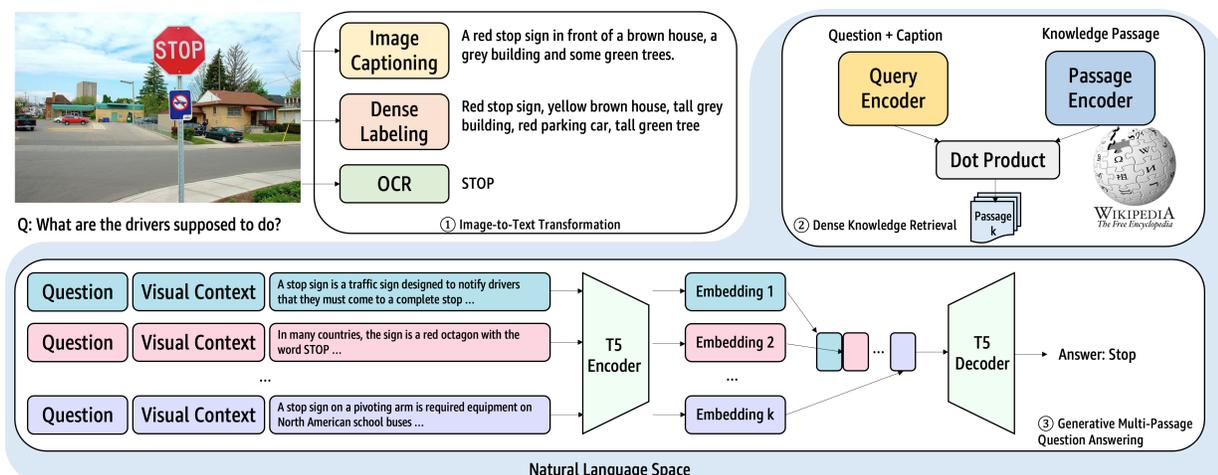
Figure 2. The overview of our **TRiG** framework. (1) **T**: Our TRiG framework transforms all visual information into natural language space on three-levels: image-level captioning, object-level dense labeling and text OCR. (2) **R**: Our dense knowledge retriever retrieve top-k knowledge passages from Wikipedia that are relevant to the query. (3) **G**: Our generative question answering model encode all question-context-knowledge tuples and fuses the output to generate a final answer.

the entire web. Most existing work for OK-VQA rely on the pre-trained vision-language models as a major workhorse for question answering [12, 37, 39, 49, 61, 64]. In [12, 49], learned knowledge embeddings are injected into vision-language models to perform knowledge-aware question answering. Other work uses vision-language models as a knowledge-free VQA model first and later adjusts the predicted answers by fusion with knowledge graphs [39] or answer validation with knowledge text [61]. Some also propose to directly learn vision-language representation for dense knowledge retrieval [37]. Different from the above, one recent work proposes to first convert the image into text caption and tags and then perform prompt-based QA on GPT-3 model purely in the language space [64]. In addition, a concurrent work [13] takes advantages of GPT-3 to retrieve implicit knowledge. However, the accessibility to this super-large-scale pre-trained language model is restricted, and it is challenging to interpret the QA result from the generative GPT-3 model.

Open-Domain Question Answering in NLP Open-domain question answering (Open-Domain QA) has been popular in the NLP community in recent years. The task is to answer a question with external knowledge bases without any given context paragraphs [44]. There are mainly two streams of approaches, namely knowledge graph-based question answering [10, 32, 38, 52, 59, 65] and knowledge retrieval-based question answering [5, 7, 19, 29, 44, 45, 60, 62]. For retrieval-based methods, both elastic-search such as BM25 [46] and semantic search such as Dense Passage Retrieval (DPR) [24] are utilized to retrieve most relevant

knowledge snippets from knowledge bases. For question answering, most existing work adopt extractive methods to predict the span of an answer in knowledge snippets [5, 7, 29, 44, 45, 60, 62]. One most recent work proposes to use generative language models for knowledge-based QA, which achieves state-of-the-art performances [19].

3. Methodology

In this section, we introduce the details of our **T**ransform-**R**etrieve-**G**enerate (**TRiG**) framework. Shown in Figure 2, our framework contains three stages: (i) image-to-text transformation, (ii) knowledge passage retrieval, (iii) multi-passages open-domain question answer generation.

