Noname manuscript No.
(will be inserted by the editor)

FVQA: Fact-based Visual Question Answering

Peng Wang* · **Qi Wu*** · **Chunhua Shen** · **Anton van den Hengel** · **Anthony Dick**

Received: date / Accepted: date

Abstract Visual Question Answering (VQA) has attracted a lot of attention in both Computer Vision and Natural Language Processing communities, not least because it offers insight into the relationships between two important sources of information. Current datasets, and the models built upon them, have focused on questions which are answerable by direct analysis of the question and image alone. The set of such questions that require no external information to answer is interesting, but very limited. It excludes questions which require common sense, or basic factual knowledge to answer, for example. Here we introduce FVQA, a VQA dataset which requires, and supports, much deeper reasoning. FVQA is mainly made up with questions that require external information to answer. We thus extend a conventional visual question answering dataset, which contains {*image-question-answer*} triplets, through additional {*image-question-answer-supporting fact*} tuples. The supporting fact is represented as a structural triplet, such as `<Cat,CapableOf,ClimbingTrees>`. We evaluate several baseline models on the FVQA dataset, and describe a novel model which is capable of reasoning about an image on the basis of supporting facts.

Keywords Visual Question Answering · Knowledge Base · LSTM

P. Wang
University of Adelaide, SA 5005, Australia
E-mail: p.wang@adelaide.edu.au

Q. Wu
University of Adelaide, SA 5005, Australia
E-mail: qi.wu01@adelaide.edu.au

C. Shen (✉)
University of Adelaide, SA 5005, Australia
E-mail: chunhua.shen@adelaide.edu.au

A. van den Hengel
University of Adelaide, SA 5005, Australia
E-mail: anton.vandenhengel@adelaide.edu.au

A. Dick
University of Adelaide, SA 5005, Australia
E-mail: anthony.dick@adelaide.edu.au
*The first two authors contributed to this work equally.

1 Introduction

Visual Question Answering (VQA) can be seen as a proxy task for evaluating a vision system's capacity for deeper image understanding. It requires elements of image analysis, natural language processing, and a means by which to relate images and text. Distinct from many perceptual visual tasks such as image classification, object detection and recognition [1–4], however, VQA requires that a method be prepared to answer a question it has never seen before. In object detection the set of objects of interest is specified at training time, for example, whereas in VQA the set of questions which may be asked inevitably extends beyond those in the training set.

The set of questions that a VQA method is able to answer is one of its key features, and limitations. Asking a method a question that is outside its scope will lead to a failure to answer, or worse, to a random answer. Much of the existing VQA effort has been focused on questions which can be answered by the direct analysis of the question and image, on the basis of a large training set [5–10]. This is quite a restricted set of questions, which require only relatively shallow image understanding to answer. It is quite possible, for example, to answer 'How many giraffes are in the image?' without understanding anything about giraffes.

The number of VQA datasets available has grown as the field progresses [5–10]. They have contributed

Fig. 1: An example visual-based question from our FVQA dataset that requires both visual and common-sense knowledge to answer. The answer and mined knowledge are generated by our proposed method.

valuable large-scale data for training neural-network based VQA models and introduced various question types, and tasks, from global association between QA pairs and images [5, 6, 9] to grounded QA in image regions [10]; from free-from answer generation [5, 7, 9, 10] to multiple-choice picking [5, 6] and blank filling [8]. For example, The questions defined in DAQUAR [6] are almost exclusively "Visual" questions, referring to "color", "number" and "physical location of the object". In the Toronto-QA dataset [9], questions are generated automatically from image captions which describe the major visible content of the image. For the VQA dataset [5], only 5.5% of questions require adult-level (18+) background knowledge (28.4% and 11.2% questions require older child (9-12) and teenager (13-17) knowledge). We cannot claim that this is a truly "AI-complete" problem, because this is not a real scenario for human beings. Humans inevitably use their background knowledge to answer questions, even visual ones, in many cases. For example, to answer the question given in Fig. 1, one not only needs to visually recognize the 'red object' is a 'fire hydrant', but also to know that '*a fire hydrant can be used for fighting fires*'.

Developing methods that are capable of deeper image understanding demands a more challenging set of questions. We consider here the set of questions which may be answered on the basis of an external source of information, such as Wikipedia. This reflects our belief that reference to an external source of knowledge is essential to general VQA. This belief is based on the observation that the number of {*image-question-answer*} training examples that would be required to provide the background information necessary to answer general questions about images would be completely prohibitive. The number of concepts that would need to be illustrated is too high, and scales combinatorially.

In contrast to previous VQA datasets which only contain *question-answer* pairs for an image, we additionally provide a *supporting-fact* for each *question-answer* pair. The *supporting-fact* is a structural representation of information that is necessary for answering the given question. For example, given an image with a cat and a dog and the question '`Which animal in the image is able to climb trees?`', then the answer is '`cat`'. The required supporting fact for answering this question is `<Cat,CapableOf,ClimbingTrees>`, which is extracted from an existing knowledge base. By providing supporting facts, the dataset supports answering complex questions, even if all of the information required to answer the question is not depicted in the image. Moreover, it supports explicit reasoning in visual question answering, *i.e.* it gives an indication as to how a method might derive an answer. This information can be used in answer inference, to search for other appropriate facts, or to evaluate answers which include an inference chain.

In demonstrating the value of the dataset in driving deeper levels of image understanding in VQA we examine the performance of the state-of-the-art LSTM (Long-Short Term Memory) models [5, 6, 9] on our FVQA dataset. We find that there are a number of limitations with this approach. The first is that there is no explicit reasoning process in these methods. This means that it is impossible to tell whether it is answering the question based on image information or just the prevalence of a particular answer in the training set. The second problem is that, because the model is trained on individual question-answer pairs, the range of questions that can be accurately answered is limited. It only can answer questions about concepts that have been observed in the training set, and there are millions of possible concepts and hundreds of millions relationships between them. Capturing this amount of information would require an implausibly large LSTM, and a completely impractical amount of training data.

Our main contributions are as follows. A new VQA dataset (FVQA) with additional supporting facts is introduced in Sec. 3, which requires and supports deeper reasoning. In response to this observed limitation of the current LSTM-based approach, we propose a method which is based on explicit reasoning about the visual concepts detected from images in Sec. 4. The proposed method first detects relevant content in the image, and relates it to information available in a pre-constructed knowledge base (we combine several publicly available

large-scale knowledge base). A natural language question is then automatically classified and mapped to a query which runs over the combined image and knowledge base information. The response of the query leads to the supporting fact and which is then processed so as to form the final answer to the question. We achieve the state-of-the-art performance with 56.91% on the Top-1 accuracy (see Sec. 5).

2 Related Work

2.1 Visual Question Answering Datasets

Several datasets designed for Visual Question Answering have been proposed. The DAQUAR [6] dataset is the first small benchmark dataset built upon indoor scene RGB-D images, which is mostly composed of questions requiring only visual knowledge. Most of the other datasets [5, 7–10] collected question-answer pairs on Microsoft COCO images [2], either generated automatically by NLP tools [9] or written by human workers [5, 7]. The Visual Genome [11] contains 1.7 million questions, which are asked by human workers based on region descriptions. The MadLibs dataset [8] provides a large number of template based text descriptions of images, which are used to answer multiple choice questions about the images. Visual 7W [10] established a semantic link-between textual descriptions and image regions by object-level grounding and the questions are asked based on groundings.

2.2 Visual Question Answering Methods

Malinowski *et al.* [12] may be the first to study the VQA problem. They proposed a method that combines image segmentation and semantic parsing with a Bayesian approach to sampling from nearest neighbors in the training set. This approach requires human defined predicates, which are inevitably dataset-specific. Tu *et al.* [13] built a query answering system based on a joint parse graph from text and videos. Geman *et al.* [14] proposed an automatic 'query generator' that is trained on annotated images and produces a sequence of binary questions from any given test image.

The current dominant trend of VQA is to combine Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) to learn the mapping from input images and questions to answers. Both Gao *et al.* [7] and Malinowski *et al.* [15] used RNNs to encode the question and output the answer. Whereas Gao *et al.* [7] used two networks, a separate encoder and decoder, Malinowski *et al.* [15] used a single network for both encoding and decoding. Ren *et al.* [9] focused on questions with a single-word answer and formulated the task as a classification problem using an LSTM. Inspired by Xu *et al.* [16] who encoded visual attention in the Image Captioning, [10, 17–20] proposed to use the spatial attention to help answering visual questions. [20, 21] formulated the VQA as a classification problem and restrict the answer only can be drawn from a fixed answer space. In other words, they can not generate open-ended answers. Zhu *et al.* [22] investigated the video question answering problem using the question form of 'fill-in-the-blank'. However, either LSTM or GRU (Gated Recurrent Unit, similar to LSTM) is still applied in these methods to model the questions. Irrespective of the finer details, we call this the LSTM approach.

2.3 Knowledge-base in the VQA

Answering general questions posed by humans about images inevitably requires reference to information not contained in the image itself. To an extent this information may be provided by an existing training set such as ImageNet [3], or MS COCO [2] as class labels or image captions. More challenging still, the description could even refer to objects that are not depicted (*e.g.*, it can talk about people waiting for a train, even when the train is not visible because it has not arrived yet) and provide external knowledge that cannot be derived directly from the image (*e.g.*, the person depicted is Mona Lisa).

Large-scale *structured* Knowledge Bases (KBs) [23–29] are one means of capturing such external information. In structured KBs, knowledge is typically represented by a large number of triples of the form `(arg1,rel,arg2)`, where `arg1` and `arg2` denote two concepts in the KB and `rel` is a predicate representing the relationship between them. A collection of such triples forms a large interlinked graph. Such triples are often described according to a Resource Description Framework [30] (RDF) specification, and housed in a relational database management system (RDBMS), or triple-store, which allows queries over the data. The information in KBs can be accessed efficiently using a query language. In this work we use SPARQL Protocol [31] to query the OpenLink Virtuoso [32] RDBMS.

