Weakly-Supervised Visual-Retriever-Reader for Knowledge-based Question Answering

Man Luo* **Yankai Zeng*** **Pratyay Banerjee** **Chitta Baral**
Arizona State University
{mluo26, yzeng55, pbanerj6, chitta}@asu.edu

Abstract

Knowledge-based visual question answering (VQA) requires answering questions with external knowledge in addition to the content of images. One dataset that is mostly used in evaluating knowledge-based VQA is OK-VQA, but it lacks a gold standard knowledge corpus for retrieval. Existing work leverage different knowledge bases (e.g., ConceptNet and Wikipedia) to obtain external knowledge. Because of varying knowledge bases, it is hard to fairly compare models' performance. To address this issue, we collect a natural language knowledge base that can be used for any VQA system. Moreover, we propose a Visual Retriever-Reader pipeline to approach knowledge-based VQA. The visual retriever aims to retrieve relevant knowledge, and the visual reader seeks to predict answers based on given knowledge. We introduce various ways to retrieve knowledge using text and images and two reader styles: classification and extraction. Both the retriever and reader are trained with weak supervision. Our experimental results show that a good retriever can significantly improve the reader's performance on the OK-VQA challenge. The code and corpus are provided in this link.

1 Introduction

Knowledge-based VQA is a challenging task, where knowledge present in an image is not sufficient to answer a question. It requires a method to seek external knowledge. Figure 1 shows two examples from the OK-VQA benchmark (Marino et al., 2019), which is normally used to study knowledge-based VQA. In each of the two examples, external knowledge is needed to answer the question. For instance, in the first example, to identify the vehicle used in the item shown in the image (top-left), a system needs to first ground the referred item as a fire hydrant and then seek external knowledge presented top-right of the image. The challenge is to ground the referred object in the image and retrieve relevant knowledge where the answer is present.

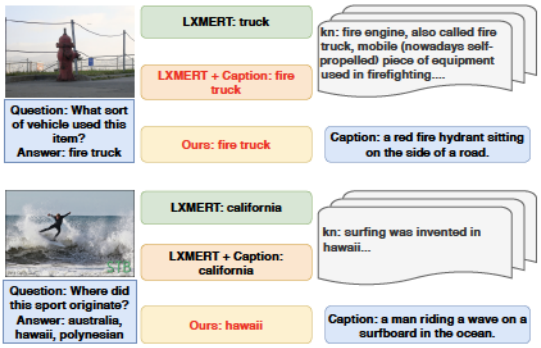

Figure 1: Two examples from OK-VQA: the middle column are predictions by two baselines and one by our proposed Visual-Retriever-Reader pipeline. The left column is relevant knowledge and the corresponding captioning of images.

Although the OK-VQA benchmark encourages a VQA system to rely on external resources to answer the question, it does not provide a knowledge corpus for a QA system to use. As such, existing methods rely on different resources such as ConceptNet (Speer et al., 2017), WordNet (Miller, 1992), and Wikidata (Vrandečić and Krötzsch, 2014), resulting in the following issues:

1. It is difficult to fairly compare different VQA systems as it is unclear whether the difference in performance arises from differing model architectures or the different knowledge sources.
2. The different formats of the knowledge sources, such as the structured ConceptNet and the unstructured Wikipedia, demand different modules to retrieve knowledge, consequently making a knowledge-based VQA system complicated.
3. External resources like ConceptNet and WordNet have limitations. First, they only cover a

*Equal contribution

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6417–6431
November 7–11, 2021.

limited amount of knowledge. For example, ConceptNet provides only 34 relation types, and there is a vast amount of knowledge that is hard to be described by a relation in a knowledge graph, such as, *describe the logo of Apple Inc.* Second, constructing a structured knowledge base requires heavy human annotation and is not available in every domain. Thus, it limits the application of a knowledge-based VQA system that relies on a structured knowledge base.

Therefore, there is a need for a general and easy-to-use knowledge base. Motivated by this, we collect a knowledge corpus for the OK-VQA benchmark. Our corpus is automatically collected via Google Search* by using the training-split question and the corresponding answers, and we provide a training corpus with 112,724 knowledge sentences and a full testing corpus with 168,306 knowledge sentences. The knowledge corpus is in a uniform format, i.e., natural language. Thus, it is easy to use by other OK-VQA methods. As we will show in §6, the knowledge base provides rich information to answer OK-VQA questions.

Utilizing the curated corpus, we further develop a weakly-supervised Visual-Retriever-Reader and evaluate it on the OK-VQA challenge. It consists of two stages, as seen in Figure 2. In the first stage, the visual retriever retrieves relevant knowledge from the corpus. In the second stage, the visual reader predicts an answer based on the given knowledge. Such a pipeline is well-studied in text-only open-domain QA (Chen et al., 2017a; Karpukhin et al., 2020). We apply its principles to the multi-modal vision and language domain with novel adaptations. On the retriever side, we introduce visual information and evaluate a cross-modality model and a text-only caption-driven model (§4.1). On the reader side, we build two visual readers, a classification and an extraction type, with both utilizing visual information (§4.2). We observe in §6, our Visual-Retriever-Reader pipeline performs strongly on the OK-VQA challenge and establishes a new state-of-the-art.

Our experiments reveal multiple insights. First, we find that the image captions are very useful for both visual retriever and visual reader, which demonstrates the application of image captioning generator on knowledge-based VQA tasks. Second, a neural retriever has much better performance than a term-based retriever. This observation is quite interesting as in the NLP domain, typically, a term-based retriever (e.g., TF-IDF and BM25) is a hard-to-beat baseline (Lee et al., 2019a; Lewis et al., 2020; Ma et al., 2021), suggesting an essential role of neural retrievers in the vision-&-language domain. Third, similar to the NLP domain, where a reader can perform well if the given knowledge contains relevant information, we discover that our visual reader has a significant leap when using noisy knowledge and high-quality knowledge. It motivates the need for developing a more efficient visual retriever for knowledge-based VQA tasks.

Our contributions are three folds. First, we build a general easy-to-use knowledge corpus for the OK-VQA benchmark, which makes model evaluation fair. Second, we propose a Visual-Retriever-Reader pipeline adapted from the NLP domain for the knowledge-based VQA task. Our model establishes a new state-of-the-art. Third, our experiments reveal several insights as mentioned above, and open a new research direction.