3.1. Image-to-Text Transformation

Contrary to existing work, we first transform the image into text and then perform all downstream tasks in the language space. In order to minimize the information loss in the process of transforming the image into plain text, three-levels of transformations are performed (Equation 1). First, image-level information is transformed to caption text with a state-of-the-art image captioning model [31]. Second, object-level information is translated to object and attribute labels [1, 15]. Lastly, according to [20], some VQA questions can only be answered with optical character recognition (OCR). We use an off-the-shelf OCR model to detect all possible texts in the images [2].

We denote C_i, L_i, and O_i as the generated caption text, attribute and object text, and OCR text from image I_i respectively. In the rest of the paper, we will denote the visual

---
[2]https://github.com/JaidedAI/EasyOCR

context $v_i = (C_i, L_i, O_i)$ for the corresponding image I_i. Please note that our proposed framework does not necessitate the use of the above-mentioned image-to-text transformation models only. One could choose to plug-and-play alternative methods into the framework.

$$C_i = (w_0^c, \ldots, w_j^c) \leftarrow f_{ImageCaptioning}(I_i)$$
$$L_i = \{(w_0^{attr}, w_0^{obj}), \ldots, (w_n^{attr}, w_m^{obj})\} \leftarrow f_{tagging}(I_i)$$
$$O_i = \{w_0^{ocr}, \ldots, w_k^{ocr}\} \leftarrow f_{ocr}(I_i)$$
(1)

3.2. Knowledge Passage Retrieval

After the image is transformed into plain-text representation, we use the text representation as the query to retrieve knowledge passages in the natural language space. In this paper, we use the Wikipedia dump as the knowledge base, which contains over 21 million Wikipedia passages [29]. We ensure that our framework is designed to be generic enough to support other textual KBs such as GenericsKB [3] or the surface forms of graph knowledge bases such as ConceptNet [51].

More specifically, given a textual query q_i of an image I_i and a knowledge base $\mathbb{K} = \{p_j\}$ where each p_j is a knowledge passage, the task is to retrieve top k knowledge passages $P_k = [p_1, p_2, \ldots, p_k]$ from $\mathbb{K}$ that are most relevant to the query q_i, where $k \ll |\mathbb{K}|$. In this paper, we empirically use the query $q_i = (Q_i, C_i)$, where Q_i is the original question and C_i is the the generated caption of corresponding image I_i.

We use dense passage retrieval (DPR) to retrieve the knowledge passages [24]. DPR encodes both query and passage with BERT layers that could better capture the semantic similarity between them than term-based retrieval methods such as TF*IDF and BM25 [24]. First, the query q_i and a passage p_k are encoded with two independent pretrained BERT encoders [9]. We take the embedding of the [CLS] token $\mathbf{x}_{q_i}$ and $\mathbf{x}_{p_i}$ in the BERT to represent q_i and p_k respectively. Second, a similarity scores $\text{sim}(q_i, p_k)$ is calculated by taking the dot product of the two encoded dense vectors of the query q_i and a passage p_k.

$$\mathbf{x}_{q_i} = E_Q(q_i), \mathbf{x}_{p_i} = E_P(p_k) \quad (2)$$

$$\text{sim}(q_i, p_k) = \mathbf{x}_{q_i}^T \cdot \mathbf{x}_{p_k} \quad (3)$$

Because of the tremendous amount of passages in the Wikipedia knowledge base, it is time-consuming to retrieve the top k passages for each query from the knowledge base with over 21 million passages. We leverage an open-sourced indexing engine FAISS [22], an extremely efficient library to speed up the clustering and indexing of large number of dense vectors. Given a query q_i, the dense passage retrieval module will return k passages $P_k = [p_1, p_2, \ldots, p_k]$

from the entire knowledge base $\mathbb{K}$ where $\text{sim}(q_i, p_1) > \text{sim}(q_i, p_2) > \cdots > \text{sim}(q_i, p_k)$ and $k \ll |\mathbb{K}|$. The retrieved passages P_k will be later used for downstream question-answering.

