Explicit Knowledge-based Reasoning for Visual Question Answering

Peng Wang[*1,2], **Qi Wu**[*3], **Chunhua Shen**[2,3], **Anthony Dick**[2], **Anton van den Hengel**[2,3]
[1]Northwestern Polytechnical University, China
[2]The University of Adelaide, Australia, [3]Australian Centre for Robotic Vision
{peng.wang}@nwpu.edu.cn,
{qi.wu01,chunhua.shen,anthony.dick,anton.vandenhengel}@adelaide.edu.au

Abstract

We describe a method for visual question answering which is capable of reasoning about an image on the basis of information extracted from a large-scale knowledge base. The method not only answers natural language questions using concepts not contained in the image, but can explain the reasoning by which it developed its answer. It is capable of answering far more complex questions than the predominant long short-term memory-based approach, and outperforms it significantly in testing. We also provide a dataset and a protocol by which to evaluate general visual question answering methods.

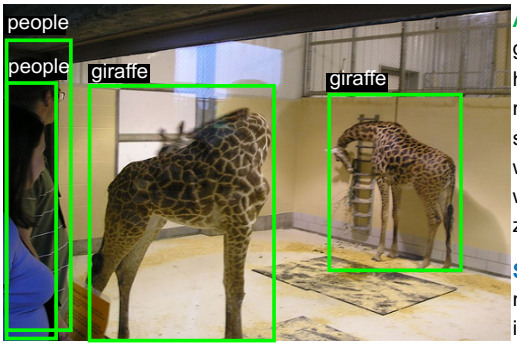
1 Introduction

Visual Question Answering (VQA) methods aim to interactively answer questions about images. The questions are typically posed in natural language, as are the answers. The problem requires image understanding, natural language processing, and a means by which to relate images and text. Importantly, the interactivity of the problem means that it cannot be determined beforehand which questions will be asked. This requirement to answer a wide range of image-based questions, on the fly, means that the problem is closely related to several ongoing challenges in Artificial Intelligence [Geman *et al.*, 2015].

Despite the apparent need to perform general reasoning about the content of images, most VQA methods perform no explicit reasoning at all. The predominant method [Antol *et al.*, 2015; Malinowski *et al.*, 2015; Lu *et al.*, 2016; Fukui *et al.*, 2016] is based on directly connecting a convolutional neural network (CNN) to perform the image analysis, and a Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] network to process the question and answer text. This approach has shown the ability to answer simple questions directly related to the content of the image, such as 'What color is the ...?', or 'How many ... are there?'.

There are a number of problems with this approach, however. The first is that the method cannot explain how it arrived at its answer. 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. Answering general questions posed by humans about images inevitably requires reference to a diverse variety of information not contained in the image itself. Capturing such large amount of information would require an implausibly large LSTM, and a completely impractical amount of training data. The third, and major, problem with the LSTM approach is that it is incapable of explicit reasoning except in very limited situations [Rocktäschel *et al.*, 2016].

Our main contribution is a method we call Ahab[1] for answering a wide variety of questions about images that require



Visual Question: How many giraffes are there in the image?
Answer: Two.

Common-Sense Question: Is this image related to zoology?
Answer: Yes. Reason: Object/Giraffe --> Herbivorous animals --> Animal --> Zoology; Attribute/Zoo --> Zoology.

KB-Knowledge Question: What are the common properties between the animal in this image and zebra?
Answer: Herbivorous animals; Animals; Megafauna of Africa.

Figure 1: A real example of the proposed KB-VQA dataset and the results given by Ahab, the proposed VQA approach. Our approach answers questions by extracting several types of visual concepts from an image and aligning them to large-scale structured knowledge bases. Apart from answers, our approach can also provide reasons and explanations for certain types of questions.

[*]The first two authors contributed to this work equally.

[1]Ahab, the captain in the novel Moby Dick, is either a brilliant visionary, or a deluded fanatic, depending on your perspective.

external information to answer. Rather than learning a mapping from image and question directly to an answer as in [Lu *et al.*, 2016; Malinowski *et al.*, 2015], Ahab maps an image and question to a *query* which is then applied to a large-scale *structured* Knowledge Base (KB) to obtain the final answer. In effect, rather than learning answers by association, we are solving *how* to find an answer for a given question and image. The use of a knowledge base means we do not need to learn or have any prior experience of concepts that may appear in questions or answers, if they can be related to known concepts through the KB. This vastly expands the range of concepts we can make use of when answering questions. It also means that answers are "traceable" as the chain of reasoning by which they were obtained is explicit.

The main limitation of the method is that there is a limited number of ways in which a KB can be queried, and this limits the types of questions that can be asked. Thus, although questions are asked in natural language, they must be reducible to one of the available query templates (in our case the 23 templates listed in Table 1). This limitation is not as severe as it might appear, since (a) each template is very general and can be used to ask a wide range of natural language questions, and (b) it is relatively straightforward to add new templates as required, although it does involve some manual creation.

