

# FVQA: Fact-Based Visual Question Answering

Peng Wang<sup>ID</sup>, Qi Wu<sup>ID</sup>, Chunhua Shen<sup>ID</sup>, Anthony Dick, and Anton van den Hengel<sup>ID</sup>

**Abstract**—Visual Question Answering (VQA) has attracted much 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 (Fact-based VQA), a VQA dataset which requires, and supports, much deeper reasoning. FVQA primarily contains 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. Each 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.

**Index Terms**—Visual question answering, knowledge base, recurrent neural networks

## 1 INTRODUCTION

**V**ISUAL 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], [2], [3], [4], however, VQA requires that a method be prepared to answer a question that has never been seen before. In object detection the set of objects of interest are specified at training time, for example, whereas in VQA the set of questions which may be asked inevitably extend beyond those in the training set.

The set of questions that a VQA method is able to answer are 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], [6], [7], [8], [9], [10]. This is a restricted set of questions, which require only relatively shallow image understanding to answer. It is possible, for example, to answer 'How many giraffes are in the image?' without understanding any non-visual knowledge about giraffes.

The number of VQA datasets available has grown as the field progresses [5], [6], [7], [8], [9], [10]. They have contributed 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 COCO-QA dataset [9], questions are generated automatically from image captions which describe the major visible content of the image.

The VQA dataset in [5], for example, has been very well studied, yet only 5.5 percent of questions require adult-level (18+) knowledge (28.4 and 11.2 percent questions require older child (9-12) and teenager (13-17) knowledge). This limitation means that this is not a truly "AI-complete" problem, because this is not a realistic test for human beings. Humans inevitably use their knowledge to answer questions, even visual ones. For example, to answer the question given in Fig. 1, one not only needs to visually recognize the 'red object' as 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, as shown in Fig. 1. The supporting-fact is a structural representation of

- P. Wang is with Northwestern Polytechnical University, Shaanxi, 710072, China. E-mail: peng.wang@nwpu.edu.cn.
- Q. Wu, C. Shen, A. Dick, and A. van den Hengel are with The University of Adelaide, Adelaide, SA 5005, Australia. E-mail: {qi.wu01, chunhua.shen, anthony.dick, anton.vandenhengel}@adelaide.edu.au.

Manuscript received 10 Dec. 2016; revised 7 Aug. 2017; accepted 10 Aug. 2017. Date of publication 18 Sept. 2017; date of current version 12 Sept. 2018. (Corresponding author: Chunhua Shen.)

Recommended for acceptance by S. Lazebnik.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.

Digital Object Identifier no. 10.1109/TPAMI.2017.2754246



**Question:** What can the red object on the ground be used for ?

**Answer:** Firefighting

**Support Fact:** Fire hydrant can be used for fighting fires.

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.

knowledge that is stored in external KBs and indispensable for answering a given visual 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?’, the answer is ‘cat’. The required supporting-fact for answering this question is  $\langle \text{Cat}, \text{CapableOf}, \text{ClimbingTrees} \rangle$ , 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 on our FVQA dataset the performance of the state-of-the-art RNN (Recurrent Neural Network) based approaches [5], [6], [9]. We find that there are a number of limitations with these approaches. First, there is no explicit reasoning process in these methods. This means that it is impossible to tell whether the method is answering the question based on image information or merely 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 can only answer questions about concepts that have been observed in the training set, but there are millions of possible concepts and hundreds of millions relationships between them.

Our main contributions are as follows. A new VQA dataset (FVQA) with additional supporting-facts is introduced in Section 3, which requires and supports deeper reasoning. In response to the observed limitations of RNN-based approaches, we propose a method which is based on explicit reasoning about the visual concepts detected from images, in Section 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 bases). 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, which is then processed so as to form the final answer to the question. Our approach achieves the Top-1 accuracy of 56.91 percent, outperforming existing baseline VQA models.

## 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], [8], [9], [10] represent question-answer pairs for Microsoft COCO images [2], either generated automatically by NLP tools [9] or written by human workers [5], [7]. The Visual Genome dataset [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] were the first to study the VQA problem. They proposed a method that combines image segmentation and semantic parsing with a Bayesian approach to sample from nearest neighbors in the training set. This approach requires human defined relationships, 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 within VQA is to combine convolutional neural networks and recurrent neural networks 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 generate 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 (Long Short Term Memory) network. Inspired by Xu et al. [16] who encoded visual attention in image captioning, authors of [10], [17], [18], [19], [20] proposed to use spatial attention to help answer visual questions. Most of existing methods formulated the VQA as a classification problem and restrict that the answer only can be drawn from a fixed answer space. In other words, they cannot generate open-ended answers. Zhu et al. [21] investigated the video question answering problem using ‘fill-in-the-blank’ questions. However, either an LSTM or a GRU (Gated Recurrent Unit) is still applied in these methods to model the questions. Irrespective of the finer details, we refer to them as the RNN approaches.

### 2.3 Knowledge-Bases and VQA

Answering general questions posed by humans about images inevitably requires reference to information that is not contained in the image itself. To an extent this information may be provided by an existing training set such as ImageNet [3], or Microsoft COCO [2] as class labels or image captions. There are a number of forms of such auxiliary information, including, for instance, question/answer pairs which refer to objects that are not depicted (e.g., which reference people waiting for a train, when the train is not visible in the image) 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) [22], [23], [24], [25], [26], [27], [28] in contrast, offer an explicit, and typically larger-scale, representation of such external information. In structured KBs, knowledge is typically represented by a large number of triples of the form  $(arg1, rel, arg2)$ , which we refer to as *Facts* in this paper.  $arg1$  and  $arg2$  denote two *Concepts* in the KB, each describing a concrete or abstract entity with specific characteristics.  $rel$  represents a specific *Relationship* between them. A collection of such triples form a large interlinked graph. Such triples are often described according to a Resource Description Framework [29] (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 [30] to query the OpenLink Virtuoso [31] RDBMS.

Large-scale structured KBs are constructed either by manual annotation (e.g., DBpedia [22], Freebase [24] and Wikidata [28]), or by automatic extraction from unstructured/semi-structured data (e.g., YAGO [27], [32], OpenIE [23], [33], [34], NELL [25], NEIL [26], WebChild [35], ConceptNet [36]). The KB that 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 over structured KBs (referred to as KB-QA). KB-QA approaches can be divided into two categories: *semantic parsing* methods [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47] that try to translate questions into accurate logical expressions and then map to KB queries, and *information retrieval* methods [48], [49], [50], [51], [52], [53], [54] that coarsely retrieve a set of answer candidates and then perform ranking. Similar to information retrieval methods, the QA systems using memory networks [54], [55], [56], [57], [58] record supporting-fact candidates in a memory module that can be read and written to. The memory networks are trained to find the supporting-fact that leads to the answer by performing lookups in memory. The ranking of stored facts can be implemented by measuring the similarity between facts and the question using attention mechanisms.

VQA systems that exploit KBs are still relatively rare. Our proposed approach, which generates KB queries to obtain answers, is similar to semantic-parsing based methods. Alternatively, information-retrieval based methods, or

memory networks, can be also used for the problem posed in this paper. We leave this for future work. The work in [37] maps natural language queries to structured expressions in the lambda calculus. Similarly, our VQA dataset also provides the supporting fact for a visual question, which, however, is not necessarily an equivalent translation of the given question.

Zhu et al. [59] 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. [60] encoded the text mined from DBpedia to a vector with the Word2Vec model which they combined with visual features to generate answers using an LSTM model. However, their proposed method only extracts discrete pieces of text from the knowledge base, thus ignoring the power of its structural representation. Neither [59] nor [60] are capable of explicit reasoning, in contrast to the method we propose here.

The approach closest to that we propose here is that of Wang et al. [61], as it is capable of reasoning about an image based on 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 used only a single manually annotated knowledge source whereas the method we propose uses this plus two additional automatically-learned knowledge bases. This is critical because manually constructing such KBs does not scale well, and using automatically generated KBs thus enables the proposed method to answer more general questions.

Krishnamurthy and Kollar [62] proposed a semantic-parsing based method to find the grounding of a natural language query in a physical environment (such as image scene or geographical data). Similar to our approach, a structured, closed-domain KB is first constructed for a specific environment, then the natural language query is parsed into a logical form and its grounding is predicted. However, our approach differs from [62] in that it focuses on visual questions requiring support from both physical environment and external commonsense knowledge, which predicts over a much larger, multi-domain KB.

Similarly to our approach, Narasimhan et al. [63] also query data from external sources when information in the existing data is incomplete. The work in [63] performs queries over unstructured Web articles, while ours performs queries over structured relational KBs. Furthermore, the query templates used in [63] are much simpler: only the articles title needs to be replaced.

