Computer Vision and Image Understanding 163 (2017) 21–40

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Visual question answering: A survey of methods and datasets

Qi Wu*, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, Anton van den Hengel

Australian Centre for Visual Technologies, School of Computer Science, The University of Adelaide, Adelaide, SA 5005, Australia

ARTICLE INFO

Article history:
Received 14 July 2016
Revised 20 January 2017
Accepted 1 May 2017
Available online 5 May 2017

Keywords:
Visual question answering
Natural language processing
Knowledge bases
Recurrent neural networks

ABSTRACT

Visual Question Answering (VQA) is a challenging task that has received increasing attention from both the computer vision and the natural language processing communities. Given an image and a question in natural language, it requires reasoning over visual elements of the image and general knowledge to infer the correct answer. In the first part of this survey, we examine the state of the art by comparing modern approaches to the problem. We classify methods by their mechanism to connect the visual and textual modalities. In particular, we examine the common approach of combining convolutional and recurrent neural networks to map images and questions to a common feature space. We also discuss memory-augmented and modular architectures that interface with structured knowledge bases. In the second part of this survey, we review the datasets available for training and evaluating VQA systems. The various datatsets contain questions at different levels of complexity, which require different capabilities and types of reasoning. We examine in depth the question/answer pairs from the Visual Genome project, and evaluate the relevance of the structured annotations of images with scene graphs for VQA. Finally, we discuss promising future directions for the field, in particular the connection to structured knowledge bases and the use of natural language processing models.

1. Introduction

Visual question answering is a task that was proposed to connect computer vision and natural language processing (NLP), to stimulate research, and push the boundaries of both fields. On the one hand, computer vision studies methods for acquiring, processing, and understanding images. In short, its aim is to teach machines *how to see*. On the other hand, NLP is the field concerned with enabling interactions between computers and humans in natural language, *i.e* teaching machines *how to read*, among other tasks. Both computer vision and NLP belong to the domain of artificial intelligence and they share similar methods rooted in machine learning. However, they have historically developed separately. Both fields have seen significant advances towards their respective goals in the past few decades, and the combined explosive growth of visual and textual data is pushing towards a marriage of efforts from both fields. For example, research in image captioning, *i.e* automatic image description (Donahue et al., 2015; Karpathy et al., 2014; Mao et al., 2015; Vinyals et al., 2014; Wu et al., 2016a; Yao et al., 2015) has produced powerful methods for jointly learning from image and text inputs to form higher-level representations. A successful approach is to combine convolutional neural networks (CNNs), trained on object recognition, with word embeddings, trained on large text corpora.

In the most common form of Visual Question Answering (VQA), the computer is presented with an image and a textual question about this image (see examples in Figs. 3–5). It must then determine the correct answer, typically a few words or a short phrase. Variants include binary (yes/no) (Antol et al., 2015; Zhang et al., 2016) and multiple-choice settings (Antol et al., 2015; Zhu et al., 2016), in which candidate answers are proposed. A closely related task is to "fill in the blank" (Yu et al., 2015), where an affirmation describing the image must be completed with one or several missing words. These affirmations essentially amount to questions phrased in declarative form. A major distinction between VQA and other tasks in computer vision is that the question to be answered is not determined until run time. In traditional problems such as segmentation or object detection, the single question to be answered by an algorithm is predetermined and only the input image changes. In VQA, in contrast, the form that the question will take is unknown, as is the set of operations required to answer it. In this sense, it more closely reflects the challenge of general image understanding. VQA is related to the task of textual question answering, in which the answer is to be found in a specific textual narrative (*i.e* reading comprehension) or in large knowledge bases (*i.e* information retrieval). Textual QA has been studied for a long

* Corresponding author.
E-mail addresses: qi.wu@adelaide.edu.au (Q. Wu), damien.teney@adelaide.edu.au (D. Teney), peng.wang@adelaide.edu.au (P. Wang), chunhua.shen@adelaide.edu.au (C. Shen), anthony.dick@adelaide.edu.au (A. Dick), anton.vandenhengel@adelaide.edu.au (A. van den Hengel).

http://dx.doi.org/10.1016/j.cviu.2017.05.001
1077-3142/

time in the NLP community, and VQA is its extension to additional visual supporting information. The added challenge is significant, as images are much higher dimensional, and typically more noisy than pure text. Moreover, images lack the structure and grammatical rules of language, and there is no direct equivalent to the NLP tools such as syntactic parsers and regular expression matching. Finally, images capture more of the richness of the real world, whereas natural language already represents a higher level of abstraction. For example, compare the phrase '`a red hat`' with the multitude of its representations that one can picture, and in which many styles could not be described in a short sentence.

Visual question answering is a significantly more complex problem than image captioning, as it frequently requires information not present in the image. The type of this extra required information may range from common sense to encyclopedic knowledge about a specific element from the image. In this respect, VQA constitutes a truly AI-complete task (Antol et al., 2015), as it requires multimodal knowledge beyond a single sub-domain. This comforts the increased interest in VQA, as it provides a proxy to evaluate our progress towards AI systems capable of advanced reasoning combined with deep language and image understanding. Note that image understanding could in principle be evaluated equally well through image captioning. Practically however, VQA has the advantage of an easier evaluation metric. Answers typically contain only a few words. The long ground truth image captions are more difficult to compare with predicted ones. Although advanced evaluation metrics have been studied, this is still an open research problem (Hodosh et al., 2013; Li et al., 2011; Vedantam et al., 2015b).

One of the first integrations of vision and language is the "SHRDLU" from system from 1972 (Winograd, 1972) which allowed users to use language to instruct a computer to move various objects around in a "blocks world". More recent attempts at creating conversational robotic agents (Cantrell et al., 2010; Kollar et al., 2013; Matuszek et al., 2012; Roy et al., 2003) are also grounded in the visual world. However, these works were often limited to specific domains and/or on restricted language forms. In comparison, VQA specifically addresses free-form open-ended questions. The increasing interest in VQA is driven by the existence of mature techniques in both computer vision and NLP and the availability of relevant large-scale datasets. Therefore, a large body of literature on VQA has appeared over the last few years. The aim of this survey is to give a comprehensive overview of the field, covering models, datasets, and to suggest promising future directions. To the best of our knowledge, this article is the first survey in the field of VQA.

In the first part of this survey (Section 2), we present a comprehensive review of VQA methods through four categories based on the nature of their main contribution. Incremental contributions means that most methods belong to multiple of these categories (see Table 1). First, the *joint embedding approaches* (Section 2.1) are motivated by the advances of deep neural networks in both computer vision and NLP. They use convolutional and recurrent neural networks (CNNs and RNNs) to learn embeddings of images and sentences in a common feature space. This allows one to subsequently feed them together to a classifier that predicts an answer (Gao et al., 2015; Ma et al., 2016; Malinowski et al., 2015). Second, *attention mechanisms* (Section 2.2) improve on the above method by focusing on specific parts of the input (image and/or question). Attention in VQA (Andreas et al., 2016b; Chen et al., 2015a; Jiang et al., 2015; Xu and Saenko, 2016; Yang et al., 2016; Zhu et al., 2016) was inspired by the success of similar techniques in the context of image captioning (Xu et al., 2015). The main idea is to replace holistic (image-wide) features with spatial feature maps, and to allow interactions between the question and specific regions of these maps. Third, *compositional models* (Section 2.3) allow to tailor the performed computations to each problem instance. For example, Andreas et al. (2016b) use a parser to decompose a given question, then build a neural network out of modules whose composition reflect the structure of the question. Fourth, *knowledge base-enhanced approaches* (Section 2.4) address the use of external data by querying structured knowledge bases. This allows retrieving information that is not present in the common visual datasets such as ImageNet (Deng et al., 2009) or COCO (Lin et al., 2014), which are only labeled with classes, bounding boxes, and/or captions. Information available from knowledge bases ranges from common sense to encyclopedic level, and can be accessed with no need for being available at training time (Wang et al., 2015; Wu et al., 2016c).

In the second part of this survey (Section 3), we examine datasets available for training and evaluating VQA systems. These datasets vary widely along three dimensions: (i) their size, *i.e* the number of images, questions, and different concepts represented. (ii) the amount of required reasoning, *e.g* whether the detection of a single object is sufficient or whether inference is required over multiple facts or concepts, and (iii) how much information beyond that present in the actual images is necessary, be it common sense or subject-specific information. Our review points out that existing datasets lean towards visual-level questions, and require little external knowledge, with few exceptions (Wang et al., 2015; 2016). These characteristics reflect the struggle with simple visual questions still faced by the current state of the art, but these characteristics must not be forgotten when VQA is presented as an AI-complete evaluation proxy. We conclude that more varied and sophisticated datasets will eventually be required.

Another significant contribution of this survey is an in-depth analysis of the question/answer pairs provided in the Visual Genome dataset (Section 4). They constitute the largest VQA dataset available at the time of this writing, and, importantly, it includes rich structured images annotations in the form of scene graphs (Krishna et al., 2017). We evaluate the relevance of these annotations for VQA, by comparing the occurrence of concepts involved in the provided questions, answers, and image annotations. We find out that only about 40% of the answers directly match elements in the scene graphs. We further show that this matching rate can be significantly increased by relating scene graphs to external knowledge bases. We conclude this paper in Section 5 by discussing the potential of better connection to such knowledge bases, together with better use of existing work from the field of NLP.

2. Methods for VQA

One of the first attempts at "open-world" visual question answering was proposed by Malinowski and Fritz (2014). They described a method combining semantic text parsing with image segmentation in a Bayesian formulation that samples from nearest neighbors in the training set. The method requires human-defined predicates, which are inevitably dataset-specific and difficult to scale. It is also very dependent on the accuracy of the image segmentation algorithm and of the estimated image depth information. Another early attempt at VQA by Tu et al. (2014) was based on a joint parse graph from text and videos. In Geman et al. (2015), Geman *et al* proposed an automatic "query generator" that is trained on annotated images and then produces a sequence of binary questions from any given test image. A common characteristic of these early approaches is to restrict questions to predefined forms. The remainder of this article focuses on modern approaches aimed at answering free-form open-ended questions. We will present methods through four categories: joint embedding approaches, attention mechanisms, compositional models, and knowledge base-enhanced approaches. As summarized in Table 1, most methods combine multiple strategies and thus belong to several categories.

Table 1
Overview of existing approaches to VQA, characterized by the use of a joint embedding of image and language features (Section 2.1), the use of an attention mechanism (Section 2.2), an explicitly compositional neural network architecture (Section 2.3), and the use of information from an external structured knowledge base (Section 2.4). We also note whether the output answer is obtained by classification over a predefined set of common words and short phrases, or generated, typically with a recurrent neural network. The last column indicates the type of convolutional network used to obtain image feature.

Method	Joint embedding	Attention mechanism	Compositional model	Knowledge base	Answer class. / gen.	Image features
Neural-Image-QA (Malinowski et al., 2015)	✓				Generation	GoogLeNet (Szegedy et al., 2015)
VIS+LSTM (Ren et al., 2015)	✓				Classification	VGG-Net (Simonyan and Zisserman, 2014)
Multimodal QA (Gao et al., 2015)	✓				Generation	GoogLeNet (Szegedy et al., 2015)
DPPnet (Noh et al., 2016)	✓				Classification	VGG-Net (Simonyan and Zisserman, 2014)
Multimodal-CNN (Ma et al., 2016)	✓				classification	VGG-Net (Simonyan and Zisserman, 2014)
iBOWING (Zhou et al., 2015)	✓				Classification	GoogLeNet (Szegedy et al., 2015)
VQA team (Antol et al., 2015)	✓				classification	VGG-Net (Simonyan and Zisserman, 2014)
Bayesian (Kafle and Kanan, 2016)	✓				Classification	ResNet (He et al., 2016)
DualNet (Saito et al., 2016)	✓				Classification	VGG-Net (Simonyan and Zisserman, 2014) & ResNet (He et al., 2016)
MLP-AQI (Jabri et al., 2016)	✓				Classification	ResNet (He et al., 2016)
MCB (Fukui et al., 2016)	✓				Classification	ResNet (He et al., 2016)
MRN (Kim et al., 2016)	✓	✓			Classification	ResNet (He et al., 2016)
MCB-Att (Fukui et al., 2016)	✓	✓			Classification	ResNet (He et al., 2016)
LSTM-Att (Zhu et al., 2016)	✓	✓			Classification	VGG-Net (Simonyan and Zisserman, 2014)
Com-Mem (Jiang et al., 2015)	✓	✓			Generation	VGG-Net (Simonyan and Zisserman, 2014)
QAM (Chen et al., 2015a)	✓	✓			Classification	VGG-Net (Simonyan and Zisserman, 2014)
SAN (Yang et al., 2016)	✓	✓			Classification	VGG-Net (Simonyan and Zisserman, 2014)
SMem (Xu and Saenko, 2016)	✓	✓			Classification	GoogLeNet (Szegedy et al., 2015)
Region-Sel (Shih et al., 2016)	✓	✓			classification	VGG-Net (Simonyan and Zisserman, 2014)
FDA (Ilievski et al., 2016)	✓	✓			Classification	ResNet (He et al., 2016)
HieCoAtt (Lu et al., 2016)	✓	✓			Classification	ResNet (He et al., 2016)
NMN (Andreas et al., 2016b)		✓	✓		Classification	VGG-Net (Simonyan and Zisserman, 2014)
DMN+ (Xiong et al., 2016)		✓	✓		Classification	VGG-Net (Simonyan and Zisserman, 2014)
Joint-Loss (Noh and Han, 2016)		✓	✓		Classification	ResNet (He et al., 2016)
Attributes-LSTM (Wu et al., 2016a)	✓			✓	Generation	VGG-Net (Simonyan and Zisserman, 2014)
ACK (Wu et al., 2016c)	✓			✓	Generation	VGG-Net (Simonyan and Zisserman, 2014)
Ahab (Wang et al., 2015)				✓	Generation	VGG-Net (Simonyan and Zisserman, 2014)
Facts-VQA (Wang et al., 2016)				✓	Generation	VGG-Net (Simonyan and Zisserman, 2014)
Multimodal KB (Zhu et al., 2015)				✓	Generation	ZeilerNet (Zeiler and Fergus, 2014)

2.1. Joint embedding approaches

Motivation. The concept of jointly embedding images and text was first explored for the task of image captioning (Donahue et al., 2015; Karpathy et al., 2014; Mao et al., 2015; Vinyals et al., 2014; Wu et al., 2016a; Yao et al., 2015). It was motivated by the success of deep learning methods in both computer vision and NLP, which allow one to learn representations in a common feature space. In comparison to the task of image captioning, this motive is further reinforced in VQA by the need to perform further reasoning over both modalities together. A representation in a common space allows learning interactions and performing inference over the question and the image contents. Practically, image representations are obtained with convolutional neural networks (CNNs) pre-trained on object recognition. Text representations are obtained with word embeddings pre-trained on large text corpora. Word embeddings practically map words to a space in which distances reflect semantic similarities (Mikolov et al., 2013; Pennington et al., 2014). The embeddings of the individual words of a question are then typically fed to a recurrent neural network to capture syntactic patterns and handle variable-length sequences.

