Damien Teney, Qi Wu, and Anton van den Hengel

Visual Question Answering

A tutorial

©ISTOCKPHOTO.COM/ZAPP2PHOTO

Digital Object Identifier 10.1109/MSP.2017.2739826
Date of publication: 13 November 2017

The task of visual question answering (VQA) is receiving increasing interest from researchers in both the computer vision and natural language processing fields. Tremendous advances have been seen in the field of computer vision due to the success of deep learning, in particular on low- and midlevel tasks, such as image segmentation or object recognition. These advances have fueled researchers' confidence for tackling more complex tasks that combine vision with language and high-level reasoning. VQA is a prime example of this trend. This article presents the ongoing work in the field and the current approaches to VQA based on deep learning. VQA constitutes a test for deep visual understanding and a benchmark for general artificial intelligence (AI). While the field of VQA has seen recent successes, it remains a largely unsolved task.

Introduction

VQA involves an image and a related text question, to which the machine must determine the correct answer. This task spans the fields of computer vision and natural language processing, since it requires both the comprehension of the question and parsing the visual elements of the image. VQA is a practical setting to evaluate deep visual understanding, itself considered the overarching goal of the field of computer vision. Deep visual understanding can be defined as the ability of algorithm to extract high-level information from images and to perform reasoning based on that information. In this regard, VQA is an alternative to other tasks proposed to evaluate this capability. Examples include the visual Turing test [23], the task of image captioning [20], [73], and recent works on visual dialogs [18].

A second parallel motivation for the study of VQA is its utility in its own right. A system capable of answering questions about images has direct practical applications, such as a personal assistant, or in robotics as aids for the visually impaired. Note, however, that current VQA data sets do not directly address this setting, because questions are typically collected in a nongoal-oriented setting. Realistic, motivated

questions would likely require information not present in the image and involve rare words and concepts. In comparison, most questions in current data sets are purely visual (e.g., about counts or colors) and centered on common concepts. For example, in one of the most popular data sets [5], a mere 1,000 different answers can correctly answer more than 90% of questions.

The recent interest in VQA [5], [45], [81] originates from the latest advances in computer vision on low- and mid-level tasks. This encouraged further research on higher-level tasks, and the combination of vision with other modalities, particularly language. Historically, one of the earliest integrations of computer vision with language was the SHRDLU system dating back to 1972 [78], which allowed the use of language to instruct a computer to move objects in a simulated "blocks world." Other attempts at creating conversational robotic agents [15], [47], [59] were also grounded in the visual world. However, these early works were often limited to specific domains and/or simple language. Deep learning has now been applied to virtually every problem imaginable in computer vision, and convolutional neural networks (CNNs) are approaching human performance in tasks such as image segmentation [39] or object recognition [19], [24]. The success of deep learning on perceptual tasks drove an increasing enthusiasm for high-level tasks. VQA particularly embodies this confidence in achieving high-level image understanding.

Task definition and data sets

An instance of VQA consists of an image and a related question given in plain text (see examples in Figure 1). The task for the machine is to determine the correct answer, which is, in current data sets, typically a few words or a short phrase. Two practical variants are usually considered, an open-ended and a multiple-choice setting [5], [92]. In the latter, a set of candidate answers are proposed. This makes the evaluation of a generated answer easier than in the open-ended setting, where the comparison between the machine's output and a ground truth (i.e., human provided) answer faces issues with synonyms and paraphrasing.

In comparison to classical tasks of computer vision such as object recognition or image segmentation, instances of VQA cover a wide range of complexity. Indeed, the question itself can take an arbitrary form, and so can the set of operations required to answer it. In this sense, VQA more closely reflects the challenges of general image understanding. VQA is also related to the task of textual question answering [10], [14], [88], in which the answer is to be found in a textual narrative (i.e., reading comprehension) or in large knowledge bases (KBs) (i.e., information retrieval). Textual QA has been studied for a long time in the natural language processing (NLP) community, and VQA is basically its extension to a visual input. The additional challenge of a visual input is significant because images are simply much higher dimensional than text. Images capture the richness of the real world in a noisy manner, whereas natural language already represents a certain level of abstraction. For example, compare the phrase "a red hat" with the multitude of its representations that one could picture, e.g., with many different styles and details that cannot be described in a short phrase.

While, to some extent, the processing of language is possible with discrete- and rule-based approaches, such as syntactic parsers and regular expression matching, the complexity of images renders such engineered methods intractable. Modern computer vision is based on statistical learning, and recent works combining vision and language (including image captioning and VQA) similarly evolved from machine-learning techniques. Finally, both language and vision are inherently compositional in their structure. This constitutes both a challenge and an opportunity when considering the generalization capabilities of learned models (see the section "Compositional Models").

Let us mention the relation of VQA with the task of automatic image captioning [20], [73], [79], i.e., generating a textual description of a given image. It has also attracted significant interest in the past few years and can be compared to VQA as they both combine vision and language. The two tasks are complementary as they evaluate different capabilities. Captioning requires mostly descriptive capabilities that involve almost purely visual information. VQA, in

FIGURE 1. The task of VQA is a significant step toward general AI and a departure from low- and mid-level tasks in classical computer vision. It requires relating visual concepts with elements of language, common-sense, and general knowledge. (Photos are examples from a major public data set [5].)

comparison, often requires reasoning with common sense and with other information not present in the given image. In this respect, VQA constitutes an AI-complete task [5] since it requires multimodal knowledge beyond specific domains. This reinforces the motivation for research on VQA, as it provides a proxy to evaluate progress toward general AI, with systems capable of advanced reasoning combined with deep image and language understanding.

Data sets for training and evaluating VQA

We now examine data sets that have been specifically compiled for research on VQA. These data sets contain, at a minimum, triples made each of an image, a question, and its correct answer. Some early data sets were generated semiautomatically (e.g., from image captions [45]) but modern data sets were created manually through crowdsourcing [5], [35]. The creation of these sets of questions with ground-truth answers is very time-consuming, and today's largest data sets of several hundreds of thousands of instances [35] represent a major effort. Those data sets are designed for both evaluating and training VQA systems in a supervised setting, and the latter demands such large amounts of data. As will be discussed in the section "Directions of Current and Future Research," this very need for large amounts of data is a significant limit of current approaches.

Despite undeniable advantages, VQA data sets of clipart images have seen little use compared to their counterparts of real images.

For the purpose of standardized comparisons and benchmarking of different algorithms, data sets are split into predetermined sets of instances for training, validation, and testing. Benchmarks typically do not provide the ground-truth answers of the test set. The evaluation is performed by an automatic online service that compares the provided answers (inferred by the algorithm to be evaluated) and the private ground truth [5]. This method typically restricts the number and frequency of submissions so as to prevent cheating or unintentional overfitting of the test set.

Existing data sets vary mainly along three dimensions 1) their size, i.e., the number and variety concepts represented in the images and questions; 2) 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 3) how much information beyond what is present in the input image is necessary to infer an answer, e.g., common sense or subject-specific information. Most data sets lean toward visual-level questions and require little external knowledge beyond common sense. These characteristics reflect the fact that current state-of-the-art methods still struggle with simple visual questions.