3.3. Generative Multi-Passages QA

After aligning the visual information, the question, and the external knowledge into the language space, we introduce our generative question-answering module. Our design of the model takes the following into consideration. First, although previous work on joint vision-language models formulates the task as an answer classification task [39, 51, 61], our preliminary studies show that language models seem to be less flexible in classifying text into such high-dimensional answer space (over 100k) given a relatively small dataset. Second, although most previous language QA models follow a span-based answer prediction paradigm [29, 44, 60, 62], it is impractical in our open-domain setting since there is no ground-truth supporting fact in our task, let alone the ground-truth answer span for prediction. On the other hand, recent work shows that a generative encoder-decoder network can achieve state-of-the-art performance on multiple open-domain QA datasets [42], and it avoids span prediction and directly generates a free-form answer.

To achieve this goal, we use a transformer-like encoder-decoder model T5 as the backbone of our generative question answering module [43]. It is impractical to include all top-k passages in one T5 model. We use T5 model to encode each $(question, visual\ context, knowledge)$ tuple independently and then fuse the k encoded representations to decode an answer following the idea in [19].

Multi-Passages Question Answer Generation First, we feed the concatenated sequence of $(Q_i, v_i, p_{i,k})$ into a self-attentive encoder to get per-position hidden embeddings $\mathbf{z}^{Q_{i,k}}$, where q_i is the question, v_i is the visual context text and p_i is one passage respectively.

$$\mathbf{z}^{Q_{i,k}} = E_{SelfAttn}(Q_i, v_i, p_{i,k})$$
$$= (z_0, \ldots, z_L) \quad (4)$$

where z_i is the hidden embedding of the i-th token in the sequence, $\mathbf{z}^{Q_{i,k}} \in \mathbb{R}^{1 \times L \times h}$ is the hidden representation of the sequence, $L = |(Q_i, v_i, p_{i,k})|$ is the length of the sequence and h is the size of the hidden embedding.

Subsequently, we perform the same encoding operation on all k passages to derive k hidden representations:

$$\mathbf{z}^{Q_i} = (\mathbf{z}^{Q_{i,1}}, \ldots, \mathbf{z}^{Q_{i,k}}) \quad (5)$$

where we concatenate the k hidden embeddings to $\mathbf{z}^{Q_i} \in \mathbb{R}^{(k \cdot L) \times h}$. This operation is to fuse all the information from different question-context-passage interactions together in

order to generate better answers. Then, we feed the concatenated hidden representation $\mathbf{z}^{Q_i}$ into a stacked self-attentive decoder to predict per-position word distribution over the vocabulary space $|V|$:

$$P(a_1), \ldots, P(a_l) = \sigma(D_{SelfAttn}(\mathbf{z}^{Q_i})) \qquad (6)$$

where σ is a non-linear function such as softmax, l is the length of the answer, and $Q_i \in \mathbb{R}^{|V|}$ is the word distribution over the vocabulary of size $|V|$. Finally, we use teacher-enforcing to train the entire model with auto-regressive cross-entropy loss:

$$L_{ans} = -\frac{1}{N \cdot l \cdot |V|} \cdot \sum_{i=1}^{N} \sum_{j=1}^{l} \sum_{w=1}^{|V|} y_{i,j,w} \cdot \log(p(a_{i,j,w})) \qquad (7)$$

Inference of the Multi-Passage Generative Model During training, teacher-enforcing is used to train the encoder-decoder model auto-regressively. During inference time, the answer tokens are generated iteratively by feeding the previous token a_{t-1} to the input of the next token a_t. We apply both greedy-decode and beam-search for the answer decoding. In greedy-decode, the best answer token is always selected with the highest probability at each decoding step. In beam search, a beam of size m is maintained during decoding, and m answer candidates are generated with ranked scores. We also take ensembles of the 6 TRiG models trained on different splits of the top-100 passages, where the best answer is selected by ranking the model answers with average log probability of all the generated tokens of the predicted answer: $a^* = \mathrm{argmax}_n\{\frac{1}{l}\sum_j^l \ln P(a_{n,j})\}$ and n is the number of ensembles.

4. Experiments

In this section, we describe the implementation details of our method and report the experimental results.

4.1. Implementation Details

OK-VQA Dataset We use the OK-VQA dataset in this research work (version v1.1[3], license CC-BY 4.0[4]). It is one of the most challenging visual question answering datasets that is open to all external knowledge usage [40]. The dataset contains 14,055 visual questions over 14,031 images from MSCOCO [33]. The dataset split is 9,009 for training and 5,046 for testing. Each entry contains an image, a question, and 10 ground-truth answers annotated by human annotators.