### 3 CREATING THE FVQA DATASET

Different from previous VQA datasets [5], [7], [8], [9], [10] that only ask annotators to provide *question-answer* pairs without any restrictions, the questions in our dataset are expected to be answered with support of some



TABLE 1  
The Relationships in Different Knowledge Bases Used for Generating Questions

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

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.

commonsense knowledge. This means that we cannot simply distribute only images to questioners like others [5], [10]. We need to provide a large number of supporting-facts (commonsense knowledge) which are linked to concepts that can be grounded in images (we refer to as *Visual Concepts*). We build our own on-line question collection system and allow users to choose images, visual concepts and candidate supporting-facts freely. Then users can ask questions based on their previous choices (all choices will be recorded). We provide users with a tutorial and restrict them to ask questions that only to be answered with both visual concept in the image and the provided external commonsense knowledge. In the following sections, we provide more details about images, visual concepts, knowledge bases and our question collection system and procedures. We also compare with other VQA datasets with data statistics.

### 3.1 Images and Visual Concepts

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

Three types of visual concepts are extracted in this work:

- *Object*: Instances of real-world object classes with certain semantic meaning (such as humans, cars, dogs) are detected by two Fast-RCNN [64] models that are trained respectively on Microsoft COCO 80-object (train split) and ImageNet 200-object datasets. The image attribute model [65] also predicts the existence of 92 objects without localisation information. Overall, there are 326 distinct object classes to be extracted.
- *Scene*: The image scene (such as office, bedroom, beach, forest) information is extracted by combining the VGG-16 model trained on MIT Places 205-class dataset [66] and the attribute classifier [65] including 25 scene classes. 221 distinct scene classes are obtained after the combination.

- *Action*: The attribute model [65] provides 24 classes of actions of humans or animals, such as walking, jumping, surfing, swimming.

A full list of extracted visual concepts can be found in the Appendix, which can be found on the Computer Society Digital Library at <http://doi.ieeecomputersociety.org/10.1109/TPAMI.2017.2754246>. In the next section, these visual concepts are further linked to a variety of external knowledge bases.

### 3.2 Knowledge Bases

The knowledge about each visual concept is extracted from a range of existing structured knowledge bases, including DBpedia [22], ConceptNet [36] and WebChild [35].

- *DBpedia*: The structured information stored in DBpedia is extracted from Wikipedia by crowd-sourcing. In this KB, concepts are linked to their categories and super-categories based on the SKOS Vocabulary.<sup>1</sup> 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.<sup>2</sup> We adopt 11 common relationships in ConceptNet to generate questions and answers.
- *WebChild*: The work in [35] considered a form of commonsense knowledge being overlooked by most of existing KBs, which involves comparative relations such as Faster, Bigger and Heavier. In [35], this form of information is extracted automatically from the Web.

The relationships 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.

1. <http://www.w3.org/2004/02/skos/>

2. <http://web.media.mit.edu/~push/Kurzweil.html>

TABLE 2  
Major Datasets for VQA and Their Main Characteristics

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 [67]	15,623	33,379	1	6.2	1.0	-	-
KB-VQA [61]	700	2,402	23	6.8	2.0	1	-
Ours (FVQA)	2,190	5,826	32	9.5	1.2	3	✓

### 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. By doing so, the selected fact becomes the supporting-fact of the asked question. 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 grounded in the questioned image (underlined in Table 1) or the concept found on the KB side.

### 3.4 Data Statistics

In total, 5826 questions (corresponding to 4216 unique facts) are collected collaboratively by 38 volunteers. The average annotation time for each question is around 1 minute. In order to report 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<sup>3</sup> for training and test respectively.

Table 2 shows summary statistics of the dataset, such as the number of question categories, average question/answer length etc. We have 32 question types in total (see Section 4.2 for more details). Compared to VQA-real [5] and Visual Genome [11], our FVQA dataset provides longer questions, with average length 9.5 words.

3. As each image contains a 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 3 show the differences.

These questions can be categorized according to the following criteria:

- *Key Visual Concept*: The visual concept that appears in the supporting-fact of the given question, which is selected in Step-1 of the question collection process. The type of Key Visual Concept is denoted by  $T_{KVC}$ , whose value can be Object, Scene or Action (see Section 3.1).
- *Key Relationship*: The relationship in the supporting-fact, whose type is denoted as  $T_{REL}$ . As shown in Table 1, there are 13 different values that can be assigned to  $T_{REL}$ .<sup>4</sup>
- *KB Source* (refer to as  $T_{KS}$ ): The external KB where the supporting-fact is stored. In this work, the value of  $T_{KS}$  can be DBpedia, ConceptNet or Webchild.
- *Answer Source* (refer to as  $T_{AS}$ ): As shown in Step3 of Section 3.3,  $T_{AS} = \text{Image}$  if answer is the key visual concept, and  $T_{AS} = \text{KB}$  otherwise.

Table 3 shows the number of training/test questions falling into different categories of  $T_{KVC}$ ,  $T_{KS}$ ,  $T_{AS}$ . We can see that most of the questions are related to the objects in images and most of the answers are from Image. As for knowledge bases, 80 percent of the collected questions rely on the supporting-facts from ConceptNet. Answering 14 and 6 percent questions depends on the knowledge from DBpedia and Webchild respectively.

The distributions of collected questions and facts over the 13 types of key relationships ( $T_{REL}$ ) are shown in Fig. 2. We can see that the questions and facts are evenly distributed over the relationships of Category, UsedFor, IsA, RelatedTo, CapableOf, AtLocation, HasProperty and HasA, although these relationships differ significantly in the total numbers of extracted facts (see Table 1).

### 3.5 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’

4. For simplicity, we consider all the comparative relationships in WebChild as one type.

TABLE 3  
The classification of Questions According to Key Visual Concept, KB Source and Answer Source

Criterion	Category	Train	Test	Total
Key Visual Concept ( $T_{KVC}$ )	Object	$2661.2 \pm 66.0$	$2621.8 \pm 66.0$	5283
	Scene	$251.2 \pm 26.2$	$260.8 \pm 26.2$	512
	Action	$14.8 \pm 2.7$	$16.2 \pm 2.7$	31
Answer Source ( $T_{AS}$ )	Image	$2437.4 \pm 63.4$	$2393.6 \pm 63.4$	4831
	KB	$489.8 \pm 27.9$	$505.2 \pm 27.9$	995
KB Source ( $T_{KS}$ )	DBpedia	$403.2 \pm 12.7$	$413.8 \pm 12.7$	817
	ConceptNet	$2348.8 \pm 71.6$	$2303.2 \pm 71.6$	4652
	Webchild	$175.2 \pm 9.5$	$181.8 \pm 9.5$	357
Total		$2927.2 \pm 69.5$	$2898.8 \pm 69.5$	5826

The number of training/test questions in each category is also demonstrated. The error bars are produced by 5 different splits.

- 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 97.6 percent of collected questions are voted as ‘require common-sense knowledge’ by at least 2 subjects. For these knowledge-requiring questions, more than 99 percent of the supporting-facts provided in our dataset are considered valuable for answering them.

### 3.6 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 percent questions in the VQA dataset require common-sense reasoning to answer (18.14 percent: 6 or more subjects). However, such external knowledge is not provided in all the existing VQA datasets. To the best knowledge of us, 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 [61] that only applied a single manually annotated knowledge source, we use two additional automatically extracted knowledge bases, which enable us to answer more general questions.

In a similar manner as ours, the Facebook bAbI [68] 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 percent questions can be answered using ‘Yes’ or ‘No’. Although ‘Yes’/‘No’ questions can also require a challenging reasoning process (see [19], [69]), they may not be a good measure of models’ reasoning abilities. A random guess process can still achieve an approximately

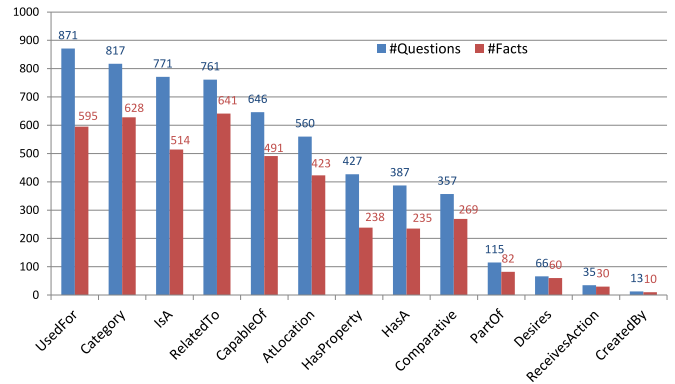


Fig. 2. The distributions of the collected 5826 questions and the corresponding 4216 facts over different relationships. The top five relationships are UsedFor, Category, IsA, RelatedTo and CapableOf. There are fewer supporting-facts than questions because one ‘fact’ can correspond to multiple ‘questions’.

50 percent accuracy on a balanced ‘Yes’/‘No’ QA dataset, but it does not perform any reasoning on the 50 percent correctly answered questions. To avoid misleading results, we simply exclude ‘Yes’/‘No’ questions.

