Multi-Modal Answer Validation for Knowledge-Based VQA

Jialin Wu[1], Jiasen Lu[2], Ashish Sabharwal[2], Roozbeh Mottaghi[2]
[1]University of Texas, Austin, [2]Allen Institute for AI

Abstract

*The problem of knowledge-based visual question answering involves answering questions that require external knowledge in addition to the content of the image. Such knowledge typically comes in a variety of forms, including visual, textual, and commonsense knowledge. The use of more knowledge sources, however, also increases the chance of retrieving more irrelevant or noisy facts, making it difficult to comprehend the facts and find the answer. To address this challenge, we propose **M**ulti-modal **A**nswer **V**alidation using **Ex**ternal knowledge (MAVEx), where the idea is to validate a set of promising answer candidates based on answer-specific knowledge retrieval. This is in contrast to existing approaches that search for the answer in a vast collection of often irrelevant facts. Our approach aims to learn which knowledge source should be trusted for each answer candidate and how to validate the candidate using that source. We consider a multi-modal setting, relying on both textual and visual knowledge resources, including images searched using Google, sentences from Wikipedia articles, and concepts from ConceptNet. Our experiments with OK-VQA, a challenging knowledge-based VQA dataset, demonstrate that MAVEx achieves new state-of-the-art results.*

1. Introduction

Over the past few years, the domain of Visual Question Answering (VQA) has witnessed significant progress [2, 41, 13, 32]. There is a recent trend towards knowledge-based VQA [37, 36, 25] which requires information beyond the content of the images. To correctly answer those challenging questions, the model requires not only the ability of visual recognition, but also logical reasoning and incorporating external knowledge about the world. These knowledge facts can be obtained from various sources, such as image search engines, encyclopedia articles, and knowledge bases about common concepts and their relations.

Figure 1 illustrates a few visual questions and the knowledge from different external sources required to answer them. Each question needs a different type of external knowledge. For example, to identify the movie that featured a man telling his life story to strangers, we need to link the image content and question to some textual facts (blue box in the figure); Vegetarian food and eating vegetables is related to the concept of health (green box); and the retrieved images for 'golden retriever' (yellow box) are visually similar to the dog in the question image. *The challenge is to effectively retrieve and correctly incorporate such external knowledge* in an open domain question answering framework.

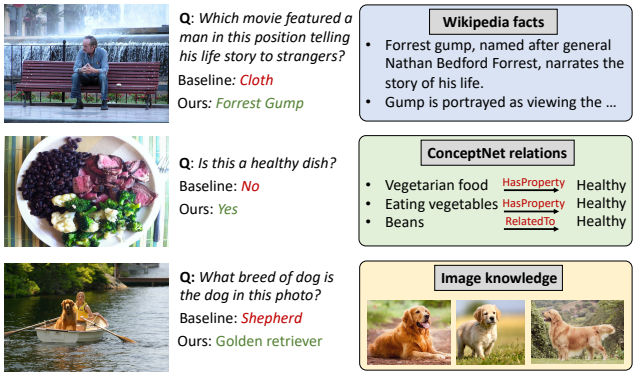

Figure 1: We address the problem of knowledge-based question answering. Retrieving relevant knowledge among diverse knowledge sources (visual knowledge, textual facts, concepts, etc.) is quite challenging. The goal in this paper is to learn what knowledge source should be used for a particular question and how to validate a set of potential answer candidates using that source.

We also witness a shift on knowledge-based VQA datasets—from structured retrieved knowledge such as triplets and dense captions [37, 36] to unstructured open knowledge [25]. Most recent knowledge-based VQA systems [25, 36, 42, 24] follow a two-stage framework, where a retriever first looks up knowledge relevant to the question and the image, and then a separate comprehension model predicts the answer.

However, knowledge retrieved directly for the question and image is often noisy and not useful for predicting the correct answer. For example, as shown in Figure 2, the sentences retrieved using only the words in questions and objects in images (top) or a wrong answer (middle) are hardly

What English city is famous for a tournament for the sport this man is playing?

Question + Image

The modern game of tennis originated in Birmingham, England, in the late 19th century as lawn tennis.

Question + Image + Incorrect Answer (Copenhagen)

It is popular for sports fixtures and hosts several annual events including a free opera concert at the opening of the opera season, other open-air concerts, carnival and labour day celebrations, and the Copenhagen historic grand prix, a race for antique cars.

Question + Image + Correct Answer (Wimbledon)

Wimbledon is notable for the longest running sponsorship in sports history due to its association with slazenger who have supplied all tennis balls for the tournament since 1902.

Figure 2: Examples of retrieved Wikipedia sentences using different sets of search words. The sentences retrieved using only the words in questions and objects in images (top) and the wrong answer (middle) are hardly helpful to answer the question. However, with the correct answer "Wimbledon" (bottom), the quality of the retrieved fact is significantly improved.

helpful to answer the question. This increases the burden on the answer predictor, leading to only marginal improvements from the use of retrieved knowledge [25]. Interestingly, with the correct answer "Wimbledon" (bottom), the quality of the retrieved fact is significantly improved, making it useful to answer the question. This observation motivates us to use retrieved knowledge for *answer validation* rather than for producing the answer.