Large-scale structured KBs are constructed either by manual annotation (*e.g.*, DBpedia [23], Freebase [25] and Wikidata [29]), or by automatic extraction from unstructured/semi-structured data (*e.g.*, YAGO [28, 33], OpenIE [24, 34, 35], NELL [26], NEIL [27, 36], WebChild [37, 38], ConceptNet [39]). The KB we use here is

the combination of DBpedia, WebChild and ConceptNet, which contains structured information extracted from Wikipedia and unstructured online articles.

In the NLP and AI communities, there is an increasing interest in the problem of natural language question answering using structured KBs (referred to as KB-QA) [40–48]. However, VQA systems exploiting KBs are still relatively rare. Zhu *et al.* [49] used a KB and RDBMS to answer image-based queries. However, in contrast to our approach, they build a KB for the purpose, using an MRF model, with image features and scene/attribute/affordance labels as nodes. The links between nodes represent mutual compatibility relationships. The KB thus relates specific images to specified image-based quantities, which are all that exists in the database schema. This prohibits question answering that relies on general knowledge about the world. Most recently, Wu *et al.* [50] encoded the mined knowledge text from the DBpedia to a vector with the word2vector model and combined with visual features together to generate answers using an LSTM model. However, their proposed method only extract discrete text pieces from the knowledge base but ignores the power of its structural representation. Both of [49] and [50] are not capable of explicit reasoning as we are.

The approach which is closet to ours is that of Wang *et al.* [51] as they described a method for visual question answering which is capable of reasoning about an image base on the information extracted from a knowledge base. However, their method largely relies on the pre-defined template, which only accepts questions in a pre-defined format. Our method does not suffer this constraint. Moreover, their proposed model only used a single manual annotated knowledge source while we use additional two self-learned knowledge bases, which enable us to answer more general questions.

3 Creating the FVQA Dataset

Different from previous VQA datasets [5, 7–10] that only ask annotators to provide *question-answer* pairs without any restrictions, we want the questions in our dataset only can be asked and answered after the annotator knowing some commonsense knowledge. This means that we can not simply distribute only images to questioners like others [5, 10], we need to provide a large number of supporting facts (commonsense knowledge) which are related to the visual concepts in the image. We build our own on-line question collection system and allow users to choose images, visual concepts and candidate supporting facts freely. Then the user can ask questions based on his/her previous choices (all choices will be recorded). We give each annotator a tutorial and restrict them to ask questions that only can be answered with both visual concept in the image and the provided external commonsense knowledge. Following sections provide more details about images, visual concepts, knowledge bases and our question collection system and procedures. We also compare with other VQA datasets with some data statistics.

3.1 Images and Visual Concepts

We sample 2190 images from the MS COCO [2] validation set and ImageNet [3] test set for collecting questions. Images from MS COCO can provide more context because they have more complicated scenes. Scenes of ImageNet images are much simpler but there are more object categories (200 in ImageNet vs. 80 in MS COCO).

Three types of visual concept extractors are applied to each image:

Object Detector: Two Fast-RCNN [52] models are trained by the authors on MS COCO 80-object (train split) and ImageNet 200-object datasets (train+val split) respectively. After combination, there are in total 234 classes of objects which can be detected in each image (see Appendix).

Scene Classifier: The scene classifier trained on MIT Places205 [53] dataset is adopted, which assigns each image with scene labels from 205 classes.

Attribute Classifier: The image attributes for training are obtained from the ground truth captions of MS COCO images, which are made up of 24 actions, 92 objects (without bounding box information) and 25 scenes (see Appendix). A deep model is trained by Wu *et al.* [54] on these training data and incorporated in this work. These 92 object and 25 scene classes are different from the concepts extracted using the above object detectors and scene classifiers, and they are combined together.

In summary, there are in total 326 object, 221 scene and 24 action classes to be extracted. These visual concepts are further linked to a variety of external knowledge, as shown in the next section.

3.2 Knowledge Bases

The knowledge about each visual concept is extracted from a range of existing structured knowledge bases, including DBpedia [23], ConceptNet [39] and WebChild [37, 38].

DBpedia: The structured information stored in DBpedia is extracted from Wikipedia by crowd-sourcing. In this KB, concepts are linked to their categories and

KB	Predicate	#Facts	Examples
DBpedia	Category	35152	(Wii,Category,VideoGameConsole)
ConceptNet	RelatedTo	79789	(Horse,RelatedTo,Zebra),(Wine,RelatedTo,Goblet)
	AtLocation	13683	(Bikini,AtLocation,Beach),(Tap,AtLocation,Bathroom)
	IsA	6011	(Broccoli,IsA,GreenVegetable)
	CapableOf	5837	(Monitor,CapableOf,DisplayImages)
	UsedFor	5363	(Lighthouse,UsedFor,SignalingDanger)
	Desires	3358	(Dog,Desires,PlayFrisbee),(Bee,Desires,Flower)
	HasProperty	2813	(Wedding,HasProperty,Romantic)
	HasA	1665	(Giraffe,HasA,LongTongue),(Cat,HasA,Claw)
	PartOf	762	(RAM,PartOf,Computer),(Tail,PartOf,Zebra)
	ReceivesAction	344	(Books,ReceivesAction,bought at a bookshop)
	CreatedBy	96	(Bread,CreatedBy,Flour),(Cheese,CreatedBy,Milk)
WebChild	Smaller, Better, Slower, Bigger, Taller, ...	38576	(Motorcycle,Smaller,Car),(Apple,Better,VitaminPill), (Train,Slower,Plane),(Watermelon,Bigger,Orange), (Giraffe,Taller,Rhino)

Table 1: The predicates in different knowledge bases used for generating questions. The '#Facts' column shows the number of facts which are related to the visual concepts described in Section 3.1. The 'Examples' column gives some examples of extracted facts, in which the visual concept is underlined.

super-categories based on the SKOS Vocabulary[1]. In this work, the categories and super-categories of all aforementioned visual concepts are extracted transitively.

ConceptNet: This KB is made up of several commonsense relations, such as `UsedFor`, `CreatedBy` and `IsA`. Much of the knowledge is automatically generated from the sentences of the Open Mind Common Sense (OMCS) project[2]. We adopt 11 common relations (predicates) in ConceptNet to generate questions and answers.

WebChild: The work in [37] considered a form of commonsense knowledge being overlooked by most of existing KBs, which involves comparative relations such as `Faster`, `Bigger` and `Heavier`. In [37], this form of information is extracted automatically from the Web.

The predicates (relations) which we extract from each KB and the corresponding number of facts can be found in Table 1. All the aforementioned structured information are stored in the form of RDF triples and can be accessed using Sparql queries.

3.3 Question Collection

In this work, we focus on collecting visual questions which need to be answered with the help of supporting-facts. To this end, we designed a specialized system, in which the procedure of asking questions is conducted in the following steps:

1) *Selecting Concept:* Annotators are given an image and a number of visual concepts (object, scene and action). They need to choose one of the visual concepts which is related to this image.
2) *Selecting Fact:* Once a visual concept is selected, the associated facts are demonstrated in the form of sentences with the two entities underlined. For example, the fact `(Train,Slower,Plane)` is expressed as 'Train is slower than plane'. Annotators should select a correct and relevant fact by themselves.
3) *Asking Question and Giving Answer:* The Annotators are required to ask a question, answering which needs the information from both of the image and the selected fact. The answer is limited to the two concepts in the supporting-fact. In other words, the source of the answer can be either the visual concept in the image (underlined in Table 1) or the concept in the KB.

3.4 Data Statistics

Dataset size and other statistics Totally, 5826 questions (corresponding to 4216 unique facts) are collected collaboratively by 38 PhD students. In order to report the significant statistics, we create 5 random splits of the dataset. In each split, we have 1100 training images and 1090 test images. Each split provides roughly 2927 and 2899 questions[3] for training and test respectively. These questions can be categorized according to the type of visual concept being asked (Object, Scene or Action), the source of the answer (Image or KB) and the knowledge base of the supporting-fact (DBpedia, ConceptNet or Webchild).

[1] http://www.w3.org/2004/02/skos/
[2] http://web.media.mit.edu/~push/Kurzweil.html

[3] Due to each image contains different number of questions, each split may contain different number of questions for training and test. Here we only report the average numbers. The error bars in the Table 2 shows the differences.

Criterion	Categories	Train	Test	Total
Visual Concept	Object	2661.2 ± 66.0	2621.8 ± 66.0	5283
	Scene	251.2 ± 26.2	260.8 ± 26.2	512
	Action	14.8 ± 2.7	16.2 ± 2.7	31
Answer-Source	Image	2437.4 ± 63.4	2393.6 ± 63.4	4831
	KB	489.8 ± 27.9	505.2 ± 27.9	995
KB-Source	DBpedia	403.2 ± 12.7	413.8 ± 12.7	817
	ConceptNet	2348.8 ± 71.6	2303.2 ± 71.6	4652
	Webchild	175.2 ± 9.5	181.8 ± 9.5	357
Total		2927.2 ± 69.5	2898.8 ± 69.5	5826

Table 2: The classification of questions according to 'the questioned visual concept', 'where the answer is from' and 'the KB-source of the supporting-fact'. The number of training/test questions in each category is also demonstrated. The error bars are produced by 5 different splits.