## 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 RNN approaches [7], [10], [15], [17], [18], [19], [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 Constructing a Unified KB

The primary step of our approach is to construct a unified KB, which links the visual concepts extracted from each image to the corresponding concepts in multiple KBs. A visual concept  $X$  of type  $T$  ( $T = \text{Object}, \text{Scene or Action}$ ) extracted from the image with id  $I$  is stored in two triples  $(X, \text{Grounded}, I)$  and  $(X, \text{VC-Type}, T)$ . Concepts in multiple KBs with the same meaning as  $X$  are directly linked to  $X$ . By doing so, the rich external knowledge about  $X$  is linked to the images that  $X$  is grounded in.

### 4.2 Question-Query Mapping

In our approach, three characteristics of visual questions are first predicted by trained LSTM models, i.e., key visual concepts ( $T_{KVC}$ ), key relationships ( $T_{REL}$ ) and answer sources ( $T_{AS}$ ). In the training data, these characteristics of a question can be obtained through the annotated supporting-fact and the given answer, and we have collected 32 different combinations of  $\langle T_{KVC}, T_{REL}, T_{AS} \rangle$  in the proposed dataset (see Appendix, available in the online supplemental material). Since both question and query are sequences, the question-query mapping problem can be treated as a sequence-to-sequence problem [70], which can be solved by Recurrent

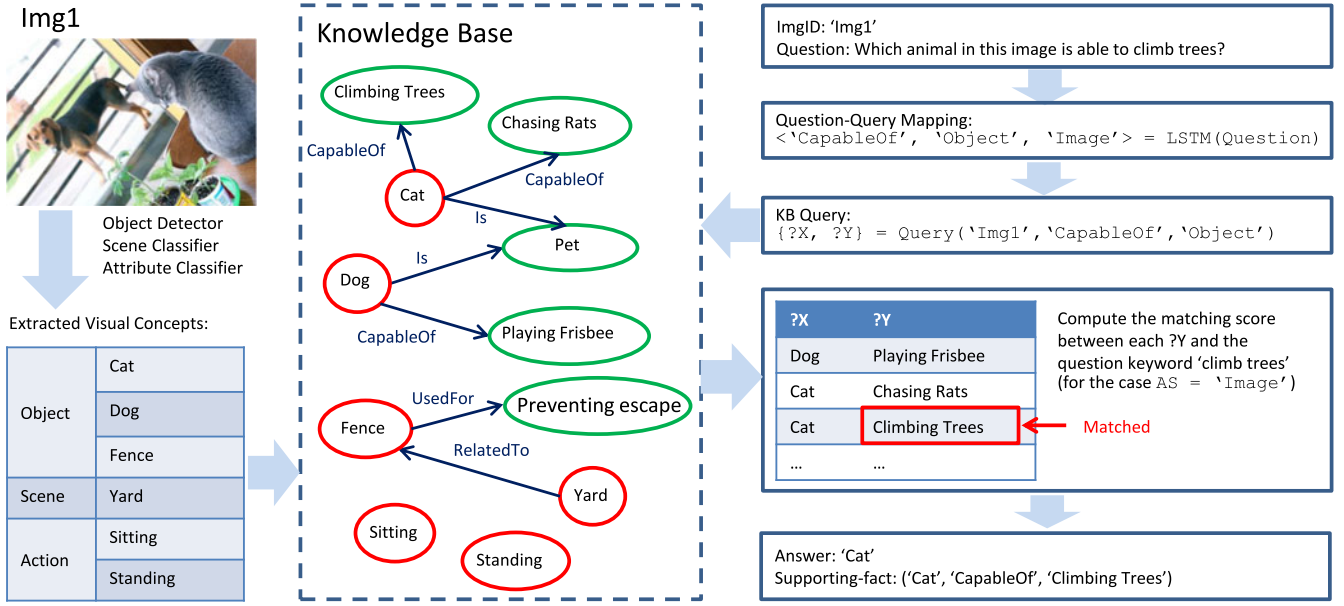


Fig. 3. An example of the reasoning process of the proposed VQA approach. The visual concepts (objects, scene, attributes) of the input image are extracted using trained models, which are further linked to the corresponding semantic entities in the knowledge base. The input question is first mapped to one of the query types using the LSTM model shown in Section 4.2. The types of key relationships, key visual concept and answer source can be determined accordingly. A specific query (see Section 4.3) 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.

Neural Network (RNN) [71]. While in this work, we consider each distinct combination of the three characteristics as a query type and learn a 32-class classifier using LSTM models [72], 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 [65]. 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  $Q = (Q_0, \dots, Q_L)$ , 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  $h_{initial} = \vec{0}$ , 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

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

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.  $\theta$  represents model parameters,  $\lambda_{\theta} \cdot \|\theta\|_2^2$  is a regularization term.

During the test phase, the question's words sequence is fed into the trained LSTM to produce the probability distribution  $p$  over all query types.

In Fig. 3, the query type of the input question 'Which animal in this image is able to climb trees?' is classified by the LSTM classifier as  $(T_{REL}, T_{KVC}, T_{AS}) = (\text{CapableOf}, \text{Object}, \text{Image})$ .

### 4.3 Querying the KB

Retrieving the correct supporting-fact is the key to answering a visual question in our proposed dataset. To this end, KB queries are constructed to search for a number of candidate supporting-facts. To be specific, given a question's key relationship ( $T_{REL}$ ) and key visual concept ( $T_{KVC}$ ) as inputs, a KB-query  $\{?X, ?Y\} = \text{Query}(I, T_{REL}, T_{KVC})$  is constructed as follows:

$$\text{Find } ?X, ?Y, \text{ subject to } \{(?X, \text{Grounded}, I) \text{ and } (?X, \text{VC-Type}, T_{KVC}) \text{ and } (?X, T_{REL}, ?Y)\},$$

where  $I$  denotes the id of questioned image; **Grounded** is a relationship representing that a specific visual concept is grounded in a specific image; **VC-Type** is another relationship used to describe the type of a visual concept;  $?X$  and  $?Y$  are two variables to be searched and returned. There are three triple templates to be matched:  $(?X, \text{Grounded}, I)$  searches all visual concepts  $?X$  that are grounded in image  $I$ ;  $(?X, \text{VC-Type}, T_{KVC})$  restricts that the type of  $?X$  should be  $T_{KVC}$ ;  $(?X, T_{REL}, ?Y)$  restricts that  $?X$  should be linked to at least one concept  $?Y$  via relationship  $T_{REL}$ . The above query will search for sets of triples that match these three triple templates, and return a list of  $\{?X, ?Y\}$  pairs residing in the associated slots of matched sets. Correspondingly, we also obtain a list of candidate supporting-facts  $(?X, T_{REL}, ?Y)$ . Note that the query is performed over the entire KB that stores all facts about visual concepts, rather than over the 4216 supporting-facts that have been used in the collected questions. As aforementioned, all comparative relationships are considered as one relation type ( $T_{REL}$ ). In this case, the returned query results will be further filtered based on the comparative words shown in questions.

In the reasoning process shown in Fig. 3, for all objects in Image  $\text{Img1}$ , the query  $\text{Query}(\text{Img1}, \text{CapableOf}, \text{Object})$



searches for the things they are capable of doing. The objects ‘Dog’ and ‘Cat’ in this image have links to concepts in the constructed KB via relationship `CapableOf`. Accordingly, a list of  $\{?X, ?Y\}$  pairs are returned by this query including  $\{\text{Dog}, \text{Playing Frisbee}\}$ ,  $\{\text{Cat}, \text{Chasing Rats}\}$  and  $\{\text{Cat}, \text{Climbing Trees}\}$ . More examples of KB-queries and their outputs can be found in Table 11.

#### 4.4 Answering

As shown in the Step-3 of Section 3.3, the answer to a visual question in our collected dataset can be either the visual concept in the supporting-fact (i.e.,  $?X$ ), or the other concept on the KB side (i.e.,  $?Y$ ). In the test phase, whether the answer is  $?X$  or  $?Y$  is determined by the value of answer source  $T_{AS}$  predicted by the LSTM classifier in Section 4.2: the answer is  $?X$  if  $T_{AS} = \text{Image}$ , and  $?Y$  if  $T_{AS} = \text{KB}$ . The last issue before arriving at the answer is how to select the most relevant supporting-fact from candidates. In this work, the supporting-fact is selected by matching between the given question and concepts  $?Y$  (or  $?X$ ) in candidate facts if answer is  $?X$  (or  $?Y$ ), as follows:

1)  $T_{AS} = \text{Image}$ : A list of high frequency words (such as ‘what’, ‘which’, ‘a’, ‘the’) is established by counting in the training examples. The keywords of a question are then extracted by removing these high frequency words. A matching score  $s$  is computed between the question keywords and the concept  $?Y$  in each candidate supporting-fact  $(?X, T_{REL}, ?Y)$ , in order to measure the relevance between the fact and question. In this work, the matching score is the simple Jaccard similarity between the normalized word sets of  $?Y$  (refer to as  $\mathcal{W}_Y$ ) and question keywords (refer to as  $\mathcal{W}_Q$ ):

$$s(\mathcal{W}_Y, \mathcal{W}_Q) = \frac{|\mathcal{W}_Y \cap \mathcal{W}_Q|}{|\mathcal{W}_Y \cup \mathcal{W}_Q|}. \quad (2)$$

The candidate fact corresponding to the highest-scored  $?Y$  is selected as the supporting-fact, and the associated visual concept  $?X$  is considered as the answer. Note that the matching score can be computed using word embedding models such as Word2Vec [73], which we leave for future work. In the example of Fig. 3, all  $?Y$ s (i.e. `Playing Frisbee`, `Chasing Rats` and `Climbing Trees`) are matched to the question keywords `climb trees` and the fact (`Cat`, `CapableOf`, `Climbing Trees`) has achieved the highest score, so the answer is `Cat`.