To address this challenge, we propose a new framework called MAVEx or **M**ulti-modal **A**nswer **V**alidation using **Ex**ternal knowledge. The key intuition behind MAVEx is that verifying the validity of an answer candidate using retrieved knowledge is more reliable compared to open knowledge search for finding the answer. Therefore, we learn a model to evaluate the validity of each answer candidate according to the retrieved facts. For this approach to work, we need a small set of answer candidates to start with. We observe that while state-of-the-art VQA models struggle with knowledge-based QA, these models are surprisingly effective at generating a small list of candidates that often contains the correct answer. Using these candidates to guide knowledge search makes retrieved facts less noisy and often more pertinent to the question, as shown in Figure 2.

MAVEx evaluates the validity of each answer candidate according to a diverse set of multi-modal knowledge facts that may be noisy or even conflicting. To address this, we propose a *consistency criterion* to assess whether each knowledge source used to retrieve facts for a specific answer candidate is actually reliable for supporting that answer. We evaluate our framework, MAVEx, on the OK-VQA dataset [25], the largest knowledge-based VQA dataset to date. Our approach achieves the state-of-the-art results on OK-VQA. This demonstrates that answer-specific knowledge retrieval results in more informative supporting evidence and a more solid knowledge-based VQA system.

In summary, our main contributions are: (a) We introduce a novel approach that uses answer candidates to guide knowledge retrieval for open-domain VQA; (b) We use multi-model knowledge retrieval by exploring visual knowledge along with textual knowledge; and (c) We propose a consistency criterion to decide when to trust knowledge retrieved from each source.

2. Related Work

Visual Question Answering. Visual Question Answering (VQA) has made significant progress over the past few years [2, 23, 1, 16, 3, 14, 4, 21, 20, 34]. More recent VQA systems [21, 34, 20, 19, 38, 17, 40, 6, 22] first extract visual features from a pre-trained object detector. Then they feed both visual and textual embeddings into a multi-modal transformer, which is pre-trained in a self-supervised way on an auxiliary task using a large-scale image captioning dataset such as [30]. Text-VQA [32] enables the VQA model to read by incorporating Optical Character Recognition (OCR) into the system. These models achieve remarkable performance on the VQA [2] dataset, however, they can only reason based on the image content and do not have a mechanism to explicitly incorporate knowledge from external sources.

Knowledge-Based VQA. Knowledge-based VQA requires acquiring commonsense or factual knowledge outside the image to answer the questions. We discuss the datasets and models developed for this task:

Datasets: KB-VQA [37] includes 2,402 questions generated by templates for 700 images. F-VQA [36] contains 5,826 questions, where each question-answer sample is annotated with a ground-truth fact triplet retrieved from the knowledge base. OK-VQA dataset [25] is a more recent dataset that covers a wide range of topics and includes 14,055 questions on 14,031 images. Our focus is on the OK-VQA dataset since it provides a larger scale dataset that requires open-domain knowledge. Knowledge-based VQA datasets to date are typically small compared to the traditional VQA datasets due to the difficulty of collecting such datasets. The small scale of the datasets adds to the challenges for learning robust models.

KB-VQA models: Recent methods for knowledge-based

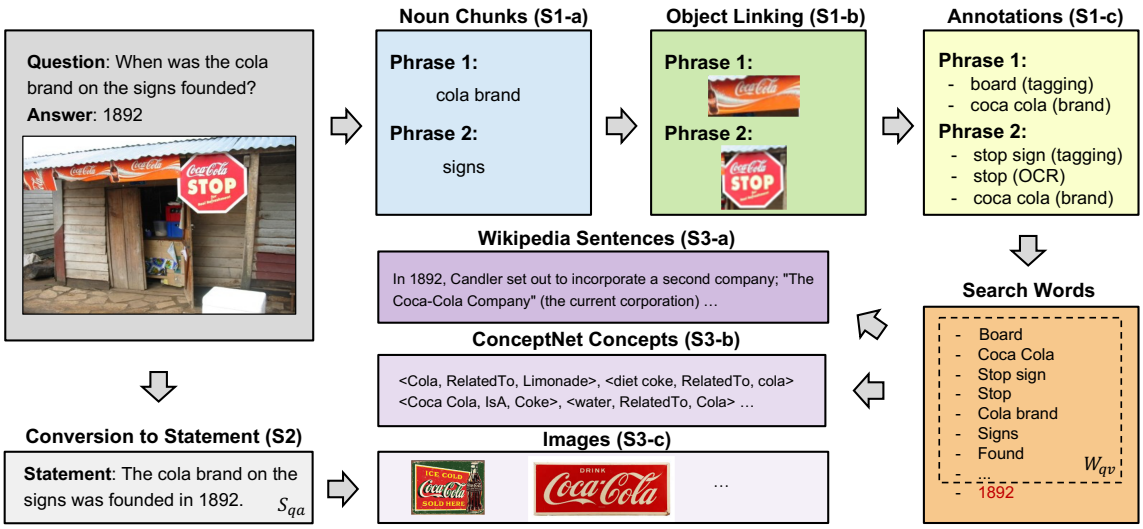

Figure 3: An example of the retrieval process for one question-answer pair.

