BREAKING DOWN QUESTIONS FOR OUTSIDE-KNOWLEDGE VQA

Anonymous authors
Paper under double-blind review

ABSTRACT

While general Visual Question Answering (VQA) focuses on querying visual content within an image, there is a recent trend towards Knowledge-Based VQA (KB-VQA) where a system needs to link some aspects of the question to different types of knowledge beyond the image, such as commonsense concepts and factual information. To address this issue, we propose a novel approach that passes knowledge from various sources between different pieces of semantic content in the question. Questions are first segmented into several chunks, and each segment is used as a key to retrieve knowledge from ConceptNet and Wikipedia. Then, a graph neural network, taking advantage of the question's syntactic structure, integrates the knowledge for different segments to jointly predict the answer. Our experiments on the OK-VQA dataset show that our approach achieves new state-of-the-art results.

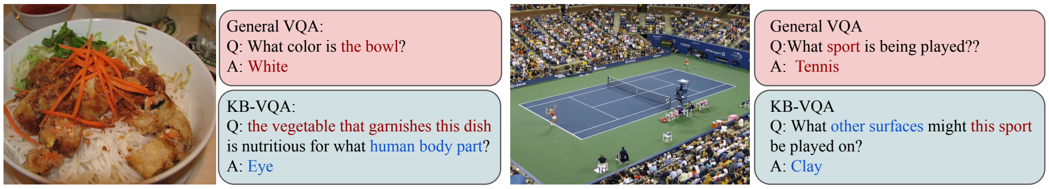
1 INTRODUCTION

Over the past few years, Visual Question Answering (VQA) has emerged as a challenging task where a machine learning system needs to recognize and analyze key visual content within the image and predict an answer to a natural language question. Most recent systems Yu et al. (2019); Lu et al. (2019); Tan & Bansal (2019); Li et al. (2019); Zhou et al. (2020); Chen et al. (2020); Lu et al. (2020) utilize multi-modal transformers to jointly encode the entire question and the visual content, achieving a strong performance on various VQA benchmarks Antol et al. (2015); Hudson & Manning (2019); Singh et al. (2019).

There is a recent trend towards knowledge-based VQA (KB-VQA) Wang et al.; Marino et al. (2019) where the information in the image is not complete for answering the visual questions. These questions cover a wide range of real-world topics, and therefore, require VQA systems to incorporate various types of external knowledge beyond the image content. For example, encyclopedia articles provide factual statements, and common-sense knowledge bases offer everyday concepts and their relations. Both knowledge sources have been proven effective and are widely used in previous work Wang et al.; Marino et al. (2019); Zhu et al. (2020); Li et al. (2020b); Marino et al. (2021); Wu et al. (2021).

While general VQA systems consider two modalities ($i.e.$ question and image), the information across more modalities has to be properly utilized by KB-VQA systems to accommodate different types of knowledge input. This key difference introduces significant challenges to achieving reasonable KB-VQA performance. First, knowledge representations can vary significantly across different knowledge sources, including factual sentences Wu et al. (2021); Marino et al. (2019), knowledge triples Wang et al., concepts Gardères et al. (2020) and images Wu et al. (2021). More importantly, a system needs to understand which knowledge should be used for different semantic segments of the question. As shown in Fig. 1, KB-VQA systems need to link the segment "the vegetable that garnishes this dish" to the carrot on the plate and then query knowledge bases to find out which "human body part" particularly benefits from the nutrients in carrots.

Simply encoding the entire question for either retrieving or filtering the knowledge, as most KB-VQA systems Wang et al.; Marino et al. (2019); Zhu et al. (2020); Li et al. (2020b); Marino et al. (2021); Wu et al. (2021) do, can cause confusion since different parts of the question focus on different aspects that can be either outside or inside the image. As depicted in Fig. 1, searching for "human body part" and "other surfaces" within the image may cause VQA systems to focus on irrelevant



General VQA:
Q: What color is the bowl?
A: White

KB-VQA:
Q: the vegetable that garnishes this dish is nutritious for what human body part?
A: Eye

General VQA
Q:What sport is being played??
A: Tennis

KB-VQA
Q: What other surfaces might this sport be played on?
A: Clay

Figure 1: Examples of general and knowledge-based (KB) visual questions. The question and answer segments that focus on visual content within the image are highlighted in red, and the segments that requires external knowledge are highlighted in blue.

aspects of the image. To address this issue, we introduce a break-down VQA approach that segments visual questions into several semantic chunks, assuming that each chunk focuses on a single aspect. Those segments serve as semantic units and are used to retrieve knowledge from various sources. Finally, using a dependency parser Honnibal & Montani (2017), a Graph Convolutional Network (GCN) Veličković et al. (2018) is constructed which assembles the retrieved knowledge to predict the answer.