2)  $T_{AS} = \text{KB}$ : In this case, we need to find out which visual concept ( $?X$ ) is the most related to the input question. If  $T_{KVC} = \text{Scene}$  or  $\text{Action}$ , the visual concept  $?X$  with the highest probability is selected and the corresponding concept  $?Y$  is considered as the answer. The probabilities of scene or action concepts are obtained from the softmax layer of the visual models shown in Section 3.1. If  $T_{KVC} = \text{Object}$ , the visual concept  $?X$  is selected based on the question keywords describing location (such as `top`, `bottom`, `left`, `right` or `center`) or size (such as `small` and `large`). Note that one visual concept  $?X$  may correspond to multiple concepts  $?Y$ , i.e., multiple answers. These answers are ordered according to their frequency in the training data: the most frequent answer appears first.

TABLE 4  
Question-Query Mapping Accuracy for Different  
KB Sources on the FVQA Testing Splits

KB Source	Question-Query Mapping Acc. $\pm$ Std (%)	
	Top-1	Top-3
DBpedia	61.73 $\pm$ 1.83	85.56 $\pm$ 2.14
ConceptNet	64.10 $\pm$ 1.10	80.85 $\pm$ 0.87
WebChild	83.15 $\pm$ 3.03	95.24 $\pm$ 1.23
Overall	64.94 $\pm$ 1.08	82.42 $\pm$ 0.56

*Top-1 and Top-3 results are reported.*

## 5 EXPERIMENTS

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

### 5.1 Question-Query Mapping

Table 4 reports the accuracy of our proposed Question-Query mapping (QQmapping) approach in Section 4.2. The mapping model is trained on the FVQA training splits and tested on the respective testing splits. **To train the model, we use the Stochastic Gradient Descent (SGD) with 100 Question-Query pairs as a mini-batch.** 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 for different KB Sources of supporting facts. Questions asked based on facts from WebChild are much easier to be mapped than questions based on other two KBs. This is mainly because 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 similar formats. For example, many questions are formulated as ‘Which object in the image is more *<a comparative adj>* ?’. Our Top-3 overall accuracy achieves  $82.42 \pm 0.56$ .

### 5.2 FVQA Experiments

Our FVQA tasks are formulated as an open-ended answer generation problem, which requires models to predict open-ended outputs that may not appear in training data. To measure the accuracy, we simply calculate the proportion of correctly answered test questions. **A predicted answer is determined as correct if and only if its string matches the corresponding ground-truth answer (all the answers have been normalized by the python INFLECT package to eliminate the singular-plurals differences etc.).** We report the top-1, 3 and 10 accuracy of the evaluated methods. Note that our proposed models may produce less than 10 candidate answers in some cases where the number of returned supporting facts is less than 10, which makes the top-10 performance of our models not as strong as on the top-1 and top-3 settings.



TABLE 5  
Overall Accuracy on Our FVQA Testing Splits for Different  
Methods Based on String Matching

Method	Overall Acc. $\pm$ Std (%)		
	Top-1	Top-3	Top-10
SVM-Question	10.37 $\pm$ 0.80	20.72 $\pm$ 0.58	34.63 $\pm$ 1.19
SVM-Image	18.41 $\pm$ 1.07	32.42 $\pm$ 1.06	47.53 $\pm$ 1.02
SVM-Question+Image	18.89 $\pm$ 0.91	32.78 $\pm$ 0.90	48.13 $\pm$ 0.73
LSTM-Question	10.45 $\pm$ 0.57	19.02 $\pm$ 0.74	31.64 $\pm$ 0.93
LSTM-Image	20.55 $\pm$ 0.81	36.01 $\pm$ 1.45	55.74 $\pm$ 2.28
LSTM-Question+Image	22.97 $\pm$ 0.64	36.76 $\pm$ 1.22	54.19 $\pm$ 2.45
LSTM-Question+Image+Pre-VQA	24.98 $\pm$ 0.60	40.40 $\pm$ 1.05	57.27 $\pm$ 1.29
Hie-Question+Image	33.70 $\pm$ 1.18	50.00 $\pm$ 0.78	64.08 $\pm$ 0.57
Hie-Question+Image+Pre-VQA	43.14 $\pm$ 0.61	59.44 $\pm$ 0.34	<b>72.20 <math>\pm</math> 0.39</b>
<b>Ours, gt-QQmapping<sup>‡</sup></b>	<b>63.63 <math>\pm</math> 0.73</b>	<b>71.30 <math>\pm</math> 0.78</b>	<b>72.55 <math>\pm</math> 0.79</b>
Ours, top-1-QQmapping	52.56 $\pm$ 1.03	59.72 $\pm$ 0.82	60.58 $\pm$ 0.86
Ours, top-3-QQmapping	<b>56.91 <math>\pm</math> 0.99</b>	<b>64.65 <math>\pm</math> 1.05</b>	65.54 $\pm$ 1.06
Ensemble	58.76 $\pm$ 0.92	-	-
Human	77.99 $\pm$ 0.75	-	-

Best single-model results are shown in bold font. <sup>‡</sup> indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

Additionally, we also report the results based on WuPalmer Similarity (WUPS) [74]. WUPS calculates the similarity between two words based on their common subsequence in the taxonomy tree. If the similarity is greater than a prescribed threshold then the predicted answer is considered as correct. In this paper, we report WUPS results at thresholds 0.9 and 0.0. All the reported results are averaged on the 5 test splits (standard deviation is also provided). It should be noted that WUPS@0.0 is calculated with a very small threshold, which offers a very loose similarity measure.

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 (or questions), we zero out the image (or question) 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 features and question features. For image features, we use the fc7 layer (4096-d) of the VggNet-16 model [4]. Questions are represented by 300-d averaged word embeddings from a pre-trained word2vec model [75]. We take the top-500 most frequent answers (93.68 percent of the training set answers) as the class labels. At test time, we select the top-1, 3 and 10 scored answer candidates. We use the LibSVM [76] and the parameter  $C$  is set to the default value 1. We found that tuning the value of  $C$  does not affect the performance significantly.

**LSTM.** We compare our system with an approach [9] based on LSTM. The LSTM outputs at the last timestep are fed into a softmax layer to predict answers over a fixed space (top-500 most frequent answers). This is 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 splits. 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, 3 and 10 scored answer candidates at test time. We additionally implement a model that is first pre-trained on the VQA dataset, and fine-tuned on the FVQA dataset, which is denoted as ‘LSTM-Question+Image+Pre-VQA’.

**State-of-the-Art.** We evaluate the model in [77] on FVQA with and without pre-training on the VQA dataset. The models are denoted as ‘Hie-Question+Image’ and ‘Hie-Question+Image+Pre-VQA’. The hierarchical co-attention model of [77] provides the state-of-the-art performance on the popular VQA dataset.

**Ensemble Model.** We combine two LSTM-based models (which are ‘Hie-Question+Image+Pre-VQA’ and ‘LSTM-Question+Image+Pre-VQA’) with our proposed model (‘Ours, top3-QQmapping’) by selecting the answer with the maximum score over the three models. The scores of LSTM models are taken from the Softmax layer, and the score of our model is 0 when it is not able to find any answer and 1 otherwise).

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

**Ours.** Our KB-query based model is introduced in Section 4. To verify the effectiveness of our method, we implement three variants. ‘gt-QQmapping’ uses the ground truth question-query mapping, while ‘top-1-QQmapping’ and ‘top-3-QQmapping’ use the top-1 and top-3 predicted question-query mapping (see Section 4.2), respectively.