2 Related Work

Knowledge-based VQA. Many benchmarks have been proposed to facilitate the research in knowledge-based VQA. FVQA (Wang et al., 2017a) is a fact-based VQA dataset that provides image-question-answer-supporting fact tuples, where the supporting fact is a structured triple, e.g., ⟨Cat, CapableOf,ClimbingTrees⟩. KB-VQA (Wang et al., 2017b) dataset consists of three types of questions: "Visual" question answerable using the visual concept in an image, "Common-sense" questions answerable by adults without looking for an external source, and "KB-knowledge" questions requiring higher-level knowledge, explicit reasoning, and external resource. KVQA (Shah et al., 2019) consists of questions requiring world knowledge of named entities in images. Specifically, the questions require multi-entities, multi-relation, multi-hop reasoning over Wikidata. KVQA is challenging, as linking the named entities in an image to the knowledge base is hard on a large scale. OK-VQA (Marino et al., 2019) covers 11 types of knowledge than previous tasks, such as cooking and food, science and technology, plants and animals, etc. VLQA (Sampat et al., 2020) consists of data points of image-passage-question-answer, it is proposed recently to facilitate the research on jointly reasoning with

*https://developers.google.com/custom-search/v1/

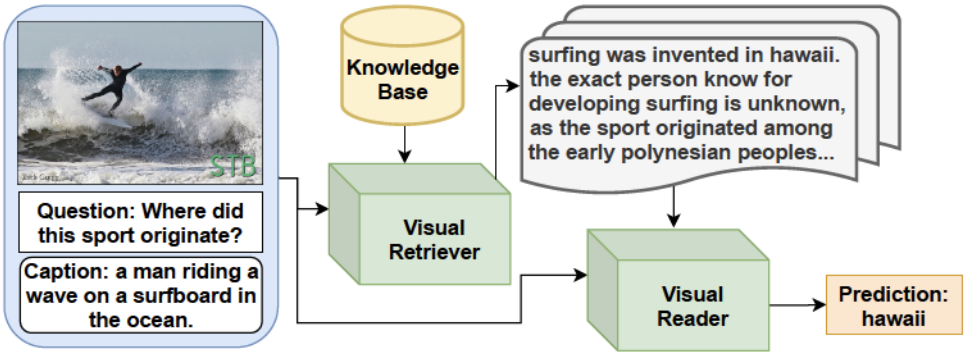

Figure 2: Visual Retriever Reader Pipeline: given an image and a question, a visual retriever is first to retrieve relevant knowledge, and then a visual reader is to predict an answer based on the given knowledge.

both image and text.

OK-VQA Systems. Out of the Box (Narasimhan et al., 2018) utilizes the Graph Convolution Networks (Kipf and Welling, 2017) to reason on the knowledge graph (KG), wherein each node image and semantic embeddings are attached. Mucko (Zhu et al., 2020) goes a step further, reasoning on visual, fact, and semantic graphs separately, and uses cross-modal networks to aggregate them together. ConceptBert (Gardères et al., 2020) combines the BERT-pretrained model (Devlin et al., 2019) with KG. It encodes the KG using a transformer with a BERT embedding query. KRISP (Marino et al., 2020) involves a BERT-pretrained transformer model to make a better semantic understanding and utilize the implicit knowledge and reasons on a GCN model. Span-Selector (Jain et al., 2021) extracts spans from the question to search most relative knowledge from Google, whereas MAVEx (Wu et al., 2021) votes among textual and visual knowledge from Wikipedia, ConceptNet, and Google Image. Besides knowledge collection, knowledge alignment (Shevchenko et al., 2021) also helps acquire a correct answer from knowledge.

Open-Domain Question Answering or ODQA tasks target collecting information from a large corpus to answer a question. The advanced reading comprehension model (Chen et al., 2017a; Banerjee et al., 2019) split this complex task into two steps: a retriever selects some most relevant documents from a corpus to a question, and a reader produces answer according to the documents from retriever. Some previous work (Kratzwald and Feuerriegel, 2018; Lee et al., 2018; Das et al., 2019; Wang et al., 2018) train the end-to-end models to rerank in a closed set. Although these models are better at retrieval, they can hardly scale to larger corpora. Open-Retrieval Question Answering (ORQA) (Lee et al., 2019b) and Dense Passage Retriever (DPR) (Karpukhin et al., 2020) constructed a dual-encoder architecture with BERT pre-trained model. This dense retrieval model shows a better performance than classic TF-IDF or BM25-based ODQA models on several natural language benchmarks.

3 Knowledge Corpus Creation

The overall process of knowledge corpus creation (Figure 3) consists of following four steps.

Step 1: Query Preparation Based on the assumption that the knowledge used for answering training set questions can also help in testing, the OK-VQA training questions are used with their answers to collect related knowledge from a search engine. We concatenate each question with each answer to get a "Question, Answer" pair. For example, in Figure 3, the question "What is the natural habitat of these animals?" has four answers, and each answer is attached to the question one by one to construct four queries.

Step 2: Google Search Webpage The generated queries are sent to Google Search API to obtain knowledge. As presented in Figure 3, a good search result web page contains a title, a link, and a snippet that consists of multiple complete or incomplete sentences and shows the most relevant part to the query. The top *ten* web pages with their snippets as the raw knowledge are chosen.

