Visual Question Answering: A Survey of Methods and Datasets

Visual Question Answering (VQA) :

Given		reasoning over		infer
an image	a question in natural language	visual elements of the image	general knowledge	the correct answer

the common approach:

combining		map	
Convolutional CNN	recurrent neural networks RNN	images	a common feature space
		questions	

1 Introduction

computer vision studies methods for：
（1）acquiring images
（2）processing images
（3）understanding images
teach machines how to see.

NLP is the field concerned with enabling：
interactions between computers and humans in natural language
teaching machines how to read

computer vision and NLP share similar methods rooted in machine learning.
a marriage of efforts from both fields:
image captioning: automatic image description [15, 35, 54, 77, 93, 85]
A successful approach: pre-trained CNNs + word embeddings

In the most common form of Visual Question Answering (VQA):
presented an image and a textual question about this image --> then determine the correct answer

a few words or a short phrase.
Variants include binary (yes/no) [3, 98]
multiple-choice settings [3, 100].
“fill in the blank” [95]

These affirmations essentially amount to questions phrased in declarative form.

VQA: the question to be answered is not determined until run time

segmentation or object detection, the single question to be answered is predetermined

VQA related to textual question answering, the answer found in a specific textual narrative (i.e. reading comprehension) or in large knowledge bases (i.e. information retrieval).

Textual QA: NLP community
VQA: extension to additional visual supporting information
images are much higher dimensional, and more noisy than pure text

images lack the structure and grammatical rules of language
no equivalent to the NLP tools such as syntactic parsers and regular expression matching
images: richness of the real world, natural language: represents a higher level of abstraction

Visual Question Answering	Image Captioning
more complex problem requires information not present in the image extra required information from common sense to encyclopedic knowledge about a specific element from the image	
AI-complete task [3] requires multi-modal knowledge beyond a single sub-domain a proxy to evaluate progress towards AI systems of advanced reasoning with deep language and image understanding	
easier evaluation metric Answers typically contain only a few words	image understanding could evaluated equally well through image captioning The long ground truth image captions are more difficult to compare with predicted ones. an open research problem [43, 26, 76]

One of the first integrations of vision and language is the “SHRDLU” system from 1972 [84]:
use language to instruct a computer to move various objects around in a “blocks world”

More recent creating conversational robotic agents [39, 9, 55, 64] also grounded in the visual world.
limited to specific domains and/or on restricted language forms.

VQA: free-form open-ended questions.
mature techniques in both computer vision and NLP and the availability of large-scale datasets.

a comprehensive overview of the field, covering:
(1) models
(2) datasets
(3) future directions

VQA methods of four categories:
（I）the joint embedding approaches
use CNNs and RNNs to learn embeddings of images and sentences in a common feature space.
feed them together to a classifier that predicts an answer [22, 52, 49].

（II）attention mechanisms focusing on specific parts of the input (image and/or question).
Attention in VQA [100, 90, 11, 32, 2, 92] inspired by similar techniques in image captioning [91].

replace holistic (image-wide) features with spatial feature maps,
allow interactions between the question and specific regions of these maps.

（III）compositional models
Andreas et al. [2] use a parser to decompose a given question
build a neural network out of modules reflect the structure of the question.

（IV） knowledge base-enhanced approaches
use external data by querying structured knowledge bases.
retrieving information present in the common visual datasets such as ImageNet [14] or COCO [45], only labeled with classes, bounding boxes, and/or captions.
Information available from knowledge bases ranges from common sense to encyclopedic level accessed with no need for available at training time [87, 78].

datasets for training and evaluating VQA systems.
vary widely along three dimensions:
(i) size: the number of images, questions, and different concepts represented.
(ii) the amount of required reasoning:
whether the detection of a single object is sufficient
whether inference is required over multiple facts or concepts
(iii) how much information beyond that present in the actual images is necessary,
be it common sense or subject-specific information.

existing datasets lean towards visual-level questions
require little external knowledge, with few exceptions [78, 79].

These characteristics reflect the struggle with simple visual questions still faced by the current state of the art
these characteristics must not be forgotten when VQA presented as an AI-complete evaluation proxy.
more varied and sophisticated datasets required.

an in-depth analysis of the question/answer pairs provided in the Visual Genome dataset.
the largest VQA dataset available
includes rich structured images annotations in the form of scene graphs [41].
We evaluate the relevance of these annotations for VQA, by comparing the occurrence of concepts involved in the provided questions, answers, and image annotations.
only about 40% of the answers directly match elements in the scene graphs.
this matching rate can significantly increased by relating scene graphs to external knowledge bases.
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

[51]
One of the first attempts at “open-world” visual question answering
combining semantic text parsing with image segmentation in a Bayesian formulation samples from nearest neighbors in the training set.
requires human-defined predicates, dataset-specific and difficult to scale.
dependent on the accuracy of image segmentation algorithm and the estimated depth information.