Methods. Malinowski et al. (2015) propose an approach named "Neural-Image-QA" with a Recurrent Neural Network (RNN) implemented with Long Short-Term Memory cells (LSTMs) (Fig. 1). The motivation behind RNNs is to handle inputs (questions) and outputs (answers) of variable size. Image features are produced by a CNN pre-trained for object recognition. Question and image features are both fed together to a first "encoder" LSTM. It produces a feature vector of fixed-size that is then passed to a second "decoder" LSTM. The decoder produces variable-length answers, one word per recurrent iteration. At each iteration, the last predicted word is fed through the recurrent loop into the LSTM until a special <END> symbol is predicted. Several variants of this approach were proposed. For example, the "VIS+LSTM" of Ren et al. (2015) directly feed the feature vector produced by the encoder LSTM into a classifier to produce single-word answers from a predefined vocabulary. In other words, they formulate the answering as a classification problem, whereas Malinowski et al. (2015) was treating it as a sequence generation procedure. Ren et al. (2015) propose other technical improvements with the "2-VIS+BLSTM" model. It uses two sources of image features as input, fed to the LSTM at the start and at the end of the question sentence. It also uses LSTMs that scan questions in both forward and backward directions. Those bidirectional LSTMs better capture relations between distant words in the question.

Gao et al. (2015) propose a slightly different method named "Multimodal QA" (mQA). It employs LSTMs to encode the question and produce the answer, with two differences from Malinowski et al. (2015). First, whereas Malinowski et al. (2015) used common shared weights between the encoder and decoder LSTMs, mQA learns distinct parameters and only shares the word embedding. This is motivated by potentially different properties (*e.g* in terms of grammar) of questions and answers. Second, the CNN features used as image representations are not fed into the encoder prior to the question, but at every time step.

Noh et al. (2016) tackle VQA by learning a CNN with a dynamic parameter layer (DPPnet) of which the weights are determined adaptively based on the question. For the adaptive parameter prediction, they employ a separate parameter prediction network, which consists of gated recurrent units (GRUs, a variant of LSTMs) taking a question as input and producing candidate weights through a fully-connected layer at its output. This arrangement was shown to significantly improve answering accu-

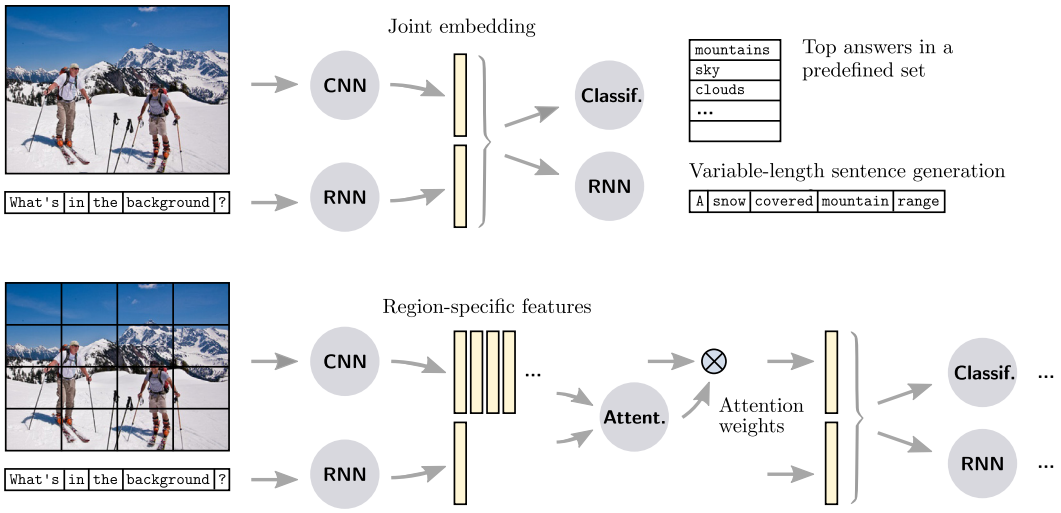


Fig. 1. *(Top)* A common approach to VQA is to map both the input image and question to a common embedding space (Section 2.1). These features are produced by deep convolutional and recurrent neural networks. They are combined in an output stage, which can take the form of a classifier (*e.g* a multilayer perceptron) to predict short answers from predefined set or a recurrent network (*e.g* an LSTM) to produce variable-length phrases. *(Bottom)* Attention mechanisms build up on this basic approach with a spatial selection of image features. Attention weights are derived from both the image and the question and allow the output stage to focus on relevant parts of the image.

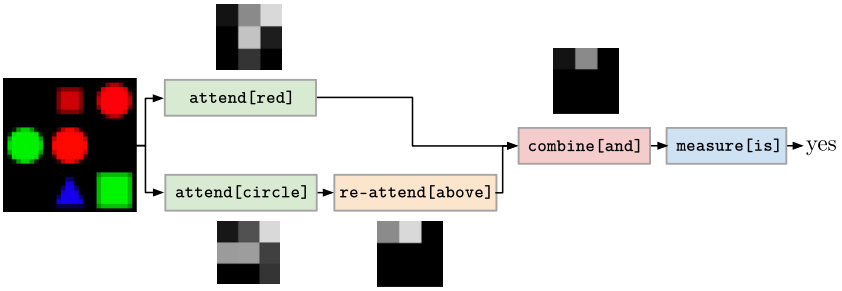


Fig. 2. The Neural Module Networks (NMN, Section 2.3.1) leverage the compositional structure of questions, *e.g* here "Is there a red shape above a circle ?" from the *Shapes* dataset (Section 3.4). The parsing of the question leads to assembling modules that operate in the space of attentions. Two `attend` modules locate red shapes and circles, `re-attend[above]` shifts the attention above the circles, `combine` computes their intersection, and `measure[is]` inspects the final attention and determines that it is non-empty (figure adapted from Andreas et al., 2016b). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

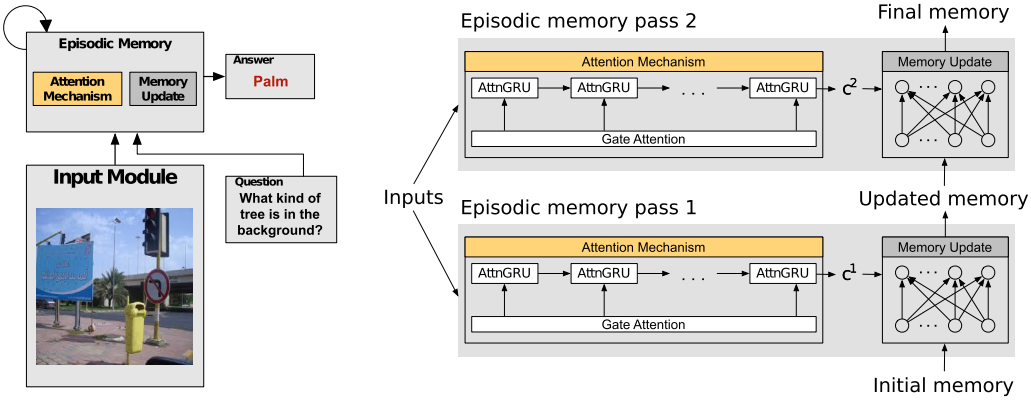


Fig. 3. Dynamic Memory Networks for VQA (figure adapted from Xiong et al., 2016). Overview (left) and details of the episodic memory module with two passes (right).

racy compared to Malinowski et al. (2015) and Ren et al. (2015). One can note a similarity in spirit with the modular approaches of Section 2.3, in the sense that the question is used to tailor the main computations to each particular instance.

Fukui et al. (2016) propose a pooling method to perform the joint embedding visual and text features. They perform their "Multimodal Compact Bilinear pooling" (MCB) by randomly projecting the image and text features to a higher-dimensional space and then convolve both vectors with multiplications in the Fourier space for efficiency. Kim et al. (2016) use a multimodal residual learning framework (MRN) to learn the joint representation of images and language. Saito et al. (2016) propose a "DualNet" which integrates two kinds of operations, namely element-wise summations and element-wise multiplications to embed their visual and textual features. Similarly as Ren et al. (2015) and Noh et al. (2016), they formulate the answering as a classification problem over a predefined set of possible answers. Kafle and Kanan (2016) integrate an explicit prediction of the type of expected answer from

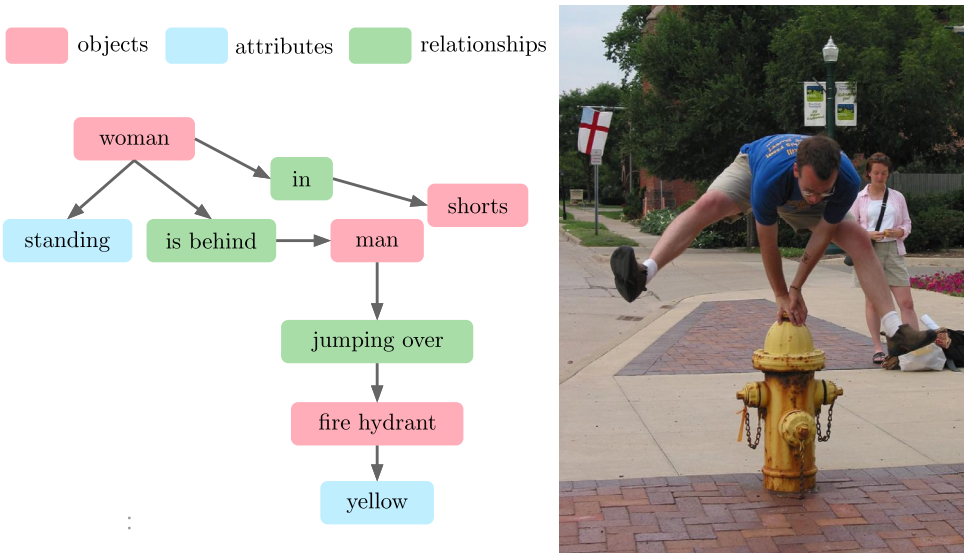


Fig. 4. Example of a scene graph (structured annotation of an image) provided in the Visual Genome dataset (Krishna et al., 2017).

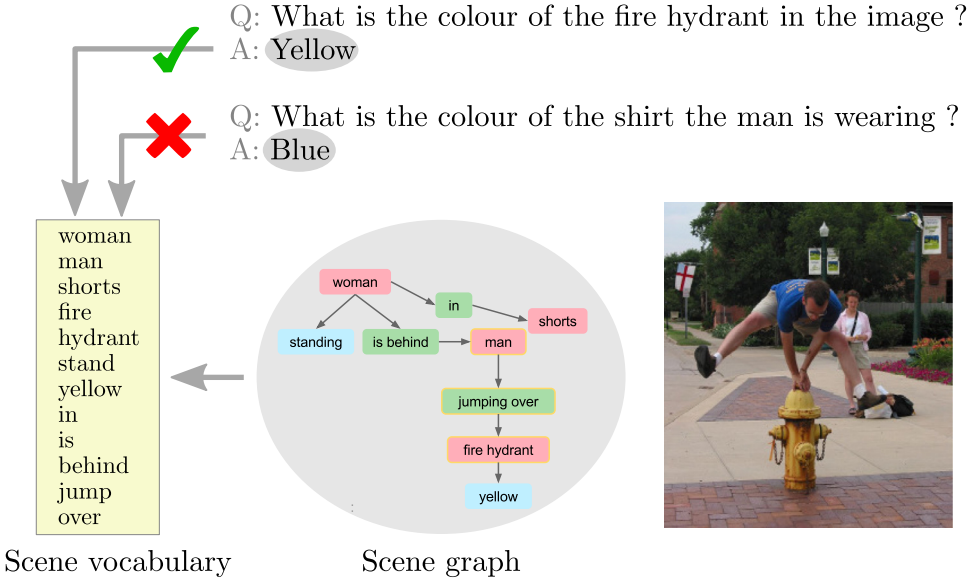


Fig. 5. We verify whether the answer to each question in the Visual Genome dataset can be found from the corresponding scene graph of the scene. Therefore, we first build the vocabulary of each image from the labels of all nodes and edges of its scene graph. Then, for each question, we check whether its answer can be found within the words or combination of words in the vocabulary of the corresponding image.

the question and formulate the answering in a Bayesian framework.