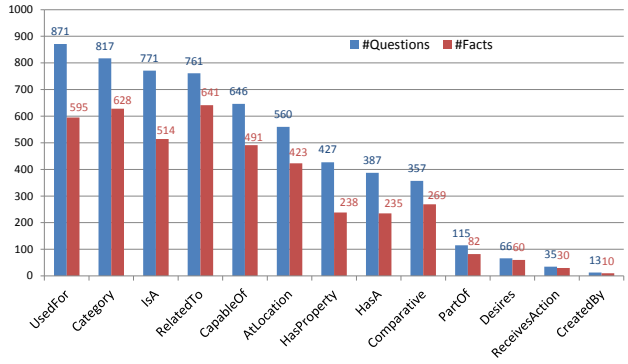

Fig. 2: The distributions of the collected 5826 questions and the corresponding 4216 facts over different predicates. The top five predicates are `UsedFor`, `Category`, `IsA`, `RelatedTo` and `CapableOf`. There are fewer supporting facts than questions because one 'fact' can correspond to multiple 'questions'.

Table 2 shows the number of training/test questions falling into each of the above categories. We can see that most of the questions are related to the objects in the images and most of the answers are visual concepts ('Answer-source' is 'Image'). As for knowledge bases, 80% of the collected questions rely on the supporting-facts from ConceptNet. Answering 14% and 6% questions depends on the knowledge from DBpedia and Webchild respectively.

Table 3 shows some other statistics of the dataset, such as number of question categories, average question/answer length *etc*. We have totally 32 question types (see Section 4.1 for more details). Compared to VQA-real [5] and Visual Genome [11], our FVQA dataset provides longer questions, which is 9.5 in average length.

Predicates distribution The distributions of collected questions and facts over different types of predicates are shown in Figure 2. The comparative predicates in WebChild are considered as one type and there are in total

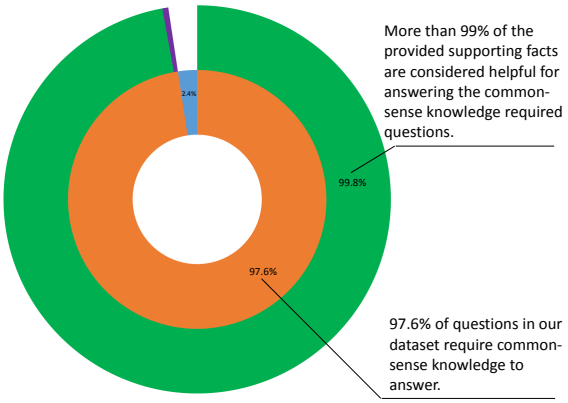

Fig. 3: Human study of how many percentage of the collected questions require common-sense knowledge and whether the collected supporting facts are critical for the reasoning.

13 types of predicates. We can see that the questions and facts are evenly distributed over the predicates of `Category`, `UsedFor`, `IsA`, `RelatedTo`, `CapableOf`, `AtLocation`, `HasProperty` and `HasA`, although these predicates differ significantly in the total numbers of extracted facts (see Table 1).

Human study of common-sense knowledge In order to verify whether our collected questions require common-sense knowledge and whether the supporting facts are helpful for answering the knowledge required questions, we conducted two human studies by asking subjects -

1. Whether or not the given question requires external common-sense knowledge to answer, and If 'yes'
2. Whether or not the given supporting fact provides the common-sense knowledge to answer the question.

The above study is repeated by 3 human subjects independently. We found that 2 or more in 3 subjects voted 'yes' to 'require common-sense' for 97.6% of questions. In the 'supporting facts' study, 99.8% of the supporting facts are considered valuable to answer the above knowledge-required questions. Figure 3 shows the distribution.

3.5 Comparison

The most significant difference between the proposed dataset and existing VQA datasets is on the provision of supporting-facts. A large portion of visual questions require not only the information from the image itself, but also the often overlooked but critical commonsense knowledge external to the image. It is shown in [5] that 3 or more subjects agreed that 47.43% questions

Dataset	Number of images	Number of questions	Num. question categories	Average quest. length	Average ans. length	Knowledge Bases	Supporting Facts
DAQUAR [12]	1,449	12,468	4	11.5	1.2	-	-
COCO-QA [9]	117,684	117,684	4	8.6	1.0	-	-
VQA-real [5]	204,721	614,163	20+	6.2	1.1	-	-
Visual Genome [11]	108,000	1,445,322	7	5.7	1.8	-	-
Visual7W [10]	47,300	327,939	7	6.9	1.1	-	-
Visual Madlibs [8]	10,738	360,001	12	6.9	2.0	-	-
VQA-abstract [5]	50,000	150,000	20+	6.2	1.1	-	-
VQA-balanced [55]	15,623	33,379	1	6.2	1.0	-	-
KB-VQA [51]	700	2,402	23	6.8	2.0	1	-
Ours (FVQA)	2,190	5,826	32	9.5	1.2	3	✓

Table 3: Major datasets for VQA and their main characteristics.

in the VQA dataset require commonsense reasoning to answer (18.14%: 6 or more subjects). However, such external knowledge is not provided in all the existing VQA datasets. To the best knowledge of the authors, this is the first VQA dataset providing supporting-facts.

In this dataset, the supporting-facts which are necessary for answering the corresponding visual questions are obtained from several large-scale structured knowledge bases. This dataset enables the development of approaches which utilize the information from both the image and the external knowledge bases. Different from [51] that only applied a single manual annotated knowledge source, we use two additional self-learned knowledge bases, which enable us to answer more general questions.

In a similar manner as ours, the Facebook bAbI [56] dataset also provides supporting-facts for pure textual questions. But the problem posed in this work is more complex than that in Facebook bAbI, as the information need to be extracted from both image and external commonsense knowledge bases.

Another feature of the proposed dataset is that the answers are restricted to the concepts from image and knowledge bases, so 'Yes'/'No' questions are excluded. In the VQA dataset [5], 38% questions can be answered using 'Yes' or 'No'. It is somewhat difficult to measure the reasoning abilities of LSTM-based approaches via these 'Yes'/'No' questions, because it is not clear how LSTM arrive at the answer.

4 Approach

As shown in Section 3, all the information extracted from images and KBs are stored as a graph of interlinked RDF triples. State-of-the-art LSTM approaches [7, 10, 15, 17–20] directly learn the mapping between questions and answers, which, however, do not scale well to the diversity of answers and cannot provide the key information that the reasoning is based on. In contrast, we propose to learn the mapping between questions and a set of KB-queries, such that there is no limitation to the vocabulary size of the answers (*i.e.*, the answer to a test question does not have to be observed ahead in the training set) and the supporting facts used for reasoning can be provided.

4.1 Question-Query Mapping

In our approach, the KB-query will be performed according to three properties of questions, *i.e.*, visual concepts (refer to as `VC`), predicates (refer to as `REL`) and answer sources (refer to as `AS`). As shown in Section 3.4, there are 3, 13 and 2 types of visual concepts, predicates and the answer sources respectively. In training data, these properties of a question can be obtained through the annotated supporting-fact and the given answer, and there are in total 32 different combinations of the three properties in the proposed dataset (see Appendix). Since both question and query are sequences, the question-query mapping problem can be treated as a sequence-2-sequence problem [57], which can be solved by Recurrent Neural Network (RNN) [58]. In this work, we consider each distinct combination as a query type and learn a 32-class classifier using LSTM models [59], in order to identify the above three properties of an input question and perform a specific query.

The LSTM is a memory cell encoding knowledge at every time step for what inputs have been observed up to this step. We follow the model used in [54]. Letting σ be the sigmoid nonlinearity, the LSTM updates for time step t given inputs x_t, h_{t-1}, c_{t-1} are:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (1)$$
$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (2)$$
$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (3)$$
$$g_t = \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (4)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$
$$h_t = o_t \odot \tanh(c_t) \quad (6)$$
$$p_{t+1} = \text{softmax}(h_t) \quad (7)$$

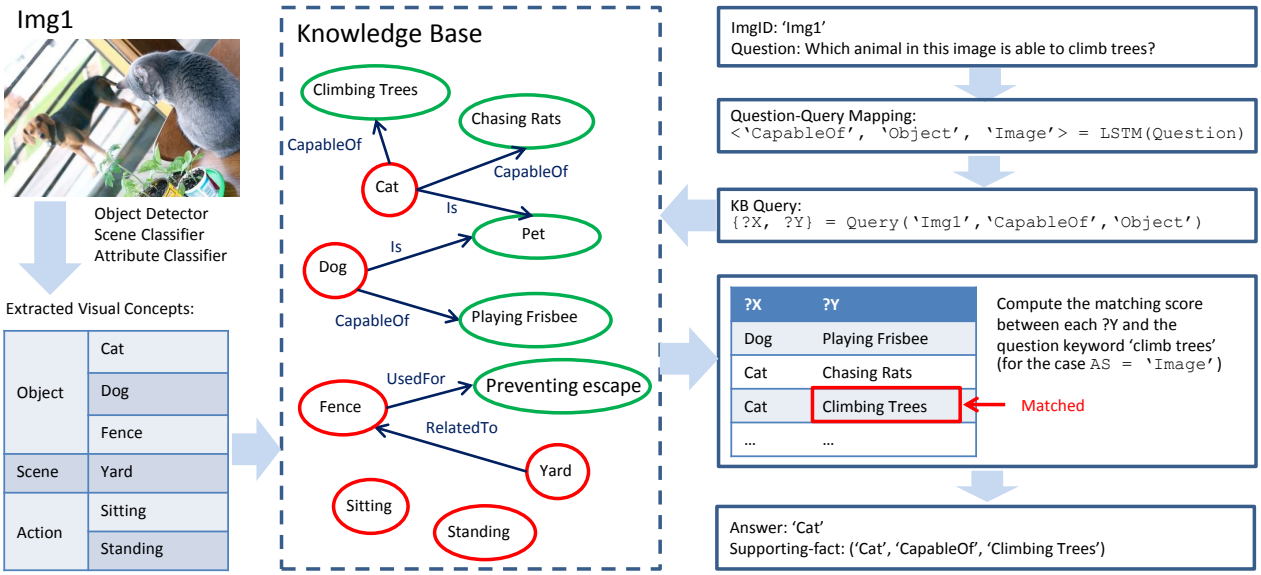

Fig. 4: An example of the reasoning process of the proposed VQA approach. The visual concepts (objects, scene, attributes) of the input image are extracted using trained models, which are further linked to the corresponding semantic entities in the knowledge base. The input question is firstly mapped to one of the query types using the LSTM model shown in Section 4.1. The types of predicate, visual concept and answer source can be determined accordingly. A specific query (see Section 4.2) is then performed to find all facts meeting the search conditions in KB. These facts are further matched to the keywords extracted from the question sentence. The fact with the highest matching score is selected and the answer is also obtained accordingly.