The first VQA data set designed as a benchmark was Data Set for Question Answering on Real World (DAQUAR) for images [45]. The most popular modern data sets [5], [35], [92] use images sourced from Microsoft Common Objects in Context (COCO), [40] a data set initially devised for image recognition, which is itself composed of images from Flickr. Those images constitute a very diverse collection of photographs.

VQA-real

The most widely used data set is currently the one proposed by a team of researchers from Virginia Tech and is commonly referred to as *VQA* [5]. It comprises two parts, one using natural images named *VQA-real*, and a second one with clipart images named *VQA-abstract* (discussed at the end of this section). VQA-real comprises 123,287 training and 81,434 test images, respectively, sourced from COCO [40]. Human annotators were encouraged to provide interesting and diverse questions and short, concise answers (typically two to three words). The data set allows evaluation both in an open-ended and in a multiple-choice setting, the latter providing 17 additional (incorrect) candidate answers for each question. Overall, the data set contains 614,163 questions. According to an analysis performed by polling annotators, most subjects (at least six out of ten) estimated that some common sense was required for 18% of the questions, and adult-level knowledge was necessary for only 5.5% of the questions. These figures show that purely visual information is likely sufficient to answer most questions.

A recent, updated version of this data set, known as *VQA v2.0*, includes two images with each question that lead to different answers [25]. This aims at addressing issues of data set biases.

Visual genome and visual7W

The Visual Genome QA data set [35] is currently the largest one designed for VQA, with 1.7 million question/answer pairs. It is built with images from the Visual Genome project [35], which includes structured annotations of scene contents in the form of scene graphs. Those scene graphs describe the visual elements of each image with their attributes and the 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 diversity of answers in the Visual Genome is larger than in VQA-real [5]. The 1,000 most-frequently given answers in the data set correspond only to the correct answers of 64% of all questions. In VQA-real, the corresponding top 1,000 answers cover more than 90% of questions. The Visual7w [92] data set is a subset of the Visual Genome that allows evaluation in a multiple-choice setting, as each question is provided with four plausible but incorrect candidate answers.

Zero-shot VQA

A special version of the Visual7W data set was proposed in [70]. The authors redefined the training and test splits such that every test instance includes one or several words that were not present in any training example. For example, a test question "How many zebras are in the image?" might arise even though the word *zebra* was never used in the training set. The evaluation of an algorithm with this data set emphasizes its capabilities for generalization beyond training examples and for using sources of information other than VQA-specific data sets. Another similar study appeared in [54].

VQA-Real [5]

Q: What shape is the bench seat?
A: Curved

Q: What color is the stripe on the train?
A: White

Q: Where are the magazines in this picture?
A: On Stool

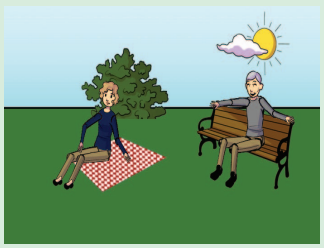
Visual Genome [35]

Q: What color is the clock?
A: Green

Q: What is the woman doing?
A: Sitting

Q: How is the ground?
A: Dry

VQA-Abstract [5]



Q: Who looks happier?
A: Man

Q: Where are the flowers?
A: Around Tree

Q: How many pillows?
A: 2

Zero-Shot VQA [70]

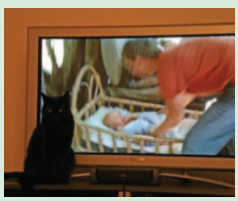
Q: What color are the *barricades*?
A: Pink

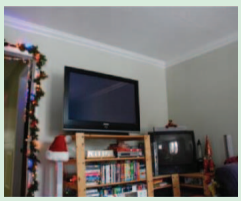
Q: What are they using to *draw*?
A: *Markers*

Q: Who is playing?
A: *Rafael Nadal*

VQA v2.0 [25]

Q: Who is wearing glasses?
A: Man A: Woman





Q: Is the TV on?
A: Yes A: No

FIGURE 2. Examples from the test splits of different VQA data sets. For the zero-shot VQA data set, the highlighted words are unknown words, i.e., not present in training examples.

Clipart images

Data sets for VQA have also been proposed with synthetic clipart images (referred to as *abstract scenes* in [5]). These images were created manually with cartoon representations of characters and objects from a predefined set. The motivation is to enable research on VQA in a controlled setting, where the computer vision part of the problem is eased by the restricted set of visual elements. Such data allows focusing on the high-level semantics of the scenes rather than on visual recognition. For this purpose, the images are provided with structured descriptions, in the form of XML files that list the objects present in the scene with their visual properties (e.g., position, scale, etc.). VQA methods can use these descriptions to completely bypass the visual parsing of the images.

Using synthetic images gives great control over the elements actually depicted, and this allowed the creation of a data set of balanced binary questions [90]. That data set contains only binary (yes/no) questions and each question appears twice in the data set, with two different images that give rise to opposite answers. This removes conditional biases that are common in other data sets, for example, a predominance of "yes" answers to questions of the form "Is there … in the image?" Those biases otherwise allow to blindly guess correct answers, which hinders a meaningful evaluation of VQA systems. Despite undeniable advantages, VQA data sets of clipart images have seen little use [5], [69], [90] compared to their counterparts of real images.

Video-based QA

In addition to the studies on image QA mentioned previously, there have been a few works on VQA with videos. Zhu et al. [91] assembled a data set of over 100,000 videos and 400,000 questions, using existing collections of videos from different domains, from cooking scenarios to movies and web videos. Tapaswi et al. [67] proposed a setting named MovieQA, where questions have to be answered using multiple sources of information including he full-length movies, but also subtitles, scripts, and plot summaries. Zeng et al. [89] proposed the generation of questions from video descriptions.

Evaluation

VQA systems are evaluated by inferring the answers on the test split of a given data set. Recent data sets [92] recommend the multiple-choice setting, since there is only one correct answer among the multiple choices. The evaluation is thus straightforward, as one can simply measure the mean accuracy over test questions. In an open-ended setting, several answers could be equally valid, because of synonyms and paraphrasing. This makes a fair evaluation nontrivial. The usual workaround is to restrict answers, at the time of the creation of the data sets, to short phrases, typically one to three words. This restriction limits ambiguities by forcing questions and answers to be more specific, and allows evaluation by exact string-matching. Most data sets partition the test questions into subsets depending on the type of answer (e.g., yes/no, number, etc.) such that performance can be reported on each subset (see Table 1).

Deep neural networks for VQA

The common approach to VQA is to train a deep neural network with supervision which maps the given image and question to a relative scoring of candidate answers. The main idea is to learn a joint embedding of the visual and textual inputs. First, the image and the question are processed independently to obtain separate vector representations (see Figure 3). Those features are then are mapped with learned functions to a joint space, then combined and fed to an output stage. We examine each of those elements next. The section "Advanced Techniques" will then look at those techniques that build onto this model.

Image encoding

On the computer vision side, the input image x^I is processed with a deep convolutional neural network (CNN) to extract image features described as a vector y^I. This large fixed-size

Table 1. A selection of results on the VQA-real data set (test-std split) in both the open-ended and multiple-choice settings. Performance has incrementally improved over the past few years. The highest accuracies per column are in boldface.