The type of questions that Ahab is designed to answer is quite different to previous VQA systems. We therefore propose a new dataset, and protocol for measuring performance, for this type of visual question answering. The questions in the dataset are asked by people based on a number of pre-defined templates. Questions are given one of three labels reflecting the information required to answer them: "Visual", "Common-sense" and "KB-knowledge" (see Fig. 1). Compared to other VQA datasets [Antol *et al.*, 2015; Ren *et al.*, 2015; Malinowski and Fritz, 2014b], our dataset is necessarily smaller because of the human effort required to create it. However the questions in the KB-VQA dataset, as a whole, require a higher level of external knowledge to answer. The evaluation protocol requires human evaluation of question answers, as this does not limit the questions which can be asked.

1.1 Background

The first VQA approach [Malinowski and Fritz, 2014a] processed questions using semantic parsingand obtained answers through Bayesian reasoning. The work in [Malinowski *et al.*, 2015] used CNNs to extract image features and relied on LSTMs to encode questions and decode answers. Inspired by Xu *et al.* [Xu *et al.*, 2015] who encode visual attention in the Image Captioning, [Xu and Saenko, 2016; Yang *et al.*, 2016] propose to use the spatial attention to help answering visual questions. However, either LSTM or GRU (Gated Recurrent Unit) is still applied in these methods to model the questions. Irrespective of the finer details, we call this the LSTM approach. Several VQA datasets [Malinowski and Fritz, 2014b; Ren *et al.*, 2015; Antol *et al.*, 2015; Yu *et al.*, 2015; Krishna *et al.*, 2017] have also been constructed for making and measuring progress. The VQA [Antol *et al.*, 2015] and Visual Genome [Krishna *et al.*, 2017] datasets are currently the largest two VQA datasets.

As discussed previously, it is difficult for LSTM approaches to encode all the background information that is required by answering general visual questions. Large-scale *structured* KBs are one means of capturing such background information. In these KBs, knowledge is typically represented by a large number of triples of the form `(arg1,rel,arg2)`, where `arg1` and `arg2` denote two entities 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 using a Resource Description Framework [Cyganiak *et al.*, 2014] (RDF) specification, and housed in a triple-store, which allows queries over the data. Large-scale structured KBs are constructed either by manual annotation or crowdsourcing (*e.g.*, Freebase [Bollacker *et al.*, 2008] and DBpedia [Auer *et al.*, 2007]) or by automatic extraction from unstructured/semi-structured data (*e.g.*, YAGO [Mahdisoltani *et al.*, 2015] and OpenIE [Etzioni *et al.*, 2011]). The KB we use here is DBpedia, which contains 2.6 million concepts and 360 million relationships that are extracted from Wikipedia.

There are a number of works focusing on the problem of question answering over structured KBs (KB-QA) [Berant *et al.*, 2013; Bordes *et al.*, 2015; Höffner *et al.*, 2016]. The VQA approach which is closest to KB-QA (and to our approach) is that of [Zhu *et al.*, 2015] as they use a KB and RDBMS to answer image-based questions. 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 information about the world. Wu *et al.* [Wu *et al.*, 2016] encode the mined knowledge text from the DBpedia to a vector and combined with visual features together to generate answers using an LSTM model. However, their proposed method only extracts discrete text pieces from the knowledge base but ignores the power of its structural representation. Moreover, both of [Zhu *et al.*, 2015] and [Wu *et al.*, 2016] are not capable of explicit reasoning. Wang *et al.* [Wang *et al.*, 2017] proposed a VQA method that performs reasoning on structured representation of images, which, however, does not utilize external structured KBs.

2 The KB-VQA Dataset

The KB-VQA dataset[2] has been constructed for the purpose of evaluating the performance of VQA algorithms on questions requiring higher level knowledge, and explicit reasoning about image contents using external information.

2.1 Data Collection

Images. We select 700 of the validation images from the MS COCO [Lin *et al.*, 2014] dataset due to the rich contextual information and diverse object classes therein. The images are selected so as to cover around 150 object classes and 100 scene classes, and typically exhibit 6 to 7 objects each.