Step 3: Snippet Processing The snippets from Google searching results consist of multiple sentences, some are complete but some are not. One option is to split snippets into multiple sen-

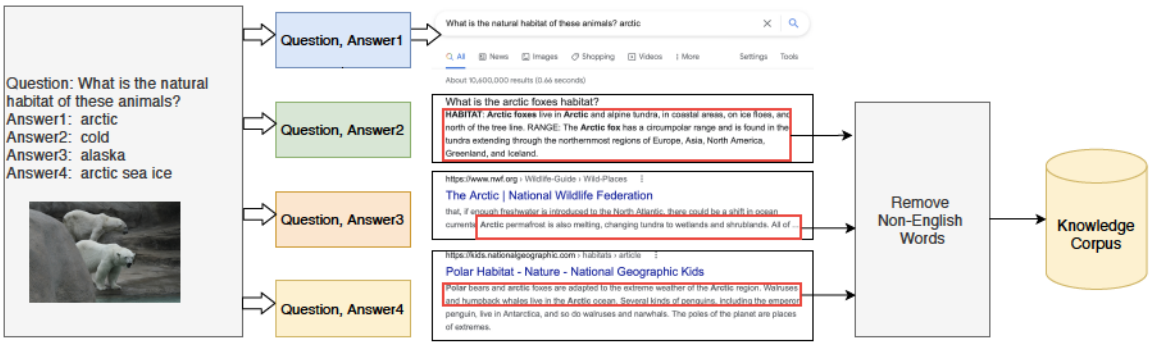

Figure 3: The overall process of Knowledge Corpus Creation. The question first combines the answers one by one to form a query, and then the query is sent to the Google Search API to retrieve the top *10* webpages. The knowledge is obtained from the snippet with further processing. Finally, we integrate the knowledge into the corpus. As shown in the searching result page, the black boxes represent webpages, and red boxes represent snippets.

tences, but experimental result shows sentence-level knowledge is worse than snippet-level. Thus, we choose to use snippet as a knowledge. To address incomplete sentence issue, we find and grab the complete sentence present in the webpage. After this pre-processing, ten snippet-knowledge from each "Question, Answer" query are selected.

Step 4: Knowledge Processing We first remove the duplicated data among each "Question, Answer" pair. Then long knowledge (more than 300 words) or short knowledge (less than ten words) are removed. Pycld2[†] is applied in this step to detect and remove the non-English part of each knowledge. Each knowledge is assigned a unique ID and duplicate knowledge sentences are removed. We curate in total 112,724 knowledge sentences for the OK-VQA training set.

4 Visual Retriever-Reader Pipeline

We present our Visual Retriever-Reader pipeline for the OK-VQA challenge, where the visual retriever aims to retrieve relevant knowledge, and the visual reader aims to predict answers given knowledge sentences. This scheme has been widely used in NLP (Chen et al., 2017b; Karpukhin et al., 2020). While previous work focuses on pure text-domain, we extend this to the visual domain with novel adaptation.

4.1 Retriever

We introduce two styles of visual retriever: term-based and neural-network-based. In the neural style, we further introduce two variants. Following the convention, we use the standard terms in next subsection, for example, in §4.1, we use *documents* and in §4.1, we use *context*, both of them are *knowledge* in our task.

Term-based Retriever. In BM25 (Robertson and Zaragoza, 2009), each query and document is represented by sparse vectors in d dimension space, where d is the vocabulary size. Then the score of a query and a document is computed based on the inverse term's frequency. BM25 can only retrieve documents for a query in text format, but an image is a part of a query in our task. To tackle this issue, we first generate image captions using a caption generation model. Then we concatenate the question and the caption as a query and obtain a list of documents by BM25.

Neural Retriever. Unlike BM25, neural retrievers extract the dense representations for a query and a context from the neural model(s). We use DPR (Karpukhin et al., 2020) as a neural retriever, which employs two BERT (Devlin et al., 2019) models to encode the query and context respectfully, then applies inner-dot product to estimate the relevancy between a query and a context. Similar to BM25, the DPR model considers the query in text format. To adapt DPR in the visual domain, we propose two methods. *Image-DPR*: we use LXMERT (Tan and Bansal, 2019) as the question encoder, which takes image and question as input and outputs a cross-modal representation. *Caption-DPR*: similar to the strategy we use in term-based retriever, we concatenate the question with the caption of an image as a query and use standard BERT as a query encoder to get the representation. In both *Image-DPR* and *Caption-DPR*, we use stan-

[†]https://pypi.org/project/pycld2/

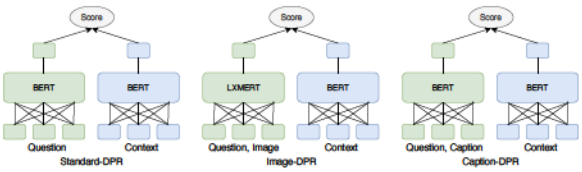

Figure 4: Comparison between standard DPR, Image-DPR and Caption-DPR: while the context encoder is the same for three models, in standard BERT(left), the question encoder only takes question as input, the Image-DPR(middle) takes both question and image as input, the Caption-DPR (right) takes the question and the caption as input.

dard BERT as context encoder. Figure 4 shows the architectures of standard DPR, *Image-DPR* and *Caption-DPR*. To train neural retriever, we use inner-dot product function to get the similarity score of relevant and irrelevant knowledge to a question, and optimize the negative log-likelihood of the relevant knowledge.

4.2 Reader

Classification Reader (CReader). Current state-of-the-art VQA systems are classification models (Tan and Bansal, 2019; Li et al., 2019; Gokhale et al., 2020b,a; Banerjee et al., 2021), where a list of answer candidates are pre-defined (from the training set), i.e., a fixed answer vocabulary, then a model classifies one of the answers as the final prediction. We build a reader in this style but incorporate external knowledge. In particular, given a question, an image, and a piece of knowledge, we first concatenate the question with the knowledge and then apply a cross-modality model to encode the text with the image and generate a cross-modal representation. We feed this representation to a Multiple Layer Perceptron (MLP) which finally predicts one of the pre-defined answers. We apply Cross-Entropy Loss to optimize the model. In this work, we use LXMERT (Tan and Bansal, 2019), while any other cross-modality models like VisualBERT(Li et al., 2019) can be adapted.

Extraction Reader (EReader). The classification model fails to generalize to out-of-domain answers, i.e., questions whose answers are not in the pre-defined answer vocabulary. To tackle this issue, we use an extraction model which is adapted from machine reading comprehension model (Chen et al., 2017b; Karpukhin et al., 2020). The model extracts a span (i.e., a start token and an end token) from the knowledge to answer the question. The image caption is given to the model as well to incorporate the image information. We also inject a special word "unanswerable" before the caption so that the model can predict "unanswerable" if the given knowledge can not be relied on to answer the question. This strategy is helpful since the retrieved knowledge might be noisy. We use a RoBERTa-large (Liu et al., 2019) as the text encoder, whose inputs are {[SEP] question [SEP] ["unanswerable"], caption, knowledge [SEP]}. Then each token representation is fed to two linear layers: one predicts a score for a token being the start token, and the other predicts a score for the end token. We apply the softmax function to get the probability of each token being a start and end token. The training objective is to maximize the probability of the ground truth start and end token.