Some other works do not make use of RNNs to encode questions. Ma et al. (2016) use CNNs to process the questions. Features from the image CNN and the text CNN are embedded in a common space through additional layers (a "multimodal CNN") forming an overall homogeneous convolutional architecture. Zhou et al. (2015) and Antol et al. (2015) both use a traditional bag-of-words representation of the questions.

Performance and limitations. We summarize performances of all discussed methods and datasets in Tables 6–9. The "Neural-Image-QA", as one of the earliest introduced methods, is considered the *de facto* baseline result. The "2-VIS+BLSTM" improves slightly on the DAQUAR dataset, mostly thanks to the bidirectional LSTM used to encode the questions. The "mQA" model was unfortunately not tested on publicly available datasets and is therefore not comparable. The DPPnet (Noh et al., 2016) showed significant benefit from tailoring computations adaptively to each question through the dynamic parameter layer. At the time of its publication, it outperformed other joint embeddings methods (Antol et al., 2015; Malinowski et al., 2015; Ren et al., 2015; Zhou et al., 2015). The more recent MCB pooling (Fukui et al., 2016) and multimodal residual learning (MRN) bring further improvements and achieve the top performances at the time of this writing.

The joint embedding approaches are straightforward in their principle and constitute the base of most current approaches to VQA. The latest improvements, exemplified by MCB and MRN, still showed potential room for improvement on both the extraction of features and their projection to the embedding space.

2.2. Attention mechanisms

Motivation. A limitation of most models presented above is to use global (image-wide) features to represent the visual input. This may feed irrelevant or noisy information to the prediction stage.

The aim of attention mechanisms is to address this issue by using local image features, and allowing the model to assign different importance to features from different regions. An early application of attention to visual tasks was proposed in the context of image captioning by Xu et al. (2015). The attentional component of their model identifies salient regions in an image, and further processing then focuses the caption generation on those regions. This concept translates readily to the task of VQA for focusing on image regions relevant to the question. In some respect, the attention process forces an explicit additional step in the reasoning process that identifies "where to look" before performing further computations.

Although attention models were inspired by computational models of human vision, the apparent resemblance with biological systems can be misleading. Attention in artificial neural networks likely helps by allowing additional non-linearities and/or types of interactions (*e.g* multiplicative interaction through attention weights), whereas attention in biological visual systems is more likely a consequence of limited resources such as resolution, field of view, and processing capacity.

Methods. Zhu et al. (2016) described how to add spatial attention to the standard LSTM model. The attention mechanism is introduced by the term z_t , which is a weighted average of convolutional features that depends upon the previous hidden state and the convolutional features:

$$e_t = w_a^T \tanh(W_{he} h_{t-1} + W_{ce} C(I)) + b_a \quad (1)$$

$$a_t = \text{softmax}(e_t) \quad (2)$$

$$z_t = a_t^T C(I) \quad (3)$$

where $C(I)$ represents the convolutional feature map of image I. The *attention term* a_t sets the contribution of each convolutional feature at the tth step. Large values in a_t indicate more relevance of the corresponding region to the question. In this formulation, a standard LSTM can be considered as a special case with values in a_t set uniformly, *i.e* each region contributing equally. A similar mechanism as above was employed by Jiang et al. (2015).

Chen et al. (2015a) use a mechanism different from the above word-guided attention. They generate a "question-guided attention map" (QAM) by searching for visual features that correspond to the semantics of the input question in the spatial image feature map. The search is achieved by convolving the visual feature map with a configurable convolutional kernel. This kernel is generated by transforming the question embeddings from the semantic space into the visual space, which contains the visual information determined by the intent of the question. Yang et al. (2016) also employ this scheme with "stacked attention networks" (SAN) that infer the answer iteratively. Xu and Saenko (2016) propose a "multi-hop image attention scheme" (SMem). The first hop is a word-guided attention, while a second hop is question-guided. In Shih et al. (2016), the authors generate image regions with object proposals and then select regions relevant to the question and possible answer choices. Similarly, Ilievski et al. (2016) employ off-the-shelf object detectors to identify regions related to the key words of the question and then fuse information from those regions with global features with an LSTM. Lu et al. (2016) present a "hierarchical co-attention model" (HieCoAtt) that jointly reasons about image and question attention. Whereas the works described above focus only on visual attention, HieCoAtt processes image and question symmetrically, in the sense that the image representation guides attention over the question and vice versa. Most recently, Fukui et al. (2016) combine the attention mechanism into their "Multimodal Compact Bilinear pooling" (MCB) already mentioned in Section 2.1.

Andreas et al. (2016b) employ attention mechanisms in a different manner. They propose a compositional model that builds a neural network from modules tailored to each question, as described in more details in Section 2.3.1. Most of these modules operate in the space of attentions, either producing an attention map from an image (*i.e* identifying a salient object), performing unary operations (*e.g* inverting the attention from an object to the context around it), or interactions between attentions (*e.g* a subtracting an attention map from another).

Performance and limitations. The reported uses of attention mechanisms always improve over models that use global image features. For example, the authors in Zhu et al. (2016) show that the attention-enhanced LSTM described above outperforms the "VIS+LSTM" model (Ren et al., 2015) in both "Telling" and "Grounding" tasks of the 'Visual7W' dataset (see Section 3.1). The multiple attention layers of SAN (Yang et al., 2016) bring further improvements over only one layer of attention (Chen et al., 2015a; Jiang et al., 2015), especially on the VQA dataset. The HieCoAtt model (Lu et al., 2016) shows benefit from the hierarchical representation of the question and also from the co-attention mechanism (question-guided visual attention and image-guided question attention).

Interestingly, attention mechanisms improve the overall accuracy on all VQA datasets, but closer inspection by question type show little or no benefit on binary (yes/no) questions. One hypothesis is that binary questions typically require longer chains of reasoning, whereas open-ended questions often require identifying and naming only one concept from the image. Therefore, improving on binary questions will likely require other innovations than visual attention. The output in end-to-end joint embedding approaches – regardless of the use of attention – is produced by a simple mapping from the co-embedded visual and textual features to the answer, learned over a large number of training examples. Little insight is available as to how an output answer arises. It can be debated whether any "reasoning" is performed and/or encoded in the mapping. Another important issue is raised by asking whether questions can be answered from the given visual input alone. Oftentimes, they require prior knowledge ranging common sense to subject-specific and even expert-level knowledge. How such information can be provided to VQA systems and incorporated into the reasoning is still an open question (see Section 2.4).

2.3. Compositional models

The methods discussed so far present limitations related to the monolithic nature of the CNNs and RNNs used to extract representations of images and sentences. An increasingly popular research direction in the design of artificial neural networks is to consider modular architectures. This approach involves connecting distinct modules designed for specific desired capabilities such as memory or specific types of reasoning. A potential advantage is a better use of supervision. On the one hand, it facilitates transfer learning, since a same module can be used and trained within different overall architectures and tasks (see Section 2.3.1). On the other hand, it allows to use "deep supervision", *i.e* optimizing an objective that depends on the outputs of internal modules (*e.g* which supporting facts an attention mechanism should focus on Weston et al., 2014). Other models discussed in Sections 2.2 (attention models) and 2.4 (connections to knowledge bases) also fall in the category of modular architectures. We focus here on two specific models whose main contribution is in the modular aspect, namely the Neural Module Networks (NMN) and the Dynamic Memory Networks (DMN).

2.3.1. Neural module networks

Motivation. The Neural Module Networks (NMN) are introduced by in Andreas et al. (2016b) and extended in Andreas et al. (2016a). They are specifically designed for VQA, with the intention of exploiting the compositional linguistic structure of the questions. Questions vary greatly in the level of complexity. For example, *Is this a truck ?* only requires retrieving one piece of information from the image, whereas *How many objects are to the left of the toaster ?* requires multiple processing steps, such as recognition and counting. NMNs reflect the complexity of a question in a network that is assembled on-the-fly for each instance of the problem. The tactic is related to approaches in textual QA (Liang et al., 2013) that use semantic parsers to turn questions into logical expressions. A significant contribution of NMNs is to apply this logical reasoning over continuous visual features, instead of discrete or logical predicates.

Method. The method is based on a semantic parsing of the question using a well-known tool in the NLP community. The parse tree is turned into an assembly of modules from a predefined set, which are then used together to answer the question. Crucially, all modules are independent and composable (Fig. 2). In other words, the computation performed will be different for each problem instance, and a problem instance at test time may use a set of modules that were not seen together during training.

The inputs and outputs of the modules can be of three types: image features, attentions (regions) over images, and labels (classification decisions). A set of possible modules is predefined, each by its type of input and output, but their exact behavior will be acquired through end-to-end training on specific problem instances. The training therefore does not need additional supervision than triples of images, questions, and answers.

The parsing of the question is a crucial step, which is performed with the Standford dependency parser (de Marneffe and Manning, 2008) which basically identifies grammatical relations between parts of the sentence. The authors of the NMNs then use ad hoc hand-written rules to deterministically transform parse trees into structured queries, in the form of compositions of modules (Andreas et al., 2016b). In their second paper (Andreas et al., 2016a), they additionally learn a ranking function to select the best structures from candidate parses. The whole procedure still uses strong simplifying assumptions about language in the question. The visual features are provided by a fixed, pre-trained VGG CNN (Simonyan and Zisserman, 2014).

Performance and limitations. The Neural Module Networks were evaluated on the VQA benchmark, and shows different strengths and weaknesses than competing approaches. It generally outperforms competitors on questions with a compositional structure, *e.g* requiring an object to be located and one of its attributes described. However, many of questions in the VQA dataset are quite simple, and require little composition or reasoning. The authors introduced a new dataset, named "Shapes", of synthetic images (Section 3.4) paired with complex questions involving spatial relations, set-theoretic reasoning, and shape and attribute recognition.

The limitations of the method are inherent to the bottleneck formed during the parsing of the question. This stage fixes the network structure and errors cannot be recovered from. Moreover, the assembly of modules uses aggressive simplification of the questions that discards some grammatical cues. As a workaround, the authors obtain the final answer by averaging their output with the one from a classical LSTM question encoder.

The potential of the NMNs is dimmed in practice by the lack of truly complex questions in the VQA benchmark. The results reported on this dataset use a restricted subset of possbible modules, presumably to avoid over-fitting. Results on the synthetic Shapes dataset show that semantic structure prediction does improve generalization in deep networks. The overall approach presents the potential of addressing the combinatorial explosion of concepts and relations that can arise in open-world VQA. Finally, note that the general formulation of NMNs can encompass other approaches, including the memory networks presented below, which may be formulated as a composition of "attention" and "classifier" modules.

2.3.2. Dynamic memory networks

Motivation. The Dynamic Memory Networks (DMN) are neural networks with a particular modular architecture. They were described in Kumar et al. (2016), with a number of variants proposed concurrently (Bordes et al., 2015; Peng et al., 2015; Sukhbaatar et al., 2015; Weston et al., 2014). Most of these were applied to textual QA. We focus here on the work of Xiong et al. (2016), who adapted them to VQA. DMNs fall into the broader category of memory-augmented networks, which perform read and write operations on an internal representation of the input. This mechanism, similarly to attention (see Section 2.2), is designed to address tasks that require complex logical reasoning by modeling interaction between multiple parts of the data over several passes.

Method. The dynamic memory networks are composed of 4 main modules (Kumar et al., 2016) that allow independence in their particular implementation (see Fig. 3). The input module transforms the input data into a set of vectors called "facts". Its implementation (described below) varies depending on the type of input data. The question module computes a vector representation of the question, using a gated recurrent unit (GRU, a variant of LSTM). The episodic memory module retrieves the facts required to answer the question. The key is to allow the episodic memory module to pass multiple times over the facts to allow transitive reasoning. It incorporates an attention mechanism that selects relevant facts and an update mechanism that generates a new memory representation from interactions between its current state and the retrieved facts. The first state is initialized with the representation from the question module. Finally, the answer module uses the final state of the memory and the question to predict the output with a multinomial classification for single words, or a GRU for datasets where a longer sentence is required.

The input module for VQA (Xiong et al., 2016) extracts image features with a VGG CNN (Simonyan and Zisserman, 2014) over small image patches. These features are fed to a GRU in the manner of the words of a sentence, traversing the image in a snake-like fashion. This is an ad hoc adaptation of the original input module in Kumar et al. (2016) that used a GRU to process words of sentences. The episodic memory module also includes an attention mechanism to focus on particular image regions.

Let us also mention here the work of Noh and Han (2016). Their approach has similarities with memory networks in the use of an internal memory-like unit, over which multiple passes are performed. The main novelty is the use of a loss over each of these passes, instead of a single one on the final results. After training, the inference at test time is performed using only one such pass.

Performance and limitations. The dynamic memory networks were evaluated on the DAQUAR and VQA benchmarks, and show competitive performance for all types of questions. Compared to Neural Module Networks (Section 2.3.1), they perform similarly on yes/no questions, slightly worse on numerical questions, but markedly better on all other types of questions. The issue with counting likely arises from the limited granularity of the fixed image patches, which may cross object boundaries.

Interestingly, the paper presents competitive results on both VQA and text-based QA (Xiong et al., 2016) using essentially a

same method, except for the input module. The text QA dataset used (Weston et al., 2015) requires inference over multiple facts, which is a positive indicator of the reasoning capabilities of this model. A potential criticism for applying a same model to text and images stems from the intrinsically different nature of sequences of words, and sequences of image patches. The temporal dimension of a textual narrative is different than relative geometrical positions, although both seem to be handled adequately by GRUs in practice.