[2] `https://bitbucket.org/sxjzwq1987/kb-vqa-dataset`

Name	Template	Num.
IsThereAny	Is there any ⟨*concept*⟩?	419
IsImgRelate	Is the image related to ⟨*concept*⟩?	381
WhatIs	What is the ⟨*obj*⟩?	275
ImgScene	What scene does this image describe?	263
ColorOf	What color is the ⟨*obj*⟩?	205
HowMany	How many ⟨*concept*⟩ in this image?	157
ObjAction	What is the ⟨*person/animal*⟩ doing?	147
IsSameThing	Are the ⟨*obj1*⟩ and the ⟨*obj2*⟩ the same thing?	71
MostRelObj	Which ⟨*obj*⟩ is most related to ⟨*concept*⟩?	56
ListObj	List objects found in this image.	54
IsTheA	Is the ⟨*obj*⟩ a ⟨*concept*⟩?	51
SportEquip	List all equipment I might use to play this sport.	48
AnimalClass	What is the ⟨*taxonomy*⟩ of the ⟨*animal*⟩?	46
LocIntro	Where was the ⟨*obj*⟩ invented?	40
YearIntro	When was the ⟨*obj*⟩ introduced?	32
FoodIngredient	List the ingredient of the ⟨food⟩.	31
LargestObj	What is the largest/smallest ⟨*concept*⟩?	27
AreAllThe	Are all the ⟨*obj*⟩ ⟨*concept*⟩?	27
CommProp	List the common properties of the ⟨*obj1*⟩ and ⟨*concept/obj2*⟩.	26
AnimalRelative	List the close relatives of the ⟨*animal*⟩.	17
AnimalSame	Are ⟨*animal1*⟩ and ⟨*animal2*⟩ in the same ⟨*taxonomy*⟩?	17
FirstIntro	Which object was introduced earlier, ⟨*obj1*⟩ or ⟨*concept/obj2*⟩?	8
ListSameYear	List things introduced in the same year as the ⟨*obj*⟩.	4

Table 1: Question templates in descending order of number of instantiations. The total number of questions is 2402. Note that some templates allow questioners to ask the same question in different forms.

Templates. Five postgraduate students (questioners) are asked to generate 3 to 5 question/answer pairs for each of the 700 images, with each question following one of the 23 templates shown in Table 1. Each individual questioner was asked to do this for 20 randomly selected images, and to choose the templates that were most appropriate. There are several slots (⟨*obj*⟩, ⟨*concept*⟩, . . .) to be filled in these templates, where ⟨*obj*⟩ is used to refer to a specific object in the image and ⟨*concept*⟩ can be filled by any word or phrase which probably corresponds to an entity in DBpedia.

Questions. The questions of primary interest here are those requiring knowledge external to the image to answer. Each question has a label reflecting the human-estimated level of knowledge required to answer it correctly. The "Visual" questions can be answered directly using visual concepts gleaned from ImageNet and MS COCO (such as "Is there a dog in this image?"); "Common-sense" questions should not require an adult to refer to an external source ("How many road vehicles are in this image?"); while answering "KB-knowledge" questions is expected to require Wikipedia or similar ("When was the home appliance in this image invented?").

2.2 Data Analysis

Question Analysis

From Table 1, we see that the top 5 most frequently used templates are *IsThereAny*, *IsImgRelate*, *WhatIs*, *ImgScene* and *ColorOf*. Some templates lead to questions that can be answered without any external knowledge, such as *ImgScene* and *ListObj*. But "Common-sense" or "KB-knowledge" is required to analyze the relationship between visual objects and concepts in questions like *IsTheA*, *AreAllThe* and *IsThereAny*. More complex questions like *YearInvent* and *AnimalClass* demand "KB-knowledge". The total number of questions labelled "Visual", "Common-sense" and "KB-knowledge" are 1256, 883 and 263 respectively. Fig. 2 shows the distribution of the 23 templates for each question type. Templates *IsImgRelate* and *IsThereAny* cover almost half of the "Common-sense" questions. There are 18 templates shown for "KB-knowledge" questions (as *WhatIs*, *IsSameThing*, *ListObj*, *LargestObj* and *ColorOf* do not appear), which exhibit a more balanced distribution. The average lengths of "KB-knowledge" questions and their answers are 7.6 and 2.5 respectively, which are significantly longer than the overall average level of our dataset (6.8 and 2.0) and the VQA dataset (6.2 and 1.1).

In total, 330 different phrases were used by questioners in filling the ⟨*concept*⟩ slot. There were 254 phrases used in questions requiring external knowledge (*i.e.*, "Common-sense" and "KB-knowledge"), 55% of which are mentioned very rarely (less than 20 times) in the 300K questions of the VQA [Antol *et al.*, 2015] dataset. For the "KB-knowledge" level, 67 phrases are used and greater than 85% of them have less than 20 mentions in VQA. Examples of concepts not occurring in VQA include "logistics", "herbivorous animal", "animal-powered vehicle", "road infrastructure" and "portable electronics".

Compared to other VQA datasets, a large proportion of the questions in KB-VQA require external knowledge to answer. The questions defined in DAQUAR [Malinowski and Fritz, 2014b] are almost exclusively "Visual" questions, referring to "color", "number" and "physical location of the object". In the COCO-QA dataset [Ren *et al.*, 2015], questions are generated automatically from image captions which describe the major visible content of the image. For the VQA dataset [Antol *et al.*, 2015], only 5.5% of questions require adult-level common-sense, and none requires "KB-knowledge" (by observation).