	VQA-Real Open Ended				Multiple Choice
Method	**Yes/No**	**Numbers**	**Other**	**All**	**All**
Baseline: Deeper LSTM Q norm. I [42]	80.6	36.5	43.7	58.2	63.1
Neural modules networks [4]	81.2	37.7	44	58.7	–
Stacked attention networks [87]	–	–	–	58.9	–
Dynamic memory networks (DMNs+) [83]	–	–	–	60.4	–
DualNet [60]	81.9	37.8	49.7	61.7	66.7
Hierarchical coattention (HieCoAtt) [43]	–	–	–	62.1	66.1
VQA-machine [74]	81.4	38.2	53.2	63.3	67.8
MLB [34]	**84**	37.9	54.8	65.1	68.9
MCB ensemble 7 models [21]	83.2	**39.5**	**58**	**66.5**	**70.1**

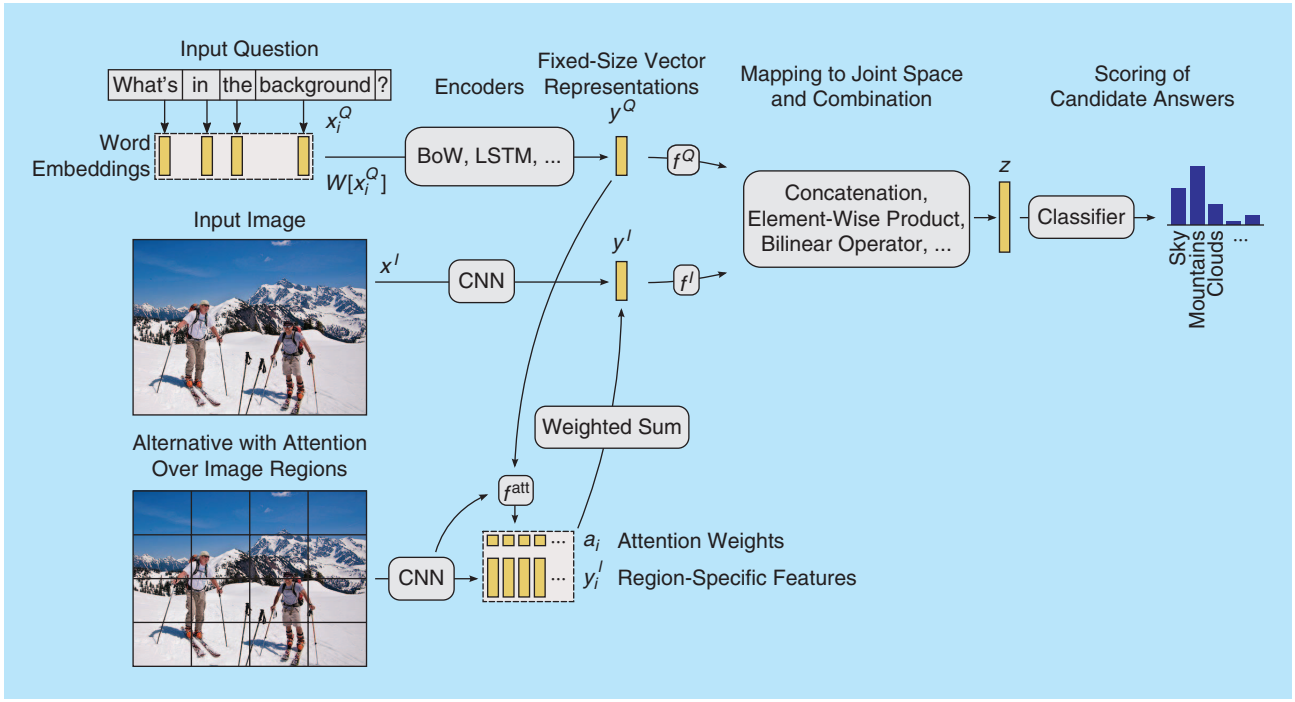


FIGURE 3. The common approach to VQA is to train a deep neural network for classification over a large set of candidate answers (see the section "Deep Neural Networks for VQA"). The input question and image are encoded into fixed-size feature vectors (orange bars), using the word *embeddings* and a CNN, respectively. The resulting representations are mapped into a joint space, then combined and passed on to the classifier. It assigns scores to a large set of candidate answers. The top-ranking candidate is returned as the final answer. An attention mechanism (see the section "Attention Mechanisms") can improve this model and allows the model to focus on relevant parts of the image. In that case, the CNN extracts region-specific image features and aggregates them using scalar weights (orange squares).

vector encodes the contents of the image. This CNN is typically a standard network architecture that has been pretrained to perform image recognition [36]. The motivation for a pretrained network is to take advantage of the vast amounts of training data available for image recognition, relative to the amounts of data annotated for VQA. The pretrained network is used as a generic feature extractor, by discarding the final classification layers, and using the features produced within the CNN prior to this classification [55]. In comparison to classical handcrafted image features such as scale-invariant feature transform (commonly known as *SIFT*) [41] or histogram of oriented gradients (commonly known as *HOG*) [16], CNN features provide higher-level representations of the contents of the image, and are naturally produced as a fixed-size vector. The size of this vector is typically in the order of 1,024 or 2,048.

Question encoding

On the language side, the input question is also processed to obtain a fixed-size representation of its contents. Initially, the ith word of the question is represented by an index x_i^Q in the input vocabulary. Each word is then turned into a vector. This uses a mapping implemented as a lookup table $W[\cdot]$ that associates the index of any word of the input vocabulary to a learned vector. An alternative implementation initially represents each word with a one-hot vector (a vector of all zeros, except for a one at the location of the word index in the vocabulary), which is then multiplied with a dense weight matrix that contains the embeddings of all words. The vectors of all words $W[x_1^Q], W[x_2^Q], \ldots, W[x_N^Q]$ are then collapsed into a single vector. A simple option for this purpose is to make a bag-of-words (BoW), which corresponds to simply averaging the word vectors, i.e., $y^Q = (1/N)\sum_i W[x_i^Q]$. Another popular option is to feed the word vectors into a recurrent neural network (RNN) such as a long short-term memory (LSTM). An RNN processes words sequentially and can capture the sequential relationships between them. In comparison, a BoW does not account for word order, and, for example, would produce a same representation for "this man eats a hot dog" and "a hot man eats this dog."

Combination of image and question features

The feature vectors y^I and y^Q represent the image and the questions, respectively. They are each passed through a learned function before being combined. The intuition here is to map the features to a joint space, in which distances between both modalities become comparable. The learned functions $f^I(\cdot)$ and $f^Q(\cdot)$ are typically implemented as additional layers of the neural network, e.g., $f(y) = \mathrm{ReLU}(Wy + b)$, where W and b are learned weights and biases, and ReLU is a rectified linear unit that serves as a nonlinearity. The mapped features are then combined before being fed to the output stage. A simple option for this combination is to simply concatenate them as $z = [f^I(y^I); f^Q(y^Q)]$. Alternatively, it is popular to include multiplicative interactions within the neural network

to increase its capacity and use $z = f^I(y^I) \cdot f^Q(y^Q)$, where $\cdot$ is the Hadamard (element-wise) product.