We evaluate our framework, break-down VQA, on the OK-VQA dataset Marino et al. (2019), the largest KB-VQA dataset to date. Our approach achieves state-of-the-art results on this benchmark. This demonstrates that breaking down questions and understanding the role of each segment is especially important in answering knowledge-based visual questions.

2 Related Work

2.1 Visual Question Answering

Visual Question Answering (VQA) has witnessed significant progress with the introduction of multi-modal transformers Yu et al. (2019); Zhou et al. (2020); Lu et al. (2020; 2019); Tan & Bansal (2019); Liu et al. (2019); Li et al. (2019; 2020a); Chen et al. (2020). These transformers are pretrained on auxiliary tasks, including VQA, referring-expression interpretation, image captioning, *etc*., using various multi-modal datasets Sharma et al. (2018); Antol et al. (2015); Hudson & Manning (2019); Suhr et al. (2017); Yu et al. (2016); Young et al. (2014). Cross attention modules are built over the textual and visual modalities to learn a joint representation for the entire question and the detected objects. With a large amount of training data and a wide range of pretraining tasks, these models achieve promising performance on various VQA benchmarks Antol et al. (2015); Hudson & Manning (2019); Singh et al. (2019).

2.2 Knowledge-Based Visual Question Answering

While VQA involves visual questions whose answers can be directly found within the image, there is a recent trend toward Knowledge-Based Visual Question Answering (KB-VQA) that requires VQA systems to incorporate knowledge from various external sources.

Recent high-performing KB-VQA systems are mainly learning-based following general VQA systems, and incorporate additional modules to retrieve external knowledge. One Narasimhan & Schwing (2018) learns to retrieve facts from a knowledge base. Another Narasimhan et al. (2018) utilizes a GCN Tompson et al. (2014) over the fact graph where each node is a representation of an image-question-entity triplet. A third Li et al. (2020b) introduces a knowledge-graph augmentation model to retrieve context-aware knowledge sub-graphs, and then learns to aggregate the useful visual and question relevant knowledge. Finally, KRISP Marino et al. (2021) combines knowledge from both implicit question-image embedding and explicit symbolic information from knowledge bases.

Although the knowledge is obtained from a wide range of sources and encoded in different formats, most previous systems simply learn to mine relevant facts based on the entire question, which, as mentioned above, could cause confusion. In contrast to previous work, we present an approach that breaks the question down into several segments and then uses each of these segments to retrieve the appropriate knowledge, which is then integrated to answer the overall question.

2.3 Breaking Down Visual Questions.

Previous work has explored both rule-based Andreas et al. (2016); Wolfson et al. (2020) and learning-based Hu et al. (2017; 2018); Mao et al. (2019); Wolfson et al. (2020) approaches to break down visual questions. Rule-based approaches typically define a set of decomposition rules and a full decomposition is obtained by recursively applying those rules until no rule is matched. In particular, one method Andreas et al. (2016) parses the questions and breaks it into a sequence of programs to execute. Another Wolfson et al. (2020) breaks the question into several steps each of which is encoded as a natural language expression. Learning-based approaches either learn to recursively rank some predefined modules to synthesize the entire network layout for solving a visual question Hu et al. (2017; 2018) or directly learn to generate the steps using a seq2seq method Mao et al. (2019); Wolfson et al. (2020). These approaches work especially well for datasets that represent queries as programs, including CLVER Johnson et al. (2017) and GQA Hudson & Manning (2019).

2.4 Neural Module Networks

A Neural Module Network (NMN) Andreas et al. (2016) consists of a layout generator and an executor. The layout generator synthesizes an instance-specific network from a predefined set of basic modules by passing arguments parsed from the question. The executor then evaluates the network to predict the answer. Existing work has explored both rule-based and learning-based layout generators. One approach Andreas et al. (2016) generates the layout using a dependency parser. Another Hu et al. (2017) adopts a learning-based approach that first defines a limited set of modules and ranks them based on the question parse tree. A third Mao et al. (2019) directly generates the layout using a seq2seq method, without the need to parse the question.

2.5 Graph Convolutional Networks

Graph Convolutional Networks (GCNs) Kipf & Welling (2017) generalize Convolutional Networks (CNN) to accommodate graph-structured input. Various types of graph input for VQA have been explored including scene graphs generated by an object and relation detector Ren et al. (2015); Yang et al. (2018), and knowledge graphs retrieved from a wide range of sources, such as DB-Pedia Auer et al. (2007), ConceptNet Liu & Singh (2004), VisualGenome Krishna et al. (2017) and hasPart KB Bhakthavatsalam et al. (2020). Most KB-VQA systems Ramnath & Hasegawa-Johnson (2021); Narasimhan et al. (2018); Li et al. (2020b); Marino et al. (2021) build their GCNs on top of these knowledge graphs and extract relevant evidence using the entire question representation. Here, we explore an approach that constructs a reasoning graph from the question, where each node is a semantic segment of the question. Our graph utilizes the syntactic structure of the questions to better integrate the question segments that utilize both the visual content in the image and relevant external knowledge.