Here, i_t, f_t, c_t, o_t are the input, forget, memory, output state of the LSTM. The various W matrices are trained parameters and $\odot$ represents the product with a gate value. h_t is the hidden state at time step t and is fed to a Softmax, which will produce a probability distribution p_{t+1} over all candidate labels.

The LSTM model for the question to query type mapping is trained in an unrolled form. More formally, the LSTM takes the sequence of words in the given question $Q = (Q_0, ..., Q_L)$, where Q_0 is a special start word. Each word has been represented as a one-hot vector S_t. At time step $t = 0$, we set $x_0 = W_{es}S_0$ and $h_{initial} = \mathbf{0}$, where W_{es} is the learnable word embedding weights. From $t = 1$ to $t = L$, we set $x_t = W_{es}S_t$ and the input hidden state h_{t-1} is given by the previous step. The cost function is

$$\mathcal{C} = -\frac{1}{N}\sum_{i=1}^{N} \log p(T^{(i)}) + \lambda_{\boldsymbol{\theta}} \cdot ||\boldsymbol{\theta}||_2^2 \tag{8}$$

where N is the number of training examples. $T^{(i)}$ is the ground truth query types of the i-th training question. $\log p(T^{(i)})$ is the log-probability distribution over all candidate query types that is computed by the last LSTM cell, given the previous hidden state and the last word of question. $\boldsymbol{\theta}$ represents model parameters, $\lambda_{\boldsymbol{\theta}} \cdot ||\boldsymbol{\theta}||_2^2$ is a regularization term.

During the testing, the testing question words sequence will be feed-forward to the trained LSTM to produce the probability distribution p over all query types, via equations (1) to (7).

In Figure 4, the query type of the input question `'Which animal in this image is able to climb trees?'` is classified by the LSTM classifier as ⟨`REL='CapableOf',VC='Object',AS='Image'`⟩.

4.2 Answering by Querying KB

Query Construction Given that an input question's types of predicate (`REL`) and visual concept (`VC`) are obtained using the aforementioned LSTM-based query-type classifier, a KB-query `{?X,?Y} = Query(ImgID,REL,VC)` is accordingly constructed as shown in the following:

Find `?X`, `?Y`, subject to {`(ImgID,Contain,?X)` and
`(?X,VC-Type,VC)` and
`(?X,REL,?Y)` },

where `?X` denotes the visual concept of type `VC` in image `ImgID`, and `?Y` stands for the concept in KB which is linked to `?X` via predicate `REL`. All pairs of {`?X,?Y`}, which satisfy the search conditions shown in the query,

will be extracted from KBs. Note that the query is performed over all the facts related to visual concepts in the KB (see Table 1), not only on the 4216 facts in the proposed dataset.

As shown in the example of Figure 4, the query `Query('Img1','CapableOf','Image')` returns all objects in Image `'Img1'` which are linked to some KB concepts via predicate `'CapableOf'`.

Answering The answer source (refer to as `AS`) of a given question can be again obtained via the LSTM classifier in Section 4.1, and different answering strategies will be used for each particular source (*i.e.*, `AS` = `'Image'` or `'KB'`).

For answers from images (*i.e.*, one of the visual concepts {`?X`}), the KB concepts {`?Y`} will be matched to the keywords which are extracted from the question sentence by removing high frequency words (such as 'what', 'which', 'a', 'the'). The matching score is computed as the Jaccard similarity between the word sets of `?Y` (refer to as $\mathcal{W}_{\text{KB}}$) and question keywords (refer to as $\mathcal{W}_{\text{Q}}$): Score = $|\mathcal{W}_{\text{KB}} \cap \mathcal{W}_{\text{Q}}|$ / $|\mathcal{W}_{\text{KB}} \cup \mathcal{W}_{\text{Q}}|$. The visual concept `?X` corresponding to the highest-scored `?Y` will be considered as the answer. In the example of Figure 4, all `?Y`s are matched to the keywords 'climb trees' and the fact `('Cat','CapableOf','Climbing Trees')` has achieved the highest score, so the answer is `'Cat'`.

For answers from KBs (*i.e.*, one of the KB concepts {`?Y`}), we need to find out which visual concept (`?X`) is related to the input question. For questions asking about scene or action (*i.e.*, `VC` = `'Scene'` or `'Action'`), the scene/action concept with the highest probability (obtained from the scene/attribute classifier shown in Section 3.1) will be selected and the corresponding KB concept `?Y` will be considered as the answer. For questions asking about objects (*i.e.*, `VC` = `'Object'`), the visual concept `?X` is selected based on the location (such as `'top'`, `'bottom'`, `'left'`, `'right'` or `'center'`) or size (such as `'small'` and `'large'`) keywords in the question. Note that a single visual concept `?X` may correspond to multiple KB entities (`?Y`), *i.e.*, multiple answers. These answers are ordered according to their frequency in the answers of the training data.

5 Experiment

In this section, we first evaluate our question to KB query mapping performance. As a key component in our model, its performance impacts the final visual question answering (VQA) accuracy. We then report the performance of several baseline models, comparing with our own proposed method. Different from all the baseline models, our method is able to do the explicit reasoning for the VQA, *i.e.* we can select the supporting fact from the knowledge base that leads to the answer. We also report the supporting facts selection accuracy.

Knowledge Base	Q-Q Mapping Acc. (%)	
	Top-1	**Top-3**
DBpedia	61.73 ± 1.83	85.56 ± 2.14
ConceptNet	64.10 ± 1.10	80.85 ± 0.87
WebChild	83.15 ± 3.03	95.24 ± 1.23
Overall	64.94 ± 1.08	82.42 ± 0.56

Table 4: Question-Query mapping (QQMapping) accuracy on different Knowledge Base on the FVQA testing splits. Top-1 and Top-3 returned mappings are evaluated.

5.1 Question-Query Mapping Experiment

Table 4 reports the accuracy of our proposed question-Query mapping (QQmaping) model in Sec 4.1. The model is trained on the FVQA training splits and tested on the testing splits. To train the model, we use the Stochastic Gradient Descent (SGD) with mini-batches of 100 question-KB query type pairs. Both the word embedding size and the LSTM memory cell size are 128. The learning rate is set to 0.001 and clip gradient is 10. The dropout rate is set to 0.5. It converged after 50 epochs of training. We also provide the results on different KB sources. Questions asked based on the facts from the WebChild knowledge base are much easier to be mapped than questions based on other two KBs. This is mainly because much of the facts in WebChild are related to the 'comparative' relationship, such as 'car is faster than bike', which further lead to user-generated questions are more repeated in the format, for example, many questions are formulated as 'Which object in the image is more $<$*a comparative adj*$>$?'. Our Top-3 overall accuracy achieves 82.42 ± 0.56.

5.2 FVQA Experiment

Our FVQA tasks are formulated as the open-ended answer generation, which means the model is required to predict open-ended text outputs. To measure the accuracy, we simply calculate the proportion of correctly answered test questions. And the predicted answer is determined as correct if and only if it matches with the ground-truth answer (all the answers have been pre-processed by the python INFLECT package to eliminate the singular-plurals differences *etc.*). We also report the accuracy when the top-3 and top-10 answers are provided by the given methods.

Method	Overall Acc. ± Std (%)		
	Top-1	**Top-3**	**Top-10**
SVM-Question	10.37 ± 0.80	20.72 ± 0.58	34.63 ± 1.19
SVM-Image	18.41 ± 1.07	32.42 ± 1.06	47.53 ± 1.02
SVM-Question+Image	18.89 ± 0.91	32.78 ± 0.90	48.13 ± 0.73
LSTM-Question	10.45 ± 0.57	19.02 ± 0.74	31.64 ± 0.93
LSTM-Image	20.55 ± 0.81	36.01 ± 1.45	55.74 ± 2.28
LSTM-Question+Image	22.97 ± 0.64	36.76 ± 1.22	54.19 ± 2.45
Ours, gt-QQmaping‡	63.63 ± 0.73	71.30 ± 0.78	72.55 ± 0.79
Ours, top-1-QQmaping	52.56 ± 1.03	59.72 ± 0.82	60.58 ± 0.86
Ours, top-3-QQmaping	**56.91 ± 0.99**	**64.65 ± 1.05**	**65.54 ± 1.06**
Human	75.76 ± 0.83	-	-

Table 5: Overall accuracy on our FVQA testing splits for different methods. ‡ indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