2.4. Models using external knowledge bases

Motivation. The task of VQA involves understanding the contents of images, but often requires prior non-visual, information, which can range from "common sense" to topic-specific or even encyclopedic knowledge. For example, to answer the question "How many mammals appear in this image ?", one must understand the word "mammal" and know which animals belong to this category. This observation allows pinpointing two major weaknesses of the joint embedding approaches (Section 2.1). First, they can only capture knowledge that is present in the training set, and it is obvious than efforts at scaling up datasets will never reach a complete coverage of the real world. Second, the neural networks trained in such approaches have a limited capacity, which is also inevitably deemed to be surpassed by the amount of information we wish to learn.

An alternative is to decouple the reasoning (*e.g* as a neural network) from the actual storage of data or knowledge. A substantial amount of research has been devoted to structured representations of knowledge. This led to the development of large-scale Knowledge Bases (KB) such as DBpedia (Auer et al., 2007), Freebase (Bollacker et al., 2008), YAGO (Hoffart et al., 2013; Mahdisoltani et al., 2015), OpenIE (Banko et al., 2007; Etzioni et al., 2011; Fader et al., 2011), NELL (Carlson et al., 2010), WebChild (Tandon et al., 2014a; 2014b), and ConceptNet (Liu and Singh, 2004). These databases store common sense and factual knowledge in a machine readable fashion. Each piece of knowledge, referred to as a fact, is typically represented as a triple `(arg1,rel,arg2)`, where `arg1` and `arg2` represent two concepts and `rel` represents a relationship between them. The collection of such facts forms a interlinked graph, which is often described according to a Resource Description Framework (RDF Group et al., 2014) specification and can be accessed by query languages such as SPARQL (Prud'Hommeaux et al., 2008). Linking such knowledge bases to VQA methods allows separating the reasoning from the representation of prior knowledge in a practical and scalable manner.

Method. Wang et al. (2015) propose a VQA framework named "Ahab" that uses DBpedia, one of the largest structured knowledge bases. Visual concepts are first extracted from the given image with CNNs, and they are then associated with nodes from DBpedia that represent similar concepts. Whereas the joint embedding approaches (Section 2.1) learn a mapping from images/questions to answers, the authors propose here to learn a mapping images/questions to *queries* over the constructed knowledge graph. The final answer is obtained by summarizing the results of this query. The main limitation of Wang et al. (2015) is to handle limited types of questions. Although the questions can be provided in natual language, they are parsed using manually designed templates. An improved method named FVQA (Wang et al., 2016) uses an LSTM and a data-driven approach to learn the mapping of images/questions to queries. This work also uses two additional knowledge bases, ConceptNet and WebChild.

An interesting byproduct of the explicit representation of knowledge is that the above methods can indicate how they arrived to the answer by providing the chain of reasoning (Wang et al., 2015) or the supporting facts (Wang et al., 2016) used in the inference process. This contrasts with monolithic neural networks which provide little insight into the computations performed to produce their final answer.

Wu et al. (2016c) proposed a joint embedding approach that also benefits from external knowledge bases. Given an image, they first extract semantic attributes with a CNN. External knowledge related to these attributes is then retrieved from a version of DBpedia containing short descriptions, which are embedded into fixed-size vectors with Doc2Vec. The embedded vectors are fed into an LSTM model that interprets them with the question and finally generates an answer. This method still learns a mapping from questions to answers (as other joint embedding methods) and cannot provide information about the reasoning process.

Performance and limitations. Both Ahab and FVQA focus specifically on visual questions requiring external knowledge. Most existing VQA datasets include a majority of questions that require little amount of prior knowledge, and performance on these datasets thus poorly reflect the particular capabilities of these methods. The authors of those two methods thus include evaluation on new small-scale datasets (see Section 3.3). Ahab (Wang et al., 2015) significantly outperforms joint embedding approaches on its KB-VQA dataset (Wang et al., 2015) in terms of overall accuracy (69.6% *vs* 44.5%). In particular, Ahab becomes significantly better than joint embedding approaches on visual questions requiring a higher level of knowledge. Similarly, the FVQA approach (Wang et al., 2016) also performs much better than conventional approaches (Wang et al., 2016) in terms of overall top-1 accuracy (58.19% *vs* 23.37%). An issue in the evaluation of both of these methods is the limited number of question types and the small scale of the datasets.

The approach of Wu et al. (2016c) is evaluated on the Toronto COCO-QA and VQA datasets, and shows an advantage in using the external KB in terms of average accuracy.

3. Datasets and evaluation

A number of datasets have been proposed specifically for research on VQA. They contain, at the minimum, triples made of an image, a question, and its correct answer. Additional annotations are sometimes provided, such as image captions, image regions supporting the answers, or multiple-choice candidate answers. Datasets and questions within the datasets vary widely in their complexity, the amount of reasoning and of non-visual (*e.g* "common sense") information required to infer the correct answer. This section contains a comprehensive comparison of the available datasets and discusses their suitability for evaluating different aspects of VQA systems.

One major characteristic that differentiates the various datasets is the type of their images, which we broadly classify into natural, clip art, and synthetic. The most widely used datasets such as DAQUAR (Malinowski and Fritz, 2014), COCO-QA (Ren et al., 2015) and VQA-real (Antol et al., 2015) all use natural (*i.e* real) images. The VQA-abstract and its balanced version (Zhang et al., 2016) are based on clip art (*i.e* cartoon) images. A second key difference between datasets is the question-answer format: open-ended vs multiple-choice. The former means there is no pre-defined set of answers, and is most common (a used for DAQUAR Malinowski and Fritz, 2014, COCO-QA Ren et al., 2015, FM-IQA Gao et al., 2015, Visual Genome Krishna et al., 2017). The multiple-choice setting provides a limited set of possible answers with each question, and is used for example in the Visual Madlibs (Yu et al., 2015). The VQA-real (Antol et al., 2015) and the Visual7W (Zhu et al., 2016) datasets each allow evaluation with either open-ended or multiple-choice questions. Results from the two settings cannot be compared, and the open-ended setting is considered more challenging, while harder to quantitatively evaluate. Most authors

address the VQA-real dataset in the open-ended setting, while the authors of the Visual7W conversely recommend the multiple-choice setting for a more interpretable evaluation.

A single dataset is usually used both for training and evaluating a VQA system. The different nature and characteristics of the datasets make their combined use non-trivial. Some recent works (Jabri et al., 2016) present positive results from training on multiple datasets. General advances in transfer learning may make this a promising avenue for future research on VQA. Another direction for scaling up VQA is to use other sources of information. The KB-VQA (Wang et al., 2015) and FVQA (Wang et al., 2016) datasets specifically address this aspect through annotations of supporting facts in structured non-visual knowledge bases (Section 3.3).

Other key differences between datasets concern their size, the type of their questions, the distribution of the lengths of the questions, their collection procedure (human input vs automatic generation), and the recommended evaluation metrics. These differences between datasets are summarized in Table 2. See Figs. 3–5 for examples from various datasets.

Other datasets non-specific to VQA are worth mentioning. They target other tasks involving vision and language, such as image captioning (*e.g* Chen et al., 2015b; Hodosh et al., 2013; Lin et al., 2014; Young et al., 2014), generating (Mao et al., 2016) and understanding (Hu et al., 2016; Kazemzadeh et al., 2014) referring expressions for retrieval of images and objects in natural language. Those datasets go beyond the scope of this article (see Ferraro et al., 2015 for a review), but are a potential source of additional training data for VQA since they combine images with textual annotations.

3.1. Datasets of natural images

An early effort at compiling a dataset specifically for VQA was presented by Geman et al. (2015). The dataset comprises questions generated from templates from a fixed vocabulary of objects, attributes, and relationships between objects. Another early dataset was presented in Tu et al. (2014). They study the joint parsing of videos and text to answer queries, and consider two datasets containing 15 video clips each. These two examples are restricted to limited settings and are of relatively small size. We discuss below the open-world large-scale datasets in use today.

DAQUAR. The first VQA dataset designed as benchmark is the DAQUAR, for DAtaset for QUestion Answering on Real-world images (Malinowski and Fritz, 2014). It was built with images from the NYU-Depth v2 dataset (Silberman et al., 2012), which contains 1449 RGBD images of indoor scenes, together with annotated semantic segmentations. The images of DAQUAR are split to 795 training and 654 test images. Two types of question/answer pairs are collected. First, *synthetic questions/answers* are generated automatically using 8 predefined templates and the existing annotations of the NYU dataset. Second, *human questions/answers* are collected from 5 annotators. They were instructed to focus on basic colors, numbers, objects (894 categories), and sets of those. Overall, 12,468 question/answer pairs were collected, of which 6794 are to be used for training and 5674 for testing. The large size of DAQUAR was key to enable the development and training of the early methods for VQA with deep neural networks (Ma et al., 2016; Malinowski et al., 2015; Ren et al., 2015). The main disadvantage of DAQUAR is the restriction of answers to a predefined set of 16 colors and 894 object categories. The dataset also presents strong biases showing that humans tend to focus on a few prominent objects, such as tables and chairs (Malinowski and Fritz, 2014).

COCO-QA. The COCO-QA dataset (Ren et al., 2015) represents a substantial effort to increase the scale of training data for VQA.

Table 2
Major datasets for VQA and their main characteristics. See Section 3.1 for discussion.

Dataset	Source of Images	Number of Images	Number of Questions	Num. questions / Num. images	Num. question Categories	Question Collection	Average quest. Length	Average ans. Length	Evaluation Metrics
DAQUAR (Malinowski and Fritz, 2014)	NYU-Depth V2	1449	12,468	8.6	4	Human	11.5	1.2	Acc. & WUPS
COCO-QA (Ren et al., 2015)	COCO	117,684	117,684	1.0	4	Automatic	8.6	1.0	Acc. & WUPS
FM-IQA (Gao et al., 2015)	COCO	120,360	–	–	–	Human	–	–	Human
VQA-real (Antol et al., 2015)	COCO	204,721	614,163	3.0	20+	Human	6.2	1.1	Acc. against 10 humans
Visual Genome (Krishna et al., 2017)	COCO	108,000	1,445,322	13.4	7	Human	5.7	1.8	Acc.
Visual7W (Zhu et al., 2016)	COCO	47,300	327,939	6.9	7	Human	6.9	1.1	Acc.
Visual Madlibs (Yu et al., 2015)	COCO	10,738	360,001	33.5	12	Human	6.9	2.0	Acc.
VQA-abstract (Antol et al., 2015)	Clipart	50,000	150,000	3.0	20+	Human	6.2	1.1	Acc.
VQA-balanced (Zhang et al., 2016)	Clipart	15,623	33,379	2.1	1	Human	6.2	1.0	Acc.
KB-VQA (Wang et al., 2015)	COCO	700	2402	3.4	23	Human	6.8	2.0	Human
FVQA (Wang et al., 2016)	COCO & ImageNet	1906	4608	2.5	12	Human	9.7	1.2	Acc.

This dataset uses images from the Microsoft Common Objects in Context data (COCO) dataset (Lin et al., 2014). COCO-QA includes 123,287 images (72,783 for training and 38,948 for testing) and each image has one question/answer pair. They were automatically generated by turning the image descriptions part of the original COCO dataset into question/answer form. The questions are categorized into four types based on the type of expected answer: object, number, color, and location. A side-effect of the automatic conversion of captions is a high repetition rate of the questions. Among the 38,948 questions of the test set, 9072 (23.29%) of them also appear as training questions.

FM-IQA. The FM-IQA (Freestyle Multilingual Image Question Answering) dataset (Gao et al., 2015) uses 123,287 images, also sourced from the COCO dataset. The difference with COCO-QA is that the questions/answers are provided here by humans through the Amazon Mechanical Turk crowd-sourcing platform. The annotators were free to give any type of questions, as long as they relate to the contents of each given image. This lead to a much greater diversity of questions than in previously-available datasets. Answering the questions typically requires both understanding the visual contents of the image and incorporating prior "common sense" information. The dataset contains 120,360 images and 250,560 question/answer pairs, which were originally provided in Chinese, then converted into English by human translators.

VQA-real. One of the most widely used dataset comes from the VQA team at Virginia Tech, commonly referred to simply as VQA (Antol et al., 2015). It comprises two parts, one using natural images named VQA-real, and a second one with cartoon images named VQA-abstract, which we will discuss in Section 3.2. VQA-real comprises 123,287 training and 81,434 test images, respectively, sourced from COCO (Lin et al., 2014). Human annotators were encouraged to provide interesting and diverse questions. In contrast to the datasets mentioned above, binary (*i.e* yes/no) questions were also allowed. The dataset also allows evaluation in a multiple-choice setting, by providing 17 additional (incorrect) candidate answers for each question. Overall, it contains 614,163 questions, each having 10 answers from 10 different annotators. The authors performed a very detailed analysis of the dataset (Antol et al., 2015) in terms of questions types, question/answer lengths, *etc*. They also conducted a study to investigate whether questions required prior non-visual knowledge, judged by polling humans. A majority of subjects (at least 6 out of 10) estimated that common sense was required for 18% of the questions. Only 5.5% of the questions were estimated to require adult-level knowledge. These modest figures show that little more than purely visual information is required to answer most questions.

Visual Genome and Visual7W. The Visual Genome QA dataset (Krishna et al., 2017) is, at the time of this writing, the largest available dataset for VQA, with 1.7 million question/answer pairs. It is built with images from the Visual Genome project (Krishna et al., 2017), which includes unique structured annotations of scene contents in the form of scene graphs. These scene graphs describe the visual elements of the scenes, their attributes, and relationships between them. Human subjects provided questions that must start with one of the 'seven Ws', *i.e* who, what, where, when, why, how, and which (the 'which' questions have not been released at the time of writing this paper). The questions must also relate to the image so as to be clearly answerable if and only if the image is shown. Two types of questions are considered: free-form and region-based. In the free-form setting, the annotator is shown an image and asked to provide 8 question/answer pairs. To encourage diversity, the annotator is forced to use 3 different "Ws" out of the 7 mentioned above. In the region-based setting, the annotator must provide questions/answers related to a specific, given region of the image. The diversity of the answers in the Visual Genome is larger than in VQA-real (Antol et al., 2015), as shown by the top-1000 most frequent answers only covering about 64% of the correct answers. In VQA-real, the corresponding top-1000 answers cover as much as 80% of the test set answers. A major advantage of the Visual Genome dataset for VQA is the potential for using the structured scene annotations, which we examine further in Section 4. The use of this information to help designing and training VQA systems is however still an open research question.

