Visual Question Answering: Datasets, Algorithms, and Future Challenges

Visual Question Answering (VQA): an algorithm needs to answer text-based questions about images.		
(1)	(2)	(3)
deep learning	computer vision	natural language processing

examine the current state of VQA in terms of			
(1) problem formulation	(2) existing datasets	(3) evaluation metrics	(4) algorithms

limitations of datasets with regard to their ability to train and assess VQA algorithms.
existing algorithms for VQA.
future directions for VQA and image understanding research.

1 Introduction

computer vision tasks：
（1） image classification [1, 2]
（2）object detection [3, 4]
（3） activity recognition [5, 6, 7]
these problems are：
（1） narrow in scope
（2）require no holistic understanding of images.
humans can：
（1） identify the objects in an image
（2）understand the spatial positions of these objects
（3）infer their attributes and relationships to each other
（4）reason about the purpose of each object given the surrounding context
（5）ask arbitrary questions about images
（6）communicate the information gleaned from them.

Visual Question Answering (VQA)：
a computer vision task where a system is given a text-based question about an image, and it must infer the answer.

Questions can be arbitrary

encompass computer vision sub-problems：		
(1)	Object recognition	What is in the image?
(2)	Object detection	Are there any cats in the image?
(3)	Attribute classification	What color is the cat?
(4)	Scene classification	Is it sunny?
(5)	Counting	How many cats are in the image?
(6)	the spatial relationships among objects	What is between the cat and the sofa?
(7)	common sense reasoning	Why is the the girl crying?

A robust VQA system must be capable of
（1）solving a wide range of classical computer vision tasks
（2）needing the ability to reason about images.

potential applications for VQA：
（1）an aid to blind and visually impaired individuals, enabling them to get information about images both on the web and in the real world.
（2）used to improve human-computer interaction as a natural way to query visual content.
（3）used for image retrieval, without using image meta-data or tags.

VQA is an important basic research problem.
VQA system： a component of a Turing Test for image understanding [8, 9].
A Visual Turing Test： whether a computer vision system is capable of human-level semantic analysis of images [8, 9].
VQA：a kind of Visual Turing Test requires the ability to understand questions, but not necessarily more sophisticated natural language processing.
If an algorithm performs as well as or better than humans on arbitrary questions about images, then arguably much of computer vision would be solved.
this is only true if the benchmarks and evaluation tools are sufficient to make such bold claims.

place particular emphasis on：
（1） whether current VQA benchmarks are suitable for evaluating whether a system is capable of robust image understanding.
（2）compare VQA with other computer vision tasks
（3）the strengths and weaknesses of currently available datasets for VQA .
（4）biases in some of these datasets severely limit their ability to assess algorithms.
（5）the evaluation metrics for VQA.
（6）existing algorithms for VQA and analyze their efficacy
（7）future developments in VQA and open questions.

2 Vision and Language Tasks Related to VQA

The overarching goal of VQA:
extract question-relevant semantic information from the images, from the detection of minute details to the inference of abstract scene attributes for the whole image.

many computer vision problems limited in scope and generality compared to VQA.

classification tasks：
（1）Object recognition : rival humans in accuracy [2].
（2）activity recognition
（3）scene classification
today's best methods doing this using CNNs trained to classify images into particular semantic categories.

object recognition: classifying the dominant object in an image without
(1) its spatial position
(2) its role within the larger scene.

Object detection：
involves the localization of specific semantic concepts (e.g., cars or people) by placing a bounding box around each instance of the object in an image.
The best object detection methods all use deep CNNs [11, 4, 3].

Semantic segmentation：
takes the task of localization a step further by classifying each pixel as belonging to a particular semantic class [12, 13].

Instance segmentation：
further builds upon localization by differentiating between separate instances of the same semantic class [14, 15, 16].
While semantic and instance segmentation are important computer vision problems that:
(1) generalize object detection and recognition,
(2) not sufficient for holistic scene understanding.

major problems: label ambiguity.

VQA required to answer arbitrary questions about images, which may require reasoning about the relationships of objects with each other and the overall scene.

The appropriate label is specified by the question.
a significant amount of recent work that combines vision with language：
（1）image captioning [17, 5, 18, 19, 20]:
goal is to produce a natural language description of a given image.
broad task involves describing complex attributes and object relationships to provide a detailed description of an image.

several problems with the visual captioning task:
(1) evaluation of captions is a challenge.
The most widely used caption evaluation schemes:
(1.1) BLEU [21], originally developed for machine translation evaluation

the most widely used metric,

have the same score for large variations in sentence structure with largely varying semantic content [25].