VQA mainly follow two trends, template fitting and learning-based approaches. [37] fit the query to several predefined query templates and explicitly reason about the answer using the templates. The main limitation of the template fitting approaches is that the template is hand designed and it is hard to accommodate rich knowledge required to answer the questions using templates. Therefore, learning-based approaches are proposed to fetch helpful facts and commonsense knowledge for better performance. [27] learn to retrieve relevant facts from a knowledge base. [36] learn to find the mappings from the question to a query triplet. [26] propose to apply GCN [35] on the fact graph where each node is a representation of an image-question-entity triplet. [42] propose a modality-aware heterogeneous GCN capturing the most supporting evidence. [18] introduce a knowledge graph augmentation model to retrieve context-aware knowledge subgraphs, and then learn to aggregate the useful visual and question relevant knowledge. [24] use knowledge implicit in the embeddings and explicit symbolic knowledge. In contrast to these approaches, we formulate our problem as an *answer validation* problem, where the idea is to learn to validate a set of potential answers using multi-modal noisy knowledge sources.

3. The MAVEx Framework

We now present our MAVEx framework, a two-stage scheme that first retrieves knowledge and then predicts the answer. The scheme has been widely adopted in knowledge-based QA tasks in both NLP [5, 33] and computer vision communities [25, 18]. Different from previous works, beyond retrieving textual knowledge potentially relevant to the question, we propose to mine multi-modal *answer-specific knowledge* for each answer candidate. In particular, we consider three knowledge sources: Wikipedia and ConceptNet as textual knowledge resources, and Google images as the image knowledge resource, for providing factual, commonsense, and visual knowledge, respectively. Then, an answer validation module tests each answer candidate using the retrieved multi-modal knowledge.

3.1. Answer Guided Knowledge Retrieval

Given a question q about an image I and an answer candidate a from a set of possible answers (see Section 3.2 for details of answer candidate set generation), we retrieve external knowledge in support of a in three main steps. Figure 3 shows the entire process for an example question and a candidate answer.

S1: Answer-Agnostic Search Word Extraction. We first generate short phrases in q and concepts represented in I as a starting point for retrieving external information. This involves the following sub-steps:

Extract Noun Chunks from q: We parse the question using a constituency parser to compute the parse tree. Then, we extract all the nouns on the leaves of the parse tree together with the words that describe the nouns and belong to one of the types from 'ADJP', 'ADVP', 'PP', 'SBAR', 'DT' or 'JJ'. Those words help us to link the mentioned objects to the images. We use AllenNLP [10] constituency parser. See Figure 3 (S1-a).

Link Nouns to Objects: As images usually contain plenty of question-irrelevant contents, making the retrieval process hard to operate, we propose to narrow down the search field

to the objects referred to by the question. In particular, we use ViLBERT-multi-task [22] as the object linker, where it outputs scores given the noun phrases from the questions. We approve the linking when the linker's score is higher than 0.5 and extract the linked objects. See Figure 3 (S1-b).

Annotate Objects: We automatically provide the category labels, OCR readings and logo information for the linked objects using Google APIs to enrich the retrieved knowledge. See Figure 3 (S1-c).

The set of answer-agnostic search words, W_{qv}, consists of all of noun chunks and verbs in q, OCR, tagging (detection), and logo annotation of the referred objects, if any.

S2: Conversion to a Natural Language Statement. In order to use the answer candidate a to inform the retrieval step, we convert q and a into a natural language statement S_{qa} using a rule-based approach [7]. Such conversion has been found to be effective as statements occur much more frequently than questions in textual knowledge sources [15].

S3: Answer Candidate Guided Retrieval. We now use the search words W_{qv} from step S1, along with the answer candidate a and the statement S_{qa} from step S2, to retrieve relevant information as follows:

Retrieval of textual facts: We query each search word w $\in$ W_{qv} and collect all sentences from the retrieved Wikipedia articles.[1] For each answer candidate a, we first collect answer-specific sentences that contain a (ignoring stop words and yes/no). Then we rank those sentences based on the BERTScore [39] between the statement S_{qa} and the sentences. We then encode each of the top k^w_{sp} sentences using a pre-trained BERT [8] model and extract the final layer representation of the [CLS] token. This results in an answer-specific (denoted sp) feature matrix $\mathbf{K}^w_{sp}(a) \in \mathbb{R}^{k^w_{sp} \times 768}$ for each question-answer pair. We also store the retrieved sentences and their corresponding BERTScores for all answer candidates. We then choose the top k^w_{ag} non-repeated sentences according to the stored scores as the answer-agnostic knowledge. Those sentences are also encoded using pre-trained BERT, resulting in an answer-agnostic (denoted ag) feature matrix $\mathbf{K}^w_{ag} \in \mathbb{R}^{k^w_{ag} \times 768}$ for each question.