The Visual7w (Zhu et al., 2016) dataset is a subset of the Visual Genome that contains additional annotations. The questions are evaluated in a multiple-choice setting, each question being provided with 4 candidate answers, of which only one is correct. In addition, all the objects mentioned in the questions are visually grounded, *i.e* associated with bounding boxes of their depictions in the images.

Visual Madlibs. The Visual Madlibs dataset (Yu et al., 2015) is designed to evaluate systems on a "fill in the blank" task. The objective is to determine words to complete an affirmation that describes a given image. For example, the description "two ___ are playing ___ in the park", provided along a corresponding image, may have to be filled in with "men" and "frisbee". These sentences essentially amount to questions phrased in declarative form, and most VQA systems could be easily adapted to this setting. The dataset comprises 10,738 images from COCO (Lin et al., 2014) and 360,001 focused natural language descriptions. Incomplete sentences were generated automatically from these descriptions using templates. Both open-ended and multiple-choice evaluation are possible.

Evaluation measures. The evaluation of computer-generated natural language sentences is an inherently complex task. Both the syntactic (grammatical) and semantic correctness should be taken into account. Comparing generated with ground truth sentences is akin to evaluating paraphrases, which is still an open research problem studied in the NLP community. Most datasets for VQA allow to bypass this issue by restricting answers to single words or short phrases, typically of 1 to 3 words. This allows automatic evaluation, and limits ambiguities during annotation since it forces questions and answers to be more specific. (Tables 3, 5, 7,8 and 10)

The seminal paper of Malinowski and Fritz (2014) proposed two evaluation metrics for VQA. The first is to simply measure the accuracy with respect to the ground truth using string matching. Only exact matches are considered as correct. The second uses the Wu-Palmer similarity (WUPS) (Wu and Palmer, 1994) which evaluates the similarity between their common subsequence in a taxonomy tree. The candidate answer is considered as correct when the similarity between two words exceeds a threshold. In Malinowski and Fritz (2014), the metric is evaluated against two thresholds, 0.9 and 0.0. Gao et al. (2015) conduct an actual *Visual Turing Test* using human judges. Subjects are presented with an image, a question and a candidate answer, from either a VQA system or another human. He or she then needs to determine, based on the answer, whether it was more likely to have been generated by a human (*i.e* pass the test) or a machine (*i.e* fail the test). They also rate each candidate answer with a score.

The VQA-real dataset (Antol et al., 2015) recognize the issue of ambiguous questions and collect, for each question, 10 ground truth answers from 10 different subjects. Evaluation on this dataset must compare a generated answer with these 10 human-generated ones as follows:

$$\text{accuracy} = \min\left(\frac{\#\text{humans provided that answer}}{3}, 1\right) \quad (4)$$

Table 3
Examples from datasets of natural images. The questions in different datasets span a wide range of complexity, involving purely visual attributes and single objects, or more complex relations, actions, and global scene structure. Note that in "VQA-real" (Antol et al., 2015), every question is provided with 10 answers, each proposed by a different human annotator.

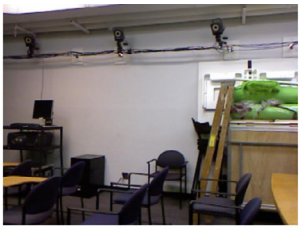
DAQUAR [34]			
	Q: How many white objects in this picture ? A: 9	Q: What color is the chair in front of the wall on the left side of the stacked chairs ? A: blue	Q: What is the largest white object on the left side of the picture ? A: printer
COCO-QA [38]			
	Q: How many giraffes walking near a hut in an enclosure ? A: two	Q: What is the color of the bus ? A: yellow	Q: What next to darkened display with telltale blue ? A: keyboard
VQA-real [7]			
	Q: What shape is the bench seat ? A: oval, semi circle, curved, curved, double curve, banana, curved, wavy, twisting, curved	Q: What color is the stripe on the train ? A: white, white, white, white, white, white, white, white, white, white	Q: Where are the magazines in this picture ? A: On stool, stool, on stool, on bar stool, on table, stool, on stool, on chair, on bar stool, stool
Visual Genome [33]			
	Q: What color is the clock ? A: Green	Q: What is the woman doing ? A: Sitting	Q: How is the ground ? A: dry

Table 4
Examples from the "VQA-abstract" (Antol et al., 2015) dataset of clip art images. Every question is provided with 10 answers, each proposed by a different human annotator.

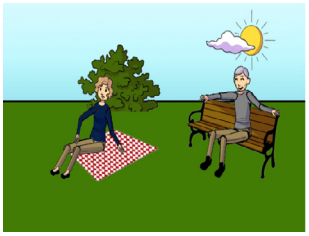
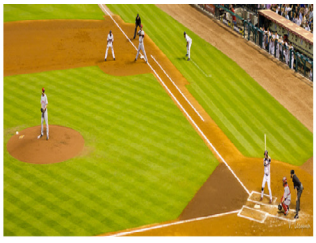
VQA-abstract [7]			
	Q: Who looks happier ?. A: old person, man, man, man, old man, man, man, man, grandpa	Q: Where are the flowers ? A: near tree, tree, around tree, tree, by tree, around tree, around tree, grass, beneath tree, base of tree	Q: How many pillows ? A: 1, 2, 2, 2, 2, 2, 2, 2, 2, 2

Table 5
Examples from knowledge base-enhanced datasets.

KB-VQA [31]			
	Q: Tell me the common property of the animal in this image and elephant. A: mammal, animals in Africa	Q: List all equipment I might use to play this sport. A: baseball bat, baseball, baseball glove, baseball field	Q: Is the image related to tourism ? A: yes
FVQA [32]			
	Q: What things in this image are eatable ? A: Apples	Q: What is the order of the animal described in this image ? A: Odd toed ungulate	Q: What thing in this image is helpful for a romantic dinner ? A: Wine

In other words, an answer is deemed 100% accurate if at least 3 annotators provided that exact answer.

Other datasets such as Krishna et al. (2017) and Zhu et al. (2016) simply measure accuracy through the ratio of exact matches between predictions and answers, which is sensible when answers are short and therefore mostly unambiguous. Evaluation in a multiple-choice setting (*e.g* Yu et al., 2015) is straightforward. It makes the task of VQA easier by constraining the output space to a few discrete points, and it eliminates any artifact in the evaluation that could arise from a chosen metric.

Results of existing methods. Most modern methods for VQA have been evaluated on the VQA-real (Antol et al., 2015), DAQUAR (Malinowski and Fritz, 2014), and COCO-QA (Ren et al., 2015) datasets. We summarize results on these three main datasets in Tables 6–9.

3.2. Datasets of clipart images

This section discusses datasets of synthetic images created manually from clipart illustrations. They are often referred to as

"abstract scenes" (Antol et al., 2015), although this denomination is confusing since they supposedly depict *realistic* situations, albeit in minimalistic representations. Such "cartoon" images allow studying connections between vision and language by focusing on high-level semantics rather than on the visual recognition. This type of images has been used before for capturing common sense (Fouhey and Zitnick, 2014; Lin and Parikh, 2015; Vedantam et al., 2015a), learning models of interactions between people (Antol et al., 2014), generating scenes from natural language descriptions (Zitnick et al., 2013), and learning the semantic relevance of visual features (Zitnick and Parikh, 2013; Zitnick et al., 2016).

VQA abstract scenes. The VQA benchmark (Section 3.1) contains clipart scenes with questions/answer pairs as a separate and complimentary set to the real images. The aim is to enable research focused on high-level reasoning, removing the need to parse real images. As such, the scenes are provided as structured (XML) descriptions, in addition to the actual images. The scenes were created manually. Annotators were instructed to represent realistic situations through a drag-and-drop interface. Two types of scenes are possible, indoor and outdoor, each allowing a different set of elements, including animals, objects, and humans with adjustable poses. A total of 50,000 scenes were generated, and 3 questions per scene (*i.e* a total of 150,000 questions) were collected, in a similar manner as for the real images of the VQA dataset (Section 3.1). Each question was answered by 10 subjects who also provided a confidence score. Questions are labeled with an answer type: "yes/no", "number", and "other". Interestingly, the distribution of question lengths and question types (based on the first four words of the questions) is similar to those of real images. However, the number of unique one-word answers is significantly lower (3,770 vs 23,234), reflecting the smaller variations and limited set of objects in the scenes. Ambiguity in the ground truth answers is also lower with abstract scenes, as reflected by a better inter-human agreement (87.5% vs 83.3%). Results on these abstract scenes have so far only reported in Antol et al. (2015) and Zhang et al. (2016).

Balanced dataset. Another version of the dataset discussed above is presented in Zhang et al. (2016). Most VQA datasets present strong biases such that a language-only "blind" model (*i.e* using no visual input) can often guess correct answers. This seriously hampers the original objective of VQA of acting as a proxy to evaluate deep image understanding. Synthetic scenes allow better control over the distribution in the dataset. The authors in Zhang et al. (2016) balance the existing abstract binary VQA dataset (discussed above) with additional complementary scenes so that each question has both "yes" and "no" answers for two very similar scenes.

As examples on strong biases can be in the VQA dataset (Antol et al., 2015), any question starting with "What sport is" can be answered correctly with "tennis" 41% of the time. Similarly, "What color are the" is answered correctly with "white" 23% of the time (Zhang et al., 2016). Overall, half of all questions can be answered correctly by a blind neural network, *i.e* using the question alone. This rises to more than 78% for the binary questions.

The resulting balanced dataset contains 10,295 and 5328 pairs of complementary scenes for the training and test set respectively. Evaluation should use the VQA evaluation metric (Antol et al., 2015). Results were reported using combinations of balanced and unbalanced training and test sets (Zhang et al., 2016), of which we summarize the interesting observations. First, when testing on unbalanced data (*i.e* the setting of prior work), it is better to train on similarly unbalanced, so as to learn and exploit dataset biases (*e.g* that 69% of answers are "yes"). Second, testing on the new balanced data, it is now better to train on similarly balanced data. It forces models to use visual information, being unable to exploit language biases in the training set. In this setting, blind models perform, as expected, close to chance. The particular model evaluated is a method for visual verification that relies on language parsing and a number of hand designed rules. The authors also provide results in an even harder form, where a prediction is considered correct only when the model can answer correctly both versions (with yes and no answers) of a scene. In this setting, a language-only model gives zero performance, and this arguably constitutes one of the most rigorous metrics to quantify actual deep scene understanding.

One criticism of forcing the removal of biases in a dataset supposedly depicting realistic scenes is that it artificially shifts the distribution away from the real world. Statistical biases that appear in datasets reflect those inherent to the world, and it is arguable how much these should enter the learning process of a general VQA system.

3.3. Knowledge base-enhanced datasets

The datasets discussed above contain various ratio of purely visual questions, and questions that require external knowledge. For example, most questions in DAQUAR (Malinowski and Fritz, 2014) are purely visual in nature, referring to colors, numbers, and physical locations of objects. In the COCO-QA dataset (Ren et al., 2015), questions are generated automatically from image captions which describe the major visual elements of the image. In the VQA dataset (Antol et al., 2015), 5.5% of questions require "adult-level common sense", but none require "knowledge base-level" knowledge (Wang et al., 2015). We discussed in Section 2.4 methods for VQA that make use of external knowledge bases. The authors of two such methods (Wang et al., 2015; 2016) proposed two datasets that allow highlighting this particular capability. The scope of these datasets is different than the general-purpose VQA datasets discussed above, and they are also smaller in scale.

KB-VQA. The KB-VQA dataset (Wang et al., 2015) was constructed to evaluate the performance of the Ahab VQA system (Wang et al., 2015). It contains questions requiring topic-specific knowledge that is present in DBpedia. 700 images were selected from the COCO image dataset (Lin et al., 2014) and 3 to 5 question/answer pairs were collected for each, for a total of 2402 questions. Each question follows one of 23 predefined templates. The questions require different levels of knowledge, from common sense to encyclopedic knowledge.

FVQA. The FVQA dataset (Wang et al., 2016) contains only questions which involve external (non-visual) information. It was designed to include additional annotations to ease the supervised training of methods using knowledge bases. In contrast with most VQA datasets (Antol et al., 2015; Gao et al., 2015; Ren et al., 2015; Zhu et al., 2016) which only provide question/answer pairs, FVQA includes, with each question/answer, a supporting fact. These facts are represented as triple `(arg1,rel,arg2)`. For example, consider the question/answer "Why are these people wearing yellow jackets ? For Safety". It will include the supporting fact `(wearing bright clothes,aids,safety)`. To collect this dataset, a large number of such facts (triples) related to visual concepts were extracted from the knowledge bases DBpedia (Auer et al., 2007), ConceptNet (Liu and Singh, 2004), and Webchild (Tandon et al., 2014a; 2014b). Annotators chose an image and a visual element of the image, and then had to select one of those pre-extracted supporting facts related to the visual concept. They finally had to propose a question/answer that specifically involves the selected supporting fact. The dataset contains 193,005 candidate supporting facts related to 580 visual concepts (234 objects, 205 scenes and 141 attributes) for a total of 4608 questions.