Dense Passage Retrieval We use BERT-base encoders, E_Q and E_P, in the retrieval module and initialize them with

---

[3] https://okvqa.allenai.org/download.html
[4] http://creativecommons.org/licenses/by/4.0/

the checkpoints pre-trained on the NQ dataset [28]. Due to the extremely large size of the Wikipedia knowledge base, we choose the HNSW indexing algorithm instead of flat indexing for a much faster speed of queries with acceptable accuracy trade-off. For more details, please refer to the implementation of [22]. Each query is composed of the question Q_i and the corresponding caption C_i. The number of retrieved passages $k = 100$ for the best possible QA performance.

Generative Multi-Passages QA We use a transformer-based [55] encoder-decoder T5-large [43] model as the backbone. By default, the embedding size of the encoder is 768. The maximum length of the input tokens is restricted to be 300. Padding to the maximum length is applied for multiple questions batch training. Because the training of the generative model with 100 passages is memory-intensive, the batch size is set to be 1 for each GPU. To optimize the QA model, we apply the following techniques: (i) AdamW as the optimizer with a linearly scheduled learning rate starting from $1e-4$; (ii) Warm-up of 2000 steps as the learning rate scheduler. We train the multi-passages QA model for 20000 optimization steps on an 8xA100 GPU cluster for 12 hours. During inference, both greedy-decode and beam-search are applied to get the best answers. Before evaluation, a normalization step is performed on the generated answers, including lower-casing and removing articles, punctuation, and duplicated white space.

4.2. Empirical Results on OK-VQA

4.2.1 Performance of Knowledge Retrieval

To evaluate the performance of the knowledge passage retrieval module, we consider a question that has a hit in its retrieved knowledge passages if at least one of its ground-truth answers appears in the retrieved passages. Then the $hit@k$ is defined as the percentage of questions in the entire dataset who get a hit in their top k retrieved knowledge passages.

	OK-VQA Train	OK-VQA Test
Top-K	hit@k	hit@k
Top-5	42.72%	45.83%
Top-10	54.66%	57.88%
Top-20	68.76%	72.11%
Top-50	72.27%	80.49%
Top-100	83.76%	86.56%

Table 1. Hit@k of the dense passage retrieval (k = the number of retrieved knowledge passages).

From Table 1, we can observe that the answer retrieval rate $hit@k$ increases along with the number of passages k from 42.7% to 83.7% as k increases from 5 to 100. A larger

k increases the probability that each question has access to at least one relevant knowledge passage during inference. We experiment with different k for the downstream QA model, which will be discussed in subsection 4.3.

4.2.2 Performance of the Generative QA Model

Exact Match and VQA Score The OK-VQA dataset has 10 annotated answers for each question, and we consider both Exact Match and VQA Score as metrics to evaluate the generative QA model. The Exact Match (EM) is defined as the percentage of questions whose predicted answer exactly matches any of the 10 annotated answers. EM metric considers every answer as equally ground-truth the same. On the other hand, VQA score defines a voting mechanism so that each annotated answer a_i is assigned a score s_i between 0 and 1 [2].

A generated answer $\hat{a}_i$ would get s_i score if it matches the annotated a_i. The VQA metric is an average of the weighted scores over the entire test set. Arguably, the voting mechanism of the VQA score may promote some ground-truth answers over others based on the annotators' consensus subjectively.

Comparison with Supervised-Learning SOTAs The performance of our proposed TRiG framework with state-of-the-art models is reported in Table 2. Please note that all the models in comparison are supervised-learning models. Several observations can be made from the table. First, most previous methods utilize the vision-language model as the backbone for question answering and then integrate it with external knowledge. Some represent the knowledge in the form of graph (KRISP [39], Concept-Bert [12], RVLESK [49]) while others fuse the output of the vision-language model with textual knowledge representation (MAVEx [61]) or implicit knowledge from a language QA mdoel [47]. Second, a concurrent work, VRR [37], transforms the image into caption text and performs span-based question answering on a trimmed knowledge base using Google search engine. Last and most importantly, all of the above methods achieve very similar VQA scores between 38.60 and 39.4, despite usage of diverse sources of knowledge bases such as ConceptNet [12, 39, 49, 61], Google Image [61], Google Web Search [37] and Wikipiedia [61] and pretraining on other datasets such as VQA [12, 39, 49, 61] and Visual Genome [49].