Table 5 shows the overall accuracy of all the baseline methods and our proposed models. In the case of Top-1 accuracy, our ‘top-3-QQmapping’ is the best-performing single model, which doubles the accuracy of the baseline ‘LSTM-Question+Image’. Our proposed model also outperforms ‘LSTM-Question+Image+Pre-VQA’ and ‘Hie-Question+Image+Pre-VQA’, although the latter use additional training data. The ‘Ensemble’ model achieves the best accuracy of 58.76 percent, showing that the conventional LSTM-based model and our knowledge based model are complementary. The ‘top-3-QQmapping model’ outperforms ‘top-1-QQmapping’ because it produces better Question-Query mapping results, as shown in the Table 4. However, it is still not as good as ‘gt-QQmapping’ due to Question-Query mapping errors. 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’ plays a more important role than ‘image’ in the VQA [5] or COCO-QA datasets [9], our ‘{SVM,LSTM}-Q’ performs worse than ‘{SVM,LSTM}-I’, meaning that FVQA questions rely more heavily on the image content than the existing VQA datasets. Actually, if ‘{SVM,LSTM}-Q’ achieves too high performance, the corresponding questions may be not *Visual Questions* and they may be actually *Textual Questions*.

TABLE 6  
WUPS@0.9 on our FVQA Testing Splits for Different Methods

Method	WUPS@0.9. $\pm$ Std (%)		
	Top-1	Top-3	Top-10
SVM-Question	17.06 $\pm$ 1.09	29.43 $\pm$ 0.62	44.73 $\pm$ 1.07
SVM-Image	24.73 $\pm$ 1.29	40.95 $\pm$ 1.34	56.33 $\pm$ 0.63
SVM-Question+Image	25.30 $\pm$ 1.09	41.37 $\pm$ 1.19	56.78 $\pm$ 0.64
LSTM-Question	15.82 $\pm$ 0.57	26.45 $\pm$ 0.61	40.99 $\pm$ 1.02
LSTM-Image	26.78 $\pm$ 1.02	44.00 $\pm$ 1.61	62.86 $\pm$ 2.23
LSTM-Question+Image	29.08 $\pm$ 0.91	44.36 $\pm$ 1.29	61.71 $\pm$ 2.82
LSTM-Question+Image+Pre-VQA	31.96 $\pm$ 0.65	48.55 $\pm$ 0.97	64.73 $\pm$ 1.38
Hie-Question+Image	39.75 $\pm$ 0.78	56.48 $\pm$ 0.68	69.78 $\pm$ 0.51
Hie-Question+Image+Pre-VQA	48.93 $\pm$ 0.71	64.75 $\pm$ 0.44	<b>76.73 <math>\pm</math> 0.38</b>
Ours, gt-QQmapping <sup>‡</sup>	65.51 $\pm$ 0.82	72.37 $\pm$ 0.89	73.55 $\pm$ 0.87
Ours, top-1-QQmapping	54.79 $\pm$ 0.91	61.41 $\pm$ 0.71	62.22 $\pm$ 0.70
Ours, top-3-QQmapping	<b>59.67 <math>\pm</math> 0.90</b>	<b>66.89 <math>\pm</math> 1.01</b>	67.77 $\pm$ 1.04
Human	82.47 $\pm$ 0.71	-	-

Best single-model results are shown in bold font. <sup>‡</sup> indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

The LSTM models that are pretrained on the VQA dataset performs slightly better than training from scratch, but not as well as our models. Using the Top-3 measure, our ‘top-3-QQmapping’ model also performs best. We produce slightly lower Top-10 performance, because our proposed methods may produce less than 10 answers in some cases.

Table 11 shows some example results generated by our final model. Tables 6 and 7 report the WUPS@0.9 and WUPS@0.0 accuracy for different methods.

Table 8 reports the accuracy for different Knowledge Base sources. Our ‘top-3-QQmapping’ model performs better than other baselines for DBpedia and ConceptNet KBs. For Webchild, ‘top-3-QQmapping’ is only worse than the state-of-the-art ‘Hie-Question+Image+Pre-VQA’ model, which, however, uses extra data.

Table 9 illustrates the performance on questions that focus on different types of visual concepts, which are object, scene and action. The performance on object-related

TABLE 7  
WUPS@0.0 on our FVQA Testing Splits for Different Methods

Method	WUPS@0.0. $\pm$ Std (%)		
	Top-1	Top-3	Top-10
SVM-Question	56.88 $\pm$ 2.57	68.45 $\pm$ 0.67	76.93 $\pm$ 0.75
SVM-Image	59.64 $\pm$ 0.72	73.30 $\pm$ 0.96	81.77 $\pm$ 0.33
SVM-Question+Image	59.97 $\pm$ 0.63	73.39 $\pm$ 0.88	81.93 $\pm$ 0.39
LSTM-Question	51.45 $\pm$ 1.41	65.65 $\pm$ 0.66	75.76 $\pm$ 0.62
LSTM-Image	59.53 $\pm$ 1.52	73.83 $\pm$ 0.50	83.53 $\pm$ 0.89
LSTM-Question+Image	61.86 $\pm$ 0.81	74.45 $\pm$ 0.59	83.54 $\pm$ 0.94
LSTM-Question+Image+Pre-VQA	63.42 $\pm$ 0.35	76.63 $\pm$ 0.50	84.94 $\pm$ 0.57
Hie-Question+Image	66.11 $\pm$ 0.62	78.90 $\pm$ 0.50	86.32 $\pm$ 0.55
Hie-Question+Image+Pre-VQA	71.51 $\pm$ 0.16	<b>82.71 <math>\pm</math> 0.39</b>	<b>89.01 <math>\pm</math> 0.50</b>
Ours, gt-QQmapping <sup>‡</sup>	73.98 $\pm$ 0.77	78.67 $\pm$ 0.78	79.98 $\pm$ 0.79
Ours, top-1-QQmapping	64.96 $\pm$ 0.70	69.57 $\pm$ 0.65	70.64 $\pm$ 0.59
Ours, top-3-QQmapping	<b>72.34 <math>\pm</math> 0.65</b>	77.52 $\pm$ 0.70	78.69 $\pm$ 0.63
Human	87.30 $\pm$ 0.54	-	-

Best single-model results are shown in bold font. <sup>‡</sup> indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

questions is significantly higher than the other two types, especially when image features are given. This is not surprising since 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 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 presents the accuracy for different methods according to different Answer Sources. If the answer is a visual concept in the image, we categorize the answer source into ‘Image’. Otherwise, it is categorized into ‘KB’. From the table, we can see that the accuracy is much higher

TABLE 8  
Accuracies on the Questions That Asked Based on Different Knowledge Base Sources

Method	KB-Source, Acc. $\pm$ Std (%)								
	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 $\pm$ 0.89	9.94 $\pm$ 1.44	20.91 $\pm$ 2.08	11.03 $\pm$ 0.76	21.84 $\pm$ 0.83	35.89 $\pm$ 1.31	16.38 $\pm$ 4.38	31.40 $\pm$ 0.93	50.17 $\pm$ 2.98
SVM-Image	7.11 $\pm$ 0.73	19.40 $\pm$ 2.50	39.75 $\pm$ 2.23	20.23 $\pm$ 1.14	33.91 $\pm$ 1.57	48.15 $\pm$ 1.18	20.95 $\pm$ 2.17	43.32 $\pm$ 1.23	57.31 $\pm$ 2.71
SVM-Question+Image	7.35 $\pm$ 0.73	20.43 $\pm$ 2.32	40.38 $\pm$ 2.46	20.76 $\pm$ 0.95	34.18 $\pm$ 1.35	48.64 $\pm$ 1.08	21.40 $\pm$ 2.01	43.33 $\pm$ 1.45	59.58 $\pm$ 2.92
LSTM-Question	5.08 $\pm$ 0.54	11.17 $\pm$ 0.21	23.68 $\pm$ 2.42	10.96 $\pm$ 0.60	19.70 $\pm$ 0.87	32.02 $\pm$ 0.96	16.37 $\pm$ 2.87	28.31 $\pm$ 2.11	44.91 $\pm$ 1.32
LSTM-Image	14.62 $\pm$ 2.38	29.68 $\pm$ 4.35	49.74 $\pm$ 2.56	20.96 $\pm$ 0.96	36.54 $\pm$ 1.21	56.32 $\pm$ 2.31	28.82 $\pm$ 2.79	43.64 $\pm$ 5.49	62.04 $\pm$ 3.87
LSTM-Question+Image	15.77 $\pm$ 2.07	28.30 $\pm$ 3.56	49.45 $\pm$ 7.25	23.57 $\pm$ 0.52	37.36 $\pm$ 1.43	54.30 $\pm$ 2.99	31.74 $\pm$ 3.69	48.18 $\pm$ 5.80	63.59 $\pm$ 5.60
LSTM-Question+Image+Pre-VQA	15.38 $\pm$ 1.11	32.64 $\pm$ 2.58	51.80 $\pm$ 3.12	25.97 $\pm$ 0.70	41.02 $\pm$ 1.31	57.35 $\pm$ 1.40	34.42 $\pm$ 2.68	50.27 $\pm$ 2.22	68.56 $\pm$ 3.19
Hie-Question+Image	27.65 $\pm$ 1.34	45.71 $\pm$ 1.16	62.44 $\pm$ 1.63	34.00 $\pm$ 1.33	50.01 $\pm$ 0.53	64.05 $\pm$ 0.46	43.74 $\pm$ 4.48	59.46 $\pm$ 4.60	67.95 $\pm$ 4.43
Hie-Question+Image+Pre-VQA	38.39 $\pm$ 1.33	56.07 $\pm$ 1.05	<b>71.60 <math>\pm</math> 1.82</b>	43.23 $\pm$ 0.76	59.47 $\pm$ 0.43	<b>72.21 <math>\pm</math> 0.77</b>	<b>52.85 <math>\pm</math> 5.10</b>	<b>66.49 <math>\pm</math> 3.77</b>	<b>73.40 <math>\pm</math> 2.35</b>
Ours, gt-QQmapping <sup>‡</sup>	65.96 $\pm$ 2.06	78.80 $\pm$ 1.08	79.43 $\pm$ 1.11	64.62 $\pm$ 0.82	71.75 $\pm$ 0.94	73.17 $\pm$ 0.94	45.55 $\pm$ 1.14	48.29 $\pm$ 1.08	48.86 $\pm$ 1.29
Ours, top-1-QQmapping	51.25 $\pm$ 2.21	63.07 $\pm$ 1.23	63.13 $\pm$ 1.21	53.50 $\pm$ 1.14	60.16 $\pm$ 0.87	61.20 $\pm$ 0.96	43.54 $\pm$ 2.14	46.58 $\pm$ 1.80	47.03 $\pm$ 2.08
Ours, top-3-QQmapping	<b>56.67 <math>\pm</math> 1.68</b>	<b>69.31 <math>\pm</math> 1.10</b>	69.36 $\pm$ 1.03	<b>57.60 <math>\pm</math> 1.29</b>	<b>64.70 <math>\pm</math> 1.24</b>	65.77 $\pm$ 1.28	48.74 $\pm$ 2.47	53.45 $\pm$ 1.99	53.90 $\pm$ 2.26
Ensemble	57.08 $\pm$ 2.02	-	-	58.98 $\pm$ 1.04	-	-	59.77 $\pm$ 5.52	-	-
Human	74.41 $\pm$ 1.13	-	-	78.32 $\pm$ 0.76	-	-	81.95 $\pm$ 1.83	-	-