Answer Analysis

Questions starting with "Is ..." and "Are ..." require logical answers (61% "yes" and 39% "no"). Questions starting with "How many", "What color", "Where" and "When" need to be answered with number, color, location and time respectively. Of the human answers for "How many" questions, 74% are less than or equal to 5. The most frequent number answers are "1"(58 occurrences), "2"(66) and "3"(44). We also have 16 "How many" questions with human answer "0". The answers for "What ..." and "List ..." vary significantly, covering a wide range of concepts.

3 The Ahab VQA approach

3.1 RDF Graph Construction

In order to reason about the content of an image we first need to extract the relevant information from it. This is done by detecting concepts in the query image and linking them to the relevant parts of the KB.

Visual Concepts. For each image, three types of visual concepts are detected by different CNN models, including objects, scenes and attributes.

Linking to the KB. Having extracted a set of concepts of interest from the image, we now need to relate them to the KB. As shown in the top side of Fig. 3, the visual concepts are stored as RDF triples. For example, the information that "The image contains a giraffe object" is expressed as: `(Img,contain,Obj-1)` and `(Obj-1,name,ObjCat-giraffe)`. Each visual

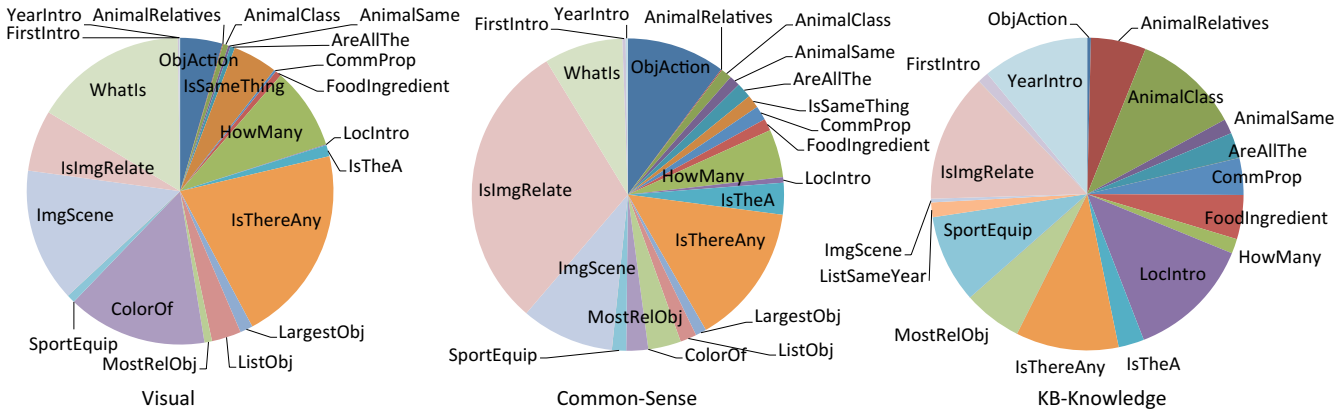


Figure 2: Question template frequencies for different knowledge levels.

concept is linked to DBpedia entities with the same semantic meaning (for example `(ObjCat-giraffe, same-concept, KB:Giraffe)`), and then linked to all of the relevant information in DBpedia.

3.2 Answering Questions

Having gathered all of the relevant information from the image and DBpedia, we now use them to answer questions.

Parsing NLQs. Given a question posed in natural language, we first need to translate it to a format which can be used to query the KB. Quepy[3] is a Python framework designed within the NLP community to achieve exactly this task. Quepy begins by tagging each word in the question using NLTK [Bird *et al.*, 2009], which is composed of a tokenizer, a part-of-speech tagger and a lemmatizer. The tagged question is then parsed by a set of regular expressions, which are carefully built to increase the flexibility of question expression as much as possible. Once a regular expression matches the question, it will extract the slot-phrases and forward them for further processing. For example, the question in Fig. 3 bottomo is matched to template *CommProp* and the slot-phrases for $\langle obj \rangle$ and $\langle concept \rangle$ are "right animal" and "zebra" respectively.

Mapping Slot-Phrases to KB-entities. Note that the slot-phrases are still expressed in natural language. The next step is to find the correct correspondences between the slot-phrases and entities in the constructed graph. Phrases in slot $\langle obj \rangle$ can be identified by provided locations, sizes and/or names. While phrases in slot $\langle concept \rangle$ can be mapped to DBpedia entities using predicate `wikiPageRedirects` that handle synonyms, capitalizations, punctuation, tenses, abbreviation and misspellings of concepts.