Our proposed TRiG framework significantly outperforms all state-of-the-art supervised-learning methods with at least a 11.1% margin. Our TRiG framework differs from the existing methods as (i) instead of aligning representation of the vision-language QA model with external knowledge in the multimodal space, TRiG transforms the image into text information as accurately as possible and aligns all

Model	EM	VQA Score
SOTA Methods		
KRISP [39]		32.31
ConceptBert [12]		33.66
CBM [47]		38.60
KRISP w/ VQA2.0 pretrained		38.70
MAVEx [61]		38.70
RVLESK [49]		39.04
Weakly Supervised VRR [37]		39.20
MAVEx w/(Ensemble 5) [61]		39.40
Ours		
TRiG w/ Q+C+DL+O, G	**53.62%**	**49.24**
TRiG w/ Q+C+DL+O, BS	**53.59%**	**49.35**
TRiG w/ Q+C+DL+O, G, **E***	**54.73%**	**50.50**

Table 2. Comparison of supervised-learning methods on the OK-VQA dataset. In TRiG Model, **Q**: Question, **C**: Caption, **DL**: Dense Labels, **O**: OCR Text, **G**: Greedy Decode, **BS**: Beam-Search, **E***: Ensembles of the 6 TRiG models.

the information of the image, question, and knowledge in language space; (ii) the generative QA model in TRiG is not pre-trained on other multimodal datasets, which helps the model to start learning to reason over external knowledge, rather than inducing data bias from other multimodal datasets.

We would like to also highlight the Exact Match (EM) score of our TRiG models, which are higher than the VQA scores. As in Figure 4, we observe that sometimes the generative QA model predicts a reasonable answer but is not credited with the highest VQA score or not even any score according to annotators' voting.

Model	#Params	VQA Score
SOTA Prompt Method [64]		
PICa w/16 RP C+T	175B	43.30
PICa w/16 SP C+T	175B	46.50
PICa w/16 SP C+T, 3×E	175B	47.70
PICa w/16 SP C+T, 5×E	175B	48.00
Ours		
TRiG w/ Q+C+DL+O, G	0.77B	**49.24**
TRiG w/ Q+C+DL+O, BS	0.77B	**49.35**
TRiG w/ Q+C+DL+O, G, **E***	0.77B	**50.50**

Table 3. Comparison of Proposed TRiG with SOTA Prompt-Based Method on the OK-VQA Dataset. In [64], **RP**: Random Prompt, **SP** Selected Prompt, **C**: Caption, **T**: Image-Tagging, **E**: Prompt Ensemble. In TRiG model, **Q**: Question, **C**: Caption, **DL**: Dense Labels, **O**: OCR Text, **G**: Greedy Decode, **BS**: Beam-Search, **E***: Ensembles of the 6 TRiG models.

Comparison with Prompt-Based SOTA We also compare our method with one very recent prompt-based method

on the OK-VQA problem [64]. By taking advantage of the super large-scale language model GPT-3 [4], the proposed prompt-based method (PICa) surpasses all existing supervised methods with sophisticated prompting. As shown in Table 3, PICa achieves 43.3 VQA score with 16 prompts randomly selected from the training data. By carefully selecting 16 prompts based on the similarity between testing and training questions, PICa further achieves 46.5. With 5 ensembles of 16 prompts, PICa reaches 48.0 VQA score.

Our method (TRiG) outperforms PICa with greedy-decode 49.24, beam-search decoding 49.35 and ensemble 50.50. Both PICa and our method share the same idea of unifying the image, the visual question, and knowledge in language space and then performing question answering with language models. The significant performance gain of both methods (9-11.1% over SOTA) highlights the potential of this idea – if the image could be transformed into plain text information faithfully, then one could take advantage of the vast volume of external knowledge in text form and advanced language models pre-trained on rich variations of human natural language to yield better answer prediction.

We would like to also highlight that our method outperforms PICa by a margin of 2.50%, especially considering the among of parameters (175 billion over 0.77 billion of our model) and accessibility of the GPT-3 model. Moreover, we argue that our prediction results are relatively easier to interpret by selecting supporting knowledge passages, whereas in PICa the explanation is generated by GPT-3 in a black-box manner. We use the averaged cross-attention score of the generative model to select supporting facts [18]. For concrete examples of such interpretability, please see the examples in Figure 4.

4.3. Ablation Study

Variant Visual Context Input   We investigate the empirical differences among the combination of the visual contexts inputs to the generative QA model, namely image caption (C), object label (L and DL), and OCR (O).