4.3 Training and Inference

Weak Supervision. The retriever is trained using weak supervision, as the ground-truth knowledge context is unknown for a given question-image pair. Particularly, given a query and an image, we assume that knowledge that contains any of the answers is relevant, and we use the in-batch negative samples (Karpukhin et al., 2020) for training, i.e., in the training time, any relevant knowledge for other questions in the same batch are considered as irrelevant. For the reader, we use the same relevant knowledge as the retriever and when given such knowledge, the target is the answer. In addition, we use the same other collected knowledge which does not contain any answer as the irrelevant knowledge and in such a case, the reader should predict "unanswerable", a special word added to every knowledge. The reader may also be considered to be trained by weak supervision as the input knowledge is noisy, i.e., the assumed relevant knowledge is not guaranteed to be relevant.

Inference Strategy. We use the retriever to retrieve K knowledge (the value and effects of K will be presented in §7), and the reader predicts an answer based on each knowledge. We propose two strategies to predict the final answer. *Highest-Score*: the answer which has the highest score is the final prediction. *Highest-Frequency*: the answer which appears most frequently is the final prediction.

5 Evaluation

5.1 Retriever Evaluation

We evaluate the performance of a retriever based on Precision and Recall. The two metrics are based on the assumption that any retrieved knowledge that contains any of the answers annotated in the OK-VQA dataset is relevant. This assumption is because it is unknown which knowledge is relevant to a question-image pair. Therefore, the computation of Precision and Recall in our case is different from the traditional definition and illustrated as follows:

Precision Precision reveals the proportion of retrieved knowledge that contains any of the answers to a question-image pair. Mean Precision is the mean of Precision of all question-image pairs. Mathematically,

$$P(Q, A, KN) = \frac{1}{K} \sum_{i=1}^{i=K} min(\sum_{\substack{j=1 \\ A_j \in KN_i}}^{j=M} 1, 1),$$

where Q is a question, KN is a list of retrieved knowledge, A is a list of correct answers, K is the number of KN, M is the number of A.

Recall Recall reveals if at least one knowledge sentence in the retrieved Knowledge contains any answers to a question-image pair. Mean Recall is the mean of the Recall of all question-image pairs. Mathematically,

$$R(Q, A, KN) = min(\sum_{i=1}^{i=K} \sum_{\substack{j=1 \\ A_j \in KN_i}}^{j=M} 1, 1),$$

where the meaning of the symbols are the same described in Precision.

5.2 Answer Evaluation.

Original Evaluation In OK-VQA, each image-question pair has five answers annotated by humans. To apply a similar evaluation as VQA (Antol et al., 2015a), OK-VQA counts per answer twice so that each image-question pair has ten answers, the same as VQA. The score is computed as follows.

$$score(A) = min(\frac{\#\text{human that said A}}{3}, 1)$$

We use the above equation to compute the score of each answer for training and testing.

Open Domain Evaluation Luo et al. (2021) propose an evaluation method for VQA, especially mitigating the issues of Synonym/Hypernym and Singular/Plural. Here, we adapt their method to the open domain setting, where Sentence Textual Entailment (STE) is used to find semantically similar answers. In STE, given a premise and a hypothesis, a score is generated to indicate whether a premise entails the hypothesis. In our case, a premise is a sentence that contains a gold answer, and a hypothesis is the same sentence while the gold answer is replaced by a predicted answer given by a model. Suppose a high STE score[‡] is generated for such a pair of premise and hypothesis. In that case, it implies that the predicted answer is semantically close to the gold answer and thus deserves a partial score. We provide the detailed steps of evaluation in Appendix A.

Mathematically, the open-domain accuracy is given by the follows,

$$\mathrm{S}_j(Q, I, a_j, a') = \frac{1}{|\mathrm{P}_{a_j}|} \sum_{g_i \in \mathrm{P}_{a_j}} \mathrm{E}(a_j, a', g_i)$$

$$\mathrm{S}'(Q, I, a') = \underset{a_j \in Ans}{\operatorname{argmax}} \mathrm{S}_j(Q, I, a_j, a) \times \mathrm{S}(Q, I, a_j)$$

where a' is a predicted answer, Ans is a set of ground truth answer of a pair of question(Q) and image(I), P_{a_j} is a set of sentences with placeholder, $\mathrm{E}(a_j, a', g_i)$ is the entailment score of a premise (sentence g_i grounded by gold answer a_j) and hypothesis (sentence g_i grounded by predicted answer a'), and $\mathrm{S}(Q, I, a_j)$ is the ground truth score of a_j which has highest entailment score with predicted answer a'.

The main difference between our evaluation and (Luo et al., 2021) is that in their evaluation, they extend each answer with a set of alternative answers which are selected from the list of pre-defined answers in the training set. While in our setting, we remove this step since in the open domain setting, many semantically similar answers can be found in the open corpus which is not necessary in the training set (see examples in Appendix B).

6 Experiments and Results

6.1 Baselines

We use a state-of-the-art vision-language model, LXMERT (Tan and Bansal, 2019), as the baselines

[‡]We only credit those predicted answers which obtain an STE score higher than 0.5.

and apply Captioning and Optical Character Recognition (OCR) results to the OK-VQA dataset to the original LXMERT model.

LXMERT LXMERT is a BERT-based cross-modality model pretrained on five different VQA datasets: MS COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), VQA v2.0 (Antol et al., 2015b), GQA balanced version (Hudson and Manning, 2019) and VG-QA (Zhu et al., 2016). We fine-tune LXMERT on OK-VQA and surprisingly find that LXMERT ranks higher than most of the SOTA models, for which reason we set LXMERT as our baseline model.

LXMERT with OCR The OCR technique captures the textual contents from the image and transfers them into characters. Here we use Google Vision API[§] to extract the texts from images. After the noise deduction step filtering all non-English words, we attach the OCR results after the question and then sent them into the LXMERT model. Our experiment shows that the OCR result helps to address the OK-VQA task.

LXMERT with Captioning Similar to OCR, we also experiment with adding captioning when training the LXMERT model. The captions are generated by the advanced model Oscar (Li et al., 2020) and attached to each question when sent into the LXMERT model. Our result shows that captioning improves the performance of the LXMERT model, and therefore, we put the LXMERT with captioning as a baseline as well.