Output

The output stage of a VQA system can be seen either as a generation or as a classification task. The generation of a free-form answer has the advantage of being able to compose complex sentences. In practice however, such a model is difficult to learn [22], [46], [80]. Current data sets are limited to short answers, and a practical alternative is to rather learn a classifier over candidate answers [22], [44], [46], [57]. For this purpose, a large set of candidate answers is predetermined from the most common ones in the training set (typically in the order of 2,000). This inevitably leaves out some infrequent words, but such a set is typically sufficient to answer correctly more than 90% of test questions [5]. This is a nonlimiting issue since this figure is well above the accuracy of current systems. The combined features z are passed to a classifier over those candidate answers (a linear layer followed by a softmax [21] or sigmoid transformation [30]). The classifier assigns score to each candidate answer, and the top-ranked one is returned as the final output. In a multiple-choice setting, only the scores assigned to proposed choices are considered. For training the model, the classifier is followed by a cross-entropy loss, and the whole network is trained end-to-end by backpropagation to minimize this loss over the set of training examples.

Variations

A vast array of variations on the method presented previously have been proposed in the literature. Here are some examples.

- Encoding the question and the image with a single recurrent neural network (an LSTM) by passing the image features together with each word embedding [22] or only once prior to the question words [46], [57].
- Encoding the question with a bidirectional RNN, i.e., two LSTMs that process the words in forward and backward order, respectively. This aims at capturing the language structure with more uniform importance on the beginning and the end of the question [57].
- Adding additional multiplicative interactions within the network and between the features of the image and of the question. For example in [51], the authors present their "DPPnet" model as a way of dynamically adapting the computations applied on the image features based on the question (one branch of the network computes weights that are then multiplied with the inputs in another branch). Such interpretations are typical of deep-learning models, but have little concrete support. Performance benefits usually stem simply from the additional capacity of the network.
- Alternative schemes for combining image and question representations, such as element-wise sums and products [33], bilinear operations [30] such as multimodal compact bilinear pooling (MCB) [21], etc.
- Gradual increases in performance of the state of the art is also explained by increasingly better pretrained CNNs to provide image features, and by the application of general enhancements for neural network architectures, such highway networks and residual networks [33], dropout, batch normalization, etc.

Advanced techniques

In this section, we review popular improvements to the general approach described so far.

Attention mechanisms

One of the most effective improvements to the joint embedding model is to use visual attention. Humans have the ability to quickly understand visual representations by attending to regions of the image instead of processing the entire scene at once [58]. Mimicking human attention in deep neural networks has been applied with success to machine translation [8], reading comprehension [63], textual question answering [84], object recognition [64] and image captioning [86], and is also used in most modern VQA models (e.g., in [43] and [87]).

The main idea behind attention mechanisms is to allow the model to focus on certain regions of the image. The technique involves 1) using region-specific image features and 2) including multiplicative interactions within the neural network. The aforementioned basic VQA model described uses a CNN to extract a global feature vector y^I that describes the whole image. This can contain irrelevant or noisy information. Instead, we now extract local features $\{y_i^I\}_i$ for different regions $i = 1 \dots M$ of the image. Those features are obtained from an earlier layer in the pretrained CNN, prior to the last spatial pooling. The network computes a scalar attention weight a_i for each region using both the region and the question features, i.e., $a_i = f^{\text{att}}(y_i^I, y^Q)$. The function $f^{\text{att}}(\cdot)$ is learned and implemented as additional layers of the network. The attention weights can be interpreted as the relevance of a given region, and the image is finally represented by a weighted sum of the region features, i.e., $y^I = \sum_i a_i y_i^I$.

The attention weights computed for a given question/image can be visualized in the form of "attention maps" for purposes of introspection into the VQA model. Each a_i corresponds to a specific region of the input image, and those values are overlaid onto the image canvas (see Figure 4).

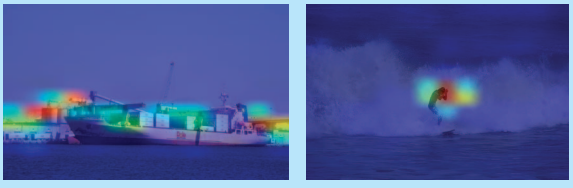


FIGURE 4. Attention weights are often visualized as spatial maps overlaid on the input image (warmer colors correspond to higher weights). They are interpreted as the importance given by the model to different regions of the image (examples used with permission from [74]).

They are interpreted as the importance given by the model to each image region.

The use of an attention mechanism has shown to be very beneficial and is now common practice. Variations on this principle have been proposed. For example, [85] and [87] use multiple rounds of visual attention to allow focusing on several regions. In [85], a two-step process performs a word-guided attention, then a question-guided one. In [65], the authors define image regions with object proposals and then select the regions most to the question and to given answer choices. In [43], the authors propose a "hierarchical coattention" (HieCoAtt) that performs a question-guided attention on the image and an image-guided attention on the question.

The overall idea of attention in neural networks was initially motivated by an analogy to the human visual system. Even though the model is capable of modeling a behavior similar to human attention, this only constitutes an interpretation. In a neural network-trained end to end, nothing enforces the attention mechanism to actually reflect human-like behavior. In a recent study [17], Das et al. compared the attention used by human subjects presented with VQA problems, and VQA models with attention [43], [87]. Their conclusion was a systematically low correlation.

One of the most effective improvements to the joint embedding model is to use visual attention.

Pretraining language representations

As described in the section "Deep Neural Networks for VQA," the first step for encoding the question is to map words to vector representations called word *embeddings*. Each word of the input vocabulary (i.e., any word appearing in the training set) is associated with its own embedding, and those embeddings are normally learned alongside the other parameters of the network via backpropagation. Two potential issues can arise, however. First, word occurrences in any data set typically follow a long-tailed distribution, meaning that a majority of words occur infrequently. It is thus difficult to learn stable and meaningful embeddings for those rare words. Second, the long-tail property, at its extreme, means that it words commonly appear in test questions that were not seen in any training example. Embeddings for those words cannot be learned from those examples, and they are typically associated with an special vector (of zeros or of a special "unknown" token), and their meaning is practically discarded from the questions.

A solution to these issues is to pretrain word embeddings on a larger auxiliary data set. This practice is known in the field of natural language processing and has shown benefit in many tasks besides VQA. Popular methods for pretraining word embeddings include Global Vectors for Word Representation [53] (GloVe) and word2vec [48], which we outline next. The general principle is to use a large, auxiliary training set of unannotated text, such as news articles and *Wikipedia* pages. Those methods require no specific annotations. That data can thus be much larger than the training set used for VQA and involve a much larger vocabulary.

The idea in the skip-gram model of *word2vec* is to train a model which, using the representation (i.e., an embedding) of a given word, is predictive of the context, i.e., the neighboring words in which it frequently appears [49]. As a consequence, words that are interchangeable or appear in similar contexts become associated with similar embeddings. Distances between embeddings thus naturally capture semantic relatedness between the words they represent.