Inputs	VQA Score
Question + K + C	42.54
Question + K + C + L	42.94
Question + K + C + L + O	43.53
Question + K + C + DL + O	**49.35**

Table 4. Ablation Study of the Different Variants of Text Input into the Generative QA Model (**K**: Knowledge passages, **C**: Caption, **L** = Bottom-Up Labels [1], **DL** = Dense Labels, **O** = OCR Text).

As in Table 4, we find that adding caption (C) to the input yields decent performance (42.5), suggesting that caption conveys basic information of the image. Adding sparse object labels and attributes (L) also helps a little (42.9). By adding OCR, the performance is further improved (43.5),
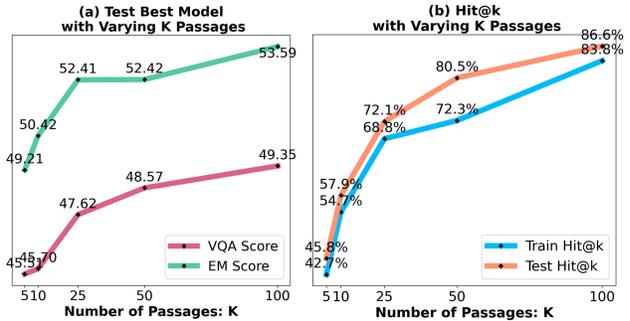

Figure 3. Testing the QA model with varying number of passages.

which is in accord with previous findings that some questions in OK-VQA require understanding the text in the image through OCR [20]. Interestingly, the largest gain is achieved by replacing sparse object and attribute labels with more semantically rich dense object labels (49.4), which again highlights that the faithfulness of image-to-text transformation is a crucial prerequisite for downstream QA in the language space.

Generative Multi-Passages QA with Varying K passages
We also investigate how the generative QA model behaves with a different number of passages k. We apply our best model trained on 100 passages and test it with varying k passages. From Figure 3-(a), we can see that the testing performance of this model steadily increases along with the growing number of passages k. However, the improvement becomes marginal after k=25 (47.62 to 49.35), while the coverage $Hit@k$ still increases by 15% as Figure 3-(b). This also supports our hypothesis that there may be a long-tail effect of the retrieval. Yet it is difficult to quantify as to which passages are essentially relevant to the question-answering.

4.4. Discussion

Error Analysis   To investigate the behavior of our TRiG model, we conduct error analysis with our best model using greedy-decoded predictions. The quantitative results are illustrated in Figure 5. We observe that answers with numerical values are harder to predict, where the model could get into a blunt generation (Figure 5-(a)). Furthermore, as the length of the answer increases, it is harder for the generative model to predict every token in the phrase correctly (Figure 5-(b)).

We also manually reviewed 50 examples where TRiG makes wrong predictions. Among these random examples, 50% of the errors are due to the information loss during image-to-text transformation, such as in Figure 4-(h), where the caption and dense labels failed to characterize the special features of the bird. We also found that 24% of the error are due to the failure in retrieving highly-relevant passages.
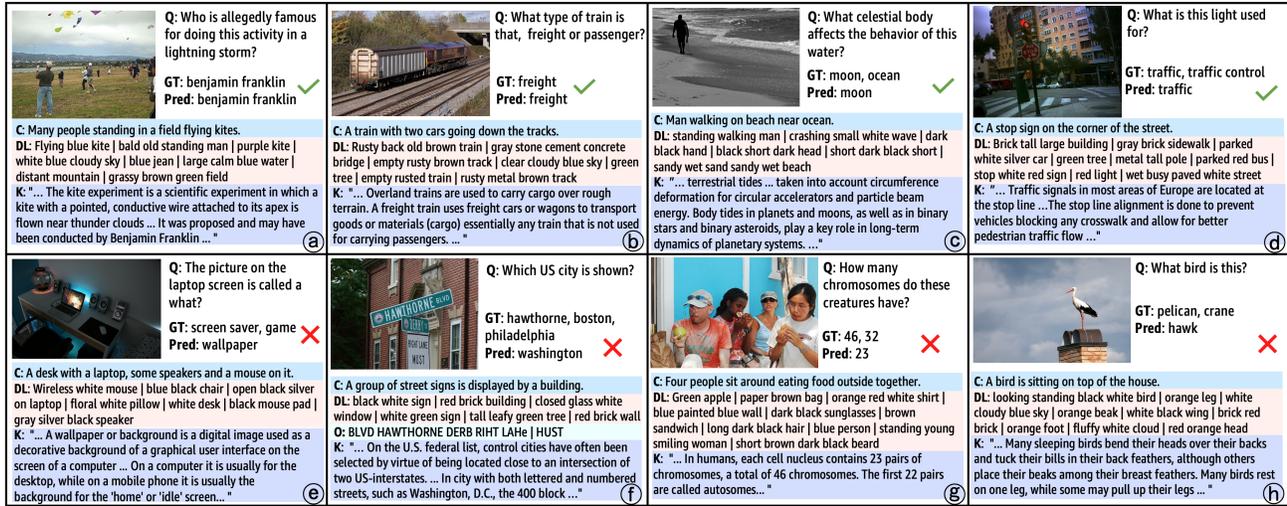
Figure 4. Examples of our TRiG model prediction together with the supporting passage. **Top**: four examples where TRiG model makes correct predictions. **Bottom**: four examples where TRiG model makes incorrect predictions. In each example: **Q**: question, **GT**: ground-truth answers, **Pred**: predicted answer, **C**: image caption, **DL**: dense labels, **O**: OCR text, **K**: top-1 supporting knowledge passage.