Query Generation. With all concepts mapped to KB entities, the next step is to form the appropriate SPARQL queries, depending on the question template. Several DBpedia predicates are used extensively for generating queries and further analysis, including `Infoboxes`, `Wikilinks` and `Transitive Categories`. To answer the questions *IsThereAny*, *HowMany*, *IsTheA*, *LargestObj* and *AreAllThe*, we need to determine if there is a hyponymy relationship between two entities. This is done by checking if one entity is a transitive category of the other. For question *CommProp*, we collect the transitive categories shared by two entities (see Fig. 3). For questions *IsImgRelate* and *MostRelObj*, the correlation between a visual concept and the concept given in the question is measured based on checking the hyponymy relationship and counting the number of `Wikilinks`. Answering other templates (*e.g.*, *FoodIngredient*) needs specific types of predicates extracted from Wikipedia infoboxes.

Answer and Reason. The last step is to generate answers according to the results of the queries. Post-processing operations are needed for some questions, such as *IsImgRelate* and *MostRelObj*. Note that our system performs searches along the paths from visual concepts to KB concepts. These paths can be used to give "logical reasons" as to how the answer is generated. Especially for questions requiring external knowledge, the predicates and entities on the path give a better understanding of how the relationships are established between visual concepts and KB concepts (see Fig. 5 for examples).

[3] `http://quepy.readthedocs.org/en/latest/`

4 Experiments

Metrics

We firstly evaluate different approaches automatically using simple string matching and Wu-Palmer similarity (WUPS) [Malinowski and Fritz, 2014a], in which the human answers are considered as ground truth. But in our case, these automatic evaluation metrics may be unsuitable because most of the questions in our dataset are open-ended, especially for the "KB-knowledge" questions. In addition, there is no automated method for assessing the reasons provided by our system. Hence, we also ask 5 human subjects (examiners) to evaluate the results manually. In order to understand the generated answers better, we ask the examiners to give each answer or reason a correctness score as follows: 1: "Totally wrong"; 2: "Slightly wrong"; 3: "Borderline"; 4: "OK"; 5: "Perfect". An answer or reason scored higher than "Borderline" is considered as "right"; otherwise, it is considered as "wrong". We perform this evaluation double-blind, *i.e.* examiners are different from questions/answers providers and do not know the answers source.

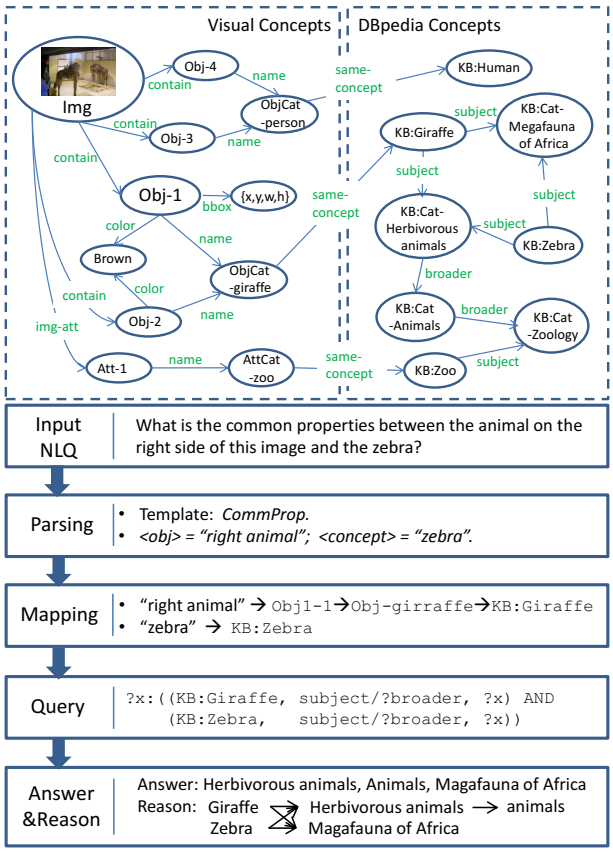


Figure 3: **Top:** An RDF graph such as might be constructed by Ahab. For simplicity, we only show entities that are relevant to answering the questions in Fig. 1. Each arrow corresponds to one triple in the graph, with circles representing entities and green text reflecting predicate type. The graph of extracted visual concepts (left side) is linked to DBpedia (right side) by mapping object/attribute/scene to DBpedia entities using the predicate `same-concept`. **Bottom:** The question processing pipeline. The input question is parsed using a set of NLP tools to identify the appropriate template. The extracted slot-phrases are then mapped to entities in the KB. Next, KB queries are generated to mine the relevant relationships for the KB-entities. Finally, the answer and reason are generated based on the query results. The predicate `category/?broader` is used to obtain the categories transitively.

Accuracy (%)	String Matching	WUPS0.9	WUPS0.0
Ours	**56.00**	**58.37**	**85.35**
LSTM	31.74	32.97	77.04

Table 2: Automatic evaluation results of different approaches, against human answers.