Retrieval of concepts: While Wikipedia articles provide factual knowledge that people need to look up when they answer a question, ConceptNet offers structured knowledge of concepts. Similar to Wikipedia article retrieval, we also query each search word in W_{qv} and collect all retrieved concepts. For each answer candidate a, we extract the concepts whose subject, relation, or object contains the candidate a, and push all retrieved concepts to the answer-agnostic concept pool. We rank those extracted concepts based on the maximum cosine similarity between the Glove embedding [28] of the words in W_{qv} and those in the concept, and select the top k^c_{sp} concepts as answer-specific knowledge. We also select the top k^c_{ag} concepts similarly from the answer-agnostic concept pool. The subjects, relations, and objects in the selected concepts are first converted into a sentence by handcrafted rules, and then encoded using pre-trained BERT model. Finally, the last layers' representation vectors are concatenated, resulting in a feature matrix $\mathbf{K}^c_{sp}(a) \in \mathbb{R}^{k^c_{sp} \times 768}$ for each question-answer pair, and a feature matrix $\mathbf{K}^c_{ag} \in \mathbb{R}^{k^c_{ag} \times 768}$ for each question.

Retrieval of visual knowledge: Pure textual knowledge is often insufficient due to two main reasons: (1) textual knowledge might be too general and not specific to the question image, (2) it might be hard to describe some concepts using text, and an image might be more informative (e.g., the 3rd question in Figure 1). Hence, visual knowledge can complement textual information, further enriching the outside knowledge feature space. We use Google image search to retrieve the top k_i images using the statement S_{qa} as the query. The images are then fed into a MaskRCNN [11] finetuned on the Visual Genome dataset [41] to extract at most 100 object features. We average the object features of visual detection results as the answer-specific visual knowledge representation, resulting in a feature matrix $\mathbf{K}^i_{sp}(a) \in \mathbb{R}^{k^i_{sp} \times 768}$ for each question-answer pair. For answer-agnostic knowledge, we simply use the zero vector.

3.2. Answer Candidate Validation

The answer validation module takes as input a question q, its visual features $\mathbf{v}$, an answer candidate a, and the supporting knowledge $\mathbf{K}^j_{ag}$ and $\mathbf{K}^j_{sp}(a)$ retrieved for a from each knowledge source j. It outputs a scalar score indicating how well the knowledge supports a.

Answer Candidate Generation. In order to use answer candidates to inform knowledge retrieval, we use ViLBERT [21], a state-of-the-art VQA model, to generate answer candidates. Note that any VQA model can be used for this purpose. As discussed in the experiments section, we found ViLBERT to be particularly effective at generating a small set of promising candidates.

3.2.1 Knowledge Embedding Module

We use cross-modal attention [38] in the knowledge embedding module, that treats the question-image embedding as a query to mine supportive knowledge from each source.

We first briefly introduce the Self-Attention (SA) and Guided-Attention (GA) units[2] as the building blocks. The SA unit takes as input a group of feature vectors $\mathbf{X} = [x_1; ...; x_m] \in \mathbb{R}^{m \times d}$ and learns the pairwise relationship between each sample pair within $\mathbf{X}$ using a multi-head attention layer by treating all possible combinations of x_i and x_j as queries and keys. Different from SA, the GA unit uses another group of features $\mathbf{Y} = [y_1; ...; y_n] \in \mathbb{R}^{n \times d}$ to guide the attention learning in $\mathbf{X}$. In particular, the GA unit learns

[1] We use the python API https://github.com/goldsmith/Wikipedia.

[2] Please refer to [38] for detailed model architectures.

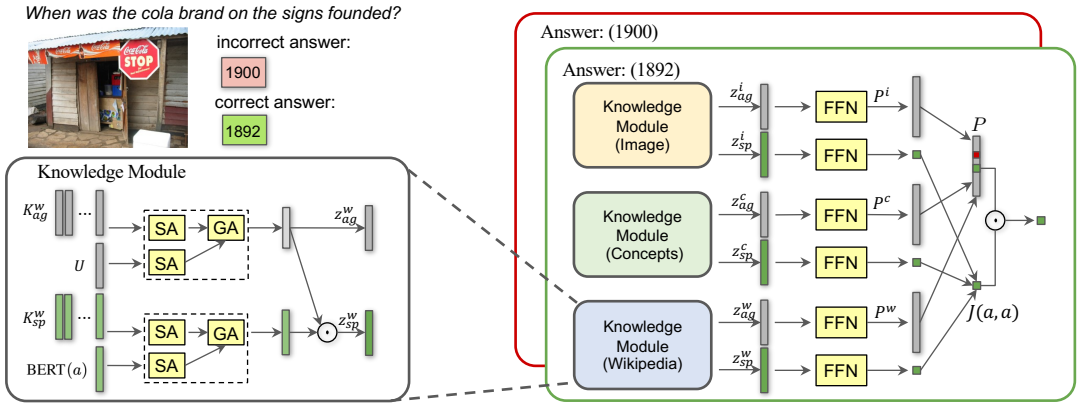

Figure 4: Model overview for validating two candidate answers. We explore three sources of external knowledge, *i.e.* Wikipedia, ConceptNet, and Google Images presented by the three parallel knowledge embedding modules. The grey blocks denote answer-agnostic features shared by all answer candidates and the green blocks denote answer-specific features.

the pairwise relationship between each pair across **X** and **Y** and treats each y_i as query and each x_i as keys. The values of the keys are weighted summed to produce an attended output features $\mathbf{T} \in \mathbb{R}^{m \times d}$ for both SA and GA. Finally, a feed-forward layer with residual links are built upon **T** to transform the output features to a new features space.