3 Approach

We present the break-down VQA approach, a three-step framework. First, it segments visual questions into semantic chunks. Next, each segment, serving as a semantic unit, is used to retrieve knowledge from different external sources. Finally, a Graph Neural Network (GCN) integrates this retrieved knowledge to predict an answer. Fig. 2 illustrates the approach.

We instantiate our approach on top of the high-performing ViLBERT-multi-task as a base system Lu et al. (2020) that provides a set of answer candidates $A = \{a_1, ..., a_n\}$ for each question-image pair. We also extract the product of its pooled features for the textual and visual BERT output, **z**, as a joint representation of the question and the image.

3.1 Breaking Down Visual Questions

Given a visual question q that consists of l tokens $(q_1, ..., q_l)$ where a token is either a word or a WordPiece produced by a tokenizer Vaswani et al. (2017), and its question segmentation is a set of token chunks $X = (x^1, ..., x^m)$ where each x^i consists of a sub-sequence of q, $i.e.$ $x^i = (q^i_1, ..., q^i_{l_i})$.

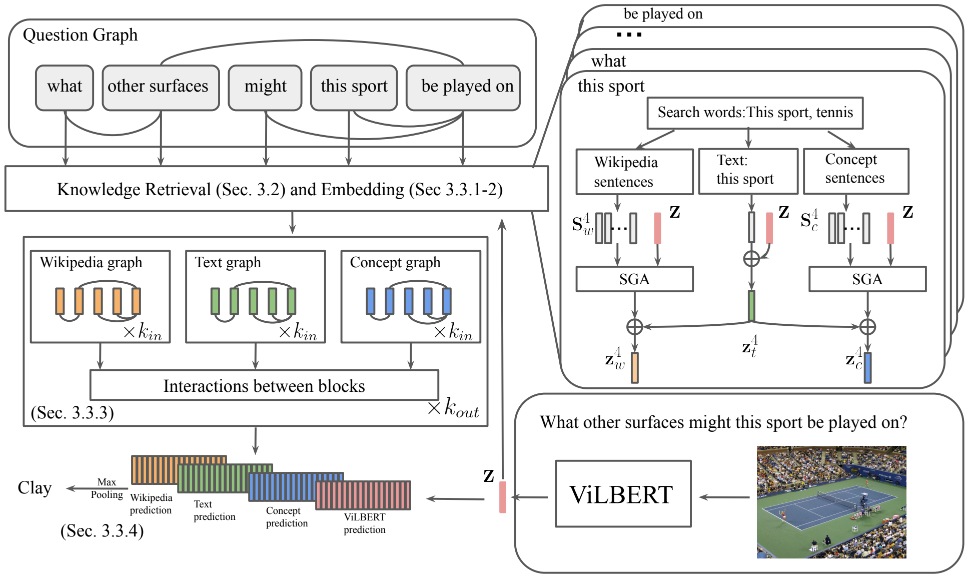


Figure 2: Model overview of the BreakDown VQA approach. The question is segmented into semantic chunks (left top). These chunks are used to retrieve external information from Wikipedia and ConceptNet. Each retrieved piece of knowledge is then encoded as a vector (right top), and fed to a graph neural net (left middle) to predict an answer for each knowledge source. The individual results are then max-pooled to get the final prediction (left bottom).

Since different parts of a knowledge-based visual question need to focus on different sources of knowledge, encoding the entire question as a whole leads to inefficiency in both knowledge retrieval and answer prediction. To address this issue, our approach breaks down the question into segments where each segment contains only one semantic unit that can either be grounded in the image, or linked to external knowledge bases. Note that, we do not restrict each segment to only retrieve from a single knowledge source but let the VQA model choose the right source.

To this end, we first extract nouns, noun chunks, and verbs in the question as knowledge segments. For example, 'other surfaces', 'this sport', and 'play' are extracted for the second example in Fig. 1. Specifically, we utilize the 'en_core_web_sm' sPacy parser Honnibal & Montani (2017) to dependency parse the question and POS-tag each word. Then, we extract the noun chunks (i.e. flattened phrases with a noun head in the parse tree) and lemmatized verbs. We also group any tokens between those extracted knowledge segments as additional segments to ensure completeness.

3.2 Knowledge Retrieval

Given the extracted segments, we retrieve knowledge from Wikipedia and ConceptNet in two main steps. Answer-guided knowledge retrieval Wu et al. (2021) is adopted to ensure the relevancy of the external knowledge.