More precisely, the skip-gram model seeks to maximize the ability to predict, from each word embedding, the occurrences of other words in a small surrounding window. The objective function to be maximized is

$$J = \frac{1}{N}\sum_{i}^{N} \frac{1}{|\Omega(i)|} \sum_{j \in \Omega(i)} \log p(x_j \mid x_i), \tag{1}$$

where i indexes the N-ordered words in the training corpus, x_i is the index in the vocabulary of word i, $\Omega(i)$ is a context window of fixed size around word i in the corpus [49]. The conditional probability $\log p(x_j \mid x_i)$ is modeled as a compatibility measure between embeddings such as a dot product followed by a sigmoid, i.e.,

$$p(x_j \mid x_i) = 1/\left(1 + e^{-W[x_i]\ W[x_j]}\right), \tag{2}$$

where $W[\cdot]$ is a lookup table containing the embeddings of all words in the vocabulary, reusing the notation of the section "Deep Neural Networks for VQA." After the training, the context-prediction part of the model is discarded, and the embeddings associated with the words are retained (i.e., the table $W[\cdot]$) and used as word embeddings in the downstream application such as VQA. The embeddings can be used as "frozen weights," i.e., static representations associated with the words, or they can serve as initial values to be subsequently fine-tuned, i.e., optimized with a lower learning rate relative to the other network parameters.

Using pretrained embeddings helps the generalization capabilities of a VQA model. Since semantically similar words are mapped to close points in the word embedding space, the processing by the subsequent layers of the network can more easily 1) interpolate across concepts and 2) generalize to words absent from training questions but for which embeddings were pretrained.

Memory-augmented neural networks

An active research area is the design of deep neural networks that include an internal memory [13], [52], [66], [77]. Memory-augmented networks have shown success on tasks such as textual question answering [28], reading comprehension [37], and VQA [83]. The general idea of memory-augmented networks is to maintain an internal representation of the input data, on which multiple read and write operations can be applied. The composition of multiple operations can potentially execute complex chains of inference on the data. A "controller" part of the network is responsible for

controlling those operations. The mechanism is comparable to multiple rounds of an attention mechanism, in that it also enables the modeling of interactions between specific section of the input data.

The variant proposed in [37] and [83], named *dynamic memory networks* (*DMNs*), was successfully applied to VQA. It is built around four modules (see Figure 5). The input module transforms the input data into a set of discrete vectors called *facts*. A question module computes a vector representation of the question, using a gated recurrent unit [(GRU), a variant of LSTM]. An episodic memory module retrieves the facts required to answer the question. A key element is to allow the episodic memory module to perform multiple passes over the facts to allow transitive reasoning. An attention mechanism selects the relevant facts and an update mechanism iteratively generates new memory representations from the current state and the retrieved facts. The initial state is set as the representation produced by the question module. Finally, the answer module uses the final state of the memory and the question to predict the final output, using a classic classifier over candidate answers.

Run time retrieval of additional information

Interfacing a VQA method with external sources of information allows one to separate the reasoning from the representation of prior knowledge in a scalable manner. One limitation of the basic joint embedding approach is to attempt to capture all of the information of training examples within the parameters of a neural network. This cannot scale arbitrarily, however. On one hand, any network has a finite capacity and, on the other hand, training examples also provide finite information. Several works explored the idea of connecting a VQA system with external sources of information that can be virtually infinite (e.g., web searches) or extensible without needing to retrain the VQA model (e.g., structured KBs).

In [75] and [82], the authors train a model to interface with a KB. Such KBs, like DBpedia [7] and Freebase [12], are databases compiled with facts ranging from common sense to encyclopedic knowledge. Such nonvisual information can be helpful for VQA. For example, the question "How many mammals appear in this image?" requires understanding the word "mammal" and which animals belong to this category. The VQA system of [75] and [82] is trained to map the input question/image to queries to be executed on KBs. The queries retrieve information relevant to the concepts involved in the question and/or image, which is fed as an additional input to the output stage of the system. The overall principle has shown limited benefits on existing VQA data sets, since most questions do not require such specific, nonvisual information. The idea remains a promising direction for developing scalable VQA systems.

In [70], the authors propose the retrieval of visual information from web searches in the form of exemplar images of question words. Rare and novel words, for example, the name of an uncommon animal or of an up-and-coming celebrity, are not likely to appear or be even known during training. The retrieval of images from the web allows the method to expand its domain of applicability as needed. The implementation of [70] simply retrieves the top five images from Google for every word of the question, from which CNN features are extracted and fed alongside the input question/image to the VQA system. This mechanism, however crude, showed an advantage to questions involving unknown words (i.e., "zero-shot VQA") while leaving substantial room for future developments; see the section "Issues with Unknown and Novel Words."

Directions of current and future research

Most modern methods for VQA have been evaluated on the data set of Antol. et al., which has served as the de facto standard benchmark. State-of-the-art methods have consistently improved performance on this data set over the past few years, from an

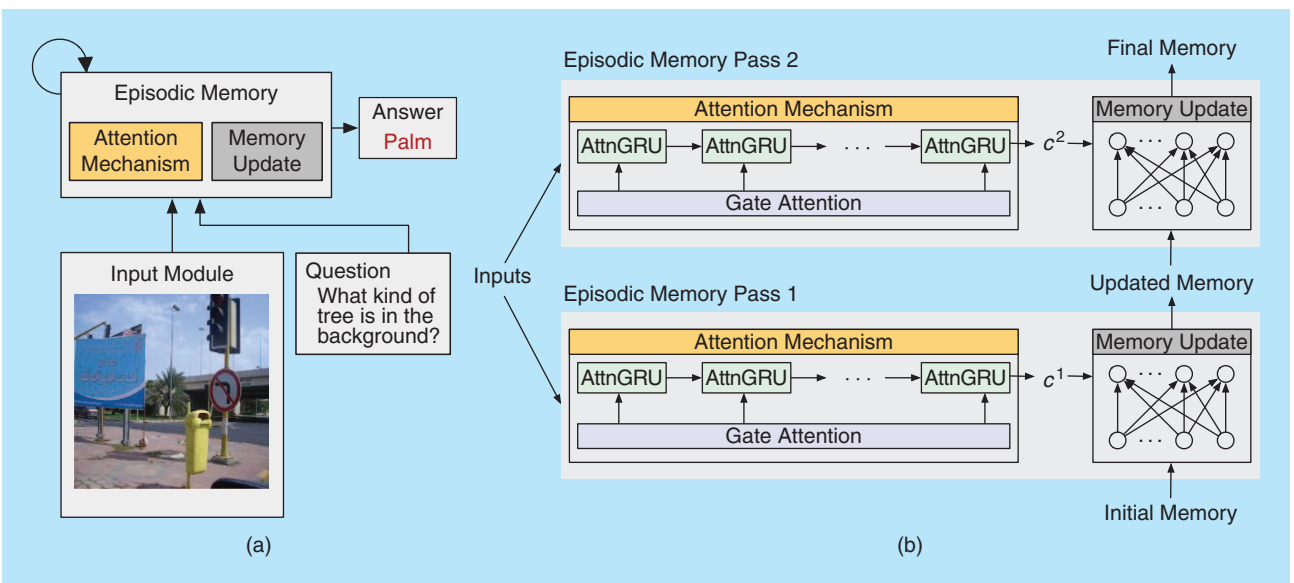


FIGURE 5. DMNs for VQA. (a) The overview and (b) details of the episodic memory module with two passes. (Figure adapted with permission from [83].)

accuracy of about 58% to over 70% today (see Tables 1 and 2 for a selection of results). These improvements have been incremental and have now seemed to plateau. In the following, we examine how current evaluations can mask some inherent issues of today's approaches and examine promising directions to bring future breakthroughs.