Given an image and the corresponding question, we first use ViLBERT to extract visual features $\mathbf{v} \in \mathbb{R}^{1024}$ and question features $\mathbf{q} \in \mathbb{R}^{1024}$ from the last layer of ViLBERT's [IMG] and [CLS] tokens, respectively. We then compute a joint feature **U** by element-wise multiplication of **q** and **v**. **U** is used as a query to mine answer-agnostic features $\mathbf{z}^j_{ag}$. **U** and the BERT embeddings of the answer candidates are used to mine answer-specific features $\mathbf{z}^j_{sp}(a, a')$ for the answer candidate a from each one of the three knowledge sources j as described in Eqs. (1) and (2):

$$\mathbf{z}^j_{ag} = \text{GA}(\text{SA}(\mathbf{U}), \text{SA}(\mathbf{K}^j_{ag})) \quad (1)$$

$$\mathbf{z}^j_{sp}(a, a') = \mathbf{z}^j_{ag} \odot \text{GA}(\text{SA}(\text{BERT}(a)), \text{SA}(\mathbf{K}^j_{sp}(a'))) \quad (2)$$

where a and a' are two answer candidates and the index j denotes one of the knowledge sources (Wikipedia w, ConceptNet c, or Google images i). Specifically, the answer-specific features $\mathbf{z}^j_{sp}(a, a')$ encode the joint features of a and the knowledge retrieved using a', and are further used to predict how well the knowledge retrieved by a' supports a.

3.2.2 Answer Validation Module

The validation module uses the attended knowledge features $\mathbf{z}^j_{sp}$ and $\mathbf{z}^j_{ag}$ from the three sources to validate the answer candidates. We introduce two approaches, early fusion and late fusion, to compute the validation score for each answer.

Early Fusion. This approach first merges the representations from the three knowledge sources, and then predicts the supportiveness score for each answer. Since not all knowledge sources are necessarily helpful, we encourage that at least one knowledge source provide helpful information to verify the answer by max pooling the answer-specific knowledge vectors ($\mathbf{z}^w_{sp}(a, a')$, $\mathbf{z}^c_{sp}(a, a')$ and $\mathbf{z}^i_{sp}(a, a')$) from the three sources retrieved by the answer a', producing a single vector $\mathbf{z}_{sp}(a, a')$ that contains the joint information.

Then, a feed-forward network, taking this joint representation as input, computes the validation score $J(a, a')$ that indicates how well the knowledge retrieved by a' supports a, as shown below:

$$J(a, a') = \text{FFN}(\max_{j \in \{w,c,i\}} \mathbf{z}^j_{sp}(a, a')), \quad (3)$$

where FFN denotes a feed-forward network that contains two FC layers (specifically, FC-GeLU-LayerNorm-FC). We also use the answer-agnostic features to predict a VQA score P for all answers in the set as $P = \text{FFN}(\max_j \{\mathbf{z}^j_{ag}\})$.

Late Fusion. Different from early fusion, where the decision is made according to the joint features from the three sources, the late fusion approach lets each knowledge source predict its own supportiveness score. The goal of this setting is to prevent misleading knowledge from contaminating valid knowledge from other sources. In particular, we compute the supportiveness score J^j for each source as $J^j(a, a') = \text{FFN}(\mathbf{z}^j_{sp}(a, a'))$, where FFN denotes a feed-forward layer. Then, the final score is computed by taking the maximum support score across the three sources as $J(a, a') = \max_j \{J^j(a, a')\}$, where $j \in \{w,\ c,\ i\}$ denotes the source index. We use the answer-agnostic features to predict single source VQA scores P^j for all answers in the set as $P^j = \text{FFN}(\mathbf{z}^j_{ag})$, and the final VQA score P is computed

as $P = \max_j\{P^j\}$. The overall architecture of the model is shown in Figure 4.

Consistency Criteria. The intuition behind our consistency criteria is that for the correct answer a, the knowledge retrieved for a from the most confident source (the one with the highest supportiveness score for a) should support a more than it supports other answer candidates, and it should also support a more than knowledge retrieved for other answer candidates. Specifically, we approve the answer validation score $J(a, a)$ only if it is higher than the scores computed using this knowledge for all other answers as well as the score for a when using knowledge retrieved for other answers. Mathematically, the consistency criteria checks that $J(a, a) > J(a', a)$ and $J(a, a) > J(a, a')$ for all $a' \neq a$. If the above condition is not met, we output the answer with the maximum VQA prediction score $P(a)$; otherwise we output the answer with the maximum VQA-weighted validation score $J(a, a)P(a)$.