Evaluation

We compare our Ahab system with an approach (which we label LSTM) that encodes both normalized CNN extracted features and questions with an encoder LSTM and generates answers with a decoder LSTM. Specifically, we use the second fully-connected layer (4096-d) of a pre-trained VGG model as the image features, and the LSTM is trained on the training set of VQA data [Antol *et al.*, 2015][4]. The LSTM layer contains 512 memory cells in each unit. We also relate "Human" performance for reference.

[4]This baseline LSTM achieves 58.16% accuracy on the VQA test set, under its evaluation protocol. We note that training LSTM on another dataset provides it an unfair advantage. However, the current KB-VQA dataset is still relatively small and so does not support training large models.

Question Type	Accuracy(%)			Correctness (Avg.)		
	LSTM	Ours	Human	LSTM	Ours	Human
IsThereAny	60.4	86.9	93.6	3.6	4.5	4.7
IsImgRelate	63.3	82.2	97.1	3.3	4.2	4.9
WhatIs	52.7	66.9	94.5	2.1	3.7	4.8
ImgScene	49.0	69.6	85.9	2.3	3.8	4.5
ColorOf	36.6	29.8	93.2	1.7	2.5	4.7
HowMany	33.8	56.1	90.4	2.3	3.3	4.6
ObjAction	41.5	57.1	90.5	1.8	3.5	4.7
IsSameThing	50.7	77.5	91.5	3.2	4.2	4.6
MostRelObj	32.1	80.4	92.9	2.3	4.2	4.6
ListObj	0.0	63.0	100	1.1	3.6	4.8
IsTheA	76.5	80.4	92.2	3.9	4.2	4.7
SportEquip	0.0	70.8	79.2	1.2	3.9	4.2
AnimalClass	0.0	87.0	95.7	1.0	4.5	4.8
LocIntro	2.5	67.5	95.0	1.1	3.6	4.8
YearIntro	0.0	46.9	93.8	1.0	2.9	4.8
FoodIngredient	0.0	58.1	74.2	1.0	3.4	4.3
LargestObj	0.0	66.7	96.3	1.0	3.8	4.8
AreAllThe	40.7	63.0	81.5	2.3	3.7	4.3
CommProp	0.0	76.9	76.9	1.0	4.1	4.2
AnimalRelative	0.0	88.2	76.5	1.1	4.4	4.1
AnimalSame	29.4	70.6	94.1	2.6	3.8	4.8
FirstIntro	37.5	25.0	75.0	2.0	1.5	4.1
ListSameYear	0.0	75.0	50.0	1.8	4.2	3.0
Overall	44.5	69.6	92.0	2.5	3.8	4.7

Table 3: Human evaluation results of different methods for different question types. Accuracy is the percentage of correctly answered questions (*i.e.*, correctness scored higher than "Borderline"). The average answer correctness ($\in [1, 5]$, the higher the better) for each question type is also listed. We also evaluated human provided answers as a reference.

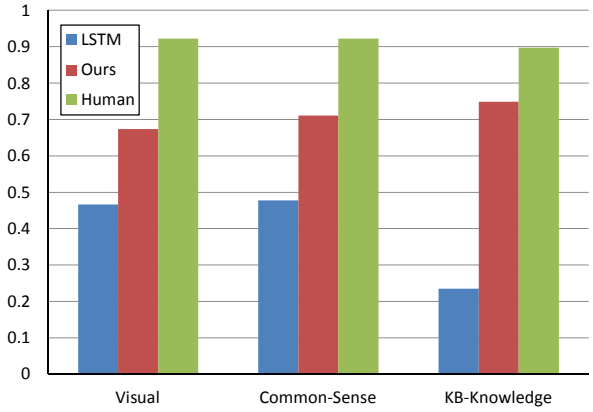


Figure 4: Accuracy of different methods for different knowledge levels. Humans perform almost equally over all three levels. LSTM performs worse for questions requiring higher-level knowledge, whereas Ahab performs better.

The results of automatic evaluation are shown in Table 2, we can see that our method performs significantly better than LSTM under different criteria. Table 3 provides the human evaluation results for different question types. Our system outperforms the LSTM on most of question types with a final accuracy of 69.6%. For question types particularly dependent on KB-knowledge, such as *AnimalClass*, *YearIntro*, *FoodIngredient*, *CommProp* and *AnimalRelative*, all LSTM-generated answers were marked as "wrong" by examiners. In contrast, our system performs very well on these questions. For example, we achieve 88.2% on questions of type *AnimalRelative*, which is better than human performance. Our