Issues of data set biases

Several studies have recently pointed out a fundamental issue with VQA data sets [25], [30], [90]. The text questions alone often provide strong cues that can be sufficient to answer them correctly, with no regards to the contents of the input image. These cues can be obvious. For example, questions starting with "Do you see a …" can be correctly answered with a "yes" almost nine times out of ten [25]. These cue can also stem from an imbalance among possible answers. For example, questions starting with "How many …" often have a correct answer of "one" or "two" but rarely "17." This issue can also be more subtle and manifest in the form of conditional biases. For example, we could imagine that questions starting with "What is the color …" can often be answered correctly with "gray" if it also contains the word "car" and "red" if it contains the word "flower." Biases conditioned on image contents are also likely and yet more subtle. Biases are inherent to the real world, and it is desirable for a VQA model to capture and exploit them to some extent. However, today's methods have been shown to overly rely on data set biases and essentially be reduced to rote-learning of training questions. This is counterproductive to the objective of evaluating visual understanding. A blinded VQA model (i.e., not being shown the input image, and only guessing from the question) still achieves an accuracy of 56% versus 65% in the nonblinded case [30].

The issue of data set biases has been recognized. Attempts at addressing it include balanced data sets. Zhang et al. [90] first proposed a data set of clipart images where each binary question is accompanied by two different images that elicit "yes" and "no" answers, respectively. Goyal et al. applied the idea to real images, associating two images with each question that lead to different answers (see example in Figure 2). An appropriate performance metric in this case is to measure accuracy on pairs of scenes. Blind models in this case would obtain an accuracy of 0%, and random guessing 25%. The use of balanced data sets encourages VQA models, to a larger extent, to utilize visual information instead of relying on language cues and data set biases. It is expected that future evaluations of algorithms on those data sets will be more representative of actual progress on visual understanding.

Issues with unknown and novel words

A VQA method to be used in a real-world setting, e.g., in robotics or as personal AI assistants, must be applicable to open, unrestricted domains. The current paradigm of training VQA systems with supervision, i.e., with data sets of questions and their ground-truth answers, can only cover a limited set of objects and concepts. Although VQA data sets have grown in size, no finite set of exemplars will ever cover the diversity of objects, actions, relations, etc. in the real world, for which an ideal VQA system should be prepared. A secondary issue with the current approach is the incentive for published methods to perform well on benchmark data sets. These benchmarks do not encourage addressing rare words and concepts, but rather focus on the concepts most frequent in the data set. Current methods are therefore designed to best learn—and often overfit—data set biases.

Recent works have argued for addressing a setting named *zero-shot VQA* [54], [70], where questions (or the proposed multiple-choice answers) specifically involve words that have not been seen in any training question. For example, a question "How many zebras are in the image?" may arise, even though no zebra was involved in the training set. This setting requires strong generalization capabilities. For example, a related training question "How many giraffes are in the image?" should be taken as an opportunity to learn to count, although not giraffes specifically. In parallel of works on VQA, the learning of high-level reasoning is addressed in the more abstract setting of program induction (see, e.g., [56]). We expect that VQA will ultimately require similar principled approaches, such as differentiable computing [26], [50], rather than brute-force learning from limited sets of examples.

Table 2. A selection of results on the newer VQA v2 data set (test-std split; open-ended questions). Baseline methods score lower on this harder data set, but the state of the art now reaches more than 70% of accuracy on open-ended questions. The highest accuracies per column are in boldface.

	VQA v2 Open Ended			
Method	**Yes/No**	**Numbers**	**Other**	**All**
Baseline: deeper LSTM Q norm. I [42]	73.46	35.18	41.83	54.22
MCB [21]	78.82	38.28	53.36	62.27
UPMC-LIP6 [9]	82.07	41.06	57.12	65.71
Athena [1]	82.50	44.19	59.97	67.59
LV-NUS [1]	81.89	46.29	58.30	66.77
HDU-USYD-UNCC [1]	**86.65**	**51.13**	**61.75**	**70.92**
Tips and Tricks VQA [2], [68]	86.60	48.64	61.15	70.34

External knowledge

The setting of the previously mentioned zero-shot VQA exposes the need for VQA systems to apply to concepts not present in training question/answers. This motivates the use of other kinds of data for training, and for retrieving additional information as needed at test time. This requires the system not only to capture actual information from training examples, but to learn to retrieve and use novel information, i.e., learn to learn. That capability of metalearning receives increased attention [11], [61], [72]. In the context of VQA, [70] showed the benefit of retrieving on-the-fly, exemplar images of unknown words from an online search engine. In [75] and [76], the authors showed the benefit of answering questions requiring background knowledge of retrieving additional information from a structured KB. The extension of these ideas is a promising research direction.

Modular approaches

Most current VQA models use a monolithic neural network and end-to-end supervision to learn the representations of data, the reasoning process, and to capture background knowledge from training examples. Alternatively, modular approaches have been explored [74], [80] with the goal of explicitly factoring the overall process of VQA into distinct subtasks. The principle of modularity allows one to decouple subtasks to some extent, and to use intermediate supervision and leverage several types of training data, as opposed to only "end-to-end" question/answer pairs. The use of pretrained word embeddings (see the section "Pretraining Language Representations") is a very successful example of this general principle. Word embeddings are pretrained to capture language-based semantic similarities, and, in a similar spirit, other representations could be pretrained from auxiliary data to capture visual similarities [38] and other kinds of background information [71].

Most current VQA models use a monolithic neural network and end-to-end supervision to learn the representations of data, the reasoning process, and to capture background knowledge from training examples.

Modular systems for VQA also allow decoupling, to some degree, the visual perception from the high-level reasoning. For example, Wang et al. [74] proposed a VQA model on top of a collection of computer vision algorithms that detect visual elements such as objects, persons, and relations between them. Thereby, the VQA model only has to reason over this explicit high-level representation of the contents of the image.

Compositional models

The compositional nature of images and language lends itself to learning similarly compositional models [6]. The approach aims at addressing the challenge of generalization, i.e., applying the learned model to novel compositions of words and visual elements. Compositional models were proposed by Hendricks et al. on the task of image captioning [27]. Andreas et al. [4], [3], [29] were the first to propose a compositional architecture for VQA, named *neural module networks*. In their approach, the input question is processed to determine the set of operations required to answer the question. A deep neural network is assembled with trained modules, each corresponding to one of those operations. A custom network is thus tailored specifically to each question, and finally applied on the image to infer the answer.