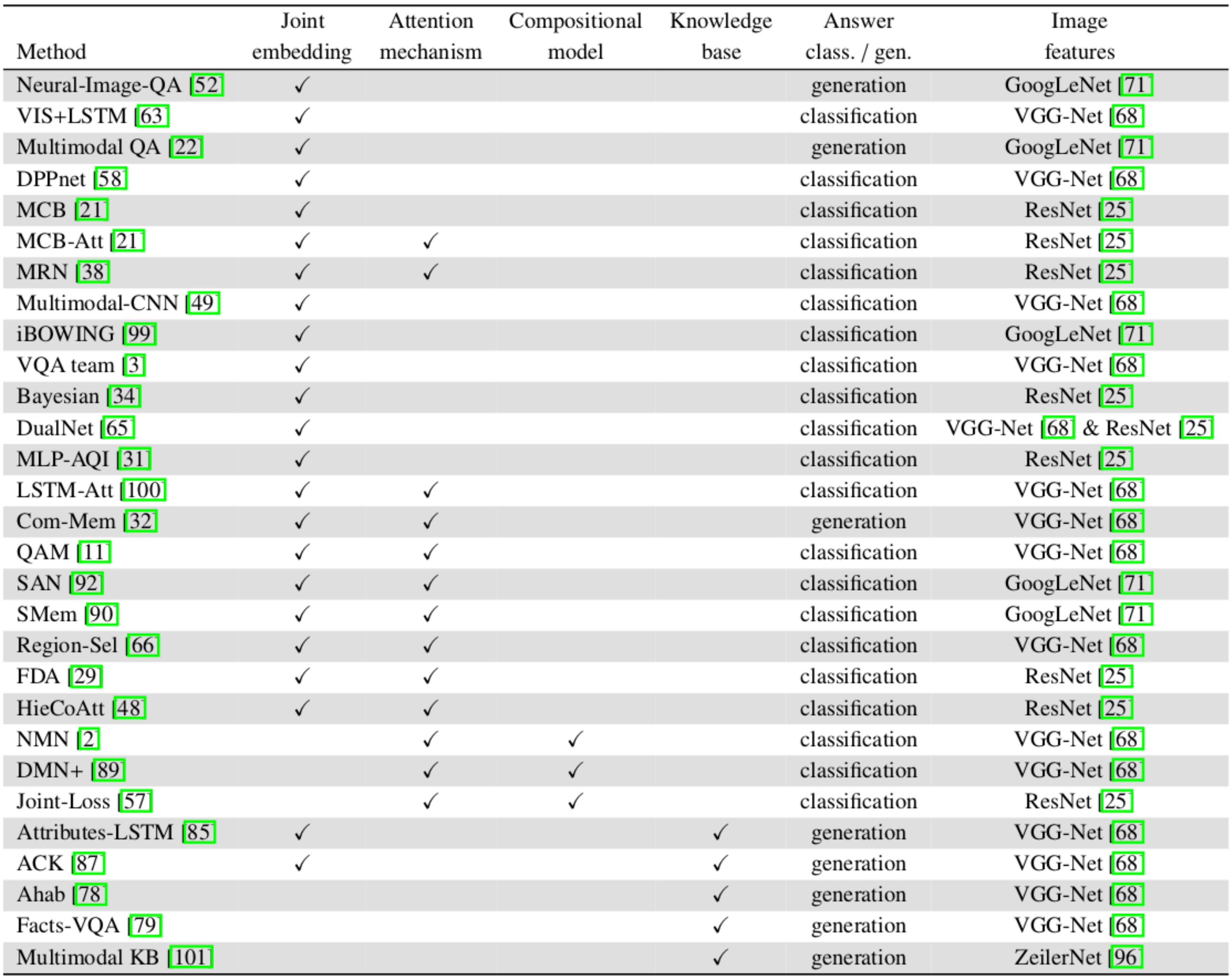
[74]
based on a joint parse graph from text and videos.

[23]
an automatic “query generator” trained on annotated images
produces a sequence of binary questions from any given test image.

these early approaches restrict questions to predefined forms.
modern approaches aimed at answering free-form open-ended questions.

methods of four categories:
（1）joint embedding approaches
（2）attention mechanisms
（3）compositional models
（4）knowledge-base-enhanced approaches

Table 1: Overview of existing approaches to VQA



Method	Joint embedding	Attention mechanism	Compositional model	Knowledge base	Answer class. / gen.	Image features
Neural-Image-QA [52]	✓				generation	GoogLeNet [71]
VIS+LSTM [63]	✓				classification	VGG-Net [68]
Multimodal QA [22]	✓				generation	GoogLeNet [71]
DPPnet [58]	✓				classification	VGG-Net [68]
MCB [21]	✓				classification	ResNet [25]
MCB-Att [21]	✓	✓			classification	ResNet [25]
MRN [38]	✓	✓			classification	ResNet [25]
Multimodal-CNN [49]	✓				classification	VGG-Net [68]
iBOWING [99]	✓				classification	GoogLeNet [71]
VQA team [3]	✓				classification	VGG-Net [68]
Bayesian [34]	✓				classification	ResNet [25]
DualNet [65]	✓				classification	VGG-Net [68] & ResNet [25]
MLP-AQI [31]	✓				classification	ResNet [25]
LSTM-Att [100]	✓	✓			classification	VGG-Net [68]
Com-Mem [32]	✓	✓			generation	VGG-Net [68]
QAM [11]	✓	✓			classification	VGG-Net [68]
SAN [92]	✓	✓			classification	GoogLeNet [71]
SMem [90]	✓	✓			classification	GoogLeNet [71]
Region-Sel [66]	✓	✓			classification	VGG-Net [68]
FDA [29]	✓	✓			classification	ResNet [25]
HieCoAtt [48]	✓	✓			classification	ResNet [25]
NMN [2]		✓	✓		classification	VGG-Net [68]
DMN+ [89]		✓	✓		classification	VGG-Net [68]
Joint-Loss [57]		✓	✓		classification	ResNet [25]
Attributes-LSTM [85]	✓			✓	generation	VGG-Net [68]
ACK [87]	✓			✓	generation	VGG-Net [68]
Ahab [78]				✓	generation	VGG-Net [68]
Facts-VQA [79]				✓	generation	VGG-Net [68]
Multimodal KB [101]				✓	generation	ZeilerNet [96]

most methods combine multiple strategies and thus belong to several categories.

（1） Joint embedding approaches

m	jointly embedding images and text was first explored for image captioning [15, 35, 54, 77, 93,

otIvatIon	85]. motivated by the success of deep learning methods in both computer vision and NLP, allow to learn representations in a common feature space. this motive reinforced in VQA by the need to perform reasoning over both modalities together. A representation in a common space allows learning interactions and performing inference over the question and the image contents. image representations obtained with CNNs pre-trained on object recognition. Text representations obtained with word embeddings pre-trained on large text corpora. Word embeddings map words to a space in which distances reflect semantic similarities [56, 61]. 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	[52] "Neural-Image-QA": RNN with Long Short-Term Memory cells (LSTMs) RNNs to handle inputs (questions) and outputs (answers) of variable size. Image features produced by a CNN pre-trained for object recognition. Question and image features fed together to a "encoder" LSTM. produces a feature vector of fixed-size then passed to a "decoder" LSTM. 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. a sequence generation procedure used common shared weights between the encoder and decoder LSTMs
	Several variants: "VIS+LSTM" [63] feed the feature vector produced by the encoder LSTM into a classifier to produce single-word answers from a predefined vocabulary. formulate the answering as a classification problem other technical improvements with the "2-VIS+BLSTM" model： uses two sources of image features to the LSTM at the start and at the end of the question sentence. uses LSTMs scan questions in both forward and backward directions. bidirectional LSTMs better capture relations between distant words in the question.
	[22] "Multimodal QA" (mQA) employs LSTMs to encode the question and produce the answer, with two differences from [52]. learns distinct parameters for the encoder and decoder LSTMs only shares the word embedding. motivated by potentially different properties (e.g. in terms of grammar) of questions and answers. the CNN features as image representations not fed into the encoder prior to the question, but at every time step.
	[58] 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, employ a separate parameter prediction network, consists of GRUs, taking a question as input and producing candidate weights through a fully-