Method	WUPS@0.9. ± Std (%)		
	Top-1	**Top-3**	**Top-10**
SVM-Question	17.06 ± 1.09	29.43 ± 0.62	44.73 ± 1.07
SVM-Image	24.73 ± 1.29	40.95 ± 1.34	56.33 ± 0.63
SVM-Question+Image	25.30 ± 1.09	41.37 ± 1.19	56.78 ± 0.64
LSTM-Question	15.82 ± 0.57	26.45 ± 0.61	40.99 ± 1.02
LSTM-Image	26.78 ± 1.02	44.00 ± 1.61	62.86 ± 2.23
LSTM-Question+Image	29.08 ± 0.91	44.36 ± 1.29	61.71 ± 2.82
Ours, gt-QQmaping‡	65.51 ± 0.82	72.37 ± 0.89	73.55 ± 0.87
Ours, top-1-QQmaping	54.79 ± 0.91	61.41 ± 0.71	62.22 ± 0.70
Ours, top-3-QQmaping	**59.67 ± 0.90**	**66.89 ± 1.01**	**67.77 ± 1.04**
Human	80.41 ± 0.78	-	-

Table 6: WUPS@0.9 on our FVQA testing splits for different methods. ‡ indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

Method	WUPS@0.0. ± Std (%)		
	Top-1	**Top-3**	**Top-10**
SVM-Question	56.88 ± 2.57	68.45 ± 0.67	76.93 ± 0.75
SVM-Image	59.64 ± 0.72	73.30 ± 0.96	81.77 ± 0.33
SVM-Question+Image	59.97 ± 0.63	73.39 ± 0.88	81.93 ± 0.39
LSTM-Question	51.45 ± 1.41	65.65 ± 0.66	75.76 ± 0.62
LSTM-Image	59.53 ± 1.52	73.83 ± 0.50	83.53 ± 0.89
LSTM-Question+Image	61.86 ± 0.81	74.45 ± 0.59	83.54 ± 0.94
Ours, gt-QQmaping‡	73.98 ± 0.77	78.67 ± 0.78	79.98 ± 0.79
Ours, top-1-QQmaping	64.96 ± 0.70	69.57 ± 0.65	70.64 ± 0.59
Ours, top-3-QQmaping	**72.34 ± 0.65**	**77.52 ± 0.70**	**78.69 ± 0.63**
Human	85.89 ± 0.66	-	-

Table 7: WUPS@0.0 on our FVQA testing splits for different methods. ‡ indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

Additionally, we also report the Wu-Palmer similarity (WUPS) [60]. The WUPS calculates the similarity between two words based on the similarity between their common subsequence in the taxonomy tree. If the similarity between two words is greater than a threshold then the candidate answer is considered to be right. We report on thresholds 0.9 and 0.0. All the reported results are averaged on the 5 test splits (standard deviation is also provided).

We evaluate baseline models on the FVQA tasks in three sets of experiments: without images (Question), without questions (Image) and with both images and questions (Question+Image). Same as [10], in the experiments without images (questions), we zero out the image (questions) features. We briefly describe the three models we used in the experiments:

SVM A Support Vector Machine (SVM) model that predicts the answer from a concatenation of image feature and question feature. For the image features, we use the fc7 features (4096-d) from the VggNet-16 [4]. The questions are represented by 300-d averaged word embeddings from a pre-trained word2vec model [61]. We take the top-500 most frequent answers (93.68% of the training set answers) as the class labels. At test time, we select the top-1, top-3 and top-10 scoring answer candidates. We use the LibSVM [62] and parameter C is set to 1.

LSTM We compare our system with an approach followed [9] (which we label LSTM) that treats the question answering as a classification problem. The LSTM outputs are fed into a softmax layer at the last timestep to predict answers over a fixed answers space (top-500 most frequent answers). This is also very similar to the 'LSTM+MLP' method proposed in [5]. Specifically, we use the fc7 layer (4096-d) of the pre-trained VggNet-16 model as the image features, and the LSTM is trained on our training split. The LSTM layer contains 512 memory cells in each unit. The learning rate is set to 0.001 and clip gradient is 5. The dropout rate is set to 0.5. Same as SVM models, we select the top-1, top-3 and top-10 scoring answer candidates at test time.

Human We also report the human performance. Testing splits are given to 5 human subjects and they are allowed to use any media (such as books, Wikipedias, Google *etc.*) to gather the information or knowledge to answer the question. Human subjects are only allowed to provide one answer to one question, so there are no Top-3 and Top-10 evaluations for the human performace. And please note that these 5 subjects are never involved in the previous questions collection procedure.

Ours Our KB-query based model is introduced in Section 4. To verify the effectiveness of our method, we implement three variants models. **gt-QQmaping** uses the ground truth question-query mapping, while **top-1-QQmaping** and **top-3-QQmaping** use the top-1 and top-3 predicted question-query mapping (see Section 4.1), respectively.

Table 5 shows the overall accuracy of all the baseline methods and our proposed models. In the case of Top-1 accuracy, our proposed **top-3-QQmaping** model performs best, which doubles the accuracy of the best baseline (LSTM-Question+Image). The top-3-QQmaping is better than top-1-QQmaping because it produces better Question-Query mapping results, as shown in the Table 4. However, it is still not as good as gt-QQmaping since there are Question-Query map-

Method	KB-Source								
	DBpedia			ConceptNet			WebChild		
	Top-1	Top-3	Top-10	Top-1	Top-3	Top-10	Top-1	Top-3	Top-10
SVM-Question	4.13 ± 0.89	9.94 ± 1.44	20.91 ± 2.08	11.03 ± 0.76	21.84 ± 0.83	35.89 ± 1.31	16.38 ± 4.38	31.40 ± 0.93	50.17 ± 2.98
SVM-Image	7.11 ± 0.73	19.40 ± 2.50	39.75 ± 2.23	20.23 ± 1.14	33.91 ± 1.57	48.15 ± 1.18	20.95 ± 2.17	43.32 ± 1.23	57.31 ± 2.71
SVM-Question+Image	7.35 ± 0.73	20.43 ± 2.32	40.38 ± 2.46	20.76 ± 0.95	34.18 ± 1.35	48.64 ± 1.08	21.40 ± 2.01	43.33 ± 1.45	59.58 ± 2.92
LSTM-Question	5.08 ± 0.54	11.17 ± 0.21	23.68 ± 2.42	10.96 ± 0.60	19.70 ± 0.87	32.02 ± 0.96	16.37 ± 2.87	28.31 ± 2.11	44.91 ± 1.32
LSTM-Image	14.62 ± 2.38	29.68 ± 4.35	49.74 ± 2.56	20.96 ± 0.96	36.54 ± 1.21	56.32 ± 2.31	28.82 ± 2.79	43.64 ± 5.49	62.04 ± 3.87
LSTM-Question+Image	15.77 ± 2.07	28.30 ± 3.56	49.45 ± 7.25	23.57 ± 0.52	37.36 ± 1.43	54.30 ± 2.99	31.74 ± 3.69	48.18 ± 5.80	**63.59 ± 5.60**
Ours, gt-QQmaping‡	65.96 ± 2.06	78.80 ± 1.08	79.43 ± 1.11	64.62 ± 0.82	71.75 ± 0.94	73.17 ± 0.94	45.55 ± 1.14	48.29 ± 1.08	48.86 ± 1.29
Ours, top-1-QQmaping	51.25 ± 2.21	63.07 ± 1.23	63.13 ± 1.21	53.50 ± 1.14	60.16 ± 0.87	61.20 ± 0.96	43.54 ± 2.14	46.58 ± 1.80	47.03 ± 2.08
Ours, top-3-QQmaping	**56.67 ± 1.68**	**69.31 ± 1.10**	**69.36 ± 1.03**	**57.60 ± 1.29**	**64.70 ± 1.24**	**65.77 ± 1.28**	**48.74 ± 2.47**	**53.45 ± 1.99**	53.90 ± 2.26
Human	74.41 ± 1.13	-	-	75.88 ± 0.84	-	-	77.2 ± 1.91	-	-

Table 8: Accuracies on the questions that asked based on different Knowledge Base sources. ‡ indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

Method	Visual Concept								
	Object			Scene			Action		
	Top-1	Top-3	Top-10	Top-1	Top-3	Top-10	Top-1	Top-3	Top-10
SVM-Question	11.39 ± 0.84	22.72 ± 0.53	37.89 ± 1.06	0.68 ± 0.24	1.98 ± 0.83	3.97 ± 0.74	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
SVM-Image	20.08 ± 1.25	35.33 ± 1.13	51.60 ± 1.11	2.88 ± 0.46	5.22 ± 0.85	9.60 ± 0.36	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
SVM-Question+Image	20.59 ± 1.04	35.73 ± 1.06	52.24 ± 0.79	3.02 ± 0.42	5.22 ± 0.84	9.83 ± 0.19	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
LSTM-Question	11.39 ± 0.52	20.70 ± 0.63	34.34 ± 0.92	1.58 ± 1.07	3.05 ± 0.67	6.03 ± 1.09	0.00 ± 0.00	3.53 ± 4.71	4.71 ± 5.76
LSTM-Image	22.34 ± 0.98	39.09 ± 1.49	60.13 ± 2.18	3.80 ± 0.92	7.07 ± 1.49	14.94 ± 1.43	0.00 ± 0.00	1.05 ± 2.11	1.05 ± 2.11
LSTM-Question+Image	24.92 ± 0.88	39.71 ± 1.48	58.32 ± 3.82	4.56 ± 0.72	9.21 ± 2.00	15.56 ± 1.62	5.99 ± 6.64	5.99 ± 6.64	9.52 ± 6.15
Ours, gt-QQmaping‡	68.70 ± 0.77	76.50 ± 0.69	77.11 ± 0.71	14.81 ± 0.83	20.93 ± 1.54	27.75 ± 2.03	28.78 ± 5.72	39.76 ± 6.61	53.32 ± 10.47
Ours, top-1-QQmaping	56.75 ± 1.48	64.11 ± 1.33	64.47 ± 1.35	12.81 ± 1.29	18.36 ± 2.68	24.19 ± 3.52	16.58 ± 10.14	19.57 ± 12.38	22.57 ± 14.32
Ours, top-3-QQmaping	**61.53 ± 1.31**	**69.51 ± 1.35**	**69.90 ± 1.20**	**12.89 ± 1.29**	**18.51 ± 2.72**	**24.34 ± 3.57**	**19.75 ± 6.16**	**22.73 ± 8.34**	**25.72 ± 10.12**
Human	80.58 ± 0.42	-	-	29.75 ± 0.96	-	-	30.65 ± 10.20	-	-