For captions generated in [26], BLEU scores ranked machine generated captions above human captions. when human judges were used to judge the same captions, only 23.3% of the judges ranked the captions to be of equal or better quality than human captions.

(1.2) ROUGE [22], originally developed for machine translation evaluation

(1.3) METEOR [23], originally developed for machine translation evaluation

show more robustness in terms of agreement with human judges, they still often rank automatically generated captions higher than human captions [27].

(1.4) CIDEr [24], developed specifically for scoring image descriptions

show more robustness in terms of agreement with human judges, they still often rank automatically generated captions higher than human captions [27].

why evaluating captions is challenging:

(1) a given image can have many valid captions, with some being very specific and others generic in nature.

(2) captioning systems that produce generic captions that only superficially describe an image's content are often ranked high by the evaluation metrics.

(3) a simple system that returns the caption of the training image with the most similar visual features using nearest neighbor yields relatively high scores using automatic evaluation metrics [28].

Dense image captioning (DenseCap) :

（1）avoids the generic caption problem by annotating an image densely with short visual descriptions pertaining to small, but salient, image regions [29].

（2）difficult to automatically assess their quality.

（3）omit important relationships between the objects in the scene by only producing isolated descriptions for each regions.

Captioning and DenseCap：

（1）task agnostic

（2）not required to perform exhaustive image understanding.

a captioning system is

（1）at liberty to choose the level of granularity of its image analysis in contrast to VQA, where the level of granularity is specified by the nature of the question asked.

（2）many kinds of questions have specific and unambiguous answers, making VQA more amenable to automated evaluation metric than captioning.

（3）Ambiguity may still exist for some question types , but for many questions the answer produced by a VQA algorithm can be evaluated with one-to-one matching with the ground truth answer.

3 Datasets for VQA

the main datasets for VQA：

（1） DAQUAR [30],

（2） COCO-QA [31],

（3） The VQA Dataset [32],
（4） FM-IQA [33],
（5） Visual7W [34], and
（6） Visual Genome [35]

Microsoft Common Objects in Context (COCO) dataset [10]：
328,000 images
91 common object categories
2 million labeled instances
an average of 5 captions per image

Visual Genome and Visual7W use images from Flickr100M in addition to the COCO images.
SYNTH-VQA： synthetic cartoon imagery

COCO-VQA：[36, 37, 38]

statistics for each of these datasets.

	DAQUAR	COCO-QA	COCO-VQA	FM-IQA	Visual7w	Visual genome
Total Images	1,449	123,287	204,721	120,360	47,300	108,000
QAPairs	12,468	117,684	614,163	250,569	327,939	1,773,258
Distinct Answers	968	430	105,969	N/A	65,161	207,675
% covered by top-1000	100	100	82.8	N/A	56.29	60.8
% covered by top-10	25.04	19.71	51.13	N/A	17.13	13.07
Human Accuracy	50.2	N/A	83.3	N/A	96.6	N/A
Longest Question (words)	25	24	32	N/A	24	26
Longest Answer (words)	7 (list of 1 words)	1	17	N/A	20	24
Avg. Answer Length (words)	1.2	1.0	1.1	N/A	2.0	1.8
Image Source	NYUDv2	COCO	COCO	COCO	COCO	COCO, YFCC
Annotation	Manual + Auto	Auto	Manual	Manual	Manual	Manual
Evaluation Type	OE	OE	MC or OE	OE	MC or OE	OE
Question	3	4	-	-	-	-

Types						

An ideal VQA dataset needs to be:
（1）sufficiently large to capture the variability within questions, images, and concepts that occur in real world scenarios.
（2）have a fair valuation scheme that is difficult to ‘game’ and doing well on it indicates that an algorithm can answer a large variety of question types about images that have definitive answers.

DAQUAR

The DAtaset for QUestion Answering on Real-world images (DAQUAR) [30]
one of the smallest VQA datasets.
6795 training
5673 testing QA pairs
NYU-DepthV2 Dataset [39]

DAQUAR-37：
an even smaller configuration consisting of only 37 object categories,
3825 training QA pairs
297 testing QA pairs.