3.4. Other datasets

Diagrams. Kembhavi et al. (2016) propose a dataset for VQA on diagrams, named as AI2 Diagrams (AI2D). It comprises more than 5000 diagrams representing grade school science topics, such as the water cycle and the digestive system. Each diagram is annotated with segmentations and relationships between graphical elements. The dataset includes more than 15,000 multiple-choice questions and answers. In the same paper, the authors propose a method specifically designed to infer correct answers on this dataset. The method builds structured representations, named diagram parse graphs (DPG) with techniques specifically tailored to diagrams, *e.g* for recognizing arrows or text with OCR. The DPGs are then used to infer correct answers. In comparison with VQA on natural images, the visual parsing of diagrams remains challenging and the questions often require a high level of reasoning, which make the task very challenging overall.

Shapes. Andreas et al. (2016b) propose a dataset of synthetic images. It is complimentary to datasets of natural image as it provides different challenges, by emphasizing the understanding of spatial and logical relations among multiple objects. The dataset consists of complex questions about arrangements of colored shapes. The questions are built around compositions of concepts and relations, *e.g Is there a red shape above a circle ?* or *Is a red shape blue ?*. This allowed the authors to highlight the capabilities of the Neural Module Networks (see Section 2.3.1). Questions contain between two and four attributes, object types, or relationships. There are 244 questions and 15,616 images in total, with all questions having a yes and no answer (and corresponding supporting image). This eliminates the risk of learning biases, as discussed in Section 3.2. The authors provide results of their own method and of a reimplementation of a joint embedding baseline (Ren et al., 2015). No other results have been reported so far. Note that this dataset is similar in spirit to the synthetic "bAbI" dataset used in textual QA (Weston et al., 2015).

4. Structured scene annotations for VQA

The Visual Genome (Krishna et al., 2017) is currently the largest dataset available for VQA. It provides the unique advantage of human-generated structured annotations for each image in the form of scene graphs. In summary, a scene graph is formed of nodes representing visual elements of the scene, which can be objects, attributes, actions, *etc*. Nodes are linked with directed edges that represent relationships between them (see Fig. 4 for an example). A detailed description is available in Krishna et al. (2017).

The inclusion of scene graphs with images is a significant step toward rich and more comprehensive annotations, compared to the more typical object-level and image-level annotations. In this section, we investigate whether scene graphs could be used to directly answer related visual questions. In other words, if we assume a perfect vision system capable of recovering the same scene graph of the input image as the one annotated by a human, could the answer be trivially obtained from it ? We check a prerequisite for such a hypothesis which is whether the answer actually appears as an element of the graph. Practically, we first build a vocabulary for each image based on its corresponding scene graph. Words in the vocabulary are formed from all node labels of the graph. Then, for each question, we check whether its answer can be matched with words or the combination of words from the vocabulary of its image (See Fig. 5).

We apply the above procedure on all images and questions of the Visual Genome dataset. We find that only 40.02% of the answers can be directly found in the scene graph, *i.e* only 40.02% of the questions could be directly answered using the scene graph representation. Another 7% of answers are numbers (*i.e* counting questions) which we choose to leave aside from the rest of our analysis. There remains 53% of questions can not be directly answered from the scene graph. This ratio is surprisingly high considering the apparent level of detail of the descriptions provided as scene graphs.

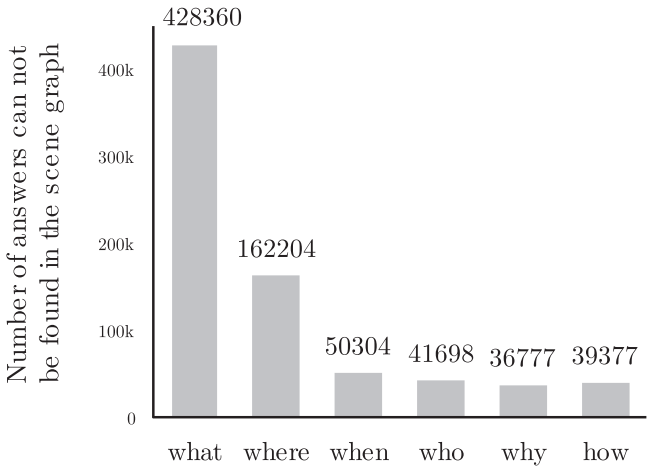


Fig. 6. Number of answers that cannot be found in the scene graph for each question type.

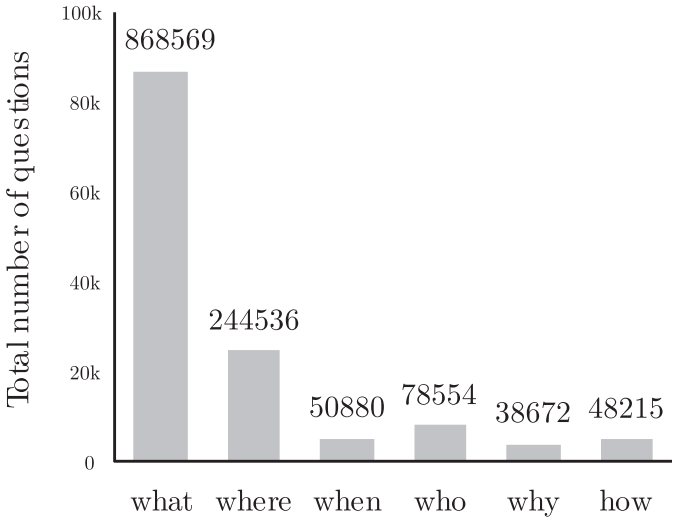


Fig. 7. Total number of questions of each type.

To characterize the remaining 53% of questions, we examine question types using their first few words (Fig. 6. A large number of questions starting with "what" are among those that cannot be directly answered by scene graphs. Note that the overall distribution of question types over the whole dataset (including those that could be answered directly from the scene graph) differs significantly (Fig. 7). We report in Fig. 8 the number of questions that cannot be answered from the scene graph as a fraction of all questions of each type separately. We find that a large fraction of the questions starting with "when", "why" and "how" have answers not be found in scene graphs. Indeed, answering such questions often involves information that does not correspond to specific visual entities, thus not represented by nodes in the scene graphs. It may be possible however to recover these answers using common sense or object-specific knowledge.

We recorded answers that could not be found in scene graphs and ranked them by frequency of occurrence (Table 4). Answers to "what" questions that can not be found in the scene graph

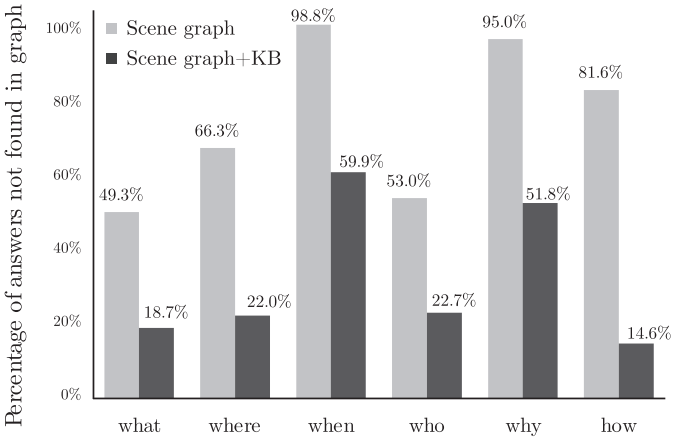


Fig. 8. Answers that can not be found in the scene graph or the knowledge-expanded scene graph, measured as a fraction of all questions of each type separately.

Table 6
Reported results on the DAQUAR-all dataset.

DAQUAR-all	Acc. (%)	WUPS @0.9	WUPS @0.0
Neural-Image-QA (Malinowski et al., 2015)	19.43	25.28	62.00
Multimodal-CNN (Ma et al., 2016)	23.40	29.59	62.95
Attributes-LSTM (Wu et al., 2016a)	24.27	30.41	62.29
QAM (Chen et al., 2015a)	25.37	31.35	65.89
DMN+ (Xiong et al., 2016)	28.79	–	–
Bayesian (Kafle and Kanan, 2016)	28.96	34.74	67.33
DPPnet (Noh et al., 2016)	28.98	34.80	67.81
ACK (Wu et al., 2016c)	29.16	35.30	68.66
ACK-S (Wu et al., 2016b)	29.23	35.37	68.72
SAN (Yang et al., 2016)	29.30	35.10	68.60

Table 7
Reported results on the DAQUAR-reduced dataset.

DAQUAR-reduced	Acc. (%)	WUPS @0.9	WUPS @0.0
GUESS (Ren et al., 2015)	18.24	29.65	77.59
VIS+BOW (Ren et al., 2015)	34.17	44.99	81.48
VIS+LSTM (Ren et al., 2015)	34.41	46.05	82.23
Neural-Image-QA (Malinowski et al., 2015)	34.68	40.76	79.54
2-VIS+BLSTM (Ren et al., 2015)	35.78	46.83	82.15
Multimodal-CNN (Ma et al., 2016)	39.66	44.86	83.06
SMem (Xu and Saenko, 2016)	40.07	–	–
Attributes-LSTM (Wu et al., 2016a)	40.07	45.43	82.67
QAM (Chen et al., 2015a)	42.76	47.62	83.04
DPPnet (Noh et al., 2016)	44.48	49.56	83.95
Bayesian (Kafle and Kanan, 2016)	45.17	49.74	85.13
SAN (Yang et al., 2016)	45.50	50.20	83.60
ACK (Wu et al., 2016c)	45.79	51.53	83.91
ACK-S (Wu et al., 2016b)	46.13	51.83	83.95

Table 8
Reported results on the COCO-QA dataset.

Toronto COCO-QA	Acc. (%)	WUPS @0.9	WUPS @0.0
GUESS(Ren et al., 2015)	6.65	17.42	73.44
VIS+LSTM(Ren et al., 2015)	53.31	63.91	88.25
Multimodal-CNN (Ma et al., 2016)	54.95	65.36	88.58
2-VIS+BLSTM(Ren et al., 2015)	55.09	65.34	88.64
VIS+BOW(Ren et al., 2015)	55.92	66.78	88.99
QAM (Chen et al., 2015a)	58.10	68.44	89.85
DPPnet (Noh et al., 2016)	61.19	70.84	90.61
Attributes-LSTM (Wu et al., 2016a)	61.38	71.15	91.58
SAN (Yang et al., 2016)	61.60	71.60	90.90
Bayesian (Kafle and Kanan, 2016)	63.18	73.14	91.32
HieCoAtt (Lu et al., 2016)	65.40	75.10	92.00
ACK (Wu et al., 2016c)	69.73	77.14	92.50
ACK-S (Wu et al., 2016b)	70.98	78.35	92.87

are mainly attributes like colors and materials, which are probably considered too fine-grained by annotators for inclusion in the scene graphs. The missing answers to the "where" questions are mostly global scene labels such as bedroom, street, zoo, *etc*. The missing answers to "when" questions are, for 90% of them, daytime, night, and variants thereof. These labels are seldom represented in the scene graph, and a large number of synonyms can represent a similar semantic concept. Finally, the "why" and "how" questions lead to higher-level concepts such as actions, reasons, weather types, *etc*.

The above analysis leads to the conclusion that the current scene graphs are rich intermediate abstractions of the scenes, but they are not comprehensive enough to capture *all* elements required for VQA. For example, low-level visual attributes such as color and material are lost in this representation and one therefore needs to access the image to answer questions involving them. Global scene attributes such as location, weather, or time of day are seldom labeled but they are often involved in "where" and "when" questions. It remains debatable whether more human effort should be invested to obtain comprehensive annotations. Another solution is to combine visual datasets with large-scale knowledge bases (KBs) that provide common sense information about visual and non-visual concepts. The type of information in those KBs is complementary to visual annotations like scene graphs. For example, although "daytime" may not be annotated for a particular scene, the annotation of a "clear blue sky" may lead to reason about daytime given some common sense knowledge. Similar reasoning could *e.g* associate "oven" to "kitchen", "food" to "eat", and "snow" to "cold".

We perform an additional experiment to estimate the potential of connecting scene graphs with a general purpose KB.

For each question, we examine whether the correct answer can be found in a first-order extension of the scene graph using relations the KB. More specifically, we use the labels of all nodes of the scene graph to query 3 large KBs (DBpedia Auer et al., 2007, WebChild Tandon et al., 2014a; 2014b, and ConceptNet Liu and Singh, 2004). The triples resulting from the query are used to expand the scene graph and its vocabulary. For example, a scene graph node labeled "cat" may return the fact `<cat,isa,mammal>`, which will be appended to the "cat" node of the scene graph. This simple procedure effectively completes the scene-specific graph with general, relevant knowledge. Similarly as above, we then examine whether questions could potentially be answered from this representation alone, checking for matches between the correct answer and words or combination of words from the expanded vocabulary. We find that this is the case for 79.58% of the questions, which is nearly the double of the same experiment without the KB expansion (40.02%). This clearly shows the potential for complementing the interpretation of visual contents with information from general-purpose KBs.

Although the answer may not be present in the scene graph of the input image, we also find that it can usually be found in scene graphs of other images in the dataset. In an additional experiment, we build an "aggregated scene graph" that combines (*i.e* merges) the scene graphs of all individual images. We then examine whether the answer for the question can be found within this aggregated scene graph. We find that 99.11% of the correct answers can be found in this large aggregated scene graph. This suggests that the correct answers correspond to reasonable node and edge labels in scene graphs, but they are not always found in the scene graph of the input image because of incomplete or ambiguous annotations (*i.e* with synonyms of the correct answers). We also look at the ratio of correct answers that can be found within the top-k words (in terms of number of occurrences) appearing in the aggregated scene graph (see Fig. 9). Using only the top 2000 words, we can still find matches for 90% of the answers. This suggests that

Table 9
Reported results on the VQA-real test set in the open-ended and multiple-choice (M.C.) settings.