Search Word Extraction. We first remove the stop words in the segment and regard the remaining tokens as the search words. Then, we enrich these search words with object annotations, including linking the segment to objects in the image, text extracted using OCR, and brand detection following Wu et al. (2021).[1] In particular, a pretrained ViLBERT-multi-task model Lu et al. (2020) is used as the object linker. This system can generate linking scores indicating the confidence of linking phrases to detected objects. The linking is approved when its score is over 0.5. With the linked objects, a Google API is used to recognize words in text regions using OCR and company brands. Besides, we

[1] See section "S1: Answer-Agnostic Search Word Extraction"

also detect common attributes of these objects using a Faster-RCNN Ren et al. (2015) on a Detectron platform pretrained on Visual-Genome data.[2] This process results in a set of search words for each segment. For example, the search word set for the segment 'the vegetable' for the first example in Fig. 1 is {'vegetable', 'carrot', 'red vegetable'}.

Knowledge Retrieval. We use two knowledge sources to extract information about the question segments in X, $i.e.$ relevant textual facts and commonsense concepts as in Wu et al. (2021). In contrast to Wu et al. (2021), we retrieve knowledge independently for each segment instead of for the entire question. This ensures that the retrieved knowledge provides information about the given segment, and allows the VQA system to determine whether a particular piece of external knowledge about this segment is important.

Retrieving from Wikipedia. For each segment x_i, we query its search words and collect all sentences from the retrieved Wikipedia articles. We use answer-guided knowledge retrieval Wu et al. (2021) to filter out irrelevant sentences. Specifically, we first keep the Wikipedia sentences that contain both at least one of the search words and one of the top 5 answer candidates predicted by ViLBERT-multi-task. Then, the remaining sentences are ranked according to the highest precision BERT-scores Zhang et al. (2020) between the sentence and statements converted from the question Demszky et al. (2018) and the top-5 answer candidates. We keep the top-80 sentences in total for each visual question and regard the other sentences as irrelevant.

Retrieving from ConceptNet. Commonsense concepts provide structured knowledge that is usually not covered in factual Wikipedia sentences. Similar to Wikipedia-article retrieval, we query the search words for each segment and collect the retrieved concepts. First, we keep all the concept triples whose subjects and objects contain the search word and one of the answer candidates from A. Then, we convert other concept triples to sentences and rank them according to the highest precision BERT-scores between the sentence and the statements from the question and answers. We also keep the top-80 sentences in total for each visual question and regard the other concepts as irrelevant.

Matching Textual Knowledge: For each query, the sentences from Wikipedia and the concepts from ConceptNet with a mean recall greater than 0.6 are matched to the search words. Mean recall is defined as the average cosine similarity between the Glove embedding of the words in the search word and their most similar word in the sentence or in the concept. To ensure knowledge relevance, we remove sentences that are matched to only a single search word. We keep the top k_w sentences $S_w^i = \{s_{w,1}^i, ..., s_{w,k_w}^i\}$ according to the mean recall as the textual facts for segment x_i from its search word set, where $s_{w,j}^i$ denotes the j-th Wikipedia sentence for the i-th segment. Similarly, for concepts, we keep the top k_c concept sentences $S_c^i = \{s_{c,1}^i, ..., s_{c,k_c}^i\}$ for segment x_i, where $s_{c,j}^i$ denotes the j-th concept sentence for the i-th segment.

3.3 VQA MODEL

This section describes the final VQA system that incorporates the retrieved knowledge for each of the semantic segments. We first generate features for each knowledge sentence from Wikipedia and ConceptNet. Then a representation of each source for each segment is computed using these sentence features. Finally, a GCN is employed that utilizes the syntactic structure of the visual question and produces joint features for predicting the answer.

3.3.1 KNOWLEDGE SENTENCE EMBEDDING

We use a word embedding matrix initialized by GloVe vectors Pennington et al. (2014) to compute a word vector for each token in the knowledge sentence. Then, a single layer LSTM with a hidden states of 768 is built on top of the word embeddings and the features for the last token are extracted. This process produces a 768-d feature vector for each sentence from both Wikipedia S_w^i and ConceptNet S_c^i, resulting in knowledge feature matrices $\mathbf{S}_w^i \in R^{k_w \times 768}$ and $\mathbf{S}_c^i \in R^{k_c \times 768}$ for segment i, respectively.

[2] We were careful to remove the OK-VQA test images from the training data for the Faster-RCNN system.

3.3.2 Segment Embedding

We produce an embedding for each segment by integrating three representations, a content representation of the text of the segment in the question and two representations of external knowledge (Wikipedia + ConceptNet).

Content Embedding. To preserve all of the information in the question, we employ the text of each segment as input to the VQA model. We use the GloVe embedding approach to encode segments. Similar to the knowledge sentence embedding, an LSTM is used to sequentially encode the GloVe vectors and the hidden state of the last token is extracted as the content representation $\mathbf{s}_t^i$. The final content embedding of segment i is computed as the element-wise summation of $\mathbf{s}_t^i$ and the projection of $\mathbf{z}$, $i.e.$ $\mathbf{z}_t^i = \mathbf{s}_t + \texttt{fc}(\mathbf{z})$, where $\texttt{fc}$ denotes a fully connected layer.