DAQUAR-consensus[40]：
additional ground truth answers
an alternative evaluation metric

LIMITS：
（1）too small to successfully train and evaluate more complex models.
（2）contains exclusively indoor scenes, constrains the variety of questions available.
（3）The images tend to have significant clutter and in some cases extreme lighting conditions .
This makes many questions difficult to answer, and even humans are only able to achieve 50.2% accuracy on the full dataset.

COCO-QA

COCO-QA [31],
QA pairs are created for images using an Natural Language Processing (NLP) algorithm that derives them from the COCO image captions.
78,736 training
38,948 testing QA pairs
questions about：
1）the object in the image (69.84%)
2）color (16.59%)
3）counting (7.47%)
4）location (6.10%)
single word answer
only 435 unique answers.
makes evaluation relatively straightforward

SHORTCOMING：
（1）flaws in the NLP algorithm used to generate the QA pairs

(2) only has four kinds of questions
(3) limited to the kinds of things described in COCO's captions

VQA Dataset [32]

real images from COCO + abstract cartoon images

COCO-VQA	SYNTH-VQA	Yin and Yang [42]
the portion containing real world imagery from COCO	the synthetic portion	from SYNTH-VQA
three questions per image ten answers per question 614,163 total questions 248,349 for training 121,512 for validation 244,302 for testing consensus-based evaluation metric	50,000 synthetic scenes that depict cartoon images in different simulated scenarios 100 different objects 30 different animal models 20 human cartoon models [41]: deformable limbs eight different facial expressions different age, gender, and races to provide variation in appearance. 150,000 QA pairs 3 questions per scene 10 ground truth answers per question possible to create a more varied and balanced dataset	tried to eliminate biases in the answers people have to questions

open-ended
multiple-choice formats
contains all the same QA pairs
contains 18 different choices

18 choices	The Correct Answer	the most frequent answer given by the ten annotators
	Plausible Answers	3 answers collected from annotators without looking at the image
	Popular Answers	the top 10 most popular answers in the dataset
	Random Answers	randomly selected correct answers for other questions

problems:
(1) many questions can be accurately answered without using the image due to language biases
(2) image-blind algorithms achieved 49.6% accuracy using the question alone [36].
(2) contains many subjective, opinion-seeking questions not have a single objective answer
(3) many questions seek explanations or verbose descriptions
(4) the dataset's biases
'yes/no' answers span about 38%
59% of them are answered with 'yes.'
(5) difficult to assess whether an algorithm is truly solving the VQA problem

FM-IQA: Freestyle Multilingual Image Question Answering

based on COCO [33].
human generated answers and questions

originally in Chinese, English translations available.
answers can be full sentences.
automatic evaluation with common metrics intractable.
suggested using human judges for evaluation
the judges are tasked with deciding whether or not the answer is provided by a human or not
assessing the quality of an answer on a scale of 0–2
impractical for most research groups and makes developing algorithms difficult

Visual Genome [35]

108,249 images：YFCC100M [43] + COCO images
1.7 million QA pairs
average 17 QA pairs per image.
the largest VQA dataset
no methods have been evaluated on it beyond the baselines established by the authors

6 types of 'W' questions					
What	Where	How	When	Who	Why

data collection modes	free-form method	annotators were free to ask any question about an image. human annotators tend to ask similar questions about an image's holistic content, can promote bias in the kinds of questions asked.
	region-specific method	questions about specific image regions. using Visual Genome's descriptive bounding-box annotations.

greater answer diversity：long-tailed distribution in the length of the answers
makes open-ended evaluation significantly more challenging
prompting the annotators to choose more concise answers
no binary (yes/no) questions

Visual7W：a subset of Visual Genome

47,300 images from Visual Genome also in COCO.

7 categories of questions:						
What	Where	How	When	Who	Why	Which

two distinct types of questions:		
(1)	The 'telling' questions	the answer is text-based
(2)	The 'pointing' questions	'Which'
		select the correct bounding box among alternatives.

the standard evaluation：a multiple-choice answer framework
four possible answers consist of answers that are plausible for the given question
Plausible answers: answer the question without seeing the image.
For pointing questions, four plausible bounding boxes surrounding the likely answer
contains no binary questions

SHAPES [44]

shapes of varying
(1) arrangements
(2) types
(3) colors

Questions are about:
(1) the attributes
(2) relationships
(3) positions
(4) the shapes

enables the creation of a vast amount of data, free of many of the biases
244 unique questions, every question asked about each of the 64 images in the dataset.
completely balanced and free of bias
All questions are binary
Many of the questions require positional reasoning about the layout and properties of the shapes.