Best single-model results are shown in bold font. <sup>‡</sup> indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

TABLE 9  
Accuracies on Questions That Focus on Three Different Visual Concepts

Method	Visual Concept, Acc. $\pm$ Std (%)								
	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 $\pm$ 0.84	22.72 $\pm$ 0.53	37.89 $\pm$ 1.06	0.68 $\pm$ 0.24	1.98 $\pm$ 0.83	3.97 $\pm$ 0.74	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
SVM-Image	20.08 $\pm$ 1.25	35.33 $\pm$ 1.13	51.60 $\pm$ 1.11	2.88 $\pm$ 0.46	5.22 $\pm$ 0.85	9.60 $\pm$ 0.36	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
SVM-Question+Image	20.59 $\pm$ 1.04	35.73 $\pm$ 1.06	52.24 $\pm$ 0.79	3.02 $\pm$ 0.42	5.22 $\pm$ 0.84	9.83 $\pm$ 0.19	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
LSTM-Question	11.39 $\pm$ 0.52	20.70 $\pm$ 0.63	34.34 $\pm$ 0.92	1.58 $\pm$ 1.07	3.05 $\pm$ 0.67	6.03 $\pm$ 1.09	0.00 $\pm$ 0.00	3.53 $\pm$ 4.71	4.71 $\pm$ 5.76
LSTM-Image	22.34 $\pm$ 0.98	39.09 $\pm$ 1.49	60.13 $\pm$ 2.18	3.80 $\pm$ 0.92	7.07 $\pm$ 1.49	14.94 $\pm$ 1.43	0.00 $\pm$ 0.00	1.05 $\pm$ 2.11	1.05 $\pm$ 2.11
LSTM-Question+Image	24.92 $\pm$ 0.88	39.71 $\pm$ 1.48	58.32 $\pm$ 3.82	4.56 $\pm$ 0.72	9.21 $\pm$ 2.00	15.56 $\pm$ 1.62	5.99 $\pm$ 6.64	5.99 $\pm$ 6.64	9.52 $\pm$ 6.15
LSTM-Question+Image+Pre-VQA	26.97 $\pm$ 0.53	43.40 $\pm$ 0.77	61.30 $\pm$ 0.95	6.46 $\pm$ 1.01	12.27 $\pm$ 1.17	19.98 $\pm$ 2.10	2.35 $\pm$ 2.88	8.35 $\pm$ 7.20	8.34 $\pm$ 7.20
Hie-Question+Image	36.35 $\pm$ 1.25	53.83 $\pm$ 1.00	68.51 $\pm$ 0.61	9.09 $\pm$ 0.46	14.36 $\pm$ 1.30	22.54 $\pm$ 1.27	1.18 $\pm$ 2.35	6.52 $\pm$ 5.42	17.38 $\pm$ 3.81
Hie-Question+Image+Pre-VQA	46.38 $\pm$ 0.81	63.65 $\pm$ 0.55	<b>76.81 <math>\pm</math> 0.48</b>	12.72 $\pm$ 0.20	<b>19.64 <math>\pm</math> 1.53</b>	<b>28.92 <math>\pm</math> 2.74</b>	8.11 $\pm$ 8.02	19.08 $\pm$ 5.07	21.43 $\pm$ 3.67
Ours, gt-QQmapping <sup>‡</sup>	68.70 $\pm$ 0.77	76.50 $\pm$ 0.69	77.11 $\pm$ 0.71	14.81 $\pm$ 0.83	20.93 $\pm$ 1.54	27.75 $\pm$ 2.03	28.78 $\pm$ 5.72	39.76 $\pm$ 6.61	53.32 $\pm$ 10.47
Ours, top-1-QQmapping	56.75 $\pm$ 1.48	64.11 $\pm$ 1.33	64.47 $\pm$ 1.35	12.81 $\pm$ 1.29	18.36 $\pm$ 2.68	24.19 $\pm$ 3.52	16.58 $\pm$ 10.14	19.57 $\pm$ 12.38	22.57 $\pm$ 14.32
Ours, top-3-QQmapping	<b>61.53 <math>\pm</math> 1.31</b>	<b>69.51 <math>\pm</math> 1.35</b>	69.90 $\pm$ 1.20	<b>12.89 <math>\pm</math> 1.29</b>	18.51 $\pm$ 2.72	24.34 $\pm$ 3.57	<b>19.75 <math>\pm</math> 6.16</b>	<b>22.73 <math>\pm</math> 8.34</b>	<b>25.72 <math>\pm</math> 10.12</b>
Ensemble	63.38 $\pm$ 1.19	-	-	15.13 $\pm$ 1.24	-	-	14.39 $\pm$ 3.41	-	-
Human	80.58 $\pm$ 0.42	-	-	52.70 $\pm$ 2.06	-	-	68.23 $\pm$ 4.90	-	-

Best single-model results are shown in bold font. ‡ indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

when the answer is from the ‘Image’ side, nearly 5 times as much as ‘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 (‘top-3-QQmapping’) achieves the correct answer and Table 12 shows some failure cases of our approach. From Table 12, we can see that the failure reasons 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 KB facts. The approach should select the most relevant fact from multiple facts that matches query conditions. In particular for questions whose answers are from KBs (in order words, open-ended questions), our method may generate multiple answers. Sometimes, the ground truth is not the first in the ordered answers. In these cases, the top1 answer is wrong, but the topN answers 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 predicted answers. Table 13 reports the accuracy of supporting-fact prediction. We have 41 percent chance to predict the correct one from millions of facts in the incorporated Knowledge Bases.