A data set of synthetic images named *CLEVR* (which stands for *compositional language and elementary visual reasoning*) [31] was specifically designed to evaluate generalization to novel combinations in VQA. It contains photorealistic images of shapes of various colors and materials. The data set also contains annotations describing the kind of reasoning that each question requires (i.e., as functional "programs"). The data set spurred a series of works on compositional models [29], [32]. The extra annotations facilitate the training of compositional models by serving as an intermediate supervision signal. This supervision correspond to an arrangement of modules or operations to be executed for each question. All of the aforementioned works demonstrated unique capabilities on synthetic data sets. However, it is still unclear how to best apply them to real images and how to train them only using end-to-end supervision, i.e., only knowing the final answer.

An alternative approach that addresses compositionality is the relational networks [62]. The idea is to consider the input as a set of objects, such as the locations in a CNN feature map, and to learn a common predictor that is applied to pairwise combinations of those objects. The predictor basically learns the relations between parts of the input. This proved effective on the CLEVR data set without the need for the intermediate supervision mentioned previously.

Conclusions

This article presented a review of the state of the art on visual question answering. We reviewed popular approaches based on deep learning, which treat the task as a classification problem over a set of candidate answers. We described the common joint embedding model, and additional improvements that build up on this concept, such as attention mechanisms. Despite shortcomings of current practices for both training and evaluating VQA systems, we identified a number of promising research avenues that could potentially bring future breakthroughs for both VQA and for the general objective of visual scene understanding.

Acknowledgment

All correspondence regarding this article should be addressed to Qi Wu at qi.wu01@adelaide.edu.au.

Authors

Damien Teney (contact@damienteney.info) obtained his B. Sc. degree in 2007, his M.Sc. degree in 2009, and his Ph.D. degree in 2013, all in computer science from the University of Liege, Belgium. He is a postdoctoral researcher at the

Australian Centre for Visual Technologies of the University of Adelaide, where he works on computer vision and machine learning. He was previously affiliated with Carnegie Mellon University, Pittsburgh, Pennsylvania; the University of Bath, United Kingdom; and the University of Innsbruck, Austria.

Qi Wu (qi.wu01@adelaide.edu.au) received a bachelor's degree in mathematical sciences from China Jiliang University, Hangzhou, and a master's degree in computer science and a Ph.D. degree in computer vision from the University of Bath, United Kingdom, in 2012 and 2015, respectively. He is a postdoctoral researcher at the Australian Centre for Robotic Vision of the University of Adelaide. His research interests include cross-depiction object detection and classification, attributes learning, neural networks, and image captioning.

Anton van den Hengel (anton.vandenhengel@adelaide.edu.au) received his bachelor's degree in mathematical science in 1991, his bachelor of laws degree in 1993, his master's degree in computer science in 1994, and his Ph.D. degree in computer vision in 2000, all from the University of Adelaide, Australia, where he is a professor and the founding director of the Australian Centre for Visual Technologies.

References

[1] VQA challenge leaderboard. [Online]. Available: http://visualqa.org/ http://evalai.cloudcv.org

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *arXiv Preprint*, arXiv:1707.07998, 2017.

[3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," in *Proc. Annu. Conf. North American Chapter Assoc. Computational Linguistics*, San Diego, CA, 2016, pp. 1545–1554.

[4] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 39–48.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 2425–2433.

[6] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik, "Learning to generalize to new compositions in image understanding," *arXiv Preprint*, arXiv:1608.07639, 2016.

[7] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, *DBpedia: A Nucleus for a Web of Open Data*. New York: Springer, 2007.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. Int. Conf. Learning Representation (ICLR*), San Diego, CA, 2015.

[9] H. Ben-younes, R. Cadène, M. Cord, and N. Thome, "MUTAN: multimodal tucker fusion for visual question answering," *arXiv Preprint*, arXiv:1705.06676, 2017.

[10] J. Berant, A. Chou, R. Frostig, and P. Liang, "Semantic parsing on freebase from question-answer pairs," in *Proc. Conf. Empirical Methods Natural Language Processing*, 2013, pp. 1533–1544.

[11] L. Bertinetto, J. F. Henriques, J. Valmadre, P. H. S. Torr, and A. Vedaldi, "Learning feed-forward one-shot learners," in *Proc. Neural Information Processing Systems (NIPS)*, 2016, pp. 523–531.

[12] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proc. ACM SIGMOD Int. Conf. Management of Data*, 2008, pp. 1247–1250.

[13] A. Bordes, N. Usunier, S. Chopra, and J. Weston, "Large-scale simple question answering with memory networks," *arXiv Preprint*, arXiv:1506.02075, 2015.

[14] Q. Cai and A. Yates, "Large-scale semantic parsing via schema matching and lexicon extension," in *Proc. Conf. Association Computational Linguistics*, 2013, pp. 423–433.

[15] R. Cantrell, M. Scheutz, P. Schermerhorn, and X. Wu, "Robust spoken instruction understanding for hri," in *Proc. 5th ACM/IEEE Int. Conf. Human-Robot Interaction*, 2010, pp. 275–282.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2005, vol. 1, pp. 886–893.

[17] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" in *Proc. Conf. Empirical Methods Natural Language Processing*, 2016, pp. 932–937.

[18] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[20] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, and J. Platt, "From captions to visual concepts and back," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1473–1482.

[21] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," in *Proc. Conf. Empirical Methods Natural Language Processing (EMNLP),* 2016, pp. 457–468.

[22] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? Data set and methods for multilingual image question answering," in *Proc. Advances in Neural Information Processing Systems*, 2015, pp. 2296–2304.

[23] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual turing test for computer vision systems," *Proc. Natl. Acad. Sci.*, vol. 112, no. 12, pp. 3618–3623, 2015.

[24] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1440–1448.

[25] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*, 2017.

[26] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," *arXiv Preprint*, arXiv:1410.5401, 2014.

[27] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. J. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2015, pp. 1–10.

[28] F. Hill, A. Bordes, S. Chopra, and J. Weston, "The goldilocks principle: Reading children's books with explicit memory representations," *arXiv Preprint*, arXiv:1511.02301, 2015.

[29] R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," *arXiv Preprint*, arXiv:1704.05526, 2017.

[30] A. Jabri, A. Joulin, and L. van der Maaten, "Revisiting visual question answering baselines," in *Proc. European Conf. Computer Vision (ECCV)* 2016, pp. 727–739.

[31] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick, "CLEVR: A diagnostic data set for compositional language and elementary visual reasoning," in *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, F. Li, C. L. Zitnick, and R. B. Girshick, "Inferring and executing programs for visual reasoning," *CoRR*, 2017. [Online]. Available: http://arxiv.org/abs/1705.03633

[33] J.-H. Kim, S.-W. Lee, D.-H. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang, "Multimodal residual learning for visual QA," in *Proc. Advances Neural Information Processing Systems (NIPS),* 2016, pp. 361–369.

[34] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang, "Hadamard product for low-rank bilinear pooling," *arXiv Preprint*, arXiv:1610.04325, 2016.

[35] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *arXiv Preprint*, arXiv:1602.07332, 2016.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114.