Knowledge Embedding. As shown in Eqs. 1 and 2, we embed the knowledge matrices $\mathbf{S}_w^i$ and $\mathbf{S}_c^i$ for segment x_i into vector representations $\mathbf{z}_w^i$ and $\mathbf{z}_c^i$ that contain the question-relevant information from the external knowledge source sentences S_w^i and S_c^i . In particular, we utilize a Self- and Guided- Attention (SGA) module Yu et al. (2019) where the question and image representation $\mathbf{z}$ from ViLBERT is used as a query, and the knowledge matrices serve as keys and values. The $\texttt{SGA}$ modules provide a trainable method for mining question-relevant knowledge from the retrieved materials in contrast to the rule-based method used in the knowledge retrieval process. In order to prevent the case where the retrieved knowledge is empty, we add the content embedding to the knowledge embedding for each source.

$$\mathbf{z}_w^i = \texttt{SGA}(\mathbf{z}, \mathbf{S}_w^i) + \mathbf{z}_t^i \tag{1}$$

$$\mathbf{z}_c^i = \texttt{SGA}(\mathbf{z}, \mathbf{S}_c^i) + \mathbf{z}_t^i \tag{2}$$

3.3.3 Graph Neural Networks

Building the Graph Structure. We treat the segments' embeddings $\{\mathbf{z}_k^i\}$ as nodes, where i denotes the segment's index and k indexes the knowledge source, and establish an edge between each pair if there is a direct connection between tokens from the two segments in the dependency parse tree. Given the parse tree $\mathcal{E}_q$ of question q, which establishes edges between tokens in q, the edges of the segments $\mathcal{E}$ are defined in Eq. 3:

$$\mathcal{E} = \{(\mathbf{z}_k^i, \mathbf{z}_k^j)|\ \exists(q^m \in x^i, q^n \in x^j)(q^m, q^n) \in \mathcal{E}_q\} \tag{3}$$

This produce a graph structure $\mathcal{G}^k = (\{\mathbf{z}_k^i\}, \mathcal{E})$ for each modality k.

Graph Neural Networks Architectures. The networks consists of k_{out} blocks, where each block contains k_{in} graph layers. The node features within each block interact with other nodes' features from the same modality, determining its importance to solve the visual question. The knowledge from different external modalities is fused outside the blocks to build connections to other types of knowledge.

Graph Neural Networks within Blocks. We formalize the input to the graph neural networks as $\mathbf{H}_{i,0}^k$ where k denotes the source of the question segments' features, and i is the index of the block. For layer l within block i, we use a graph layer that operates a non-linear function $\mathrm{F}(\mathbf{H}_{i,l}^k, \mathcal{G}^k)$, producing the input to the next graph layer $\mathbf{H}_{i,l+1}^k$, $i.e.$ $\mathbf{H}_{i,l+1}^k = \mathrm{F}(\mathbf{H}_{i,l}^k, \mathcal{G}^k)$. The input $\mathbf{H}_{i,0}^k$ for block i is the output of the previous block $\mathbf{H}_{i-1,l}^k$ after interactions between modalities described below except for the first block that receives the segments' features as inputs, $i.e.$ $\{\mathbf{z}_k\}$.

Interactions between Modalities outside the Blocks. To give the graph neural networks access to the all types of external knowledge, we fused features from different modalities outside the blocks. The fused features serve as the inputs to the next blocks of graph neural nets. In particular, the input $\mathbf{H}_{i+1,0}^k$ to the $i+1$ block is the concatenation of the segments representation $\{\mathbf{z}_k\}$ and the summation of the output of the previous block from all modalities.

3.3.4 Answer Prediction

We build answer prediction heads for each knowledge source that compute a probability distribution over all answer candidates. The knowledge features from the last block, $i.e.$ $\mathbf{H}_{k_{out},k_{in}}^k$ are averaged

Method	**Knowledge Resources**	**Performance**
ViLBERT Lu et al. (2019)	—	36.1
MMBERT Marino et al. (2019)	—	37.1
KRISP Marino et al. (2021)	Wikipedia + ConceptNet	37.8
KRISP(incl. graph pretraining)	Wikipedia + ConceptNet	38.9
MAVEx Wu et al. (2021)	Wikipedia + ConceptNet + Google Images	38.7
Ours	Wikipedia + ConceptNet	39.1
Ours + MAVEx	Wikipedia + ConceptNet + Google Images	40.8
Ours + MAVEx (oracle)	Wikipedia + ConceptNet + Google Images	**42.5**

Table 1: Our approach outperforms current state-of-the-art approaches on the OK-VQA dataset. The middle column lists the external knowledge sources, if any, used by each VQA system.

and fed to the answer prediction head that consists of two consecutive fully-connected layers with ReLU activation. Then, we take the maximum value of these predictions for each answer candidate as the final answer predictions.