TABLE 10  
Accuracies for Different Methods According to Different Answer Sources

















Method	Answer-Source, Acc. $\pm$ Std (%)					
	Image			KB		
	Top-1	Top-3	Top-10	Top-1	Top-3	Top-10
SVM-Question	12.38 $\pm$ 0.88	24.68 $\pm$ 0.47	41.05 $\pm$ 1.05	0.78 $\pm$ 0.23	2.00 $\pm$ 0.47	4.16 $\pm$ 0.79
SVM-Image	21.81 $\pm$ 1.30	38.27 $\pm$ 1.33	55.93 $\pm$ 1.29	2.30 $\pm$ 0.25	4.73 $\pm$ 0.62	7.74 $\pm$ 0.51
SVM-Question+Image	22.38 $\pm$ 1.09	38.72 $\pm$ 1.20	56.63 $\pm$ 0.97	2.38 $\pm$ 0.30	4.65 $\pm$ 0.63	7.86 $\pm$ 0.44
LSTM-Question	12.35 $\pm$ 0.57	22.45 $\pm$ 0.68	37.06 $\pm$ 0.98	1.45 $\pm$ 0.68	2.72 $\pm$ 0.63	5.89 $\pm$ 0.95
LSTM-Image	24.19 $\pm$ 0.98	42.40 $\pm$ 1.73	64.92 $\pm$ 2.44	3.31 $\pm$ 0.74	5.69 $\pm$ 0.87	11.87 $\pm$ 0.71
LSTM-Question+Image	26.98 $\pm$ 1.08	42.90 $\pm$ 1.58	62.87 $\pm$ 4.10	4.03 $\pm$ 0.95	7.70 $\pm$ 1.30	13.11 $\pm$ 1.51
LSTM-Question+Image+Pre-VQA	28.97 $\pm$ 0.64	46.62 $\pm$ 0/85	65.83 $\pm$ 0.97	6.13 $\pm$ 1.01	10.94 $\pm$ 1.24	16.73 $\pm$ 1.23
Hie-Question+Image	39.20 $\pm$ 1.36	57.80 $\pm$ 1.07	73.41 $\pm$ 0.55	8.16 $\pm$ 0.39	13.81 $\pm$ 1.46	20.78 $\pm$ 1.11
Hie-Question+Image+Pre-VQA	49.93 $\pm$ 1.08	68.21 $\pm$ 0.67	<b>82.08 <math>\pm</math> 0.56</b>	11.61 $\pm$ 0.95	18.69 $\pm$ 1.55	<b>26.25 <math>\pm</math> 1.51</b>
Ours, gt-QQmapping <sup>‡</sup>	73.69 $\pm$ 0.67	81.04 $\pm$ 0.51	81.04 $\pm$ 0.51	15.95 $\pm$ 0.64	25.17 $\pm$ 0.89	32.39 $\pm$ 1.16
Ours, top-1-QQmapping	61.11 $\pm$ 1.48	68.31 $\pm$ 1.24	68.34 $\pm$ 1.22	12.12 $\pm$ 1.01	19.13 $\pm$ 1.30	23.91 $\pm$ 1.39
Ours, top-3-QQmapping	<b>66.32 <math>\pm</math> 1.20</b>	<b>74.11 <math>\pm</math> 1.21</b>	74.15 $\pm$ 1.18	<b>12.39 <math>\pm</math> 1.09</b>	<b>19.87 <math>\pm</math> 1.35</b>	24.81 $\pm$ 1.31
Ensemble	68.15 $\pm$ 0.97	-	-	15.18 $\pm$ 1.22	-	-
Human	82.97 $\pm$ 0.38	-	-	54.47 $\pm$ 1.77	-	-

Best single-model results are shown in bold font. ‡ indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.



TABLE 11





Some Example Results Generated by Our Methods (Pred.: Predicted, QT: Question Type, SF: Supporting-Fact, GT: Ground Truth)

			
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?
<i>Pred. QT:</i> (UsedFor, Object, Image)	(CapableOf, Object, Image)	(HasA, Object, Image)	(Cheaper, Object, Image)
<i>Keywords:</i> 'lie on'	'pulling carriage'	'stripes'	'taxi'
<i>Pred. SF:</i> a sofa is usually to sit or lie on	horses sometimes pull carriages	zebras have stripes	bus are cheaper than taxi
<i>Pred. Answer:</i> sofa	horse	zebras	bus
<i>GT Answer:</i> sofa	horse	zebras	bus
			
Which object in this image can I ride?	What 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?
<i>Pred. QT:</i> (UsedFor, Object, Image)	(HasProperty, Object, Image)	(AtLocation, Object, Image)	(AtLocation, Scene, KB)
<i>Keywords:</i> 'ride'	'romantic dinner'	'birthday party'	-
<i>Pred. SF:</i> 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
<i>Pred. Answer:</i> motorcycle	wine	cake	cow
<i>GT Answer:</i> 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 in this image is capable of hunting a mouse?
<i>Pred. QT:</i> (AtLocation, Scene, KB)	(PartOf, Object, KB)	(RelatedTo, Object, Image)	(CapableOf, Object, Image)
<i>Keywords:</i> -	-	'sail'	'hunting mouse'
<i>Pred. SF:</i> skiers can be on a ski slope	snails have shells	boat is related to sailing	a cat can hunt mice
<i>Pred. Answer:</i> skiers	shells	boat	cat
<i>GT Answer:</i> skiers	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?
<i>Pred. QT:</i> (UsedFor, Object, Image)	(IsA, Object, Image)	(RelatedTo, Object, Image)	(IsA, Object, Image)
<i>Keywords:</i> 'measure passage time'	'trainable animal'	'wool'	'jazz'
<i>Pred. SF:</i> 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
<i>Pred. Answer:</i> clock	horse	sheep	saxophone
<i>GT Answer:</i> clock	horse	sheep	saxophone

The question type is represented by a 3-tuple  $(T_{REL}, T_{KVC}, T_{AS})$ . The supporting-facts triplet have been translated to textual sentence for easy understanding. Note that no keywords are mined by our approach, if the predicted answer-source (AS) is KB.

TABLE 12

Failure Cases of Our Approach (Pred.: Predicted, GT: Ground Truth, QT: Query Type, VC: Visual Concept, SF: Supporting-Fact)

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

The question type is represented by a 3-tuple ( $T_{REL}$ ,  $T_{KVC}$ ,  $T_{AS}$ ). 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 4.4 for details).

TABLE 13  
Facts Prediction Accuracy for Our Proposed Methods

Method	Supporting-Fact Prediction Acc. $\pm$ Std (%)		
	Top-1	Top-3	Top-10
Ours, gt-QQmapping <sup>‡</sup>	56.31 $\pm$ 0.89	62.55 $\pm$ 0.96	63.55 $\pm$ 0.96
Ours, top-1-QQmapping	38.76 $\pm$ 0.88	42.96 $\pm$ 0.78	43.60 $\pm$ 0.77
Ours, top-3-QQmapping	<b>41.12 <math>\pm</math> 0.74</b>	<b>45.49 <math>\pm</math> 0.89</b>	<b>46.13 <math>\pm</math> 0.87</b>

Best results are shown in bold font. <sup>‡</sup> indicates that ground truth Question-Query mappings are used, which (in gray) will not participate in rankings.

## 6 CONCLUSION

In this work, we have built a new 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.

## ACKNOWLEDGEMENTS

This work was in part supported by ARC Future Fellowship FT120100969. The authors Peng Wang and Qi Wu contributed to this work equally. Part of the work was done when Peng Wang was with The University of Adelaide.

## REFERENCES

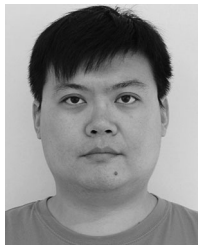
- [1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.
- [2] T.-Y. Lin, et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comp. Vis.*, 2014, pp. 740–755.
- [3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2009, pp. 248–255.
- [4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [5] S. Antol, et al., "VQA: Visual Question Answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2425–2433.
- [6] M. Malinowski and M. Fritz, "Towards a visual turing challenge," *CoRR* abs/1410.8027, 2014.
- [7] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Dataset and methods for multilingual image question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2296–2304.
- [8] L. Yu, E. Park, A. C. Berg, and T. L. Berg, "Visual Madlibs: Fill in the blank description generation and question answering," in *Proc. IEEE Int. Conf. Comp. Vis.*, Dec. 2015, pp. 2461–2469.
- [9] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2953–2961.
- [10] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, 2016, pp. 4995–5004.
- [11] R. Krishna, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comp. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.
- [12] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1682–1690.
- [13] K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu, "Joint video and text parsing for understanding events and answering queries," *IEEE MultiMedia*, vol. 21, no. 2, pp. 42–70, Apr.-Jun. 2014.
- [14] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual Turing test for computer vision systems," *Proc. Nat. Academy Sci. United States America*, vol. 112, no. 12, pp. 3618–3623, 2015.
- [15] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1–9.
- [16] K. Xu, et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
- [17] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "ABC-CNN: An attention based convolutional neural network for visual question answering," *CoRR* abs/1511.05960, 2015.
- [18] A. Jiang, F. Wang, F. Porikli, and Y. Li, "Compositional memory for visual question answering," *CoRR* abs/1511.05676, 2015.
- [19] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, 2016, pp. 39–48.