Table 9: Accuracies on questions that focus on three different visual concepts. ‡ indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

Method	Answer-Source					
	Image			KB		
	Top-1	Top-3	Top-10	Top-1	Top-3	Top-10
SVM-Question	12.38 ± 0.88	24.68 ± 0.47	41.05 ± 1.05	0.78 ± 0.23	2.00 ± 0.47	4.16 ± 0.79
SVM-Image	21.81 ± 1.30	38.27 ± 1.33	55.93 ± 1.29	2.30 ± 0.25	4.73 ± 0.62	7.74 ± 0.51
SVM-Question+Image	22.38 ± 1.09	38.72 ± 1.20	56.63 ± 0.97	2.38 ± 0.30	4.65 ± 0.63	7.86 ± 0.44
LSTM-Question	12.35 ± 0.57	22.45 ± 0.68	37.06 ± 0.98	1.45 ± 0.68	2.72 ± 0.63	5.89 ± 0.95
LSTM-Image	24.19 ± 0.98	42.40 ± 1.73	64.92 ± 2.44	3.31 ± 0.74	5.69 ± 0.87	11.87 ± 0.71
LSTM-Question+Image	26.98 ± 1.08	42.90 ± 1.58	62.87 ± 4.10	4.03 ± 0.95	7.70 ± 1.30	13.11 ± 1.51
Ours, gt-QQmaping‡	73.69 ± 0.67	81.04 ± 0.51	81.04 ± 0.51	15.95 ± 0.64	25.17 ± 0.89	32.39 ± 1.16
Ours, top-1-QQmaping	61.11 ± 1.48	68.31 ± 1.24	68.34 ± 1.22	12.12 ± 1.01	19.13 ± 1.30	23.91 ± 1.39
Ours, top-3-QQmaping	**66.32 ± 1.20**	**74.11 ± 1.21**	**74.15 ± 1.18**	**12.39 ± 1.09**	**19.87 ± 1.35**	**24.81 ± 1.31**
Human	82.97 ± 0.38	-	-	41.63 ± 1.84	-	-

Table 10: Accuracies for different methods according to different answer sources. ‡ indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

ping errors. And there is still a significant gap between our models and the human performance. Among the baseline models, LSTM methods perform slightly better than SVM. Question+Image models always predict more accurate answers than Question or Image alone, no matter in SVM or LSTM. Interestingly, contradictory with previous works [5,9] which found that 'question' played a more important role than 'image' in the visual question answering, our {SVM,LSTM}-Q does worse than {SVM,LSTM}-I means that our questions rely more heavily on the image content, compared to the VQA [5] or COCO-QA datasets [9]. Actually, If {SVM,LSTM}-Q achieves too high performance, the corresponding questions may be not *Visual questions* and they may be just *Textual questions*. According to the cases of Top-3 and Top-10, our **top-3-QQmaping** model also performs best, however, due to both of our SVM and LSTM models are optimized for the top-1 classification accuracy, not for the ranking performance, it is not accurate to compare the performance of SVM and LSTM models in terms of Top-10 accuracy. Table 11 shows some example results generated by our final model. Table 6 and 7 report the WUPS@0.9 and WUPS@0.0 accuracy for different methods.

Table 8 reports the accuracy of the methods on the questions that asked based on different Knowledge Base sources. Our methods produce the same-level accuracy on all the three KB sources, which suggests our models can generalize to many different KBs. DBpedia has highly structured knowledge from Wikipedia. ConceptNet includes many commonsense knowledge while there are many 'comparative' facts in WebChild.

Table 9 illustrates the performance on questions that focus on three different visual concepts, which are object, scene and action. The performance on object-related questions is much higher than the other two

				
	Which furniture in this image can I lie on?	What animal in this image are pulling carriage?	Which animal in this image has stripes?	Which transportation way in this image is cheaper than taxi?
Mined Facts:	a sofa is usually to sit or lie on	horses sometimes pull carriages	zebras have stripes	bus are cheaper than taxi
Predicted Answer:	sofa	horse	zebras	bus
Ground Truth:	sofa	horse	zebras	bus

				
	Which object in this image can I ride?	What thing in this image is helpful for a romantic dinner?	Which food in this image can be seen on a birthday party?	What animal can be found in this place?
Mined Facts:	motorcycle is used for riding	wine is good for a romantic dinner	cake is related to birthday party	You are likely to find a cow in a pasture
Predicted Answer:	motorcycle	wine	cake	cow
Ground Truth:	motorcycle	wine	cake	cow

				
	What kind of people can we usually find in this place?	What does the animal in the right of this image have as a part?	Which object in this image is related to sail?	What thing in this image is capable of hunting a mouse?
Mined Facts:	skiiers can be on a ski slope	snails have shells	boat is related to sailing	a cat can hunt mice
Predicted Answer:	skiiers	shells	boat	cat
Ground Truth:	skiiers	shells	boat	cat

		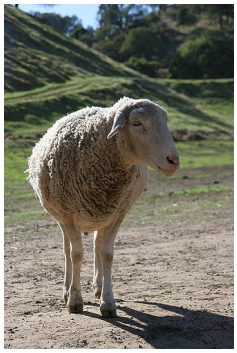		
	Which object in this image is used to measure the passage of time?	Which object in this image is a very trainable animal?	Which object in this image is related to wool?	Which instrument in this image is common in jazz?
Mined Facts:	a clock is for measuring the passage of time	horses are very trainable animals	sheep is related to wool	a saxophone is a common instrument in jazz
Predicted Answer:	clock	horse	sheep	saxophone
Ground Truth:	clock	horse	sheep	saxophone

Table 11: Some example results generated by our methods. The supporting facts triplet have been translate to textual sentence for easy understanding.

	 What animal in this image can rest standing up?	 What does the place in the image can be used for?	 Which object in this image is utilized to chill food?	 What can I do using this place?
Predicted VC:	Person, Cart, ...	Kitchen, ...	Refrigerator, Over, Stove, ...	Kitchen, Refrigerator, ...
GT VC:	Horse	Bathroom	Refrigerator	Kitchen
Predicted QT:	(CapableOf, Image, Object)	(UsedFor, KB, Scene)	(UsedFor, Image, Object)	(UsedFor, KB, Scene)
GT QT:	(CapableOf, Image, Object)	(UsedFor, KB, Scene)	(IsA, Image, Object)	(UsedFor, KB, Scene)
Mined Fact:	People can stand up for themselves	A bathroom is for washing your hands	A refrigerator is used for chilling food	A kitchenette is for cooking
GT Fact:	Horses can rest standing up	A kitchen is for cooking	An oven is a device to heat food	A kitchenette is for preparing food
Predicted Answer:	People	Cooking	Refrigerator	Cooking
Ground Truth:	Horse	Washing	Oven	Preparing food

Table 12: False examples generated by our methods (GT: ground truch, QT: query type, VC: visual concept). The false reason for the first two examples is that the visual concepts are not extracted correctly. Our method makes a mistake on the third example due to the false question-to-query mapping. The reason for the fourth example is that the question has multiple answers (our method orders these answers according to the frequency in the training data, see Section 4.2 for details).

types, especially when image features are given. This is not surprise since the image features are extracted from the VggNet which has been pre-trained on the object classification task. The accuracy of action or scene related questions is much poorer than object-related questions (even for human subjects), which is partially because the answers of many scene or action related questions can be expressed in different ways. For example, the answer to 'What can I do in this place' (the image scene is kitchen) can be 'preparing food' or 'cooking'. On the other hand, the performance of action classification is also worse than objects, which also leads to poor VQA performance.

Table 10 gives the accuracy for different methods according to different answer sources. If the answer is a visual concept in the image, we categorized the answer source into the 'Image', otherwise, it is categorized into the 'KB'. From the table, we can see the accuracy is much higher when the answer is from the 'Image' side, nearly 5 times as much as the 'KB'. This suggests that generating answers from a nearly unlimited answer space (and the answer is not directly appeared in the image) is a very challenging task. Our proposed models performs better than other baseline models.

Table 11 shows some examples in which our method achieves the right answer and Table 12 gives some false examples. From Table 12, we can see that the reasons of false examples are categorized into three aspects: 1. The visual concepts of the input image are not extracted correctly. In particular, the errors usually occur when the questioned visual concepts are missing. 2. The question-to-query mapping (via LSTM) is not correct, which means that the question text is wrongly understood. 3. Some errors occur during the stage of post-processing that generates the final answer from queried

Method	Facts Prediction Acc. ± Std (%)		
	Top-1	**Top-3**	**Top-10**
Ours, gt-QQmaping‡	56.31 ± 0.89	62.55 ± 0.96	63.55 ± 0.96
Ours, top-1-QQmaping	38.76 ± 0.88	42.96 ± 0.78	43.60 ± 0.77
Ours, top-3-QQmaping	$\mathbf{41.12 \pm 0.74}$	$\mathbf{45.49 \pm 0.89}$	$\mathbf{46.13 \pm 0.87}$

Table 13: Facts prediction accuracy for our proposed methods. ‡ indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

KB facts. The approach should selected the most relevant fact from multiple facts that matches with query conditions. In particular for questions whose answers are from KB (in order words, open-ended questions), our method may generate multiple answers (see Section 4.2). Sometimes, the ground truth is not the first in the ordered answers. In these cases, the top1 answer is wrong, but the top$N$ answer may be correct.