	connected layer at its output. significantly improve answering accuracy compared to [52, 63]. a similarity with the modular approaches, the question used to tailor the main computations to each particular instance. formulate the answering as a classification problem
	[21] a pooling method to perform the joint embedding visual and text features. "Multimodal Compact Bilinear pooling" (MCB) by randomly projecting the image and text features to a higher-dimensional space then convolve both vectors with multiplications in the Fourier space for efficiency.
	[38] use a multimodal residual learning framework (MRN) to learn the joint representation of images and language.
	[65] a "DualNet" integrates element-wise summations and element-wise multiplications to embed their visual and textual features. formulate the answering as a classification problem over a predefined set of possible answers.
	[34] integrate an explicit prediction of the type of expected answer from the question formulate the answering in a Bayesian framework.
	not make use of RNNs to encode questions: [49] 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 "multi-modal CNN") forming an overall homogeneous convolutional architecture. [99] [3] use a traditional bag-of-words representation of the questions.
performance and limit	"Neural-Image-QA" one of the earliest methods, the de facto baseline result. "2-VIS+BLSTM" improves slightly on DAQUAR, thanks to the bidirectional LSTM used to encode the questions. "mQA": not tested on publicly available datasets not comparable. DPPnet [58] benefit from tailoring computations to each question through the dynamic parameter layer. outperformed other joint embeddings methods [52, 63, 3, 99]. MCB pooling [21] and multimodal residual learning (MRN) bring further improvements achieve the top performances at the time of this writing. The joint embedding approaches： straightforward in their principle constitute the base of most current approaches to VQA.

ations	The latest improvements, exemplified by MCB and MRN, still showed potential room for improvement on both （1）the extraction of features （2）their projection to the embedding space

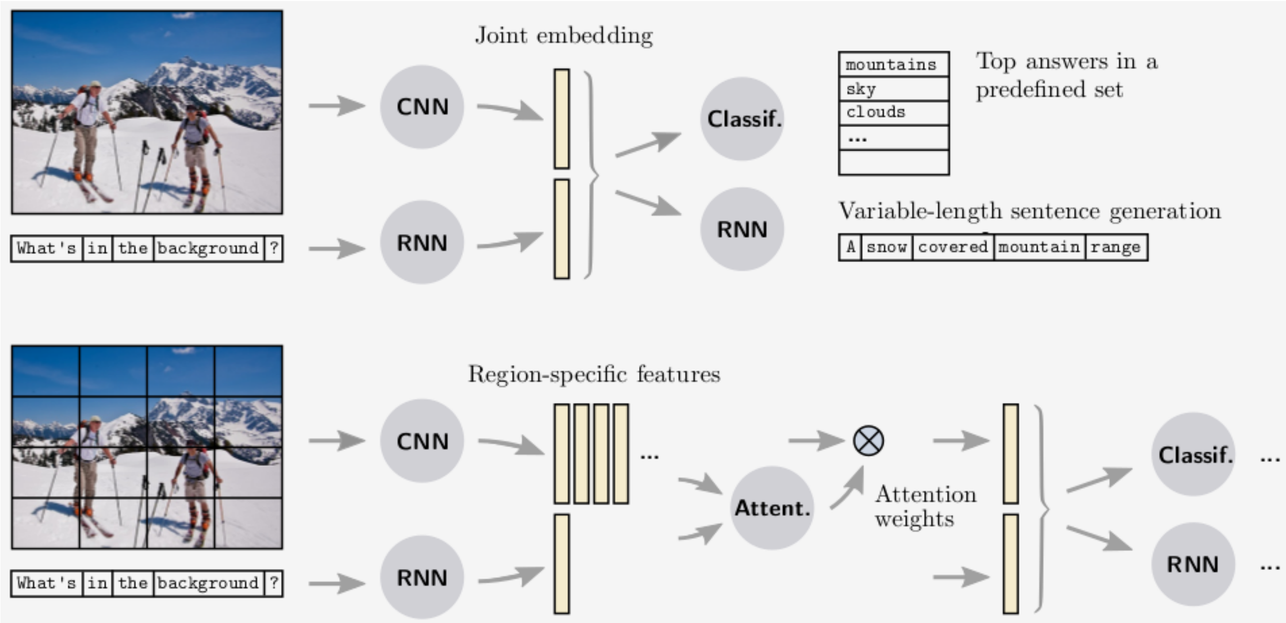

Joint embedding
CNN
RNN
What's in the background ?
Classif.
RNN
mountains
sky
clouds
...
Top answers in a predefined set
Variable-length sentence generation
A snow covered mountain range
Region-specific features
CNN
RNN
What's in the background ?
...
Attent.
Attention weights
Classif.
RNN
...
...

Figure 1: (Top) A common approach to VQA is to map both the input image and question to a common embedding space. 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 arederived from both the image and the question and allow the output stage to focus on relevant parts of the image.

（2）Attention mechanisms

Motivation	limitation：global (image-wide) features to represent the visual input feed irrelevant or noisy information to the prediction stage. attention mechanisms address this issue: (1) using local image features (2) allowing the model to assign different importance to features from different regions [91] image captioning early application of attention to visual tasks The attentional component identifies salient regions in an image further processing then focuses the caption generation on those regions.

	translates readily to VQA for focusing on image regions relevant to the question. the attention process forces an explicit additional step in the reasoning process that identifies “where to look” before performing further computations. 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	[100] described how to add spatial attention to the standard LSTM model. computations performed by an attention-enhanced LSTM: sigmoid nonlinearity： σ the input (e.g. a word of a question)： x_t hidden state at time step t： h_t element-wise products with gate values： $\circ$ rained parameters： W and b $e_t = w^T_a \tanh(W_{he}h_{t-1} + W_{ce}C(I)) + b_a$ $a_t = \mathrm{softmax}(e_t)$ attention mechanism： $z_t = a^T_t C(I)$ the convolutional feature map of image I： C(I) Input state ： $i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{zi}z_t + b_i)$ forget state ： $f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{zf}z_t + b_f)$ $g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + W_{zg}z_t + b_c)$ memory state： $c_t = f_t \circ c_{t-1} + i_t \circ g_t$ output state ： $o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{zo}z_t + b_o)$ $h_t = o_t \circ \tanh(c_t)$ standard LSTM considered as a special case with values in a t set uniformly, i.e. each region contributing equally.
	[32] similar mechanism
	[11] use a mechanism different from the above word-guided attention. 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. search is achieved by convolving the visual feature map with a configurable convolutional kernel. 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.
	[92] also employ this scheme with “stacked attention networks” (SAN) infer the answer iteratively.
	[90] “multi-hop image attention scheme” (SMem).