4 Implementation and Training Details

Implementation. Our break-down VQA approach is implemented on top of ViLBERT-multi-task Lu et al. (2019), which utilizes a Mask-RCNN head He et al. (2017) in conjunction with a ResNet-152 base network He et al. (2016) as the object detection module. Convolutional features for at most 100 objects are extracted for each image as the visual features, *i.e.* a 2,048 dimensional vector for each object.

Since the OK-VQA test dataset contains COCO images from the validation set that are used to train the officially released ViLBERT model, we retrain the system from scratch using clean datasets where we remove all of the OK-VQA test images from the Visual Genome, MSCOCO, and GQA datasets. We used the default configuration when training the object detection module, pretraining on Conceptual Captions, and finally finetuning on the 12 visual-and-language tasks used in Lu et al. (2020). We utilize a BERT tokenizer Devlin et al. (2019) to tokenize the question and use the first 23 tokens of the question. We encode the top 5 Wikipedia sentences and top 10 ConceptNet concepts for each knowledge segment, *i.e.* k_w = 5 and k_c = 10. The number of hidden units in the SGA modules in the knowledge embedding modules is set to 768. We use 4 attention heads in the SGA modules. The graph neural networks contain 2 blocks and 4 layers within each block. A SAGE Hamilton et al. (2017) layer with transformed root node features is used as the graph layer. The Pytorch Geometric toolbox Fey & Lenssen (2019) is used for the GCN implementation.

Training. For training, we optimize the answer predictions for each knowledge source using the standard VQA loss, together with the VQA loss on the final predictions. We train the system for 75 epochs using a learning rate of 2e-5 for the ViLBERT parameters and 5e-5 for the additional parameters introduced in the BreakDown VQA system. We freeze the first 10 layers of the ViLBERT base network.

5 Experiments

This section evaluates our BreakDown VQA approach on the OK-VQA dataset Marino et al. (2019). We first briefly describe the dataset, and then present results comparing to current state-of-the-art systems.

OK-VQA dataset. This is currently the largest knowledge-based VQA dataset. The questions are crowdsourced from human workers on Amazon Mechanical Turk instead of artificially synthesized from knowledge bases. Human judges are asked to ensure that outside knowledge beyond the image is required. Also, since it is not synthesized, there are no ground truth knowledge bases that can provide a VQA system all of the necessary external knowledge. Therefore, systems have to retrieve knowledge from a variety of knowledge sources. The dataset contains 14,031 images and 14,055 questions covering a variety of topics, including transportation, brands, material, sports, cooking, geography, plants, animals, science, weather, *etc.*

5.1 Main Results

We report results on version 1.1 of the OK-VQA dataset in Table 1, unlike the original version ($i.e.$ version 1.0), answers are lemmatized to improve scoring. Our BreakDown VQA approach outperforms all previous systems, achieving a new state-of-the-art accuracy score of 39.1%.

5.2 Ablation Study on Source Knowledge

This sections gives results when we ablate the external knowledge sources. In particular, we manually zero out the knowledge features $\mathbf{z}_k$ to exclude the external information obtained from knowledge source k during training and test. We use 2 blocks and 4 layers within each block in the graph neural networks. As shown in Table 2, each knowledge source helps improve the overall performance, indicating the need to access to a variety of knowledge sources for solving the KB visual questions.

Sources	**Performance**
Wikipedia	38.2
ConceptNet	38.5
Wikipedia + ConceptNet	39.1

Table 2: Ablation study of knowledge sources.

5.3 Ablation Studies on the Graph Model

Table 3 shows results on how different values for the hyper-parameters in the GCN influence VQA performance. It includes an extreme case using only one graph block ($i.e.$ k_{out} = 1), where the knowledge sources do not interact and predict the answer independently. We also tested two ablated models to test the contribution of the graph structure that exploits the parse tree of the question. We simply build the answer prediction heads on top of the knowledge embedding of each source, $\mathbf{z}_k$, where k is the knowledge source indicator. This baseline system achieves a score of 38.5, and a fully-connected graph achieves 38.7. That indicates that building the segments' graph using the question's syntactic structure helps the VQA system improve its use of the retrieved knowledge, improving the results.

Number of Blocks k_{out}	Number of layer with Blocks k_{in}	**Performance**
1	4	38.6
2	4	39.1
2	6	38.7
3	4	38.8

Table 3: Ablation study using different GCN hyper-parameter values.

5.4 Using BERT for Knowledge Embedding

We also tested a BERT-based knowledge embedding for encoding the retrieved sentences from the external knowledge sources. We used a pretrained BERT-base-uncased model Devlin et al. (2019) to compute the features for each sentence. We extract the final layer representation for the "[CLS]" token as the sentence embedding to replace the GloVe embedding used in Sec. 3.3.1. Note that this BERT model is not finetuned for VQA. The BERT Embedding approach achieves a score of 38.9 compared to 39.1 using the GloVe embedding. Our hypothesis is that though BERT features may encode richer information, fine-tuning on the down-stream task is important for the final performance.