Method	Test-dev					Test-standard				
	Open-ended				M.C.	Open-ended				M.C.
	Y/N	Num.	Other	All	All	Y/N	Num.	Other	All	All
Com-Mem (Jiang et al., 2015)	78.3	35.9	34.5	52.6	–	–	–	–	–	–
Attributes-LSTM (Wu et al., 2016a)	79.8	36.1	43.1	55.6	–	78.7	36.0	43.4	55.8	–
iBOWING (Zhou et al., 2015)	76.5	35.0	42.6	55.7	–	76.8	35.0	42.6	55.9	62.0
Region-Sel (Shih et al., 2016)	–	–	–	–	62.4	–	–	–	–	62.4
DPPnet (Noh et al., 2016)	80.7	37.2	41.7	57.2	–	80.3	36.9	42.2	57.4	–
VQA team (Antol et al., 2015)	80.5	36.8	43.1	57.8	62.7	80.6	36.5	43.7	58.2	63.1
MLP-AQI (Jabri et al., 2016)	–	–	–	–	–	–	–	–	–	65.2
SMem (Xu and Saenko, 2016)	80.9	37.3	43.1	58.0	–	80.9	37.5	43.5	58.2	–
Neural-Image-QA (Malinowski et al., 2015)	78.4	36.4	46.3	58.4	–	78.2	36.3	46.3	58.4	–
NMN (Andreas et al., 2016b)	81.2	38.0	44.0	58.6	–	81.2	37.7	44.0	58.7	–
SAN (Yang et al., 2016)	79.3	36.6	46.1	58.7	–	–	–	–	58.9	–
ACK (Wu et al., 2016c)	81.0	38.4	45.2	59.2	–	81.1	37.1	45.8	59.4	–
DNMN (Andreas et al., 2016a)	81.1	38.6	45.5	59.4	–	–	–	–	59.4	–
FDA (Ilievski et al., 2016)	81.1	36.2	45.8	59.2	–	–	–	–	59.5	–
ACK-S (Wu et al., 2016b)	81.0	38.5	45.3	59.2		81.1	37.2	45.9	59.5	–
Bayesian (Kafle and Kanan, 2016)	80.5	37.5	46.7	59.6		80.3	37.8	47.6	60.1	–
DMN+ (Xiong et al., 2016)	80.5	36.8	48.3	60.3	–	–	–	–	60.4	–
MCB (Fukui et al., 2016)	81.7	36.9	49.0	61.1	–	–	–	–	–	–
DualNet (Saito et al., 2016)	82.0	37.9	49.2	61.5	66.7	81.9	37.8	49.7	61.7	66.7
MRN (Kim et al., 2016)	82.3	39.1	48.8	61.5	66.3	82.4	38.2	49.4	61.8	66.3
HieCoAtt (Lu et al., 2016)	79.7	38.7	51.7	61.8	65.8	–	–	–	62.1	66.1
MCB-Att (Fukui et al., 2016)	82.7	37.7	54.8	64.2	–	–	–	–	–	–
Joint-Loss (Noh and Han, 2016)	81.9	39.0	53.0	63.3	67.7	81.7	38.2	52.8	63.2	67.3
Ensemble of 7 models (Fukui et al., 2016)	83.4	39.8	58.5	66.7	70.2	83.2	39.5	58.0	66.5	70.1

Table 10
Frequent answers that cannot be found in scene graphs of the input image, for each question type, ranked by rate of occurrence.

what	white, green, brown, black, blue, wood, red, gray, yellow, grey, black and white, metal, silver, orange, tree, male, tan, round, sunny, brick, grass, skateboarding, cloud, daytime, surfing, right, skiing, left, pink, dirt, female, standing, water, ...
where	ground, air, street, table, field, road, sidewalk, zoo, left, sky, right, water, beach, wall, kitchen, background, restaurant, park, bathroom, living room, ocean, in distance, tennis court, plate, airport, bedroom, city, baseball field, ...
when	daytime, during day, day time, during daytime, in daytime, afternoon, at night, night time, winter, now, nighttime, during day time, night, morning, outside during day time, during daylight hour, evening, daylight, sunny day, daylight hour, ...
who	no one, nobody, man, person, woman, photographer, boy, lady, girl, pilot, surfer, child, man on right, man on left, skateboarder, tennis player, no person, little girl, spectator, skier, conductor,guy, passenger, young man, man and woman, ...
why	sunny, to hit ball, to eat, sunlight, raining, safety, daytime, to be eaten, during day, remembrance, for fun, for balance, cold, to surf, resting, to play tennis, protection, sun, balance, winter, to catch frisbee, decoration, to play, ...
how	clear, none, sunny, cloudy, overcast, calm, open, good, closed, white, clear blue, happy, standing, short, partly cloudy, rainy, green, blue, cold, wet, black, brown, in motion, long, down, blurry, dirty, small, up, large, clean, gray, upside down, sliced, ...

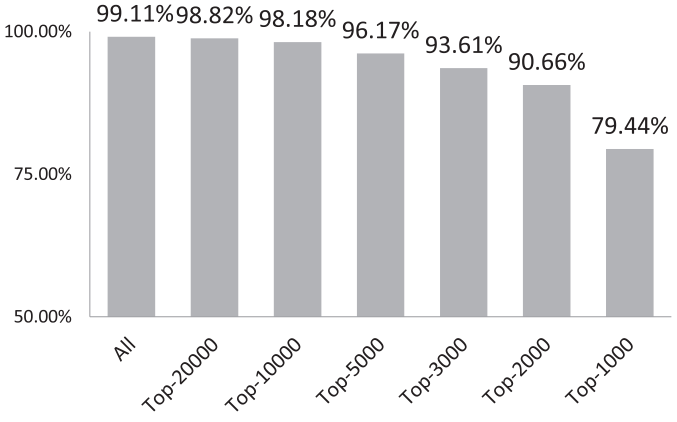


Fig. 9. Percentage of answers that can be found in the scene graph, according to the different top-*k* words in the aggregated vocabulary.

most answers correspond to relatively frequent labels used in the scene graphs, as opposed to rare word, and that those answers could probably be found *if* individual scene graph were complete and perfect in terms of coverage (including detailed local attributes such as color and material, global attributes such as location and weather, *etc.*). However, this might ultimately be achievable with advanced computer vision systems performing image classification, attributes detection, scene classification *etc.*

5. Discussion and future directions

The introduction of the task of VQA is relatively recent and it sparked significant interest and accelerating developments in just a few years. VQA is a complex task and it was initially encouraged by a certain level of maturity reached in the fundamental tasks of computer vision such as image recognition. VQA is particularly attractive because it constitutes an AI complete task in its ultimate form, *i.e* considering open-world free-form questions and answers. Recent results, although encouraging, should however not fool us, as this ultimate goal is indisputably a long way from any current technique. Reduced and limited forms of VQA, *e.g* multiple-choice format, short answer lengths, limited types of questions, *etc*, are reasonable intermediate objectives that seem attainable. Their evaluation is practically easier and may be more representative of our actual progress.

Our review of datasets (Section 3) showed a diversity of protocols for collecting data (from human annotators or semi-automatically from image captions) and imposing certain constraints (*e.g* focusing on certain image regions, objects, or types

of questions). These choices influence the collected questions and answers in many ways. First, they impact the level of complexity and number of facts involved, *i.e* whether the correct answers can be inferred after recognizing a single item/relation/attribute, or requires inference over multiple elements and characteristics of the scene. Second, they influence the ratio of visual vs textual understanding required. One extreme example is the synthetic "Shapes" dataset (Section 3.4) which only requires recognizing a handful of shapes and colors by their names, and rather places the emphasis on the reasoning over relationships between such elements. Third, they influence the amount of prior external knowledge required. External is to be understood in the sense of not inferable from the given visual and textual input. This information may however be visual in nature, *e.g* bright blue skies occur during daytime, or yellow jackets are usually worn for safety.

External knowledge. As mentioned above, VQA constitutes an AI-complete challenge since most tasks in AI can be formulated as questions over images. Note however that these questions will often require external knowledge to be answered. This is a reason for the recent interest in methods connecting VQA with structured knowledge bases, and in specific datasets of questions requiring such mechanisms. One may argue that such complex questions are a distraction from the purely visual questions that should be tackled first. We believe that both paths can be explored in parallel. Unfortunately, the current approaches that use knowledge bases for VQA present serious limitations. Wang et al. (2015); 2016) can only handle a limited number of question types predefined by hand-coded templates. Wu et al. (2016c) encode retrieved information using Doc2Vec, but the encoding process is question-independent and may include information irrelevant to the question. The concept of memory-augmented neural networks could offer a suitable and scalable framework for incorporating and adaptively selecting relevant external knowledge for VQA. This avenue has not been explored yet to our knowledge. On the dataset front, questions involving significant external knowledge are unevenly represented. Specific datasets have been proposed with such questions and additional annotations of supporting facts, but they are limited in scale. Efforts on datasets will likely stimulate research in this direction and help training suitable models. Note that these efforts could simply involve additional annotations of existing datasets.

Textual question answering. The task of textual question answering predates its visual counterpart by several decades and has produced a substantial amount of work. Two distinct types of approaches have traditionally been proposed: information retrieval, and semantic parsing coupled with knowledge bases. On the one hand, information retrieval approaches use unstructured collections of documents, in which the key words of the question are looked for to identify relevant passages and sentences (Jurafsky and Martin, 2000; Kolomiyets and Moens, 2011). A ranking function then sorts these candidates, and the answer is extracted from one or several top matches. This approach can be compared to the basic "joint embedding with attention" method of VQA, where features describing each image region are compared to a representation of the question, identifying the region(s) to focus on, then extracting the answer. A key concept is the prediction of answer type from the question (*e.g* a colour, a date, a person, *etc*) to facilitate the final extraction of the answer from candidate passages. This very concept of answer-type prediction was recently brought back to VQA by Kafle and Kanan (2016). On the other hand, the semantic parsing approaches focus on a better understanding of the question, using more sophisticated language models and parsers to turn the question into structured queries (Dong and Lapata, 2016; Iyyer et al., 2014; Singh and Dwivedi, 2014; Wen-tau et al., 2015). These queries can then be executed on domain-specific databases or general purpose structured knowledge bases. A similar process is used in the VQA methods of Wang *et al* for querying external knowledge bases (Wang et al., 2015; 2016).

As we stated before, interest in VQA grew from the maturity of deep learning on tasks of image recognition (of objects, activities, scenes, *etc*). Most current work on VQA is therefore built with tools and methods from the computer vision community. Textual question answering has traditionally been addressed in the natural language processing community, with different approaches and algorithms. A number of concepts have permeated from NLP to recent efforts on VQA, for example word embeddings, sentence representations, processing with recurrent neural networks, *etc*. Some notable successes are attributable to joint efforts from both fields (*e.g* Andreas et al., 2016a; Xiong et al., 2016). We believe that there still exists potential for better use of concepts from NLP for addressing challenges in VQA. Language models are trainable on large amounts of minimally-labeled text, independently from visual data. They can then be used in the output stage of VQA systems to generate long answers in natural language. Similarly, syntactic parsers may be pre-trained on text alone and be reused for a more principled processing of input questions. The understanding of the question does not have to be trained end-to-end as most VQA systems currently do. The interpretation of text queries into logical representations has been studied in NLP in its own right (*e.g* Dong and Lapata, 2016; Peng and Yao, 2015; Zettlemoyer and Collins, 2007).

6. Conclusion

This article presented a comprehensive review of the state-of-the-art on visual question answering. We reviewed the most popular approach that maps questions and images to vector representations in a common feature space. We described additional improvements that build up on this concept, namely attention mechanisms, modular and memory-augmented architectures. We reviewed the growing number of datasets available for training and evaluating VQA methods, highlighting differences in the type and difficulty of questions that they include. In addition to a descriptive review, we pinpointed a number of promising directions for future research. In particular, we suggest to scale up the inclusion of additional external knowledge from structured knowledge bases, as well as a continued exploration of the potential of natural language processing tools. We believe that the ongoing and future work on these particular points will benefit the specific task of VQA as well as the general objective of visual scene understanding.

Acknowledgement

This research was in part supported by the Data to Decisions Cooperative Research Centre funded by the Australian Government.

References

Andreas, J., Rohrbach, M., Darrell, T., Klein, D., 2016a. Learning to compose neural networks for question answering. In: Annual Conference of the North American Chapter of the Association for Computational Linguistics.
Andreas, J., Rohrbach, M., Darrell, T., Klein, D., 2016b. Neural module networks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.
Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D., 2015. VQA: visual question answering. In: Proc. IEEE Int. Conf. Comp. Vis.
Antol, S., Zitnick, C.L., Parikh, D., 2014. Zero-shot learning via visual abstraction. In: Proc. Eur. Conf. Comp. Vis.
Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., 2007. DBpedia: a nucleus for a web of open data. Springer.
Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O., 2007. Open information extraction for the web. In: Proc. Int. Joint Conf. Artificial Intell.
Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J., 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In: ACM SIGMOD International Conference on Management of Data. ACM, pp. 1247–1250.
Bordes, A., Usunier, N., Chopra, S., Weston, J., 2015. Large-scale simple question answering with memory networks. CoRR abs/1506.02075.

Cantrell, R., Scheutz, M., Schermerhorn, P., Wu, X., 2010. Robust spoken instruction understanding for hri. In: Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on. IEEE, pp. 275–282.

Carlson, A., Betteridge, J., Kisiel, B., Settles, B., 2010. Toward an architecture for never-ending language learning.. In: Proc. Conf. AAAI.