	The first hop is a word-guided attention, a second hop is question-guided.
	[66] generate image regions with object proposals then select regions relevant to the question and possible answer choices.
	[29] employ off-the-shelf object detectors to identify regions related to the key words of the question then fuse information from those regions with global features with an LSTM.
	[48] “hierarchical co-attention model” (HieCoAtt) jointly reasons about image and question attention. HieCoAtt processes image and question symmetrically the image representation guides attention over the question and vice versa.
	[21] combine the attention mechanism into their “Multimodal Compact Bilinear pooling” (MCB)
	[2] employ attention mechanisms in a different manner. a compositional model builds a neural network from modules tailored to each question. 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	uses of attention mechanisms always improve over models that use global image features. [100] the attention-enhanced LSTM outperforms the “VIS+LSTM” model [63] in both “Telling” and “Grounding” tasks of the ‘Visual7W’ dataset [92] The multiple attention layers of SAN bring further improvements over only one layer of attention [32, 11], especially on the VQA dataset. [48] The HieCoAtt model 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). attention mechanisms: (1) improve the overall accuracy on all VQA datasets (2) show little or no benefit on binary (yes/no) questions.

One hypothesis：
binary questions typically require longer chains of reasoning
open-ended questions often require identifying and naming only one concept from the image.
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.
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.

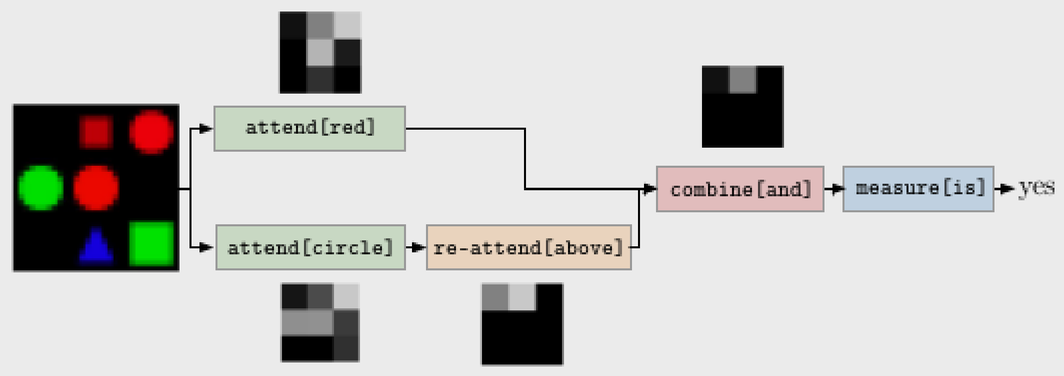


Figure 2: The Neural Module Networks (NMN) leverage the compositional structure of questions, e.g. here “Is there a red shape above a circle ?” from the Shapes dataset. The parsing of the question leads to assembling modules that operate in the space of attentions. Two attend modules locate red shapes 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.

（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.

consider modular architectures.
connecting distinct modules designed for specific desired capabilities such as
memory or
specific types of reasoning.
A potential advantage： better use of supervision.

facilitates transfer learning, since a same module can be used and trained within different overall architectures and tasks.
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 [83])

two specific models whose main contribution is in the modular aspect,
（1）Neural Module Networks (NMN)
（2）Dynamic Memory Networks (DMN).

（3-1）Neural Module Networks

Motivation	introduced in [2] extended in [1] specifically designed for VQA with the intention of exploiting the compositional linguistic structure of the questions Questions vary greatly in the level of complexity Is this a truck ? only requires retrieving one piece of information from the image How many objects are to the left of the toaster ? requires multiple processing steps, such as recognition and counting. reflect the complexity of a question in a network assembled on-the-fly for each instance of the problem. The tactic is related to approaches in textual QA [44] use semantic parsers to turn questions into logical expressions. A significant contribution of NMNs： apply this logical reasoning over continuous visual features, instead of discrete or logical predicates.
Method	based on a semantic parsing of the question using a well-known NLP tool parse tree turned into an assembly of modules from a predefined set, then used together to answer the question. all modules are independent and composable. the computation performed will be different for each problem instance 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: (1) image features (2) attentions (regions) over images (3) labels (classification decisions) A set of possible modules： （1）predefined by its type of input and output,

	（2）exact behavior acquired through end-to-end training on specific problem instances. The training does not need additional supervision than triples of images, questions, and answers. The parsing of the question is a crucial step, performed with the Standford dependency parser [13] identifies grammatical relations between parts of the sentence. use hand-written rules to transform parse trees into structured queries, in the form of compositions of modules [2]. In second paper [1], 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 [68].
Performance and limitations	Neural Module Networks evaluated on the VQA benchmark shows different strengths and weaknesses than competing approaches. generally outperforms competitors on questions with a compositional structure, e.g. requiring an object to be located and one of its attributes described. many of questions in the VQA dataset are quite simple, and require little composition or reasoning. introduced a new dataset, named “Shapes”, of synthetic images paired with complex questions involving （1）spatial relations, （2）set-theoretic reasoning, and （3）shape and （4）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. 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 possible 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. 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.
（3-2）Dynamic Memory Networks	
Motivation	Dynamic Memory Networks (DMN)：neural networks with a particular modular

	architecture. described in [42], with a number of variants proposed[83, 70, 8, 59]. Most of these were applied to textual QA. [89] adapted to VQA. DMNs：memory-augmented networks, perform read and write operations on an internal representation of the input. similarly to attention, designed to address tasks that require complex logical reasoning by modeling interaction between multiple parts of the data over several passes.
Method	composed of 4 main modules [42] allow independence in their particular implementation： （1）The input module transforms the input data into a set of vectors called “facts”. implementation (described below) varies depending on the type of input data. （2）The question module computes a vector representation of the question, using a gated recurrent unit (GRU, a variant of LSTM). （3）The episodic memory module retrieves the facts required to answer the question. allow to pass multiple times over the facts to allow transitive reasoning. an attention mechanism that selects relevant facts 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. （4）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 [89]： extracts image features with a VGG CNN [68] 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. adaptation of the original input module in [42] that used a GRU to process words of sentences. The episodic memory module also includes an attention mechanism to focus on particular image regions. [57]：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	evaluated on the DAQUAR and VQA benchmarks show competitive performance for all types of questions. Compared to Neural Module Networks, 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 [89] using essentially a same method, except for the input module. The text QA dataset used [82] requires inference over multiple facts, which is a positive indicator of the reasoning capabilities of this model. A potential criticism： 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