An algorithm that cannot perform well on SHAPES, but performs well on other VQA datasets may indicate that it is only capable of analyzing images in a limited manner.

4 Evaluation Metrics for VQA

VQA has been posed as:
(1) an open-ended task：generates a string to answer a question
(2) a multiple-choice question： simple accuracy used to evaluate
(3) open-ended VQA: simple accuracy can also be used. an algorithm's predicted answer string must exactly match the ground truth answer.

alternatives to exact accuracy for evaluating open-ended VQA algorithms.
(I) Wu-Palmer Similarity (WUPS) [45]:
an alternative to accuracy in [30].
measure semantic meaning difference
assign a value between 0 and 1 based on the similarity of a ground truth answer and a predicted answer to each other. finding the least common subsumer between two semantic senses and assigning scores based on how far back the semantic tree needs to be traversed to find the common subsumer.
semantically similar, but non-identical, words are penalized relatively less

tends to assign relatively high scores to even distant concepts
[30] proposed to threshold WUPS scores, a score below a threshold will be scaled down by a factor. A threshold of 0.9 and scaling factor of 0.1 was suggested by [30].
This modified WUPS metric is the standard measure used for evaluating performance on DAQUAR and COCO-QA, in addition to simple accuracy.

two major shortcomings:
(1) despite thresholded version of WUPS, certain pairs of words are lexically very similar but carry vastly different meaning
(2) only works with rigid semantic concepts, which are almost always single words.

cannot used for phrasal or sentence answers that are occasionally

(II) multiple independently collected ground truth answers for each question [32] [40]
DAQUAR-consensus
an average of five human annotated ground truth answers per question were collected.
two ways to use these answers:
(1) average consensus
the final score is weighted toward preferring the more popular answer provided by the annotators
(2) min consensus
the answer needs to agree with at least one annotator.

VQA Dataset, ten answers per question

$$\text{Accuracy}_{VQA} = \min(n/3, 1), \quad (1)$$

where n is the total number of annotators that had the same answer as the algorithm.
Using this metric, if the algorithm agrees with three or more annotators then it is awarded a full score for a question.
COCO-VQA: Using Accuracy_{VQA} , the inter-human agreement on COCO-VQA is only 83.3%

result in the scores being inflated
Evaluating the open-ended responses of VQA systems is made simpler when the answers consist of one word answers.
The possibility of multiple correct answers increases greatly when answers need to be multiple words. This occurs frequently in FM-IQA, Visual7W, and Visual Genome, e.g., 27% of Visual7W answers have three or more words. In this scenario, metrics such as Accuracy_{VQA} are unlikely to help score predicted answers to ground truth answers in open-ended VQA.
FM-IQA [33] suggested using human judges to assess multi-word answers, problems:
(1) an extremely demanding process in terms of time, resources, and expenses.
difficult to iteratively improve a system by measuring how changing the algorithm altered performance.
(2) need to be given criteria for judging the quality of an answer.

FM-IQA proposed two metrics for human judges:
(1) determine whether the answer was produced by a human or not, regardless of the answer's correctness
(2) The second metric is to rate an answer on a 3-point scale of totally wrong(0), partially correct(1), and perfectly correct (2).

multiple-choice paradigm - used by part of The VQA Dataset, Visual7W, and Visual Genome.
only needs to predict which of the given choices is correct
greatly simplifies evaluation
multiple-choice is ill-suited for VQA because it undermines the effort by allowing a system to peek at the correct answer.

evaluate a VQA system: an open question.
The method to use depends on:
(1) how the dataset was constructed
(2) the level of bias within it
(3) available resources