6.2 Main Results

Table 1 shows that our best model based on Caption-DPR and EReader outperforms previous methods and establishes the new state-of-the-art result on the OK-VQA challenge. Interestingly, the LXMERT baseline without utilizing any knowledge achieves better performance than KRISP (Marino et al., 2020) and ConceptBert (Gardères et al., 2020) which leverage external knowledge. Incorporating OCR and captioning further improve the baseline accuracy by 1% and 1.6%, respectively.

Among different variations of Visual Retriever-Reader, the best combination is Caption-DPR and CReader when the retrieved knowledge size is 80. We evaluate retrievers' performance based on Precision and Recall. Table 2 shows that Caption-DPR consistently outperforms BM25 and Caption-DPR on the various number of retrieved knowledge. It is interesting to see that Caption-DPR outperforms BM25 significantly since BM25 is a hard-to-beat baseline in open-domain QA (Lee et al., 2019a; Lewis et al., 2020; Ma et al., 2021). It indicates that neural retriever has better application than term-based retrieval methods in the vision domain.

Method	Knowledge Src.	Acc.	Open Acc.
Existing Method			
KRISP (Marino et al., 2020)	W & C	32.3	-
ConceptBert (Gardères et al., 2020)	C	33.7	-
MAVEx (Wu et al., 2021)	W & C & GI	38.7	-
Baselines			
LXMERT (without pretraining)	-	18.9	25.5
LXMERT	-	36.2	42.6
LXMERT + OCR	-	37.2	42.2
LXMERT + Caption	-	37.8	45.6
LXMERT + OCR + Caption	-	37.2	44.5
Visual Retriever-Reader			
BM25 + CReader	GS	35.13	43.8
BM25 + EReader	GS	32.10	40.6
Image-DPR + CReader	GS	34.64	43.2
Image-DPR + EReader	GS	33.95	41.7
Caption-DPR + CReader	GS	36.78	43.4
Caption-DPR + EReader	GS	**39.20**	**47.3**
Caption-DPR + EReader †	GS	59.22	66.6

Table 1: Performance on the OK-VQA Test-split. Our model outperforms existing methods. † means given oracle knowledge to the reader. GS-Google Search (Training Corpus). W-Wikipedia, C-ConceptNet, GI-Google Image, Acc-Accuracy.

We also present the results obtained by open domain evaluation. First, we observe that the open domain evaluation is correlated with the original accuracy evaluation, i.e., the higher the original accuracy is, the higher the open domain accuracy is. Second, by open domain evaluation, the score is higher than before as some semantic similar answers get credits. We present some examples in Appendix B.

7 Analysis

Effects of the Quality of Knowledge. A common observation in open-domain question answering in NLP is that the reader can perform well if the given knowledge is good to answer a question. Here, we are interested to see if this also holds for our reader. Specifically, before we feed the retrieved knowledge to the reader, we remove knowledge that does not contain any answer, then we send the remaining knowledge to the reader. The last row in Table 1 shows that our reader can perform much better if the quality of the knowledge is good, suggesting that a more efficient cross-modality retriever is needed.

[§]https://cloud.google.com/vision/

Model	# of Retrieved Knowledge													
	1		5		10		20		50		80		100	
	P*	R*	P*	R*	P*	R*	P*	R*	P*	R*	P*	R*	P*	R*
BM25	37.63	37.63	35.21	56.72	34.03	67.02	32.62	75.90	29.99	84.56	28.46	88.21	27.69	89.91
Image-DPR	33.04	33.04	31.80	62.52	31.09	73.96	30.25	83.04	28.55	90.84	27.40	93.80	26.75	94.67
Caption-DPR	**41.62**	**41.62**	**39.42**	**71.52**	**37.94**	**81.51**	**36.10**	**88.57**	**32.94**	**94.13**	**31.05**	**96.20**	**30.01**	**96.95**

Table 2: Evaluation of three proposed visual retrievers on Precision and Recall: Caption-DPR achieves the highest Precision and Recall on all number of retrieved knowledge. We have a * marker on the **P**recision and **R**ecall to distinguish from traditional Precision and Recall as illustrated in §5.1.

Model	# of Retrieved Knowledge						
	1	5	10	20	50	80	100
BM25	+6.00	+6.28	+4.88	+4.32	+3.83	+3.17	+2.56
Image-DPR	+2.24	+2.60	+2.93	+2.29	+1.83	+1.29	+1.25
Caption-DPR	+8.88	+8.88	+7.04	+4.65	+2.98	+2.23	+1.88

Table 3: Recall increases when the Caption-DPR method retrieves knowledge from a complete knowledge corpus created using train and test questions.

Effects of Size of Retrieving Knowledge and Prediction Strategy. The performance of reader is directly affected by the size of retrieved knowledge. A more extensive knowledge set is more likely to include the relevant knowledge to answer the question yet along with more distracting knowledge. In contrast, a small set might exclude relevant knowledge but with fewer distracting knowledge. We use Caption-DPR to retrieve the different number of pieces of knowledge and use the EReader to predict an answer given the different number of pieces of knowledge. We compare the effects on two prediction strategies mentioned in §4.3. Figure 5 shows the comparison, and we have the following observations. First, when the knowledge size is small (equal or less than 5), the Highest-Score strategy is better than the Highest-Frequency; on the other hand, when the knowledge size is large, the Highest-Frequency strategy performs better than the Highest-Score strategy. Second, for the Highest-Score strategy, the size of 5 is the best, and increasing the knowledge size reduces the performance. Third, for the Highest-Frequency strategy, when the size equal to 80, it yields the best performance. To summarize, if one uses a small set of knowledge, then Highest-Frequency negatively impacts the accuracy and the Highest-Score strategy is preferable. If one uses a larger corpus of knowledge, the Highest-Frequency strategy can achieve higher accuracy.

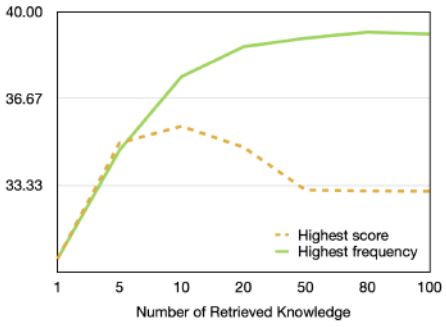

Figure 5: Highest-Score Strategy: Performance of EReader decreases when the knowledge number increase and the best is at 5. Highest-Frequency strategies: Performance of EReader increase when the knowledge number increase and the best is at 80.