（4） **Models using external knowledge bases**	
Motivation	VQA: understanding the contents of images often requires: prior non-visual, information, range from “common sense” to topic-specific or even encyclopedic knowledge. “How many mammals appear in this image ?”: understand the word “mammal” know which animals belong to this category two major weaknesses of the joint embedding approaches: (1) only capture knowledge present in the training set, never reach a complete coverage of the real world. (2) neural networks trained in such approaches have a limited capacity surpassed by the amount of information we wish to learn. decouple the reasoning (e.g. as a neural network) from the actual storage of data or knowledge. structured representations of knowledge. large-scale Knowledge Bases (KB): （1）DBpedia [5], （2）Freebase [7], （3）YAGO [27, 50], （4）OpenIE [6, 17, 18], （5）NELL [10], （6）WebChild [73, 72], （7）ConceptNet [47]. store （1）common sense and （2）factual knowledge in a machine readable fashion.

	fact： piece of knowledge represented as a triple (arg1,rel,arg2), – arg1 and arg2 represent two concepts – rel represents a relationship between them The collection of such facts forms a interlinked graph described according to a Resource Description Framework (RDF [24]) specification accessed by query languages such as SPARQL [62]. Linking such knowledge bases to VQA methods allows separating the reasoning from the representation of prior knowledge in a practical and scalable manner.
Method	[78] propose “Ahab”: a VQA framework uses DBpedia, one of the largest structured knowledge bases. Visual concepts first extracted from the given image with CNNs then associated with nodes from DBpedia that represent similar concepts. Whereas the joint embedding approaches learn a mapping from images/questions to answers propose 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: handle limited types of questions. Although the questions can be provided in natual language, parsed using manually designed templates.
	FVQA [79] An improved method uses an LSTM and a data-driven approach to learn the mapping of images/questions to queries. 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: （1）the chain of reasoning [78] （2）the supporting facts [79] used in the inference process. monolithic neural networks provide little insight into the computations performed to produce their final answer.
	[87] proposed a joint embedding approach also benefits from external knowledge bases. Given an image, first extract semantic attributes with a CNN. External knowledge related to these attributes is then retrieved from a version of DBpedia containing short descriptions, embedded into fixed-size vectors with Doc2Vec. The embedded vectors are fed into an LSTM model that interprets them with the question 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, performance on these datasets poorly reflect the capabilities of these methods. evaluation on new small-scale datasets. Ahab [78] significantly outperforms joint embedding approaches on its KB-VQA dataset [78] in terms of overall accuracy (69.6% vs. 44.5%). Ahab becomes significantly better than joint embedding approaches on visual questions requiring a higher level of knowledge. FVQA approach [79] performs much better than conventional approaches [79] 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. [87] evaluated on the Toronto COCO-QA and VQA datasets shows an advantage in using the external KB in terms of average accuracy.

3 Datasets and evaluation

triple:
(1) an image
(2) a question
(2) its correct answer

Additional annotations:
(1) image captions
(2) image regions supporting the answers
(3) multiple-choice candidate answers

Datasets and questionsvary in their complexity, the amount of reasoning and of non-visual (e.g. “common sense”) information required to infer the correct answer.

We broadly classify dataset according to their type of images (natural, clipart, synthetic).
A given dataset is typically used for both training and evaluating a VQA system.
The open-ended nature of the task suggests however that other, large-scale sources of information would be beneficial and likely necessary train practical VQA systems.
Some datasets specifically address this aspect through annotations of supporting facts in structured non-visual knowledge bases.
Other datasets non-specific to VQA are worth mentioning.
They target other tasks involving vision and language, such as image captioning (e.g. [12, 26, 45, 94]), generating [53] and understanding [36, 28] referring expressions for retrieval of images and objects in natural language.
Those datasets 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. [23]. The dataset comprises questions generated from templates from a fixed vocabulary of objects, attributes, and relationships between objects.

Another early dataset was presented in [74] by Tu et al. They study the joint parsing of videos and text to answer queries, and consider two datasets containing 15 video clips each.
restricted to limited settings
relatively small size

DAQUAR
DAtaset for QUestion Answering on Real-world images [51]
The first VQA dataset designed as benchmark.
images from the NYU-Depth v2 dataset [67],
1449 RGBD images of indoor scenes
annotated semantic segmentations.

split to 795 training and 654 test images.
Two types of question/answer pairs:
(1) synthetic questions/answers are generated automatically using 8 predefined templates and the existing annotations of the NYU dataset.
(2) human questions/answers are collected from 5 annotators.
They were instructed to focus on basic colors, numbers, objects (894 categories), and sets of those.

12,468 question/answer pairs
6,794 for training and 5,674 for testing.

enable the development and training of the early methods for VQA with deep neural networks [52, 63, 49].

The main disadvantage is the restriction of answers to a predefined set of 16 colors and 894 object categories.
presents strong biases showing that humans tend to focus on a few prominent objects, such as tables and chairs [51].

COCO-QA [63]
images from the Microsoft Common Objects in Context data (COCO) dataset [45].
123,287 images (72,783 for training and 38,948 for testing)
each image has one question/answer pair.
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:
(1) object
(2) number
(3) color
(4) 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, 9,072 (23.29%) of them also appear as training questions.

FM-IQA [22]
Freestyle Multilingual Image Question Answering dataset

uses 123,287 images, sourced from the COCO dataset
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 [3]
One of the most widely used dataset comes from the VQA team at Virginia Tech, commonly referred to simply as VQA .
VQA-real: natural images
VQA-abstract: cartoon images
VQA-real:
123,287 training and 81,434 test images
sourced from COCO [45]
Human annotators were encouraged to provide interesting and diverse questions.
binary (i.e. yes/no) questions 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 [3] 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 [41] is the largest available dataset for VQA, with 1.7 million question/answer pairs.
It is built with images from the Visual Genome project [41], 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':
who
what
where
when
why
how
which

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 [3], 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.
The use of this information to help designing and training VQA systems is however still an open research question.
The Visual7w [100] 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 [95] 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.
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 [45] 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.