The high Hit@k value doesn't guarantee the passages are indeed relevant to the question. Note that some examples failed due to multiple reasons including QA error (22%) or subjective human annotations (30%) as in Figure 4(g).
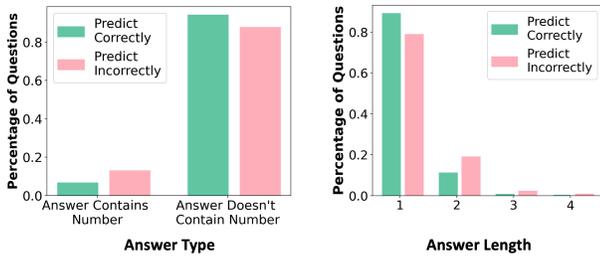

Figure 5. Performance of Generative QA model by different answer types. **Left**: whether numerical answers are harder to predict. **Right**: whether longer answers are harder to predict.

Interpretability To interpret the visual question-answering models, previous works attempt to supervise the VQA models with visual grounding annotations [8, 8, 70] or neural symbolic network [16, 56, 66]. When it comes to knowledge-based VQA, it is all the more challenging to interpret the model in multimodal space because the knowledge has been transformed into a fused representation and loses its meaning.

Our TRiG framework alleviates this problem by providing transparent explanations in the language space. In the top row in Figure 4, the image-to-text transformations provide sufficient information for both the knowledge retrieval and QA model. Meanwhile, when Figure 4(e, f, g, h) make

wrong predictions, the QA model is still predicting the answer according to the visual context and retrieved passages.

OK-VQA Evaluation Metrics Some researchers [36] also argue that the VQA score metric is subjective. In one OK-VQA example, a model will achieve 1.0 VQA score for the answer ***wetsuit*** but only 0.66 score for the answer ***wet suit***. In daily language, the usage of any of the semantically-similar answers is subtle and sometimes random. We also look at the top-3 answers of our TRiG model using beam search, and the model achieves significantly higher performance, *i.e.* **67.4** VQA score and **71.8%** EM. We call for better VQA metrics that probably compare two sets of answers instead of comparing only the top one answer or other alternatives such as AAS that automatically expands the ground-truth answer set for better matching [36].

5. Conclusion

In this paper, we approach the OK-VQA task from a new perspective, where all the visual information is aligned into the language space to take advantage of the comprehensiveness in textual knowledge bases. Moreover, we propose a robust Transform-Retrieve-Generate (TRiG) framework that outperforms state-of-the-art supervised methods by 11.1%. One can plug-and-play with different image-to-text methods and textual knowledge bases into TRiG for potential further improvement. Our work has limitations that the dense passage retrieval is not optimized for the OK-VQA task, due to the unavailability of ground-truth supporting facts. We consider this as one of our future work, as well as improving the quality of image-to-text transformation.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 3, 7

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1, 2, 6

[3] Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. Genericskb: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*, 2020. 4

[4] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 7

[5] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017. 2, 3

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 1, 2

[7] Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*, 2017. 2, 3

[8] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 8

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[10] Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. Scalable multi-hop relational reasoning for knowledge-aware question answering. *arXiv preprint arXiv:2005.00646*, 2020. 3