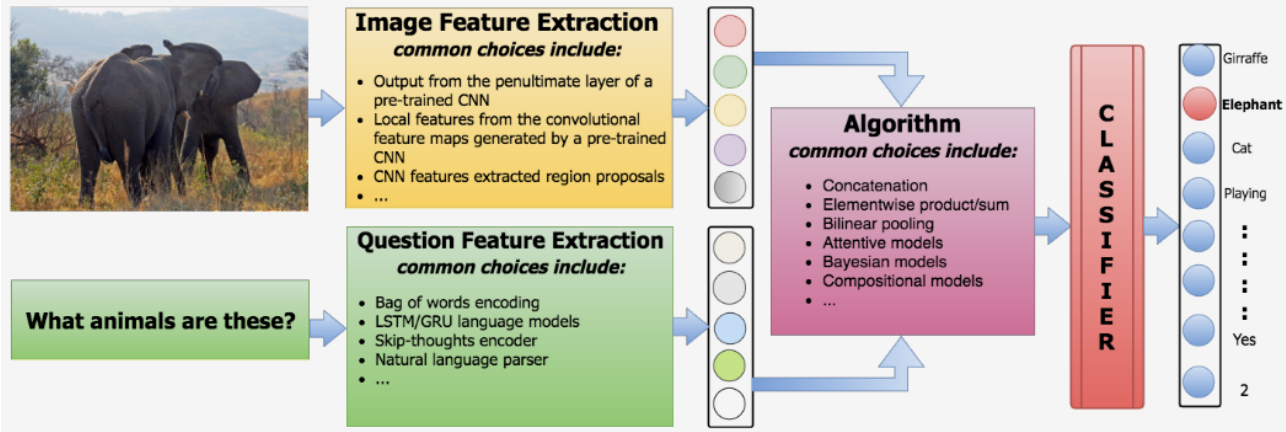
develop better tools for：

（1） measuring the semantic similarity of answers

（2）handling multi-word answers

	Pros	Cons
Simple Accuracy	Very simple to evaluate and interpret	Both minor and major errors are penalized equally
	Works well for small number of unique answers	Can lead to explosion in number of unique answers
		especially with presence of phrasal or sentence answers
Modified WUPS	More forgiving to simple variations and errors	Generates high scores for answers that are lexically related but have diametrically opposite meaning
	Does not require exact match	
	Easy to evaluate with simple script	Cannot be used for phrasal or sentence answers
Consensus Metric	Common variances of same answer could be captured	Can allow for some questions having two correct answers
	Easy to evaluate after collecting consensus data	Expensive to collect ground truth
		Difficulty due to lack of consensus
Manual Evaluation	Variances to same answer is easily captured	Can introduce subjective opinion of individual annotators
	Can work equally well for single word as well as phrase or sentence answers	Very expensive to setup and slow to evaluate, especially for larger datasets

5 Algorithms for VQA



Classification based framework for VQA

VQA algorithm consist of：

1) extracting image features (image featurization)

2) extracting question features (question featurization)
3) combines these features to produce an answer

For image features：most algorithms use CNNs that are pre-trained on ImageNet
  VGGNet [1]
  ResNet [2]
  GoogLeNet [58]
question featurizations:
  bag-of-words (BOW)
  long short term memory (LSTM) encoders [59]
  gated recurrent units (GRU) [60]
  skip-thought vectors [61]

generate an answer: the most common approach is to treat VQA as a classification problem:
  the image and question features are the input to the classification system
  each unique answer is treated as a distinct category

integrate the question and image features:
1) Combining the image and question features using simple mechanisms:
  concatenation,
  element-wise multiplication, or
  element-wise addition
  then giving them to a linear classifier or a neural network [36, 38, 32, 33]
(2) Combining the image and question features using:
  2.1 bilinear pooling
  2.2 related schemes in a neural network framework [46, 53, 62]
(3) a classifier that uses the question features to compute spatial attention maps for the visual features or that adaptively scales local features based on their relative importance [49, 51, 48, 63]
(4) Bayesian models exploiting the underlying relationships between question-image-answer feature distributions [36, 30]
(5) Using the question to break the VQA task into a series of sub-problems [50, 44]

the classification framework can only generate answers seen during training

an LSTM to produce multi-word answer [33][40]: the answer limited to words seen during training.
treating multiple-choice VQA[64][63] as a ranking problem:
  a system is trained to produced a score for each possible multiple-choice answer, question, and image trio, and then it selects the highest scoring answer choice.