[51]：wo evaluation metrics
（1）measure the accuracy with respect to the ground truth using string matching.
Only exact matches are considered as correct.
（2）Wu-Palmer similarity (WUPS) [88]：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.
[51], the metric is evaluated against two thresholds, 0.9 and 0.0.
[22], Gao et al. 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 [3] 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:

In other words, an answer is deemed 100% accurate if at least 3 annotators provided that exact answer.
Other datasets such as [41, 100] 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. [95]) 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 [3],
DAQUAR [51], and
COCO-QA [63] datasets.
We summarize results on these three main datasets in Tables 6, 7, 8 and 9.

3.2 Datasets of clip-art images

This section discusses datasets of synthetic images created manually from clip-art illustrations.

“abstract scenes” [3], 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 [20, 46, 75], learning models of interactions between people [4], generating scenes from natural language descriptions [103], and learning the semantic relevance of visual features [102, 104].

VQA abstract scenes
The VQA benchmark contains clip-art 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 [3] and [98].
Balanced dataset Another version of the dataset discussed above is presented in [98].
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 [98] 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 [3], 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 [98].
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 5,328 pairs of complementary scenes for the training and test set respectively.
Evaluation should use the VQA evaluation metric [3].
Results were reported using combinations of balanced and unbalanced training and test sets [98], 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 [51] are purely visual in nature, referring to colors, numbers, and physical locations of objects.

In the COCO-QA dataset [63], questions are generated automatically from image captions which describe the major visual elements of the image.
In the VQA dataset [3], 5.5% of questions require “adult-level common sense”, but none require “knowledge base-level” knowledge [78].
We discussed in Section 2.4 methods for VQA that make use of external knowledge bases.
The authors of two such methods [78, 79] 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 [78] was constructed to evaluate the performance of the Ahab VQA system [78].
It contains questions requiring topic-specific knowledge that is present in DBpedia.
700 images were selected from the COCO image dataset [45] and 3 to 5 question/answer pairs were collected for each, for a total of 2,402 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 [79] 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 [3, 22, 63, 100] 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 [5], ConceptNet [47], and Webchild [73, 72].
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 4,608 questions.

3.4 Other datasets

Diagrams Kembhavi et al. [37] propose a dataset for VQA on diagrams, named as AI2 Diagrams (AI2D).
It comprises more than 5,000 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. [2] 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.
he authors provide results of their own method and of a reimplementation of a joint embedding baseline [63].
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 [82].

4 Structured scene annotations for VQA

The Visual Genome [41] is currently the largest dataset available for VQA.
scene graphs: human-generated structured annotations for each image
a scene graph: 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.
a significant step toward rich and more comprehensive annotations, compared to object-level and image-level annotations.

whether the answer actually appears as an element of the graph:
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
apply the above procedure on all images and questions of the Visual Genome dataset.
only 40.02% of the answers can be directly found in the scene graph,
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.
remains 53% of questions can not be directly answered from the scene graph.
ratio is surprisingly high

To characterize the remaining 53% of questions:
examine question types using their first few words:
A large number of questions starting with “what” are among those that cannot be directly answered by scene graphs
the number of questions that cannot be answered from the scene graph as a fraction of all questions of each type separately.
a large fraction of the questions starting with “when”, “why” and “how” have answers not be found in scene graphs.
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 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 current scene graphs are rich intermediate abstractions of the scenes, but not comprehensive enough to capture all elements required for VQA.
low-level visual attributes such as color and material are lost in this representation and 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 involved in “where” and “when” questions.
It remains debatable whether more human effort should be invested to obtain comprehensive annotations.
Another solution: 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, examine whether the correct answer can be found in a first-order extension of the scene graph using relations from relations in the KB.

use the labels of all nodes of the scene graph to query 3 large Kbs:
(1) DBpedia [5]
(2) WebChild [73, 72]
(3) ConceptNet [47]

The triples resulting from the query are used to expand the scene graph and its vocabulary.
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.
completes the scene-specific graph with general, relevant knowledge.
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, nearly the double of the same experiment without the KB expansion (40.02%).

the potential for complementing the interpretation of visual contents with information from general-purpose Kbs.

5 Discussion and future directions

VQA is a complex task
VQA constitutes an AI complete task in its ultimate form
this ultimate goal is indisputably a long way from any current technique.
Reduced and limited forms of VQA are reasonable intermediate objectives that seem attainable:
(1) multiple-choice format,
(2) short answer lengths,
(3) limited types of questions,
(4) etc.
Their evaluation is practically easier and may be more representative of our actual progress.

a diversity of protocols for collecting data (from human annotators or semi-automatically from image captions)
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.
(1) 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.
(2) they influence the ratio of visual vs textual understanding required.

One extreme example is the synthetic “Shapes” dataset 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.
(3) 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
VQA constitutes an AI-complete challenge： most tasks in AI can be formulated as questions over images.
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.
[78, 79] can only handle a limited number of question types predefined by hand-coded templates.
[87] 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
Two distinct types of approaches:
(1) information retrieval
use unstructured collections of documents
the key words of the question are looked for to identify relevant passages and sentences [33, 40]
A ranking function then sorts these candidates
the answer is extracted from one or several top matches.
can be compared to the basic "joint embedding with attention" method of VQA
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 brought to VQA by Kafle and Kanan [34].

(2) semantic parsing coupled with knowledge bases.
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 [16, 30, 69, 81].
These queries then executed on domain-specific databases or general purpose structured knowledge bases.
similar process is used in the VQA methods of for querying external knowledge bases [78, 79].
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. [1, 89]).
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. [97, 60, 16]).