[11] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020. 2

[12] François Garderes, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. Conceptbert: Concept-aware representation for visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 489–498, 2020. 1, 3, 6

[13] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021. 3

[14] Dalu Guo, Chang Xu, and Dacheng Tao. Bilinear graph networks for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 1, 2

[15] Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Image scene graph generation (sgg) benchmark. *arXiv preprint arXiv:2107.12604*, 2021. 3

[16] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017. 8

[17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[18] Gautier Izacard and Edouard Grave. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584*, 2020. 7

[19] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020. 2, 3, 4

[20] Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. *arXiv preprint arXiv:2103.05568*, 2021. 3, 7

[21] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020. 1, 2

[22] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 4, 5

[23] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[24] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020. 2, 3, 4

[25] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NeurIPS*, 2018. 2

[26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016. 2

[27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome:

Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2

[28] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. 2, 5

[29] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*, 2019. 2, 3, 4

[30] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 1, 2

[31] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, 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. *ECCV 2020*, 2020. 1, 2, 3

[32] Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*, 2019. 3

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5

[34] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 1, 2

[35] Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. R-vqa: learning visual relation facts with semantic attention for visual question answering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1880–1889, 2018. 1, 2

[36] Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, and Chitta Baral. 'just because you are right, doesn't mean i am wrong': Overcoming a bottleneck in development and evaluation of open-ended vqa tasks. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2766–2771, 2021. 8

[37] Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. Weakly-supervised visual-retriever-reader for knowledge-based question answering. *arXiv preprint arXiv:2109.04014*, 2021. 2, 3, 6

[38] Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8449–8456, 2020. 3

[39] Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14111–14121, 2021. 1, 3, 4, 6

[40] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019. 2, 5

[41] Chen Qu, Hamed Zamani, Liu Yang, W Bruce Croft, and Erik Learned-Miller. Passage retrieval for outside-knowledge visual question answering. *arXiv preprint arXiv:2105.03938*, 2021. 2

[42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 4

[43] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 4, 5

[44] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018. 2, 3, 4

[45] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In *EMNLP*, 2016. 2, 3

[46] Stephen Robertson and Hugo Zaragoza. *The probabilistic relevance framework: BM25 and beyond.* Now Publishers Inc, 2009. 3

[47] Ander Salaberria, Gorka Azkune, Oier Lopez de Lacalle, Aitor Soroa, and Eneko Agirre. Image captioning for effective use of language models in knowledge-based visual question answering. *arXiv preprint arXiv:2109.08029*, 2021. 6

[48] Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8876–8884, 2019. 1, 2

[49] Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *arXiv preprint arXiv:2101.06013*, 2021. 1, 3, 6

[50] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 1, 2

[51] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 4

[52] Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William W Cohen. Open domain question answering using early fusion of knowledge bases and text. *arXiv preprint arXiv:1809.00782*, 2018. 3

[53] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[54] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 5

[56] Ramakrishna Vedantam, Karan Desai, Stefan Lee, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Probabilistic neural symbolic models for interpretable visual question answering. In *International Conference on Machine Learning*, pages 6428–6437. PMLR, 2019. 8

[57] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 1, 2

[58] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015. 1

[59] Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, et al. Improving natural language inference using external knowledge in the science questions domain. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7208–7215, 2019. 3

[60] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*, 2019. 2, 3, 4

[61] Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-modal answer validation for knowledge-based vqa. *arXiv preprint arXiv:2103.12248*, 2021. 2, 3, 4, 6

[62] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, 2019. 2, 3, 4

[63] Xiaofeng Yang, Guosheng Lin, Fengmao Lv, and Fayao Liu. Trrnet: Tiered relation reasoning for compositional visual question answering. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 414–430. Springer, 2020. 1, 2

[64] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. *arXiv preprint arXiv:2109.05014*, 2021. 3, 6, 7

[65] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. *arXiv preprint arXiv:2104.06378*, 2021. 3

[66] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *arXiv preprint arXiv:1810.02338*, 2018. 8

[67] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019. 1, 2

[68] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 29(12):5947–5959, 2018. 2

[69] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. *CVPR 2021*, 2021. 1, 2

[70] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto. Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 ieee winter conference on applications of computer vision (wacv)*, pages 349–357. IEEE, 2019. 8