Effects of Completeness of Corpus. So far, when we test the model performance, we use the knowledge corpus collected only by training questions. However, if the entire training corpus does not include relevant knowledge to testing questions, our model is under-evaluated because of the incompleteness of the knowledge corpus. To fairly see how our model performs when the knowledge corpus is complete, we use the same knowledge collection method described in §3 to collect knowledge for testing questions. Then we combine the training and testing knowledge as a complete corpus, which increases the corpus size from 112,724 to 168,306. We use Caption-DPR to retrieve knowledge from the complete corpus and ask EReader to predict

answers based on these pieces of knowledge. Table 3 shows the increase of recall. As we expected, a complete corpus is helpful for Caption-DPR even though the corpus size increased, thus yields better performance of EReader. Figure 6 compares the accuracy of EReader using knowledge retrieved from two corpora. EReader consistently achieves higher performance using the knowledge retrieved from complete corpus, where the biggest gain of 7.86% is achieved when using five knowledge. We further clean up the corpus following similar steps in (Raffel et al., 2019), and 1% of the knowledge got removed from the initial ones. We provide the details in Appendix D.

8 Discussion

Training Corpus Bias One potential concern of our knowledge corpus is that the training corpus might tend to bias to the training set, i.e., the training corpus includes knowledge for the training set and excludes knowledge for some testing set. To alleviate such concern, we analyze the training and testing sets in OKVQA and find that 74.69% of testing answers overlap with the training answers, which indicates that a large portion of common knowledge is shared by training and testing sets. To further complement the training corpus, we also provide the entire corpus for testing (discussed in §7), which also includes relevant knowledge for the testing set. The entire corpus is larger than the training corpus and includes prior unseen knowledge. Thus, such testing corpus can evaluate the generalization ability of a retriever, which is an essential skill of any AI system (Mishra et al., 2021).

Extension. Although our pipeline is evaluated on the OK-VQA benchmark, it is generic and can be adapted for other knowledge-based question answering tasks such as FVQA (Wang et al., 2017a), KB-VQA (Wang et al., 2017b), and KVQA (Shah et al., 2019). For example, in KVQA, we can first collect a named-entity knowledge corpus by the proposed knowledge collection approach and then apply our Visual-Retriever-Reader pipeline. It should be noted that our proposed Extractive reader is a more challenging problem as classification models tend to learn correlation between output classes (answers) (Agarwal et al., 2020) and input image and question. In contrast, the extractive reader extracts answer-spans which we exactly match with targets (answers).

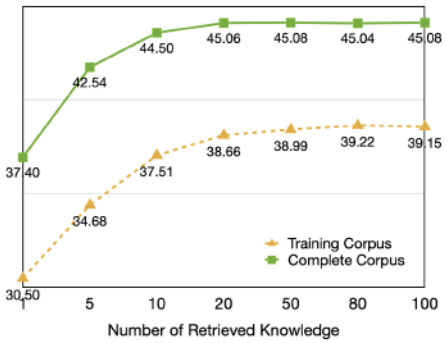

Figure 6: EReader achieves significant improvement when using knowledge retrieved from complete corpus compared to knowledge from training corpus.

9 Conclusion

This paper collects an easy-to-use free-form natural language knowledge corpus for VQA tasks with external knowledge. A weakly-supervised Visual Retriever-Reader Pipeline, where the retriever introduces dense representation, and the reader contains classification and extraction two styles, is also evaluated. The Visual Retriever-Reader Pipeline has been evaluated on the OK-VQA challenge benchmark and has established a new state-of-the-art performance. Further analysis reveals that good knowledge from the retriever makes vital progress in predicting the correct answer. Besides, the captioning and the neural retriever can both significantly improve the QA system's performance.

Acknowledgements

The authors acknowledge support from the NSF grant 1816039, DARPA grant W911NF2020006, DARPA grant FA875019C0003, and ONR award N00014-20-1-2332; and thank the reviewers for their feedback.

Ethical Considerations

All the knowledge base data is fully automatically collected through the website source open to public. We select several sentences from these pages of passages for non-commercial use, which will not violate the intellectual property and privacy rights. The whole dataset is aimed to address the external knowledge source for Knowledge-based QA, and thus each piece contains some commonsense or encyclopedic knowledge. To ensure that the sentences are properly excerpted from the source page, we also visit the source website to extract the complete sentences. (see Step 3 of §3). Our experiment

(§6) also shows that the collected knowledge is helpful to answer the OK-VQA questions.

References

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. 2020. Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9687–9695. IEEE.

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015a. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015b. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2021. WeaQA: Weak supervision via captions for visual question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3420–3435, Online. Association for Computational Linguistics.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. Careful selection of knowledge to solve open book question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6120–6129, Florence, Italy. Association for Computational Linguistics.

Benjamin Börschinger and Mark Johnson. 2011. A particle filter algorithm for Bayesian wordsegmentation. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, pages 10–18, Canberra, Australia.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017a. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017b. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2019. Multi-step retriever-reader interaction for scalable open-domain question answering. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*, pages 489–498, Online. Association for Computational Linguistics.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020a. MUTANT: A training paradigm for out-of-distribution generalization in visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 878–892, Online. Association for Computational Linguistics.

Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. 2020b. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer.

James Goodman, Andreas Vlachos, and Jason Naradowsky. 2016. Noise reduction and targeted exploration in imitation learning for Abstract Meaning Representation parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1–11, Berlin, Germany. Association for Computational Linguistics.

Mary Harper. 2014. Learning from 26 languages: Program management and science in the babel program. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, page 1, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. *ArXiv preprint*, abs/2103.05568.