6 Conclusion

the state-of-the-art on visual question answering

the approach maps questions and images to vector representations in a common feature space

described additional improvements that build up on this concept, namely attention mechanisms, modular and memory-augmented architectures

growing number of datasets for training and evaluating VQA methods, differences in the type and difficulty of questions

promising directions for future research

inclusion of additional external knowledge from structured knowledge bases

potential of natural language processing tools

References				
[1]	J. Andreas, M. Rohrbach, T. Darrell, and D. Klein	Learning to compose neural networks for question answering	Annual Conference of the North American Chapter of the Association for Computational Linguistics	2016
[2]	J. Andreas, M. Rohrbach, T. Darrell, and D. Klein	Neural Module Networks	Proc. IEEE Conf. Comp. Vis. Patt. Recogn	2016
[3]	S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh	VQA: Visual Question Answering	Proc. IEEE Int. Conf. Comp. Vis.	2015
[4]	S. Antol, C. L. Zitnick, and D. Parikh	Zero-shot learning via visual abstraction	Proc. Eur. Conf. Comp. Vis.	2014
[5]	S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives	DBpedia: A nucleus for a web of open data.		2007
[6]	M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni	Open information extraction for the web	Proc. Int. Joint Conf. Artificial Intell.	2007
[7]	K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor	Freebase: a collaboratively created graph database for structuring human knowledge	ACM SIGMOD International Conference on Management of Data	2008
[8]	A. Bordes, N. Usunier, S. Chopra, and J. Weston	Large-scale simple question answering with memory networks	arXiv	2015
[9]	R. Cantrell, M. Scheutz, P.	Robust spoken instruction understanding for hri	Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International	2010

	Schermerhorn, and X. Wu		Conference on, pages 275–282. IEEE	
[10]	A. Carlson, J. Betteridge, B. Kisiel, and B. Settles	Toward an Architecture for Never-Ending Language Learning	Proc. Conf. AAAI	2010
[11]	K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia	ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering	arXiv	2015
[12]	X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick	Microsoft COCO captions: Data collection and evaluation server	arXiv	2015
[13]	M.-C. de Marneffe and C. D. Manning	The stanford typed dependencies representation	COLING Workshop on Cross-framework and Cross-domain Parser Evaluation	2008
[14]	J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei	Imagenet: A large-scale hierarchical image database	Proc. IEEE Conf. Comp. Vis. Patt. Recogn	2009
[15]	J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell	Long-term recurrent convolutional networks for visual recognition and description	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2015
[16]	L. Dong and M. Lapata	Language to logical form with neural attention	Proc. Conf. Association for Computational Linguistics	2016
[17]	O. Etzioni, A. Fader, J. Christensen, S. Soderland, and M. Mausam	Open Information Extraction: The Second Generation	Proc. Int. Joint Conf. Artificial Intell.	2011
[18]	A. Fader, S. Soderland, and O. Etzioni	Identifying relations for open information extraction	Proc. Conf. Empirical Methods in Natural Language Processing	2011
[19]	F. Ferraro, N. Mostafazadeh, T.-H. Huang, L. Vanderwende, J. Devlin, M. Galley, and M. Mitchell	A survey of current datasets for vision and language research	Conference on Empirical Methods in Natural Language Processing pages 207–213. Association for Computational Linguistics	2015
[20]	D. F. Fouhey and C. Zitnick	Predicting object dynamics in scenes	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2014
[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	arXiv	2016

[22]	H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu	Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering	Proc. Advances in Neural Inf. Process. Syst.	2015
[23]	D. Geman, S. Geman, N. Hallonquist, and L. Younes	Visual Turing test for computer vision systems	Proceedings of the National Academy of Sciences	2015
[24]	R. W. Group et al.	Resource description framework	http://www.w3.org/standards/techs/rdf	2014
[25]	K. He, X. Zhang, S. Ren, and J. Sun	Deep residual learning for image recognition	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2016
[26]	M. Hodosh, P. Young, and J. Hockenmaier	Framing image description as a ranking task: Data, models and evaluation metrics	JAIR	2013
[27]	J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum	YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia	Proc. Int. Joint Conf. Artificial Intell	2013
[28]	R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell	Natural language object retrieval	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2016
[29]	I. Ilievski, S. Yan, and J. Feng	A focused dynamic attention model for visual question answering	arXiv	2016
[30]	M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé III	A neural network for factoid question answering over paragraphs	Empirical Methods in Natural Language Processing	2014
[31]	A. Jabri, A. Joulin, and L. van der Maaten	Revisiting visual question answering baselines	arXiv	2016
[32]	A. Jiang, F. Wang, F. Porikli, and Y. Li	Compositional Memory for Visual Question Answering	arXiv	2015
[33]	D. Jurafsky and J. H. Martin	Question answering	Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, chapter 28	2000
[34]	K. Kafle C. Kanan	Answer-type prediction for visual question answering	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2016
[35]	A. Karpathy, A. Joulin, and F. F. Li	Deep fragment embeddings for bidirectional image sentence mapping	Proc. Advances in Neural Inf. Process. Syst.	2014
[36]	S. Kazemzadeh, V.	Referitgame: Referring to	Conference on Empirical Methods	2014

	Ordonez, M. Matten, and T. L. Berg	objects in photographs of natural scenes	in Natural Language Processing	
[37]	A. Kembhavi, M. Salvato, E. Kolve, M. J. Seo, H. Hajishirzi, and A. Farhadi	A diagram is worth a dozen images	arXiv	2016
[38]	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	arXiv	2016
[39]	T. Kollar, J. Krishnamurthy, and G. P. Strimel	Toward interactive grounded language acqusition	Robotics: Science and Systems	2013
[40]	O. Kolomiyets and M.-F. Moens	A survey on question answering technology from an information retrieval perspective	Information Sciences	2011
[41]	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	2016
[42]	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	Proc. Int. Conf. Mach. Learn.	2016
[43]	S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi	Composing simple image descriptions using web-scale n-grams	The SIGNLL Conference on Computational Natural Language Learning	2011
[44]	P. Liang, M. I. Jordan, and D. Klein	Learning dependency-based compositional semantics	Computational Linguistics	2013
[45]	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	Proc. Eur. Conf. Comp. Vis	2014
[46]	X. Lin and D. Parikh	Don’t just listen, use your imagination: Leveraging visual common sense for non-visual tasks	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2015
[47]	H. Liu and P. Singh.	ConceptNet - A practical	BT technology journal	2004