3.3. Training and Implementation Details

Implementation. We implemented our approach on top of ViLBERT-multi-task [21], which utilizes a Mask-RCNN head [11] in conjunction with a ResNet-152 base network [12] as the object detection module. Convolutional features for at most 100 objects are then extracted for each image as the visual features, $i.e.$ a 2,048 dimensional vector for each object. For question embedding, following [8], our framework utilizes a BERT tokenizer to tokenize the question and use the first 23 tokens as the question tokens. We encode top 10 Wikipedia sentences, 20 concepts and 5 images as the answer-specific retrieved knowledge features, $i.e.$ k^w_{sp}=20, k^c_{sp}=20 and k^i_{sp}=5, and we use 20 sentences and 20 concepts as answer-agnostic knowledge features, $i.e.$ k^w_{ag}=20, k^c_{ag}=20. The number of hidden units in the SA and GA modules in the answer validation module is set to 1,024 to match the dimension of the ViLBERT features.

Training. The OK-VQA test images are a subset of COCO validation images which are used to pre-train most of transformer-based vision and language models [21, 34, 19]. Although the test questions never appear in the pre-training process, other questions on the test images may help the system understand the image better, leading to a higher performance. Besides, there is also data contamination from extra object annotations from Visual Genome (VG) dataset, which also contains some OK-VQA test images. As the VG dataset is used to pre-train the object detector, those test images can access the ground truth object annotations. We carefully remove all OK-VQA test images from the pre-training and re-train the ViLBERT-multi-task model and the object detector from scratch using the default configurations.

For answer candidate generation, we finetune the ViLBERT-multi-task model on OK-VQA using default configuration for 150 epochs. Binary cross-entropy loss and VQA soft score[3] are employed to optimize the system. We use the finetuned model to extract the top 5 answers for each question in the training and test set. We follow the default settings of ViLBERT. BertAdam optimizer [8] with a linear warmup learning rate is applied.

For the training of the answer validation module, we optimize the validation score $J(a, a')$ using the loss in Eq. 4 for the three knowledge sources, where $s(a)$ denotes the VQA soft scores for answer a. We also add the standard VQA loss on the VQA score P to train the answer-agnostic knowledge embedding modules. We train the system using a learning rate of 1e-5 for the ViLBERT parameters and 1e-4 for the parameters that are additionally introduced in the validation module. We freeze the first 6 layers of the ViLBERT base network. We use $\mathcal{L}_{bce}$ to denote binary cross-entropy loss.

$$\begin{aligned}\mathcal{L}_{\text{MAVEx}} = \mathcal{L}_{bce}\Big(\max_{\substack{a \\ s.t.\ a \neq a'}} J(a, a'),\ \mathbf{0}\Big) \\ + \mathcal{L}_{bce}\Big(\max_{\substack{a' \\ s.t.\ a \neq a'}} J(a, a'),\ \mathbf{0}\Big) \\ + \mathcal{L}_{bce}\Big(J(a, a), s(a)\Big) \qquad (4)\end{aligned}$$

4. Experiments

We evaluate our answer validation framework on the OK-VQA dataset [25]. We first briefly describe the dataset, and then present our result and provide comparisons to the current state-of-the-art systems.

OK-VQA dataset. It is the largest knowledge-based VQA dataset at present. The questions are crowdsourced from Amazon Mechanical Turkers, leading to two main advantages: (1) the questions indeed require outside knowledge beyond images; (2) there are no existing knowledge bases that cover all the questions, thus requiring systems to explore open-domain resources. The dataset contains 14,031 images and 14,055 questions covering a variety of knowledge categories. The metric is the VQA soft score (see footnote 3).

4.1. Intrinsic Evaluation

We begin with an intrinsic evaluation of MAVEx, assessing the quality of the answer candidate generation and knowledge retrieval modules.

Answer Candidate Accuracy. Our answer candidate generation module, which is based on the finetuned ViLBERT-multi-task model, outputs its top-5 answers as the candidates. We found that the best answer in this small candidate set achieves a VQA soft score of 59.7 on the test set, substantially higher than the top-1 answer score of this system (35.2) as well as other state-of-the-art systems without data contamination (33.7 or below).

[3] OK-VQA provides 5 annotations for each question. Soft scores are 0, 0.6, and 1 corresponding to 0, 1, more than 1 matching answer annotations.