Different from all the other state-of-art VQA methods, our proposed models are capable of explicit reasoning, *i.e.* providing the supporting facts of the predicted answer. Table 13 reports the accuracy of the facts prediction. We have 41% chance to predict the correct supporting facts. This is a surprisingly good result given the truth that there are millions of facts in the incorporated Knowledge Bases.

6 Conclusion

In this work, we have proposed a dataset and an approach for the task of visual question answering with external commonsense knowledge. The proposed FVQA dataset differs from existing VQA datasets in that it provides a supporting-fact which is critical for answering each visual question. We have also developed a novel VQA approach, which is able to automatically find the supporting fact for a visual question from large-scale

structured knowledge bases. Instead of directly learning the mapping from questions to answers, our approach learns the mapping from questions to KB-queries, so it is much more scalable to the diversity of answers. Not only give the answer to a visual question, the proposed method also provides the supporting fact based on which it arrives at the answer, which uncovers the reasoning process.

References

1. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, 2012. 1
2. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. Eur. Conf. Comp. Vis.*, 2014. 1, 3, 4
3. Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2009. 1, 3, 4
4. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 1, 10
5. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015. 1, 2, 3, 4, 6, 7, 10, 11
6. Mateusz Malinowski and Mario Fritz. Towards a Visual Turing Challenge. *arXiv:1410.8027*, 2014. 1, 2, 3
7. Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. In *Proc. Adv. Neural Inf. Process. Syst.*, 2015. 1, 2, 3, 4, 7
8. Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. Visual Madlibs: Fill in the Blank Description Generation and Question Answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, December 2015. 1, 2, 3, 4, 7
9. Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring Models and Data for Image Question Answering. In *Proc. Adv. Neural Inf. Process. Syst.*, 2015. 1, 2, 3, 4, 7, 10, 11
10. Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. *arXiv preprint arXiv:1511.03416*, 2015. 1, 2, 3, 4, 7, 10
11. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *arXiv preprint arXiv:1602.07332*, 2016. 3, 6, 7
12. Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1682–1690, 2014. 3, 7
13. Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. Joint video and text parsing for understanding events and answering queries. *MultiMedia, IEEE*, 21(2):42–70, 2014. 3
14. Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual Turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015. 3
15. Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015. 3, 7
16. Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *Proc. Int. Conf. Mach. Learn.*, 2015. 3
17. Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering. *arXiv preprint arXiv:1511.05960*, 2015. 3, 7
18. Aiwen Jiang, Fang Wang, Fatih Porikli, and Yi Li. Compositional Memory for Visual Question Answering. *arXiv preprint arXiv:1511.05676*, 2015. 3, 7
19. Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Deep Compositional Question Answering with Neural Module Networks. *arXiv preprint arXiv:1511.02799*, 2015. 3, 7
20. Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked Attention Networks for Image Question Answering. *arXiv preprint arXiv:1511.02274*, 2015. 3, 7
21. Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction. *arXiv preprint arXiv:1511.05756*, 2015. 3
22. Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. Uncovering Temporal Context for Video Question and Answering. *arXiv preprint arXiv:1511.04670*, 2015. 3
23. Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. *DBpedia: A nucleus for a web of open data*. Springer, 2007. 3, 4
24. Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. Open information extraction for the web. In *Proc. Int. Joint Conf. on Artificial Intell.*, 2007. 3
25. Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proc. ACM SIGMOD/PODS Conf.*, pages 1247–1250, 2008. 3
26. Andrew Carlson, Justin Betteridge, Bryan Kisiel, and Burr Settles. Toward an Architecture for Never-Ending Language Learning. In *Proc. National Conf. Artificial Intell.*, 2010. 3
27. Xinlei Chen, Ashish Shrivastava, and Arpan Gupta. Neil: Extracting visual knowledge from web data. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2013. 3
28. Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. YAGO3: A knowledge base from multilingual Wikipedias. In *CIDR*, 2015. 3
29. Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. 3
30. RDF Working Group et al. Resource description framework, 2014. http://www.w3.org/standards/techs/rdf. 3

31. Eric Prud'Hommeaux, Andy Seaborne, et al. SPARQL query language for RDF. *W3C recommendation*, 15, 2008. 3
32. Orri Erling. Virtuoso, a Hybrid RDBMS/Graph Column Store. *IEEE Data Eng. Bull.*, 35(1):3–8, 2012. 3
33. Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. In *Proc. Int. Joint Conf. on Artificial Intell.*, 2013. 3
34. Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open Information Extraction: The Second Generation. In *Proc. Int. Joint Conf. on Artificial Intell.*, 2011. 3
35. Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proc. Conf. Empirical Methods Natural Language Processing*, 2011. 3
36. Xinlei Chen, Ashish Shrivastava, and Arpan Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014. 3
37. Niket Tandon, Gerard De Melo, and Gerhard Weikum. Acquiring Comparative Commonsense Knowledge from the Web. In *Proc. National Conf. Artificial Intell.*, 2014. 3, 4, 5
38. Niket Tandon, Gerard de Melo, Fabian Suchanek, and Gerhard Weikum. Webchild: Harvesting and organizing commonsense knowledge from the web. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 523–532. ACM, 2014. 3, 4
39. Hugo Liu and Push Singh. ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004. 3, 4
40. Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proc. Conf. Empirical Methods Natural Language Processing*, pages 1533–1544, 2013. 4
41. Antoine Bordes, Sumit Chopra, and Jason Weston. Question answering with subgraph embeddings. *arXiv:1406.3676*, 2014. 4
42. Qingqing Cai and Alexander Yates. Large-scale Semantic Parsing via Schema Matching and Lexicon Extension. In *Proc. Conf. the Association for Computational Linguistics*, 2013. 4
43. Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. Open question answering over curated and extracted knowledge bases. In *Proc. ACM Int. Conf. Knowledge Discovery & Data Mining*, 2014. 4
44. Oleksandr Kolomiyets and Marie-Francine Moens. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434, 2011. 4
45. Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. Scaling semantic parsers with on-the-fly ontology matching. In *Proc. Conf. Empirical Methods Natural Language Processing*, 2013. 4
46. Xuchen Yao and Benjamin Van Durme. Information extraction over structured data: Question answering with Freebase. In *Proc. Conf. the Association for Computational Linguistics*, 2014. 4
47. Percy Liang, Michael I Jordan, and Dan Klein. Learning dependency-based compositional semantics. *Computational Linguistics*, 39(2):389–446, 2013. 4
48. Christina Unger, Lorenz Bühmann, Jens Lehmann, Axel-Cyrille Ngonga Ngomo, Daniel Gerber, and Philipp Cimiano. Template-based question answering over RDF data. In *WWW*, 2012. 4
49. Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal Knowledge Base for Visual Question Answering. *arXiv:1507.05670*, 2015. 4
50. Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. 4
51. Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit Knowledge-based Reasoning for Visual Question Answering. *arXiv preprint arXiv:1511.02570*, 2015. 4, 7
52. Ross Girshick. Fast r-cnn. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2015. 4
53. Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Proc. Adv. Neural Inf. Process. Syst.*, 2014. 4
54. Qi Wu, Chunhua Shen, Anton van den Hengel, Lingqiao Liu, and Anthony Dick. What Value Do Explicit High-Level Concepts Have in Vision to Language Problems? In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. 4, 7
55. Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016. 7
56. Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015. 7
57. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 3104–3112, 2014. 7
58. Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3, 2010. 7
59. Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 7
60. Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proc. Conf. the Association for Computational Linguistics*, 1994. 10
61. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 3111–3119, 2013. 10
62. Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 10

A Appendix

Table 14 shows the 32 query types classified by the types of visual concepts, predicates and answer sources, and the number of training/test questions for each query type. In Table 15, more examples of the proposed dataset are given. In Tables 16 and 17, the visual concepts extracted by object detectors and attribute classifiers are demonstrated. Figure 5 shows a snapshot of the designed system for collecting questions.

Query type (REL,VC,AS)	Train	Test	Total
(Category,Image,Object):	381.20 ± 11.72	383.80 ± 11.72	765
(IsA,Image,Object):	374.80 ± 11.94	371.20 ± 11.94	746
(RelatedTo,Image,Object):	347.80 ± 21.34	353.20 ± 21.34	701
(UsedFor,Image,Object):	330.80 ± 28.25	318.20 ± 28.25	649
(CapableOf,Image,Object):	313.60 ± 14.14	301.40 ± 14.14	615
(HasA,Image,Object):	184.80 ± 5.56	184.20 ± 5.56	369
(HasProperty,Image,Object):	173.40 ± 8.43	162.60 ± 8.43	336
(Comparative,Image,Object):	129.20 ± 10.57	131.80 ± 10.57	261
(AtLocation,Image,Object):	126.80 ± 8.98	120.20 ± 8.98	247
(AtLocation,KB,Scene):	110.00 ± 17.10	106.00 ± 17.10	216
(UsedFor,KB,Scene):	61.80 ± 5.74	61.20 ± 5.74	123
(UsedFor,KB,Object):	52.20 ± 4.21	46.80 ± 4.21	99
(Desires,Image,Object):	33.60 ± 4.88	32.40 ± 4.88	66
(RelatedTo,KB,Object):	25.60 ± 1.96	34.40 ± 1.96	60
(AtLocation,KB,Object):	30.80 ± 1.72	28.20 ± 1.72	59
(HasProperty,KB,Scene):	21.60 ± 8.19	31.40 ± 8.19	53
(Comparative,KB,Object):	23.80 ± 2.64	24.20 ± 2.64	48
(HasA,KB,Object):	27.20 ± 6.88	19.80 ± 6.88	47
(HasA,KB,Scene):	21.00 ± 2.19	23.00 ± 2.19	44
(PartOf,Image,Object):	23.20 ± 4.66	18.80 ± 4.66	42
(AtLocation,KB,Scene):	20.20 ± 1.72	17.80 ± 1.72	38
(HasProperty,KB,Object):	12.00 ± 1.10	16.00 ± 1.10	28
(Comparative,KB,Scene):	11.80 ± 2.23	15.20 ± 2.23	27
(Category,KB,Object):	22.00 ± 1.67	30.00 ± 1.67	52
(IsA,KB,Object):	12.80 ± 2.93	12.20 ± 2.93	25
(ReceivesAction,Image,Object):	11.60 ± 2.33	9.40 ± 2.33	21
(Comparative,KB,Action):	10.40 ± 0.80	10.60 ± 0.80	21
(CapableOf,KB,Object):	10.20 ± 2.14	9.80 ± 2.14	20
(ReceivesAction,KB,Object):	7.20 ± 1.72	6.80 ± 1.72	14
(CreatedBy,Image,Object):	6.60 ± 1.20	6.40 ± 1.20	13
(CapableOf,KB,Scene):	4.80 ± 1.72	6.20 ± 1.72	11
(HasProperty,KB,Action):	4.40 ± 2.33	5.60 ± 2.33	10

Table 14: The 32 query types. VC: type of visual concepts, REL: type of predicates, AS: answer source.