[37] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *Proc. Int. Conf. Machine Learning*, 2016, pp. 1378–1387.

[38] A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," in *Proc. Conf. North American Chapter Assoc. Computational Linguistics–Human Language Technologies (HLT-NAACL)*, 2015, pp. 153–163.

[39] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[40] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. European Conf. Computer Vision*, 2014, pp. 740–755.

[41] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Computer Vision*, 1999, vol. 2, pp. 1150–1157.

[42] J. Lu, X. Lin, D. Batra, and D. Parikh. (2015). Deeper lstm and normalized CNN visual question answering model [Online]. Available: https://github.com/VT-vision-lab/VQA_LSTM_CNN

[43] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Proc. Advances Neural Information Processing Systems*, 2016, pp. 289–297.

[44] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *Proc. 30th AAAI Conference on Artificial Intelligence*, 2016, pp. 3567–3573.

[45] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," in *Proc. Advances Neural Information Processing Systems*, 2014, pp. 1682–1690.

[46] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A neural-based approach to answering questions about images," in *Proc. IEEE Int. Conf. Computer Vision*, 2015, pp. 1–9.

[47] C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox, "A joint model of language and perception for grounded attribute learning," in *Proc. Int. Conf. Machine Learning*, 2012, pp. 1671–1678.

[48] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv Preprint*, arXiv:1301.3781, 2013.

[49] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[50] K. W. Murray and J. Krishnamurthy, "Probabilistic neural programs," *arXiv Preprint*, arXiv:1612.00712, 2016.

[51] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016, pp. 30–38.

[52] B. Peng, Z. Lu, H. Li, and K. Wong, "Toward neural network-based reasoning," *arXiv Preprint*, arXiv:1508.05508, 2015.

[53] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Language Processing*, 2014, pp. 1532–1543.

[54] S. K. Ramakrishnan, A. Pal, G. Sharma, and A. Mittal, "An empirical evaluation of visual question answering for novel objects," *arXiv Preprint*, arXiv:1704.02516, 2017.

[55] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[56] S. E. Reed and N. de Freitas, "Neural programmer-interpreters," in *Proc. Int. Conf. Learning Representations*, 2016.

[57] M. Ren, R. Kiros, and R. Zemel, "Image question answering: a visual semantic embedding model and a new data set," in *Proc. Advances Neural Information Processing Systems*, 2015.

[58] R. A. Rensink, "The dynamic representation of scenes," *Visual Cognition,* vol. 7, no. 1–3, pp. 17–42, 2000.

[59] D. Roy, K.-Y. Hsiao, and N. Mavridis, "Conversational robots: building blocks for grounding word meaning," in *Proc. HLT-NAACL Workshop on Learning Word Meaning Non-Linguistic Data*, 2003, pp. 70–77.

[60] K. Saito, A. Shin, Y. Ushiku, and T. Harada, "Dualnet: Domain-invariant network for visual question answering," *arXiv Preprint*, arXiv:1606.06108, 2016.

[61] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. P. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Machine Learning*, 2016, vol. 48, pp. 1842–1850.

[62] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," *arXiv Preprint*, arXiv:1706.01427, 2017.

[63] M. J. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension," *arXiv Preprint*, arXiv:1611.01603, 2016.

[64] P. Sermanet, A. Frome, and E. Real, "Attention for fine-grained categorization," *arXiv Preprint*, arXiv:1412.7054, 2014.

[65] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2016, pp. 4613–4621.

[66] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "Weakly supervised memory networks," *arXiv Preprint*, arXiv:1503.08895, 2015.

[67] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, "Movieqa: Understanding stories in movies through question-answering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4631–4640.

[68] D. Teney, P. Anderson, X. He, and A. van den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," *arXiv Preprint*, arXiv:1708.02711, 2017.

[69] D. Teney, L. Liu, and A. van den Hengel, "Graph-structured representations for visual question answering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2017.

[70] D. Teney and A. van den Hengel, "Zero-shot visual question answering," *arXiv Preprint*, arXiv: 1611.05546. 2016.

[71] I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *Proc. Int. Conf. Learning Representations*, 2016.

[72] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Proc. Neural Information Processing System (NIPS)*, 2016, pp. 3630–3638.

[73] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2014, pp. 3156–3164.

[74] P. Wang, Q. Wu, C. Shen, and A. v d. Hengel, "The VQA-machine: Learning how to use existing vision algorithms to answer new questions," *arXiv Preprint*, arXiv:1612.05386, 2016.

[75] P. Wang, Q. Wu, C. Shen, A. v d. Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," *arXiv Preprint*, arXiv:1511.02570, 2015.

[76] P. Wang, Q. Wu, C. Shen, A. v d. Hengel, and A. Dick, "FVQA: Fact-based visual question answering," *arXiv Preprint*, arXiv:1606.05433, 2016.

[77] J. Weston, S. Chopra, and A. Bordes, "Memory networks," *arXiv Preprint*, arXiv:11410.3916, 2015.

[78] T. Winograd, "Understanding natural language," *Cognit. Psychol.*, vol. 3, no. 1, pp. 1–191, 1972.

[79] Q. Wu, C. Shen, A. v. d. Hengel, L. Liu, and A. Dick, "What value do explicit high level concepts have in vision to language problems?" in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 203–212.

[80] Q. Wu, C. Shen, A. v d. Hengel, P. Wang, and A. Dick, "Image captioning and visual question answering based on attributes and their related external knowledge," *arXiv Preprint*, arXiv:1603.02814, 2016.

[81] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: a survey of methods and data sets," *Computer Vision and Image Understanding*, to be published.

[82] Q. Wu, P. Wang, C. Shen, A. Dick, and A. v. d. Hengel, "Ask me anything: Free-form visual question answering based on knowledge from external sources," *in Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4622–4630.

[83] C. Xiong, S. Merity, and R. Socher, "Dynamic memory networks for visual and textual question answering," in *Proc. Int. Conf. Machine Learning*, 2016, pp. 2397–2406.

[84] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," *arXiv Preprint*, arXiv:1611.01604, 2016.

[85] H. Xu and K. Saenko, "Ask, attend and answer: exploring question-guided spatial attention for visual question answering," *arXiv Preprint*, arXiv:1511.05234, 2015.

[86] K. Xu, J. Ba, R. Kiros, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: neural image caption generation with visual attention," in *Proc. Int. Conf. Machine Learning*, 2015, pp. 2048–2057.

[87] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 21–29.

[88] X. Yao and B. Van Durme, "Information extraction over structured data: Question answering with freebase," in *Proc. Conf. Association Computational Linguistics*, 2014, pp. 956–966.

[89] K.-H. Zeng, T.-H. Chen, C.-Y. Chuang, Y.-H. Liao, J. C. Niebles, and M. Sun, "Leveraging video descriptions to learn video question answering," in *Proc. Conf Artificial Intelligence AAAI*, 2017, pp. 4334–4340.

[90] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and yang: Balancing and answering binary visual questions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 5014–5022.

[91] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering temporal context for video question and answering," *arXiv Preprint*, arXiv:1511.04670, 2015.

[92] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7W: Grounded question answering in images," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2016, pp. 4995–5004.

SP