5.5 Combining with Answer Validation

Previous work on OK-VQA Wu et al. (2021) introduced an answer validation module (MAVEx) that reweighs the answer confidence with a verification score obtained by examining the knowledge retrieved for each of the top answer candidates. MAVEx also uses retrieved images from Google as a

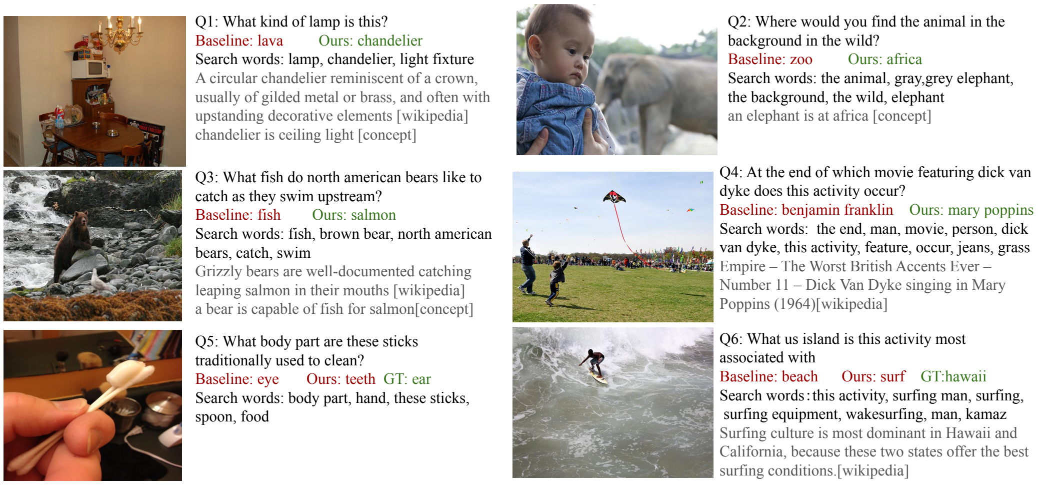

Q1: What kind of lamp is this?
Baseline: lava
Ours: chandelier
Search words: lamp, chandelier, light fixture
A circular chandelier reminiscent of a crown, usually of gilded metal or brass, and often with upstanding decorative elements [wikipedia]
chandelier is ceiling light [concept]
Q2: Where would you find the animal in the background in the wild?
Baseline: zoo
Ours: africa
Search words: the animal, gray,grey elephant, the background, the wild, elephant
an elephant is at africa [concept]
Q3: What fish do north american bears like to catch as they swim upstream?
Baseline: fish
Ours: salmon
Search words: fish, brown bear, north american bears, catch, swim
Grizzly bears are well-documented catching leaping salmon in their mouths [wikipedia]
a bear is capable of fish for salmon[concept]
Q4: At the end of which movie featuring dick van dyke does this activity occur?
Baseline: benjamin franklin
Ours: mary poppins
Search words: the end, man, movie, person, dick van dyke, this activity, feature, occur, jeans, grass
Empire – The Worst British Accents Ever – Number 11 – Dick Van Dyke singing in Mary Poppins (1964)[wikipedia]
Q5: What body part are these sticks traditionally used to clean?
Baseline: eye
Ours: teeth
GT: ear
Search words: body part, hand, these sticks, spoon, food
Q6: What us island is this activity most associated with
Baseline: beach
Ours: surf
GT:hawaii
Search words: this activity, surfing man, surfing, surfing equipment, wakesurfing, man, kamaz
Surfing culture is most dominant in Hawaii and California, because these two states offer the best surfing conditions.[wikipedia]

Figure 3: Qualitative results from our Break Down VQA and a ViLBERT baseline. Q1-Q4 show success cases and Q5 and Q6 illustrate a couple failure cases. Red and green denote wrong and right answers, respectively.

third knowledge source to provide visual external knowledge. We combined our BreakDown VQA approach with a static MAVEx system that provides the weights of the top 5 answer candidates. As shown in Table 1, we achieve a score of 40.8 when combining the MAVEx weights using predicted answer candidates and 42.5 when using an oracle answer candidate set where a ground truth answer is manually inserted into the answer candidate set during validation. This shows that our approach can be effectively combined with other recent advances in KB-VQA to further improve the state-of-the-art.

5.6 Qualitative Results

We show some representative examples of our approach versus a ViLBERT baseline system in Fig. 3. Q1 shows an example where the answer is already in the search word list ($i.e.$ chandelier), illustrating the effectiveness of enriching the segments parsed from the question with various types of annotations. Q2-Q4 show examples where our approach successfully retrieves relevant knowledge about specific segments which allows it to predict the correct answer. Q4 shows an example where Wikipedia knowledge is especially helpful and Q3 shows an example where both knowledge sources provide useful information.