Chen, K., Wang, J., Chen, L.-C., Gao, H., Xu, W., Nevatia, R., 2015a. ABC-CNN: An attention based convolutional neural network for visual question answering. CoRR abs/1511.05960.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, P., Zitnick, C. L., 2015b. Microsoft COCO captions: data collection and evaluation server. CoRR abs/1504.00325.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Dong, L., Lapata, M., 2016. Language to logical form with neural attention. In: Proc. Conf. Association for Computational Linguistics.

Etzioni, O., Fader, A., Christensen, J., Soderland, S., Mausam, M., 2011. Open information extraction: the second generation. In: Proc. Int. Joint Conf. Artificial Intell.

Fader, A., Soderland, S., Etzioni, O., 2011. Identifying relations for open information extraction. In: Proc. Conf. Empirical Methods in Natural Language Processing.

Ferraro, F., Mostafazadeh, N., Huang, T.-H., Vanderwende, L., Devlin, J., Galley, M., Mitchell, M., 2015. A survey of current datasets for vision and language research. In: Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, pp. 207–213.

Fouhey, D.F., Zitnick, C., 2014. Predicting object dynamics in scenes. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M., 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847.

Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W., 2015. Are you talking to a machine? Dataset and methods for multilingual image question answering. In: Proc. Advances in Neural Inf. Process. Syst.

Geman, D., Geman, S., Hallonquist, N., Younes, L., 2015. Visual turing test for computer vision systems. Proc. Nation. Acad. Sci. 112 (12), 3618–3623.

Group, R. W., et al., 2014. Resource description framework. http://www.w3.org/standards/techs/rdf.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Hodosh, M., Young, P., Hockenmaier, J., 2013. Framing image description as a ranking task: data, models and evaluation metrics. JAIR 853–899.

Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G., 2013. YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia. In: Proc. Int. Joint Conf. Artificial Intell.

Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T., 2016. Natural language object retrieval. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Ilievski, I., Yan, S., Feng, J., 2016. A focused dynamic attention model for visual question answering. CoRR abs/1604.01485.

Iyyer, M., Boyd-Graber, J., Claudino, L., Socher, R., Daumé III, H., 2014. A neural network for factoid question answering over paragraphs. Empirical Methods in Natural Language Processing.

Jabri, A., Joulin, A., van der Maaten, L., 2016. October. Revisiting visual question answering baselines. In: European Conference on Computer Vision. Springer International Publishing, pp. 727–739.

Jiang, A., Wang, F., Porikli, F., Li, Y., 2015. Compositional memory for visual question answering. CoRR abs/1511.05676.

Jurafsky, D., Martin, J.H., 2000. Question answering. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall PTR. chapter 28.

Kafle, K., Kanan, C., 2016. Answer-type prediction for visual question answering. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Karpathy, A., Joulin, A., Li, F.F., 2014. Deep fragment embeddings for bidirectional image sentence mapping. In: Proc. Advances in Neural Inf. Process. Syst.

Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.L., 2014. Referitgame: Referring to objects in photographs of natural scenes. In: Conference on Empirical Methods in Natural Language Processing, pp. 787–798.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A., 2016. October. A diagram is worth a dozen images. In: European Conference on Computer Vision. Springer International Publishing, pp. 235–251.

Kim, J.H., Lee, S.W., Kwak, D., Heo, M.O., Kim, J., Ha, J.W., Zhang, B.T., 2016. Multimodal residual learning for visual qa. In: Advances in Neural Information Processing Systems, pp. 361–369.

Kollar, T., Krishnamurthy, J., Strimel, G.P., 2013. Toward interactive grounded language acqusition. Robotics: Science and Systems.

Kolomiyets, O., Moens, M.-F., 2011. A survey on question answering technology from an information retrieval perspective. Inf. Sci. (Ny) 181 (24), 5412–5434.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M., Fei-Fei, L., 2017. Visual genome: connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. 123, 32. doi:10.1007/s11263-016-0981-7.

Kumar, A., Irsoy, O., Su, J., Bradbury, J., English, R., Pierce, B., Ondruska, P., Gulrajani, I., Socher, R., 2016. Ask me anything: Dynamic memory networks for natural language processing. In: Proc. Int. Conf. Mach. Learn.

Li, S., Kulkarni, G., Berg, T.L., Berg, A.C., Choi, Y., 2011. Composing simple image descriptions using web-scale n-grams. In: The SIGNLL Conference on Computational Natural Language Learning.

Liang, P., Jordan, M.I., Klein, D., 2013. Learning dependency-based compositional semantics. Comput. Ling. 39 (2), 389–446.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft COCO: common objects in context. In: Proc. Eur. Conf. Comp. Vis.

Lin, X., Parikh, D., 2015. Don't just listen, use your imagination: leveraging visual common sense for non-visual tasks. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Liu, H., Singh, P., 2004. Conceptnet - A practical commonsense reasoning toolkit. BT Technol. J. 22 (4), 211–226.

Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems, pp. 289–297.

Ma, L., Lu, Z., Li, H., 2016. Learning to answer questions from image using convolutional neural network. In: Proc. Conf. AAAI.

Mahdisoltani, F., Biega, J., Suchanek, F., 2015. YAGO3: a knowledge base from multilingual Wikipedias. CIDR.

Malinowski, M., Fritz, M., 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In: Proc. Advances in Neural Inf. Process. Syst., pp. 1682–1690.

Malinowski, M., Rohrbach, M., Fritz, M., 2015. Ask Your Neurons: A Neural-based Approach to Answering Questions about Images. In: Proc. IEEE Int. Conf. Comp. Vis.

Mao, J., Jonathan, H., Toshev, A., Camburu, O., Yuille, A., Murphy, K., 2016. Generation and comprehension of unambiguous object descriptions. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A., 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). In: Proc. Int. Conf. Learn. Representations.

de Marneffe, M.-C., Manning, C.D., 2008. The stanford typed dependencies representation. COLING Workshop on Cross-framework and Cross-domain Parser Evaluation.

Matuszek, C., FitzGerald, N., Zettlemoyer, L., Bo, L., Fox, D., 2012. A joint model of language and perception for grounded attribute learning. In: Proc. Int. Conf. Mach. Learn.

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. CoRR abs/1301.3781.

Noh, H., Han, B., 2016. Training recurrent answering units with joint loss minimization for vqa. CoRR abs/1606.03647.

Noh, H., Seo, P.H., Han, B., 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Peng, B., Lu, Z., Li, H., Wong, K., 2015. Towards neural network-based reasoning. CoRR abs/1508.05508.

Peng, B., Yao, K., 2015. Recurrent neural networks with external memory for language understanding. Proc. CCF Conference on Natural Language Processing & Chinese Computing.

Pennington, J., Socher, R., Manning, C., 2014. Glove: global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing. Doha, Qatar.

Prud'Hommeaux, E., Seaborne, A., et al., 2008. SPARQL query language for RDF. W3C Recommend. 15.

Ren, M., Kiros, R., Zemel, R., 2015. Image question answering: a visual semantic embedding model and a new dataset. In: Proc. Advances in Neural Inf. Process. Syst.

Roy, D., Hsiao, K.-Y., Mavridis, N., 2003. Conversational robots: building blocks for grounding word meaning. In: HLT-NAACL Workshop on Learning word meaning from non-linguistic data. Association for Computational Linguistics, pp. 70–77.

Saito, K., Shin, A., Ushiku, Y., Harada, T., 2016. Dualnet: domain-invariant network for visual question answering. CoRR abs/1606.06108.

Shih, K.J., Singh, S., Hoiem, D., 2016. Where to look: focus regions for visual question answering. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images. In: Proc. Eur. Conf. Comp. Vis.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Singh, V., Dwivedi, S.K., 2014. Question answering: a survey of research, techniques and issues. Int. J. Inf. Retr. Res. 4 (3).

Sukhbaatar, S., Szlam, A., Weston, J., Fergus, R., 2015. Weakly supervised memory networks. CoRR abs/1503.08895.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Tandon, N., De Melo, G., Weikum, G., 2014a. Acquiring comparative commonsense knowledge from the web. In: Proc. Conf. AAAI.

Tandon, N., de Melo, G., Suchanek, F., Weikum, G., 2014b. Webchild: harvesting and organizing commonsense knowledge from the web. In: International Conference on Web Search and Data Mining. ACM.

Tu, K., Meng, M., Lee, M.W., Choe, T.E., Zhu, S.-C., 2014. Joint video and text parsing for understanding events and answering queries. IEEE Trans. Multimed. 21 (2), 42–70.

Vedantam, R., Lin, X., Batra, T., Zitnick, C.L., Parikh, D., 2015a. Learning common sense through visual abstraction. In: Proc. IEEE Int. Conf. Comp. Vis.

Vedantam, R., Zitnick, C.L., Parikh, D., 2015b. CIDEr: consensus-based image description evaluation. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2014. Show and tell: a neural image caption generator. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.

Wang, P., Wu, Q., Shen, C., Hengel, A. v. d., Dick, A., 2015. Explicit knowledge-based reasoning for visual question answering. CoRR abs/1511.02570.
Wang, P., Wu, Q., Shen, C., Hengel, A. v. d., Dick, A., 2016. Fvqa: fact-based visual question answering. CoRR abs/1606.05433.
Wen-tau, Y., Ming-Wei, C., Xiaodong, H., Jianfeng, G., 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. In: International Joint Conference on Natural Language Processing of the AFNLP. ACL.
Weston, J., Bordes, A., Chopra, S., Mikolov, T., 2015. Towards ai-complete question answering: a set of prerequisite toy tasks. CoRR abs/1502.05698.
Weston, J., Chopra, S., Bordes, A., 2014. Memory networks. CoRR abs/1410.3916.
Winograd, T., 1972. Understanding natural language. Cogn. Psychol. 3 (1), 1–191.
Wu, Q., Shen, C., Hengel, A.v.d., Liu, L., Dick, A., 2016a. What value do explicit high level concepts have in vision to language problems? In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.
Wu, Q., Shen, C., Hengel, A. v. d., Wang, P., Dick, A., 2016b. Image captioning and visual question answering based on attributes and their related external knowledge. CoRR abs/1603.02814.
Wu, Q., Wang, P., Shen, C., Dick, A., Hengel, A.v.d., 2016c. Ask me anything: free-form visual question answering based on knowledge from external sources. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.
Wu, Z., Palmer, M., 1994. Verbs semantics and lexical selection. In: Proc. Conf. Association for Computational Linguistics.
Xiong, C., Merity, S., Socher, R., 2016. Dynamic memory networks for visual and textual question answering. In: Proc. Int. Conf. Mach. Learn.
Xu, H., Saenko, K., 2016. October. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision. Springer International Publishing, pp. 451–466.
Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y., 2015. Show, attend and tell: neural image caption generation with visual attention. In: Proc. Int. Conf. Mach. Learn.
Yang, Z., He, X., Gao, J., Deng, L., Smola, A., 2016. Stacked attention networks for image question answering. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.
Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., Courville, A., 2015. Describing videos by exploiting temporal structure. In: Proc. IEEE Int. Conf. Comp. Vis.
Young, P., Lai, A., Hodosh, M., Hockenmaier, J., 2014. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Proc. Conf. Assoc. Comput. Ling. 2.
Yu, L., Park, E., Berg, A.C., Berg, T.L., 2015. Visual madlibs: fill in the blank image generation and question answering. In: Proc. IEEE Int. Conf. Comp. Vis.
Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Proc. Eur. Conf. Comp. Vis.
Zettlemoyer, L.S., Collins, M., 2007. Online learning of relaxed ccg grammars for parsing to logical form. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning.
Zhang, P., Goyal, Y., Summers-Stay, D., Batra, D., Parikh, D., 2016. Yin and yang: balancing and answering binary visual questions. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.
Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R., 2015. Simple baseline for visual question answering. CoRR abs/1512.02167.
Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L., 2016. Visual7W: grounded question answering in images. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.
Zhu, Y., Zhang, C., Ré, C., Fei-Fei, L., 2015. Building a large-scale multimodal knowledge base system for answering visual queries. CoRR abs/1507.05670.
Zitnick, C.L., Parikh, D., 2013. Bringing semantics into focus using visual abstraction. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn.
Zitnick, C.L., Parikh, D., Vanderwende, L., 2013. Learning the visual interpretation of sentences. In: Proc. IEEE Int. Conf. Comp. Vis.
Zitnick, C.L., Vedantam, R., Parikh, D., 2016. Adopting abstract images for semantic scene understanding. IEEE Trans. Pattern Anal. Mach. Intell. 38 (4), 627–638.

Qi Wu is a postdoctoral researcher at the Australian Centre for Visual Technologies (ACVT) of the University of Adelaide. His research interests include cross-depiction object detection and classification, attributes learning, neural networks, and image captioning. He received a Bachelor in mathematical sciences from China Jiliang University, a Masters in Computer Science, and a PhD in computer vision from the University of Bath (UK) in 2012 and 2015, respectively.

Damien Teney is a postdoctoral researcher at the Australian Centre for Visual Technologies (ACVT) of the University of Adelaide, working on computer vision and machine learning. He was previously affiliated with Carnegie Mellon University, the University of Bath (UK), and the University of Innsbruck (Austria). He obtained his PhD in Computer Science from the University of Liège (Belgium) in 2013, advised by Justus Piater.

Peng Wang is a postdoctoral researcher at the Australian Centre for Visual Technologies (ACVT) of the University of Adelaide. He received a Bachelor in electrical engineering and automation, and a PhD in control science and engineering from Beihang University (China) in 2004 and 2011, respectively.

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

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

Anton van den Hengel is a Professor at the University of Adelaide and the founding Director of The Australian Centre for Visual Technologies (ACVT). He received a PhD in Computer Vision in 2000, a Master Degree in Computer Science in 1994, a Bachelor of Laws in 1993, and a Bachelor of Mathematical Science in 1991, all from The University of Adelaide.