		commonsense reasoning toolkit		
[48]	J. Lu, J. Yang, D. Batra, and D. Parikh	Hierarchical question-image co-attention for visual question answering	arXiv	2016
[49]	L. Ma, Z. Lu, and H. Li.	Learning to Answer Questions From Image using Convolutional Neural Network	Proc. Conf. AAAI	2016
[50]	F. Mahdisoltani, J. Biega, and F. Suchanek	YAGO3: A knowledge base from multilingual Wikipedias	CIDR	2015
[51]	M. Malinowski and M. Fritz	A multi-world approach to question answering about real-world scenes based on uncertain input	Proc. Advances in Neural Inf. Process. Syst.	2014
[52]	M. Malinowski, M. Rohrbach, and M. Fritz	Ask Your Neurons: A Neural-based Approach to Answering Questions about Images	Proc. IEEE Int. Conf. Comp. Vis.	2015
[53]	J. Mao, H. Jonathan, A. Toshev, O. Camburu, A. Yuille, and K. Murphy	Generation and comprehension of unambiguous object descriptions	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2016
[54]	J. Mao, W. Xu, Y. Yang, J. Wang, and A. Yuille	Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)	Proc. Int. Conf. Learn. Representations	2015
[55]	C. Matuszek, N. FitzGerald, L. Zettlemoyer, L. Bo, and D. Fox	A joint model of language and perception for grounded attribute learning	Proc. Int. Conf. Mach. Learn.	2012
[56]	T. Mikolov, K. Chen, G. Corrado, and J. Dean	Efficient estimation of word representations in vector space	arXiv	2013
[57]	H. Noh and B. Han	Training recurrent answering units with joint loss minimization for vqa	arXiv	2016
[58]	H. Noh, P. H. Seo, and B. Han	Image Question Answering using Convolutional Neural Network with Dynamic Parameter Prediction	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2016
[59]	B. Peng, Z. Lu, H. Li, and K. Wong	Towards neural network-based reasoning	arXiv	2015
[60]	B. Peng and K. Yao	Recurrent neural networks with external memory for language understanding	arXiv	2015

[61]	J. Pennington, R. Socher, and C. Manning	Glove: Global Vectors for Word Representation	Conference on Empirical Methods in Natural Language Processing	2014
[62]	E. Prud'Hommeaux, A. Seaborne, et al.	SPARQL query language for RDF	W3C recommendation	2008
[63]	M. Ren, R. Kiros, and R. Zemel	Image Question Answering: A Visual Semantic Embedding Model and a New Dataset	Proc. Advances in Neural Inf. Process. Syst.	2015
[64]	D. Roy, K.-Y. Hsiao, and N. Mavridis	Conversational robots: building blocks for grounding word meaning	HLT-NAACL Workshop on Learning word meaning from non-linguistic data. Association for Computational Linguistics	2003
[65]	K. Saito, A. Shin, Y. Ushiku, and T. Harada	Dualnet: Domain-invariant network for visual question answering	arXiv	2016
[66]	K. J. Shih, S. Singh, and D. Hoiem	Where to look: Focus regions for visual question answering	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2016
[67]	N. Silberman, D. Hoiem, P. Kohli, and R. Fergus	Indoor segmentation and support inference from rgbd images	Proc. Eur. Conf. Comp. Vis.	2012
[68]	K. Simonyan and A. Zisserman	Very deep convolutional networks for large-scale image recognition	arXiv	2014
[69]	V. Singh and S. K. Dwivedi	Question answering: A survey of research, techniques and issues	Int. J. Inf. Retr. Res.	2014
[70]	S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus	Weakly supervised memory networks	arXiv	2015
[71]	C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich	Going deeper with convolutions	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2015
[72]	N. Tandon, G. de Melo, F. Suchanek, and G. Weikum	Webchild: Harvesting and organizing commonsense knowledge from the web	International Conference on Web Search and Data Mining. ACM	2014
[73]	N. Tandon, G. De Melo, and G. Weikum	Acquiring Comparative Commonsense Knowledge from the Web	Proc. Conf. AAAI	2014
[74]	K. Tu, M. Meng, M. W. Lee, T. E. Choe, and S.-C. Zhu	Joint video and text parsing for understanding events and answering queries	Trans. Multimedia	2014

[75]	R. Vedantam, X. Lin, T. Batra, C. L. Zitnick, and D. Parikh	Learning common sense through visual abstraction	Proc. IEEE Int. Conf. Comp. Vis.	2015
[76]	R. Vedantam, C. L. Zitnick, and D. Parikh	CIDEr: Consensus-based Image Description Evaluation	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2015
[77]	O. Vinyals, A. Toshev, S. Bengio, and D. Erhan	Show and tell: A neural image caption generator.	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2014
[78]	P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick	Explicit knowledge-based reasoning for visual question answering	arXiv	2015
[79]	P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick	FVQA: Fact-based visual question answering	arXiv	2016
[80]	P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. Dick	FVQA: Fact-based visual question answering	arXiv	2016
[81]	Y. Wen-tau, C. Ming-Wei, H. Xiaodong, and G. Jianfeng	Semantic parsing via staged query graph generation: Question answering with knowledge base	International Joint Conference on Natural Language Processing of the AFNLP. ACL	2015
[82]	J. Weston, A. Bordes, S. Chopra, and T. Mikolov	Towards ai-complete question answering: A set of prerequisite toy tasks	arXiv	2015
[83]	J. Weston, S. Chopra, and A. Bordes	Memory networks	arXiv	2014
[84]	T. Winograd	Understanding natural language	Cognitive psychology	1972
[85]	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?	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2016
[86]	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	2016
[87]	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	Proc. IEEE Conf. Comp. Vis. Patt. Recogn.	2016
[88]	Z. Wu and M. Palmer.	Verbs semantics and lexical selection	Proc. Conf. Association for Computational Linguistics	1994

[89]	C. Xiong, S. Merity, and R. Socher	Dynamic memory networks for visual and textual question answering	Proc. Int. Conf. Mach. Learn.	2016
[90]	H. Xu and K. Saenko	Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering	arXiv	2015

[91] 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. Mach. Learn., 2015.

[92] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked Attention Networks for Image Question Answering. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

[93] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In Proc. IEEE Int. Conf. Comp. Vis., 2015.

[94] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Proc. Conf. Association for Computational Linguistics, 2, 2014.

[95] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank image generation and question answering. In Proc. IEEE Int. Conf. Comp. Vis., 2015.

[96] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In Proc. Eur. Conf. Comp. Vis., 2014.

[97] L. S. Zettlemoyer and M. Collins. 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, 2007.

[98] 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. Comp. Vis. Patt. Recogn., 2016.

[99] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus. Simple baseline for visual question answering. arXiv preprint arXiv:1512.02167, 2015.

[100] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2016.

[101] Y. Zhu, C. Zhang, C. Ré, and L. Fei-Fei. Building a Large-scale Multimodal Knowledge Base System for Answering Visual Queries. arXiv preprint arXiv:1507.05670, 2015.

[102] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2013.

[103] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In Proc. IEEE Int. Conf. Comp. Vis., 2013.

[104] C. L. Zitnick, R. Vedantam, and D. Parikh. Adopting abstract images for semantic scene understanding. IEEE Trans. Pattern Anal. Mach. Intell., 38(4):627–638, 2016.