- [20] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, 2016, pp. 21–29.
- [21] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering temporal context for video question and answering," CoRR abs/1511.04670, 2015.
- [22] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," *Semantic Web*, vol. 4825, pp. 722–735, 2007.
- [23] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction for the web," in *Proc. Int. Joint Conf. Artif. Intell.*, 2007, pp. 2670–2676.
- [24] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD/PODS Conf.*, 2008, pp. 1247–1250.
- [25] A. Carlson, J. Betteridge, B. Kiesel, and B. Settles, "Toward an architecture for never-ending language learning," in *Proc. Nat. Conf. Artif. Intell.*, 2010, pp. 1306–1313.
- [26] X. Chen, A. Shrivastava, and A. Gupta, "Neil: Extracting visual knowledge from web data," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1409–1416.
- [27] F. Mahdisoltani, J. Biega, and F. Suchanek, "YAGO3: A knowledge base from multilingual Wikipedias," in *Proc. Conf. Innovative Data Syst. Res.*, 2015.
- [28] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, 2014.
- [29] R. W. Group, et al., "Resource description framework," 2014. [Online]. Available: <http://www.w3.org/standards/techs/rdf>
- [30] E. Prud'Hommeaux, et al., "SPARQL query language for RDF," *W3C Recommend.*, vol. 15, 2008.
- [31] O. Erling, "Virtuoso, a hybrid RDBMS/graph column store," *IEEE Data Eng. Bull.*, vol. 35, no. 1, pp. 3–8, Jan. 2012.
- [32] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," in *Proc. Int. Joint Conf. Artificial Intell.*, 2013, pp. 3161–3165.
- [33] O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam, "Open Information Extraction: The Second Generation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 3–10.
- [34] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2011, pp. 1535–1545.
- [35] N. Tandon, G. De Melo, and G. Weikum, "Acquiring comparative commonsense knowledge from the web," in *Proc. Nat. Conf. Artif. Intell.*, 2014, pp. 166–172.
- [36] H. Liu and P. Singh, "ConceptNet—a practical commonsense reasoning tool-kit," *BT Technol. J.*, vol. 22, no. 4, pp. 211–226, 2004.
- [37] L. S. Zettlemoyer and M. Collins, "Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars," in *Proc. Uncertainty Artif. Intell.*, 2005, pp. 658–666.
- [38] L. S. Zettlemoyer and M. Collins, "Learning context-dependent mappings from sentences to logical form," in *Proc. Int. Joint Conf. Natural Language Process.*, 2005, pp. 976–984.
- [39] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on Freebase from question-answer pairs," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2013, pp. 1533–1544.
- [40] Q. Cai and A. Yates, "Large-scale semantic parsing via schema matching and lexicon extension," in *Proc. Conf. Assoc. Comput. Linguistics*, 2013, pp. 423–433.
- [41] P. Liang, M. I. Jordan, and D. Klein, "Learning dependency-based compositional semantics," *Comput. Linguistics*, vol. 39, no. 2, pp. 389–446, 2013.
- [42] T. Kwiatkowski, E. Choi, Y. Artzi, and L. Zettlemoyer, "Scaling semantic parsers with on-the-fly ontology matching," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2013, pp. 1545–1556.
- [43] J. Berant and P. Liang, "Semantic parsing via paraphrasing," in *Proc. Conf. Assoc. for Comput. Linguistics*, 2014, pp. 1415–1425.
- [44] A. Fader, L. Zettlemoyer, and O. Etzioni, "Open question answering over curated and extracted knowledge bases," in *Proc. ACM Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1156–1165.
- [45] S. W.-t. Yih, M.-W. Chang, X. He, and J. Gao, "Semantic parsing via staged query graph generation: Question answering with knowledge base," in *Proc. Int. Joint Conf. Natural Language Process.*, 2015, pp. 1321–1331.
- [46] S. Reddy, et al., "Transforming dependency structures to logical forms for semantic parsing," *Trans. Assoc. for Comput. Linguistics*, vol. 4, pp. 127–140, 2016.
- [47] C. Xiao, M. Dymetman, and C. Gardent, "Sequence-based structured prediction for semantic parsing," in *Proc. Conf. Assoc. Comput. Linguistics*, 2016, pp. 1341–1350.
- [48] C. Unger, L. Bühmann, J. Lehmann, A.-C. NgongaNgomo, D. Gerber, and P. Cimiano, "Template-based question answering over RDF data," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 639–648.
- [49] O. Kolomiyets and M.-F. Moens, "A survey on question answering technology from an information retrieval perspective," *Inform. Sci.*, vol. 181, no. 24, pp. 5412–5434, 2011.
- [50] X. Yao and B. Van Durme, "Information extraction over structured data: Question answering with Freebase," in *Proc. Conf. Assoc. Comput. Linguistics*, 2014, pp. 956–966.
- [51] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2014, pp. 615–620.
- [52] A. Bordes, J. Weston, and N. Usunier, "Open question answering with weakly supervised embedding models," in *Proc. Joint Eur. Conf. Mach. Learning Knowl. Discovery Databases*, 2014, pp. 165–180.
- [53] L. Dong, F. Wei, M. Zhou, and K. Xu, "Question answering over freebase with multi-column convolutional neural networks," in *Proc. Int. Joint Conf. Natural Language Process.*, 2015, pp. 260–269.
- [54] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [55] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [56] S. Sukhbaatar, J. Weston, R. Fergus, et al., "End-to-end memory networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2440–2448.
- [57] A. Kumar, et al., "Ask me anything: Dynamic memory networks for natural language processing," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1378–1387.
- [58] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2397–2406.
- [59] Y. Zhu, C. Zhang, C. Ré, and L. Fei-Fei, "Building a large-scale multimodal knowledge base for visual question answering," CoRR abs/1507.05670, 2015.
- [60] Q. Wu, P. Wang, C. Shen, A. van den Hengel, and A. Dick, "Ask me anything: Free-form visual question answering based on knowledge from external sources," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2016, pp. 4622–4630.
- [61] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," CoRR abs/1511.02570, 2015.
- [62] J. Krishnamurthy and T. Kollar, "Jointly learning to parse and perceive: Connecting natural language to the physical world," *Trans. Assoc. Comput. Linguistics*, vol. 1, pp. 193–206, 2013.
- [63] K. Narasimhan, A. Yala, and R. Barzilay, "Improving information extraction by acquiring external evidence with reinforcement learning," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2016, pp. 2355–2365.
- [64] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [65] Q. Wu, C. Shen, A. van den Hengel, L. Liu, and A. Dick, "What value do explicit high-level concepts have in vision to language problems?" in *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, 2016, pp. 203–212.
- [66] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.
- [67] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and yang: Balancing and answering binary visual questions," in *Proc. IEEE Conf. Comput. Vis. Patt. Recogn.*, 2016, pp. 5014–5022.
- [68] J. Weston, A. Bordes, S. Chopra, and T. Mikolov, "Towards ai-complete question answering: A set of prerequisite toy tasks," CoRR abs/1502.05698, 2015.
- [69] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," CoRR abs/1612.06890, 2016.
- [70] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
- [71] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Inter-speech*, 2010, vol. 2, Art. no. 3.



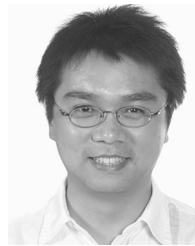
- [72] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [73] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," CoRR abs/1301.3781, 2013.
- [74] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proc. 32Nd Annu. Meet. Assoc. Comput. Linguistics*, 1994, pp. 133–138. [Online]. Available: <http://dx.doi.org/10.3115/981732.981751>
- [75] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
- [76] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011, Art. no. 27.
- [77] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 289–297.



**Peng Wang** received the bachelor's degree in electrical engineering and automation, and a PhD in control science and engineering from Beihang University (China) in 2004 and 2011, respectively. He is a professor in the School of Computer Science, Northwestern Polytechnical University, China. He was with the Australian Centre for Visual Technologies (ACVT) of the University of Adelaide for about four years. His research interests include computer vision, machine learning and artificial intelligence.



**Qi Wu** received the bachelor's degree in mathematical sciences from China Jiliang University, the master's degree in computer science, and the PhD degree in computer vision from the University of Bath, United Kingdom, in 2012 and 2015, respectively. He is a postdoctoral researcher with the University of Adelaide. His research interests include cross-depiction object detection and classification, attributes learning, neural networks, and image captioning.



**Chunhua Shen** received the PhD degree from the University of Adelaide. He is a professor of computer science with the University of Adelaide. He was with the computer vision program at NICTA (National ICT Australia) in Canberra for almost six years before moving back to Adelaide. He studied at Nanjing University (China), at the Australian National University. In 2012, he was awarded the Australian Research Council Future Fellowship.



**Anthony Dick** received the PhD degree from the University of Cambridge, in 2002, where he worked on 3D reconstruction of architecture from images. He is an associate professor with the University of Adelaide. His research interests include image-based modeling, automated video surveillance, and image search.



**Anton van den Hengel** received bachelor of mathematical science degree, in 1991, the bachelor of laws degree, in 1993, the master degree in computer science, in 1994, and the PhD degree in computer vision, in 2000, all from The University of Adelaide. He is a professor with the University of Adelaide and the founding director of The Australian Centre for Visual Technologies (ACVT).

▷ For more information on this or any other computing topic, please visit our Digital Library at [www.computer.org/publications/dlib](http://www.computer.org/publications/dlib).