				
Which object in this image is able to stop cars	Can you name the beer that we usually enjoy with the fruit in the image?	Which instrument in this image is usually used in polka music	Why do they need a bow tie?	
GT Fact:	Traffic light can stop cars	Lemon is related to corona	Accordions are used in polka music	Bow ties are worn at formal events
Ground Truth:	Traffic light	Corona	Accordion	Formal events

				
Whether this animal runs slower or faster than horse?	What drink is made with this fruit?	Whether the game is a summer or winter Olympic?	How many times you should use this stuff per day?	
GT Fact:	Camel are slower than horse	Grenadine is related to pomegranates	Balance beam belongs to the category of Summer Olympic disciplines	A toothbrush should be used twice a day
Ground Truth:	Slower	Grenadine	Summer Olympic disciplines	Used twice a day

				
What is the difference between the animal on the left and moth?	Is there present in the image any tool used for logging?	Can you identify any medical equipment in the image?	Can you describe the metal thing on the right?	
GT Fact:	Butterfly are usually more colorful than moth	Chain saw belongs to the category of Logging	Crutch belongs to the category of Medical equipment	A knife is a metal blade for cutting or as a weapon with usually one long sharp edge fixed in a handle
Ground Truth:	More colorful than moth	Chain saw	Crutches	Metal blade for cutting or as a weapon with usually one long sharp edge fixed in a handle

Table 15: More examples of the constructed dataset with a diverse set of questions, supporting facts and answers.

Category	Number	Object Name categories
person	1	person
vehicle	12	bicycle, car, motorcycle, airplane, bus, train, truck, boat, cart, snowmobile, snowplow, unicycle
outdoor	5	traffic light, fire hydrant, stop sign, parking meter, bench
animal	49	cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, ant, antelope, armadillo, bee, butterfly, camel, centipede,dragonfly, fox, frog, giant panda, goldfish, hamster, hippopotamus, isopod, jellyfish, koala bear, ladybug, lion, lizard, lobster,monkey, otter, porcupine, rabbit, ray, red panda, scorpion, seal, skunk, snail, snake, squirrel, starfish, swine, tick, tiger, turtle, whale, bird
accessory	21	backpack, umbrella, handbag, tie, suitcase, band aid, bathing cap, crutch, diaper, face powder, hat with awide brim, helmet, maillot, miniskirt, neck brace, plastic bag, stethoscope, swimming trunks, bow tie, sunglasses, brassiere
sports	25	frisbee, skis, snowboard, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket, balance beam, baseball, basketball, croquet ball, golf ball, golfcart, horizontal bar, punching bag, racket, rugby ball,soccer ball, tennis ball, volleyball, ping-pong ball, puck, dumbbell
kitchen	20	bottle, wine glass, cup, fork, knife, spoon, bowl, beaker, can opener, cocktail shaker, corkscrew, frying pan, ladle, milk can,pitcher, plate rack, salt or pepper shaker, spatula, strainer, water bottle
food	27	banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, artichoke, bagel, bell pepper, burrito, cream,cucumber, fig, guacamole, hamburger, head cabbage, lemon, mushroom, pineapple, pomegranate, popsicle, pretzel,strawberry
furniture	8	chair, couch, potted plant, bed, dining table, toilet, baby bed, filing cabinet
electronic	7	tv, laptop, mouse, remote, keyboard, cell phone, iPod
appliance	14	microwave, oven, toaster, sink, refrigerator, coffee maker, dishwasher, electric fan, printer, stove, tape player, vacuum,waffle iron, washer
indoor	19	clock, vase, scissors, teddy bear, hair drier, toothbrush, binder, bookshelf, digital clock, hair spray, lamp, lipstick, pencil box, pencil sharpener, perfume, rubber eraser, ruler, soap dispenser, book
music	17	accordion, banjo, cello, chime, drum, flute, french horn, guitar, harmonica, harp, oboe, piano, saxophone,trombone, trumpet, violin, maraca
tool	9	axe, bow, chain saw, hammer, power drill, screwdriver, stretcher, syringe, nail

Table 16: The 234 objects which can be detected by object detectors.

Super-category	Number	Attribute categories
Action	24	playing, sitting, standing, swinging, catching, cutting, dining, driving, eating, flying, hitting, jumping, laying, racing, reads, swimming, running, sleeping, smiling, taking, talking, walking, wearing, wedding
Scene	16	road, snow, airport, bathroom, beach, city, court, forest, hill, island, lake, market, park, room, sea, field zoo
Object	92	children, computer, drink, glass, monitor, tree, wood, basket, bathtub, beer, blanket, box, bread, bridge, buildings, cabinets, camera, candles, cheese, chicken, chocolate, church, clouds, coat, coffee, decker, desk, dishes, door, face, fence, fish, flag, flowers, foods, fruits, furniture, grass, hair, hands, head, hotdog, house, ice, jacket, kitten, lettuce, lights, luggage, meat, metal, mouth, onions, palm, pants, papers, pen, pillows, plants, plates, players, police, potatoes, racquet, railing, rain, rocks, salad, sand, seat, shelf, ship, shirt, shorts, shower, sofa, station, stone, suit, toddler, tomatoes, towel, tower, toys, tracks, vegetables, vehicles, wall, water, wii, windows, wine

Table 17: The 24 actions, 16 scenes and 92 objects detected by the attribute classifier.

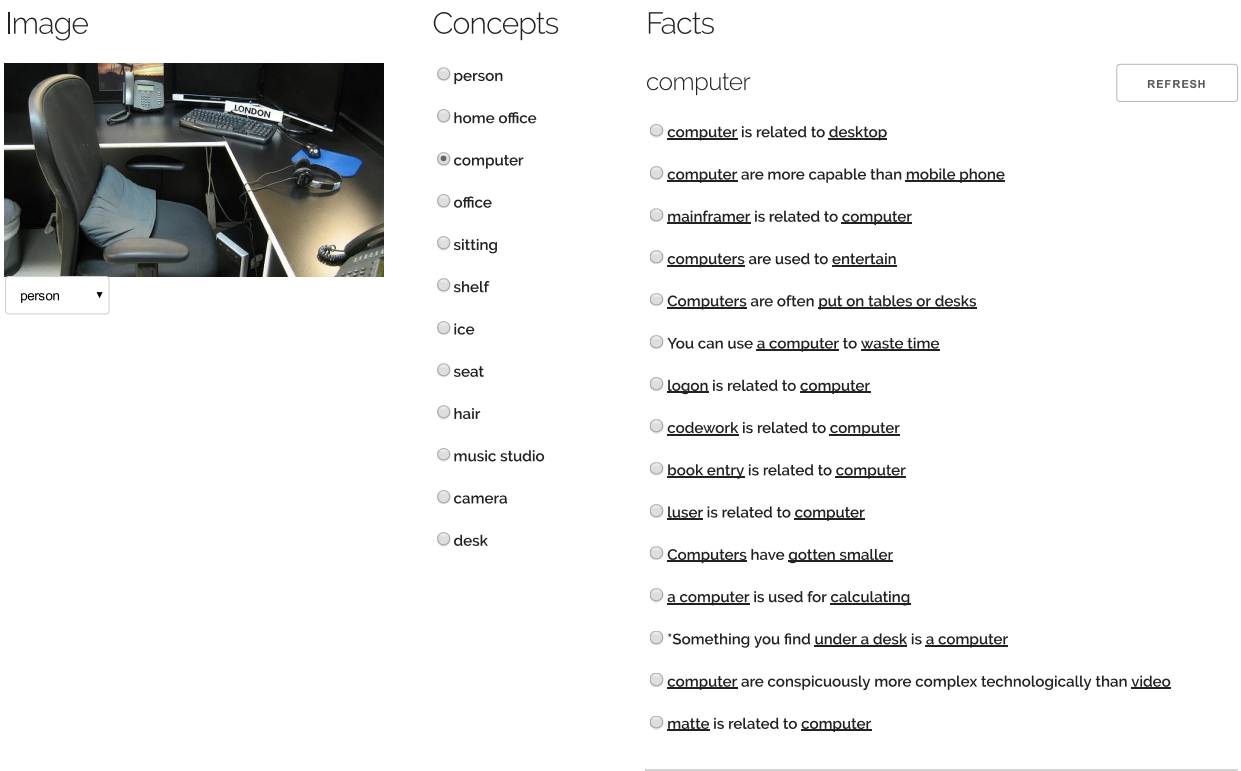

Fig. 5: The system for collecting questions.