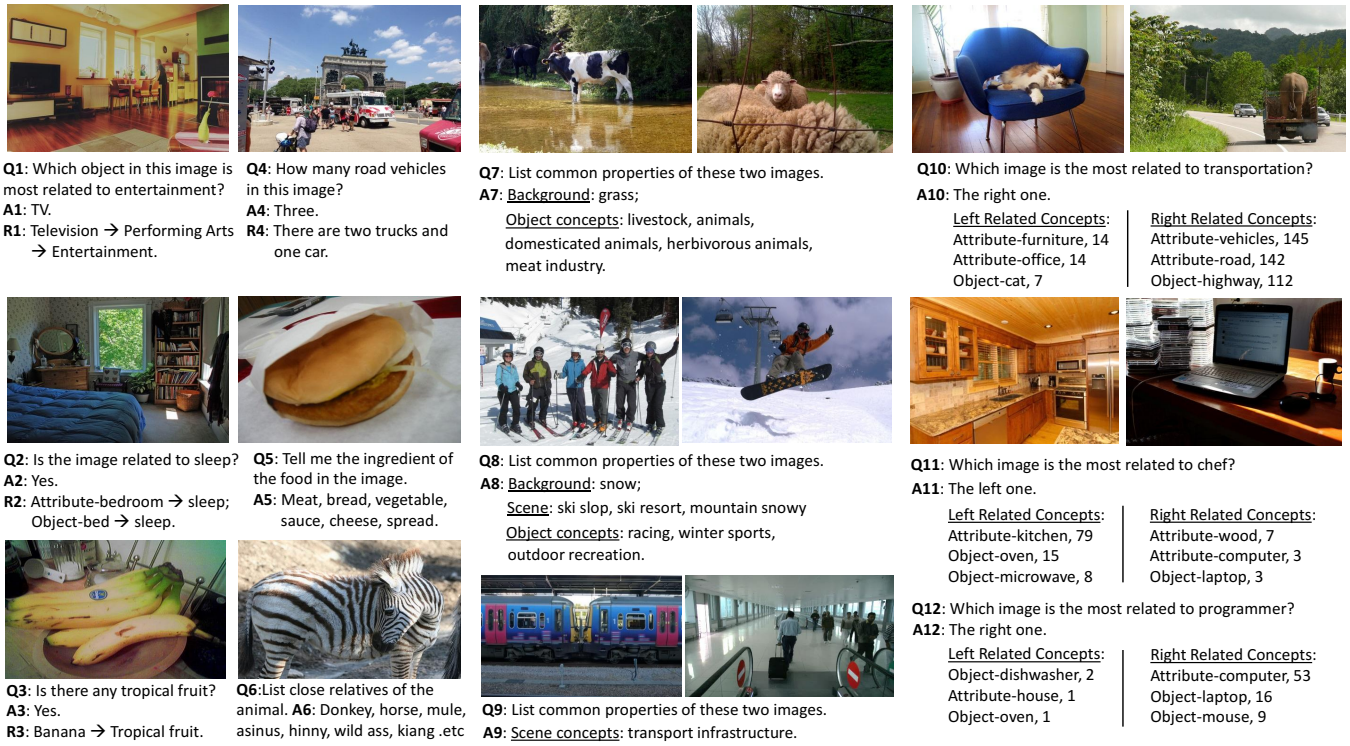


Figure 5: Examples of KB-VQA questions and the answers and reasons given by Ahab. Q1-Q6 are questions on single images, in the proposed KB-VQA dataset. Q7-Q12 demonstrate questions involving two images, which can be answered by the proposed system. The numbers after visual concepts in Q10-Q12 are scores measuring the correlation between visual concepts and concepts given in questions.

system also outperforms humans on the question type *ListSameYear*, which requires knowledge of the year a concept was introduced, and all things introduced in the same year. For the purely "visual" questions such as *WhatIs*, *HowMany* and *ColorOf*, there is still a gap between our proposed system and humans. However, this is mainly caused by object detection errors, which is not the focus of this paper. According to the overall average correctness, we achieve 3.8, which lies between "Borderline" and "OK". The LSTM scores only 2.3 while Human achieves 4.7. We also examine the distribution of the correctness scores for different approaches, and find that Ahab provides more middle-level answers (around 20%) than LSTM (around 7%) and Human (around 10%).

Fig. 4 relates the performance for "Visual", "Commonsense" and "KB-knowledge" questions. The overall trend is the same as in Table 3 — Ahab performs better than the LSTM method but not as well as humans. It is not surprising that humans perform almost equally for different knowledge levels, since we allow the human subjects to use Wikipedia to answer the "KB-knowledge" related questions. For the LSTM method, there is a significant decrease in performance as the dependency on external knowledge increases. In contrast, Ahab performs better as the level of external knowledge required increases. In summary, our system Ahab performs better than LSTM at all three knowledge levels. Furthermore, the performance gap between Ahab and LSTM is more significant for questions requiring external knowledge.