Method	Knowledge Resources	Performance
ArticleNet (AN) [25]	Wikipedia	5.3
Q-only [25]	—	14.9
MLP [25]	—	20.7
BAN [16]	—	25.2
+ AN [25]	Wikipedia	25.6
+ KG-AUG [18]	Wikipedia + ConceptNet	26.7
MUTAN [3]	—	26.4
+ AN [25]	Wikipedia	27.8
Mucko [42]	Dense Caption	29.2
KRISP [24]	Wikipedia + ConceptNet	32.3*
+ VQAv2 Pre-training	Wikipedia + ConceptNet	37.8*
+ VQAv2 (incl. graph) Pre-training	Wikipedia + ConceptNet	38.9*
ConceptBert [9]	ConceptNet	33.7
ViLBERT [21]	—	35.2*
MAVEx (ours) – w/o answer validation	Wikipedia + ConceptNet + Google Images	37.6*
MAVEx (ours) – Early Fusion	Wikipedia + ConceptNet + Google Images	37.8*
MAVEx (ours) – Late Fusion	Wikipedia + ConceptNet + Google Images	38.7*
MAVEx (ours) – Late Fusion (Ensemble 5)	Wikipedia + ConceptNet + Google Images	**39.4***
RVL† [31]	Wikipedia + ConceptNet	39.0†
MAVEx † (ours) – Late Fusion	Wikipedia + ConceptNet + Google Images	40.5†*

Table 1: MAVEx outperforms current state-of-the-art approaches on the OK-VQA dataset. The middle column lists the external knowledge sources, if any, used in each VQA system. † indicates that the system uses a pretrained model that is contaminated by OK-VQA test images. * indicates that the results have been reported on version 1.1 of the dataset. The difference between version 1.0 and 1.1 is different ways of answer stemming. As reported in [24] there is not much difference in the results obtained on these two versions.

We also evaluate the score achieved by slightly larger candidate sets, consisting of the top 6, 8 and 10 candidates. These achieve VQA soft scores of 62.1, 65.1, and 67.1, respectively. Since our answer validation framework needs to retrieve and encode answer-specific knowledge, we use only top-5 answer candidates as a reasonable trade-off between efficiency, answer coverage, and overall accuracy.

Knowledge Retrieval Accuracy. We assess the accuracy of our knowledge retrieval modules for Wikipedia and ConceptNet using the OK-VQA test set.

For *Wikipedia sentences*, we observe that 71.8% of the top-10 Wikipedia sentences retrieved for question-answer pairs contain the answer candidate used for retrieval, suggesting strong relevance of the answer-specific knowledge.

For *ConceptNet concepts*, we first define a strong relation set where both the answer candidate and at least one other search word generated from **S1** exist in the concept triplets. 29% of question-answer pairs[4] have concept triplet(s) inside the strong relation set, indicating answer relevance.

4.2. Main Results

Table 1 shows that MAVEx consistently outperforms prior approaches by a clear margin. For example, MAVEx outperforms recent state-of-the-art models Mucko [42], KRISP [24], and ConceptBert [9] by 9.5, 6.4, 5.0 points, respectively. Our approach also outperforms ViLBERT [21] base system by 3.5 points. We consider a MAVEx baseline model that uses the retrieved knowledge ($\mathbf{K}_{ag}^{j}$) as additional inputs without answer validation. This model achieves 37.6 overall score, 2.4% higher than the ViLBERT model and 1.1% lower than the late fusion model, indicating that using answer-guided retrieved knowledge is helpful and answer validation further improves the performance. An ensemble of 5 MAVEx late fusion models with different initializations improves the results to 39.4. The standard deviation of the 5 runs is 0.2. We also observe that the late fusion setting outperforms early fusion by 0.9, indicating that it is important to allow each knowledge source to make its own decision first, and then combine the information across sources.

4.3. Ablation Study of Knowledge Sources

We use the late fusion model and report, in the 2nd column of Table 2, the system's performance when only one knowledge source is used. We see that the three sources provide an improvement of 2.6, 2.2, and 2.0, respectively, compared to not using any external knowledge source. This indicates the effectiveness and value of all three sources.

The combination of the three sources achieves a net performance gain of 3.5 over the ViLBERT baseline, supporting the intuition that the three sources together provide comple-

[4]The correct answer included if not among the answer candidates.

Figure 5: Examples that the VQA model is wrong but MAVEx with the three external knowledge sources answers correctly. The correct answer is in the green box and the incorrect answer is shown in the red box. The grey box shows the question. The most influential knowledge content (judged by GradCAM [29]) is shown in the boxes under the predicted answers.

mentary pieces of knowledge.

We show some qualitative examples in Figure 5, where the VQA model is wrong but provides good answer candidates. Our MAVEx gathers the external knowledge from the three sources and predicts the correct answers.

4.4. Oracle Performance as Upper Bounds

We present two oracle settings to show the potential of our framework. The first oracle selects the best knowledge source at test time in the late fusion setting. The second oracle adds one correct answer[5] to the answer candidate set.

Oracle Source Selector. Our answer validation framework achieves an oracle score of 43.5 if we choose the best source to trust for each question. This indicates that the three knowledge sources provide complementary features, leaving further potential to improve the system.