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)*, pages 6769–6781, Online. Association for Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Bernhard Kratzwald and Stefan Feuerriegel. 2018. Adaptive document retrieval for deep question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 576–581, Brussels, Belgium. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. 2018. Ranking paragraphs for improving answer recall in open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 565–569, Brussels, Belgium. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019a. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019b. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *ArXiv preprint*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

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 *European conference on computer vision*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, and Chitta Baral. 2021. '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, Online. Association for Computational Linguistics.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088, Online. Association for Computational Linguistics.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2020. Krisp: Integrating implicit and symbolic knowledge for open-domain knowledge-based vqa. *ArXiv preprint*, abs/2012.11014.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long*

Beach, CA, USA, June 16-20, 2019, pages 3195–3204. Computer Vision Foundation / IEEE.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Natural instructions: Benchmarking generalization to new tasks from natural language instructions. *ArXiv preprint*, abs/2104.08773.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander G. Schwing. 2018. Out of the box: Reasoning with graph convolution nets for factual visual question answering. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2659–2670.

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

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *ArXiv preprint*, abs/1503.06733.

S. Robertson and H. Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Shailaja Keyur Sampat, Yezhou Yang, and Chitta Baral. 2020. Visuo-lingustic question answering (vlqa) challenge. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4606–4616.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. KVQA: knowledge-aware visual question answering. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 8876–8884. AAAI Press.

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)*, pages 1–18, Kyiv, Ukraine. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

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)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

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

Peng Wang, Qi Wu, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. 2017b. Explicit knowledge-based reasoning for visual question answering. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 1290–1296. ijcai.org.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R 3: Reinforced ranker-reader for open-domain question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2021. Multi-modal answer validation for knowledge-based vqa. *ArXiv preprint*, abs/2103.12248.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004. IEEE Computer Society.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1097–1103. ijcai.org.

A Open Domain Evaluation

Particularly, our evaluating contains three phases: *Grounding* that apply each answer and prediction to the question to ground it as a statement; *Assembling* that rearrange the statements by answers; and

Entailment that calculate the similarity of the different grounded sentences. Then the final score is calculated according to the STE results. Figure 7 gives an overview of the open-domain evaluation. Considering that the Extraction Reader predicts an answer within the open domain, probably resulting in the generated phrases not showing up in the answer field,

Grounding In the grounding phase, we convert a question to a statement using the answers and predictions. Since a good prediction should be of the similar semantic meaning as the answers, we assume that for one question, every answer and prediction acts a same role in the grounded statement, and thus we ground the question with a reserved position for any answer to fill in. For example, the original question "Who invented this device?" is grounded to "_ invented this device.", where "_" can be any of the answers to this question. An example for grounding is shown in Figure 8.

To achieve this, a simple sentence role labeling work is applied to the questions to detect different elements in the sentence (e.g. question word, object, subject, auxiliary word, etc.) After settling the role of elements, the question is then re-ordered to accord with the word order of declarative sentences. We apply the above method to the wh-questions and choice questions, which in total cover the 98.6% of questions and 98.9% of unique answers. Table 4 shows some examples of grounded sentences.

Assembling In grounding step, the statements are gathered by question. We re-arrange the these grounded statements ordered by the provided answers for the further processing. Figure 8 provides an example for this assembling step.

Entailment The grounded sentences are then sent to the Natural Language Inference (NLI) model [¶]. NLI is used widely in the NLP tasks to check whether the hypothesis can be entailed from the given premise, and here we use NLI to check whether the correct answers and the predicted answer are semantically same. To compare between a provided answer and a predicted answer, we first list all grounded statements that use the provided answer as a correct answer. Then, for each of these statements, we fill the reserved position with the provided answer as the premise, and our prediction is the hypothesis, and calculate the entailment score. We use the arithmetic mean of these scores as the final entailment score.

The threshold is set to be 0.5. We also skip the choice questions and the questions with numbers as answers, since, with only grounded statements provided, it is hard to tell whether the two numbers or two choices are similar. For each question with multiple answers, we pick the highest entailment score as the similarity score.

Figure 9 shows the steps acquiring the entailment score and calculating the final score for a predicted answer.

B Examples of Open Domain Evaluation

Here, we show four examples such that the predicted answer given by our model is in fact semantically close to one of the ground truth answer. Such predictions get 0 score given the original accuracy evaluation, but get reasonable score by our proposed open domain evaluation.

C Training Setup

Our neural retrievers were trained on eight Nvidia RTX8000 GPUs, where we set the training epoch to be 30, learning rate (lr) be 1e-5, batch size (bs) be 64, gradient accumulation step (gas) be 4. All the readers were performed at four GTX1080 and V100 NVIDIA GPUs. For both Image-DPR and Caption-DPR, In CReader, we set the training epoch as 3, lr as 2e-5, and batch-size as 16. In EReader, we set the training epoch as 3, lr as 1e-5, batch-size as 4, and gradient accumulation as 4.

D Dataset Cleaning

We cleaned our knowledge corpus following the steps in Section 2.2 of (Raffel et al., 2019). Specifically, we removed the knowledge that contains any word from "List of Dirty, Naughty, Obscene or Otherwise Bad Words". [‖] Knowledge that includes "JavaScript" and "lorem ipsum" is also removed. We also eliminate every knowledge with curly bracket "{". Such cleaning steps remove 1% of the knowledge from the corpus, leading the clean training corpus size to be 111,412. Similarly, 1% of the knowledge from the full corpus is removed to make the clean full corpus size under 166,390.

[¶] https://github.com/allenai/allennlp-models/tree/v1.0.0.rc2/training_config/nli

[‖] https://github.com/LDNOOBW/List%2Dof%2DDirty%2DNaughty%2DObscene-and-Otherwise%2DBad%2DWords