Since "Visual" questions can be answered by direct interrogation of pixels, we have not coded reasons into the question answering process for the corresponding templates (the reasons would be things like "The corresponding pixels are brown"). For the remaining knowledge-required questions, the accuracy of the provided reasons are measured by human examiners using the same protocol. It shows that more than 80% of reasons are marked as correct (*i.e.*, scored higher than 3). To be specific, 69% reasons are marked as score 5, 13% as score 4, 2% as score 3, 3% as score 2 and 13% as score 0.

5 Conclusion

We have described a method capable of reasoning about the content of images, and interactively answering a wide variety of questions about them. The method develops a structured representation of images, and relevant information about the rest of the world, on the basis of a large external knowledge base. It can explain its reasoning in terms of the entities in the knowledge base, and the connections between them. Ahab is applicable to any knowledge base for which a SPARQL interface is available. This includes any of the over a thousand RDF datasets online [Schmachtenberg *et al.*, 2014] with a host of topics. Each could be used to provide a specific VQA capability, and many can also be linked to form larger repositories. If a knowledge base capturing "common sense" were available, the method we have described could use it to draw "sensible" general conclusions about the content of images. We have also provided a dataset and methodology for testing the performance of general visual question answering techniques, and shown that Ahab substantially outperforms the currently predominant visual question answering approach.

References

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

[Auer *et al.*, 2007] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. *DBpedia: A nucleus for a web of open data*. Springer, 2007.

[Berant *et al.*, 2013] Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic Parsing on Freebase from Question-Answer Pairs. In *Proc. EMNLP*, 2013.

[Bird *et al.*, 2009] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

[Bollacker *et al.*, 2008] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD*, 2008.

[Bordes *et al.*, 2015] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. Large-scale simple question answering with memory networks. In *Proc. ICLR*, 2015.

[Cyganiak *et al.*, 2014] Richard Cyganiak, David Wood, and Markus Lanthaler. Rdf 1.1 concepts and abstract syntax, 2014. `http://www.w3.org/standards/techs/rdf`.

[Etzioni *et al.*, 2011] Oren Etzioni, Anthony Fader, Janara Christensen, Stephen Soderland, and Mausam Mausam. Open Information Extraction: The Second Generation. In *Proc. IJCAI*, 2011.

[Fukui *et al.*, 2016] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proc. EMNLP*, 2016.

[Geman *et al.*, 2015] Donald Geman, Stuart Geman, Neil Hallonquist, and Laurent Younes. Visual Turing test for computer vision systems. *Proc. NAS*, 112(12):3618–3623, 2015.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[Höffner *et al.*, 2016] Konrad Höffner, Sebastian Walter, Edgard Marx, Ricardo Usbeck, Jens Lehmann, and Axel-Cyrille Ngonga Ngomo. Survey on challenges of question answering in the semantic web. *Semantic Web*, 2016.

[Krishna *et al.*, 2017] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, 2017.

[Lin *et al.*, 2014] 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. ECCV*, 2014.

[Lu *et al.*, 2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Proc. NIPS*, 2016.

[Mahdisoltani *et al.*, 2015] Farzaneh Mahdisoltani, Joanna Biega, and Fabian Suchanek. YAGO3: A knowledge base from multilingual Wikipedias. In *CIDR*, 2015.

[Malinowski and Fritz, 2014a] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proc. NIPS*, 2014.

[Malinowski and Fritz, 2014b] Mateusz Malinowski and Mario Fritz. Towards a Visual Turing Challenge. In *NIPS Workshop on Learning Semantics*, 2014.

[Malinowski *et al.*, 2015] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In *Proc. ICCV*, 2015.

[Ren *et al.*, 2015] Mengye Ren, Ryan Kiros, and Richard Zemel. Image Question Answering: A Visual Semantic Embedding Model and a New Dataset. In *Proc. NIPS*, 2015.

[Rocktäschel *et al.*, 2016] Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. Reasoning about Entailment with Neural Attention. In *Proc. ICLR*, 2016.

[Schmachtenberg *et al.*, 2014] Max Schmachtenberg, C Bizer, and H Paulheim. State of the LOD Cloud 2014, 2014. `http://linkeddatacatalog.dws.informatik.unimannheim.de/state`.

[Wang *et al.*, 2017] Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *Proc. CVPR*, 2017.

[Wu *et al.*, 2016] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources. In *Proc. CVPR*, 2016.

[Xu and Saenko, 2016] Huijuan Xu and Kate Saenko. Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. In *Proc. ECCV*, 2016.

[Xu *et al.*, 2015] 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. ICML*, 2015.

[Yang *et al.*, 2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked Attention Networks for Image Question Answering. In *Proc. CVPR*, 2016.

[Yu *et al.*, 2015] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank image generation and question answering. In *Proc. ICCV*, 2015.

[Zhu *et al.*, 2015] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a Large-scale Multimodal Knowledge Base System for Answering Visual Queries. *arXiv preprint arXiv:1507.05670*, 2015.