We also show some common failure cases from our approach in Q5 and Q6. Q5 shows an example where object recognition fails since the cotton swabs are just annotated as sticks, making it hard to retrieve the relevant knowledge. Q6 shows a case where the retrieved knowledge seems helpful but the final prediction is wrong. It seems the VQA system failed to understand that the question is asking about a location rather than an activity.

6 Conclusion

We have introduced a novel approach to knowledge-based VQA that breaks down visual questions requiring external knowledge into multiple semantic segments which are used to drive the retrieval of relevant knowledge from multiple external sources that include both text (Wikipedia) and structured knowledge (ConceptNet). This approach achieves a new state of the art on the challenging OK-VQA benchmark, the largest available crowdsourced KB-VQA dataset. We find that segmenting questions is especially helpful for open-domain KB-VQA because different parts of the question require utilizing different types of information, such as linking to objects in the image and exploiting factual information from encyclopedias or commonsense knowledge from knowledge-bases.

REFERENCES

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural Module Networks. In *CVPR*, 2016.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A Nucleus for a Web of Open Data. In *The semantic web*. Springer, 2007.

Sumithra Bhakthavatsalam, Kyle Richardson, Niket Tandon, and Peter Clark. Do Dogs Have Whiskers? A New Knowledge Base of hasPart Relations. *arXiv preprint arXiv:2006.07510*, 2020.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: UNiversal Image-TExt Representation learning. In *ECCV*. Springer, 2020.

Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming Question Answering Datasets into Natural Language Inference datasets. *arXiv preprint arXiv:1809.02922*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ACL*, 2019.

Matthias Fey and Jan E. Lenssen. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

William L Hamilton, Rex Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs. *NeurIPS*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017.

Matthew Honnibal and Ines Montani. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. To appear, 2017.

Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to Reason: End-to-End Module Networks for Visual Question Answering. In *ICCV*, 2017.

Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable Neural Computation via Stack Neural Module Networks. In *ECCV*, 2018.

Drew A Hudson and Christopher D Manning. GQA: A New Dataset for Compositional Question Answering over Real-World Images. *CVPR*, 2019.

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 *CVPR*, 2017.

Thomas N Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *ICLR*, 2017.

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. *IJCV*, 2017.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A UNiversal encoder for Vision and Language by Cross-Modal Pre-training. In *AAAI*, 2020a.

Guohao Li, Xin Wang, and Wenwu Zhu. Boosting Visual Question Answering with Context-aware Knowledge Aggregation. In *ACM Conference on Multimedia*, 2020b.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A Simple and Performant Baseline for Vision and Language. *arXiv preprint arXiv:1908.03557*, 2019.

Bei Liu, Zhicheng Huang, Zhaoyang Zeng, Zheyu Chen, and Jianlong Fu. Learning Rich Image Region Representation for Visual Question Answering. *arXiv preprint arXiv:1910.13077*, 2019.

Hugo Liu and Push Singh. ConceptNet: a Practical Commonsense Reasoning Tool-kit. *BT technology journal*, 2004.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*, 2019.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-Task Vision and Language Representation Learning. In *CVPR*, 2020.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences from Natural Supervision. *ICLR*, 2019.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *CVPR*, 2019.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA. In *CVPR*, 2021.

Medhini Narasimhan and Alexander G Schwing. Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering. In *ECCV*, 2018.

Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out-of-The-Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering. In *NeurIPS*, 2018.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *EMNLP*, 2014.

Kiran Ramnath and Mark Hasegawa-Johnson. Seeing is Knowing! Fact-based Visual Question Answering using Knowledge Graph Embeddings. *AAAI*, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks. In *NIPS*, 2015.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*, 2018.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA Models that Can Read. In *CVPR*, 2019.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *ACL*, 2017.

Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*, 2019.

Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *NeurIPS*, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All You Need. In *NeurIPS*, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph Attention Networks. *ICLR*, 2018.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-Based Visual Question Answering. *PAMI*.

Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant. Break It Down: A Question Understanding Benchmark. *TACL*, 2020.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. Multi-Modal Answer Validation for Knowledge-Based VQA. *arXiv preprint arXiv:2103.12248*, 2021.

Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph R-CNN for Scene Graph Generation. In *ECCV*, 2018.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *TACL*, 2014.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling Context in Referring Expressions. In *ECCV*, pp. 69–85. Springer, 2016.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep Modular Co-Attention Networks for Visual Question Answering. In *CVPR*, 2019.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. *ICLR*, 2020.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified Vision-Language Pre-Training for Image Captioning and VQA. In *AAAI*, February 2020.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering. In *IJCAI*, 2020.