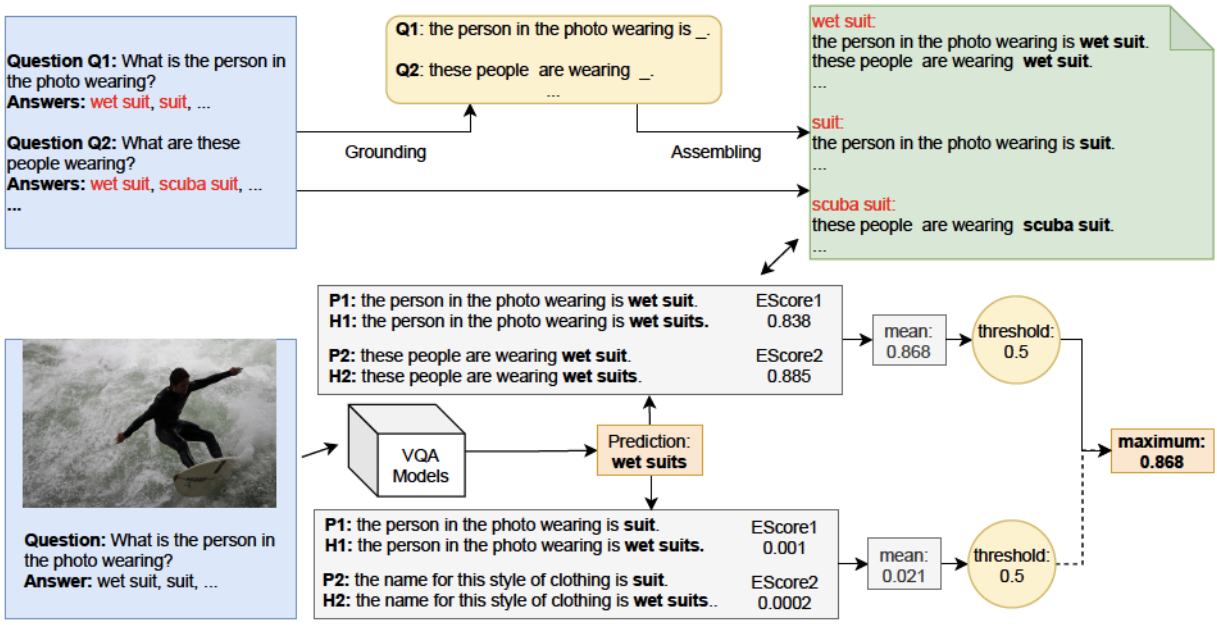

Figure 7: This example calculates the entailment score of provided answer "wet suit" and our prediction "wet suits". We first ground all questions into statements with a reserved position "_" for the answer. Then, we congregate all the grounded statements by the provided answer. We replace the "_" with the provided and predicted answer separately as the premise and hypothesis to get the entailment score. The entailment score of a provided answer and a prediction is calculated as the mean of all the entailment scores under that answer in the assembling list. We take 0.5 as the threshold, and use the maximum as the final entailment score.

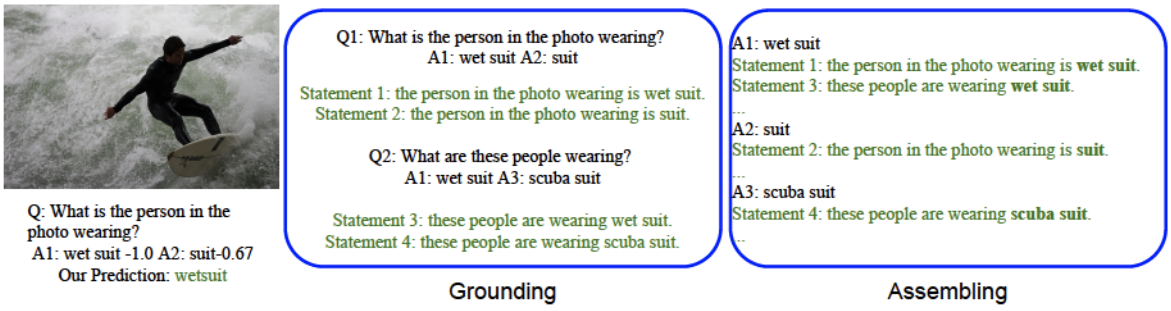

Figure 8: Example of Grounding and Assembling Step in Open-Domain Evaluation.

After obtaining the clean corpus, we apply our best visual retriever and visual reader to the OKVQA challenge. Specifically, first, we apply Caption-DPR fetch 100 knowledge from the clean corpus, then ask EReader to predict the answer. When using the clean training corpus, the accuracy is 39.15, and the accuracy of the clean full corpus is 44.98.

Original Question	Grounded Statement
What is this type of blanket called?	this type of blanket is called _.
What is the name of the board he is on?	the name of the board he is on is _.
The food in the photo contains which healthy vitamins?	The food in the photo contains _ healthy vitamins.
Is this bathroom high or low end?	this bathroom is _.
Why is the cow going to the water?	the cow is going to the water because of _.

Table 4: Examples for some grounded sentences where the hypothesis gets score over the threshold.

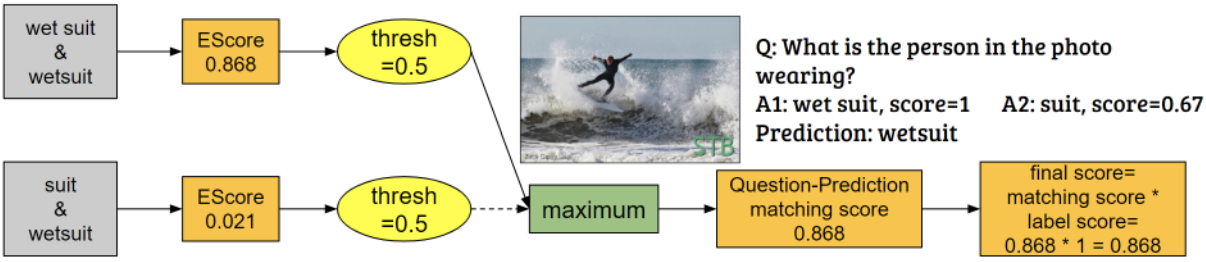

Figure 9: Example of Entailment Step in Open-Domain Evaluation.

Question: Is this a room for a boy or a girl?
Ground-Truth Answer: girl : 1.0
Our Prediction: girls
Original Score: 0
Open-Evaluation Score: 0.88

Question: Name the model of train shown in this picture?
Ground-Truth Answer: steam locomotive : 1.0;
e2 class steam locomotive : 0.67; steam engine : 0.67
Our Prediction: locomotive
Original Score: 0
Open-Evaluation Score: 0.63

Question: What is needed to use this vehicle?
Ground-Truth Answer: license : 1.0; gasoline : 0.67
Our Prediction: fuel
Original Score: 0
Open-Evaluation Score: 0.61

Question: What place of business is this?
Ground-Truth Answer: grocery store : 1.0;
supermarket : 0.67
Our Prediction: grocery
Original Score: 0
Open-Evaluation Score: 0.86

Figure 10: Examples for Open Domain Evaluation.