Oracle Answer Candidates. The top-5 answer candidate list we use in MAVEx does not always contain the correct answer. To assess the potential of a more powerful answer candidate generator, we consider the performance of MAVEx when the ground-truth answer is guaranteed to be in the candidate set. Specifically, for the questions whose extracted answer candidate set did not contain the correct answer, we use *one* correct answer with the maximum soft score to replace the least scoring answer in the list. The results are shown in the last column of Table 2. The 4.3-4.7 gain over using original extracted answers suggests that extracting a better answer candidate set can make MAVEx more effective. Figure 6 presents some examples where the VQA answer candidate set does not contain the right answer. By manually adding the right answer to the candidate set, the validation module is able to find the supportive evidence and predict correctly. The most influential evidence (as judged by GradCAM [29]) is shown under the prediction boxes.

[5]If there are more than one correct answer with soft scores larger than zero, we choose the one with largest soft score.

Figure 6: Examples where the right answer is not in the extracted answer candidate set. By manually adding the right answer to the answer candidate set, MAVEx learns to figure out the most supportive evidence and predicts correctly.

Knowledge Source	System Score	Oracle Score
—	35.2	—
Wikipedia	37.8	42.1
ConceptNet	37.4	42.0
Google Images	37.2	41.9
Wikipedia + ConceptNet + Images	38.7	43.2

Table 2: Ablation study (2nd col.) using one knowledge source at a time. Oracle (3rd col.) when the answer list is altered, if necessary, to contain the correct answer.

5. Conclusion

We presented MAVEx, a novel approach for knowledge based visual question answering. The goal is to retrieve answer-specific textual and visual knowledge from different knowledge sources and learn what sources contain the most relevant information. Searching through the vast amount of retrieved knowledge, which is often quite noisy, is challenging. Hence, we formulate the problem as answer validation, where the goal is to learn to verify the validity of a set of candidate answers according to the retrieved knowledge. More specifically, an answer candidate validation module predicts the degree of support provided by the knowledge retrieved for each candidate, and decides which sources to trust for each candidate answer. MAVEx demonstrates the clear advantages of answer-guided knowledge retrieval, achieving new state-of-the-art performance on the OK-VQA dataset.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and VQA. In *CVPR*, 2018. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1, 2

[3] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *ICCV*, 2017. 2, 7

[4] Remi Cadene, Hedi Ben-Younes, Matthieu Cord, and Nicolas Thome. Murel: Multimodal relational reasoning for visual question answering. In *CVPR*, 2019. 2

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

[6] 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*, 2020. 2

[7] Dorottya Demszky, Kelvin Guu, and Percy Liang. Transforming question answering datasets into natural language inference datasets. *arXiv*, 2018. 4

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019. 4, 6

[9] François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. Conceptbert: Concept-aware representation for visual question answering. In *EMNLP*, 2020. 7

[10] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, 2018. 3

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 4, 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016. 6

[13] Drew A Hudson and Christopher D Manning. Gqa: a new dataset for compositional question answering over real-world images. In *CVPR*, 2019. 1

[14] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the Winning Entry to the VQA Challenge 2018. *arXiv*, 2018. 2

[15] Tushar Khot, Ashish Sabharwal, and Peter Clark. Answering complex questions using open information extraction. In *ACL*, 2017. 4

[16] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear Attention Networks. In *NeurIPS*, 2018. 2, 7

[17] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 2

[18] Guohao Li, Xin Wang, and Wenwu Zhu. Boosting visual question answering with context-aware knowledge aggregation. In *ACM Conference on Multimedia*, 2020. 3, 7

[19] 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*, 2019. 2, 6

[20] Bei Liu, Zhicheng Huang, Zhaoyang Zeng, Zheyu Chen, and Jianlong Fu. Learning rich image region representation for visual question answering. *arXiv*, 2019. 2

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2, 4, 6, 7

[22] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, 2020. 2, 4

[23] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical Question-Image Co-attention for Visual Question Answering. In *NeurIPS*, 2016. 2

[24] 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. 1, 3, 7

[25] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 1, 2, 3, 6, 7

[26] Medhini Narasimhan, Svetlana Lazebnik, and Alexander Schwing. Out-of-The-Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering. In *NeurIPS*, 2018. 3

[27] Medhini Narasimhan and Alexander G Schwing. Straight to the facts: Learning knowledge base retrieval for factual visual question answering. In *ECCV*, 2018. 3

[28] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *EMNLP*, 2014. 4

[29] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al.

Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *ICCV*, 2017. 8

[30] 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. 2

[31] Violetta Shevchenko, Damien Teney, Anthony Dick, and Anton van den Hengel. Reasoning over vision and language: Exploring the benefits of supplemental knowledge. *arXiv*, 2021. 7

[32] 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. 1, 2

[33] Haitian Sun, Tania Bedrax-Weiss, and William Cohen. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text. In *EMNLP*, 2019. 3

[34] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 2, 6

[35] 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. 3

[36] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *TPAMI*, 2018. 1, 2, 3

[37] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. In *IJCAI*, 2017. 1, 2, 3

[38] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019. 2, 4

[39] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020. 4

[40] 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*, 2020. 2

[41] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded Question Answering in Images. In *CVPR*, 2016. 1, 4

[42] 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. 1, 3, 7