	DAQUAR		COCO-QA	COCO-VQA	
	FULL	**37**		**OE**	**MC**
IMG-ONLY [36]	6.19	7.93	34.36	29.59	-
QUES-ONLY [36]	25.57	39.66	39.24	49.56	-
MULTI-WORLD [30]	7.86	12.73	-	-	-
ASK-NEURON [40]	21.67	34.68	-	-	-
ENSEMBLE [31]	-	36.94	57.84	-	-
LSTM Q+I [32]	-	-	-	54.06	57.17
iBOWIMG [38]	-	-	-	55.89	61.97
DPPNet [47]	28.98	44.48	61.19	57.36	62.69
SMem [48]	-	40.07	-	58.24	-
SAN [49]	29.3	45.5	61.6	58.9	-
NMN [44]	-	-	-	58.7	-
D-NMN [50]	-	-	-	59.4	-
FDA [51]	-	-	-	59.54	64.18
HYBRID [36]	28.96	45.17	63.18	60.06	-
DMN+ [52]	-	-	-	60.4	-
MRN [53]	-	-	-	61.84	66.33
HieCoAtten [54]	-	-	65.4	62.1	66.1
RAU_ResNet [55]	-	-	-	63.2	67.3
DAN [56]	-	-	-	64.2	69.0
MCB+Att [46]	-	-	-	64.2	-
MLB [57]	-	-	-	65.07	68.89
AMA [37]	-	-	69.73	59.44	-
MCB-ensemble [46]	-	-	-	66.5	70.1
HUMAN	**50.20**	**60.27**	**-**	**83.30**	**91.54**

5.1 Baseline Models

Baseline methods:
(1) help determine the difficulty of a dataset
(2) establish the minimal level of performance that a more sophisticated algorithms should exceed.

VQA: the simplest baselines are random guessing and guessing the most repeated answers.
widely used baseline classification system: apply a linear or non-linear, e.g., multi-layer perceptron (MLP), classifier to the image and question features after they have been combined into a single vector [32, 36, 38].

Common methods to combine the features include:
1) concatenation
2) the elementwise product

3) the elementwise sum
4) Combining these schemes lead to improved results [62].

Method	Accuracy (%) (Acc_{VQA})	CNN Network	Use of Attention	Ext. Data	Compo-sitional
LSTM Q+I [32]	54.1	VGGNet	-	-	-
iBOWIMG [38]	55.9	GoogLeNet	-	-	-
DPPNet [47]	57.4	VGGNet	-	-	-
SMem [48]	58.2	GoogLeNet	✓	-	-
SAN [49]	58.9	GoogLeNet	✓	-	-
NMN [44]	58.7	VGGNet	✓	-	✓
D-NMN [50]	59.4	VGGNet	✓	-	✓
AMA [37]	59.4	VGGNet	-	✓	-
FDA [51]	59.5	ResNet	✓	-	
HYBRID [36]	60.1	ResNet	-	-	-
DMN+ [52]	60.4	ResNet	✓	-	-
MRN [53]	61.8	ResNet	✓	-	-
HieCoAtten-VGG* [54]	60.5	VGGNet	✓	-	-
HieCoAtten-ResNet [54]	62.1	ResNet	✓	-	-
RAU_VGG* [55]	61.3	VGGNet	✓	-	-
RAU_ResNet [55]	63.2	ResNet	✓	-	
MCB* [46]	61.2	ResNet	-	-	-
MCB-ATT* [46]	64.2	ResNet	✓	-	-
DAN-VGG* [56]	62.0	VGGNet	✓	-	-
DAN-ResNet [56]	64.3	ResNet	✓	-	-
MLB [57]	65.1	ResNet	✓	-	-
MLB+VG* [57]	65.8	ResNet	✓	✓	-
MCB-ensemble [46]	66.5	ResNet	✓	✓	-

featurization approaches for baseline classification frameworks	
[38]	bag-of-words to represent the question GoogLeNet for the visual features concatenation multi-class logistic regression classifier. surpassing the previous baseline on COCO-VQA (an LSTM to represent the question [32])
[36]	skip-thought vectors [61] for question features ResNet-152 to extract image features an MLP model with two hidden layers trained on these off-the-shelf features a linear classifier outperformed the MLP model on smaller datasets
[32]	an LSTM encoder of one-hot encoding of the sentenceused to represent question features GoogLeNet for image features reduce the CNN features to match the dimensionality of the LSTM encoding the Hadamard product of these two vectors used to fuse them together. The fused vector was used as input to an MLP with two hidden layers. 4) [40] an LSTM model fed an embedding of each word sequentially with CNN features concatenated to it. continued until the end of the question was reached The subsequent time-steps were used to generate a list of answers.
[31]	an LSTM was fed CNN features during the first and last time-steps, with word features in

	between. The image features acted as the first and last words in the sentence. The LSTM network was followed by a softmax classifier to predict the answer.
[33]	the CNN image features were only fed into the LSTM at the end of the question instead of a classifier, another LSTM was used to generate the answer one word at a time

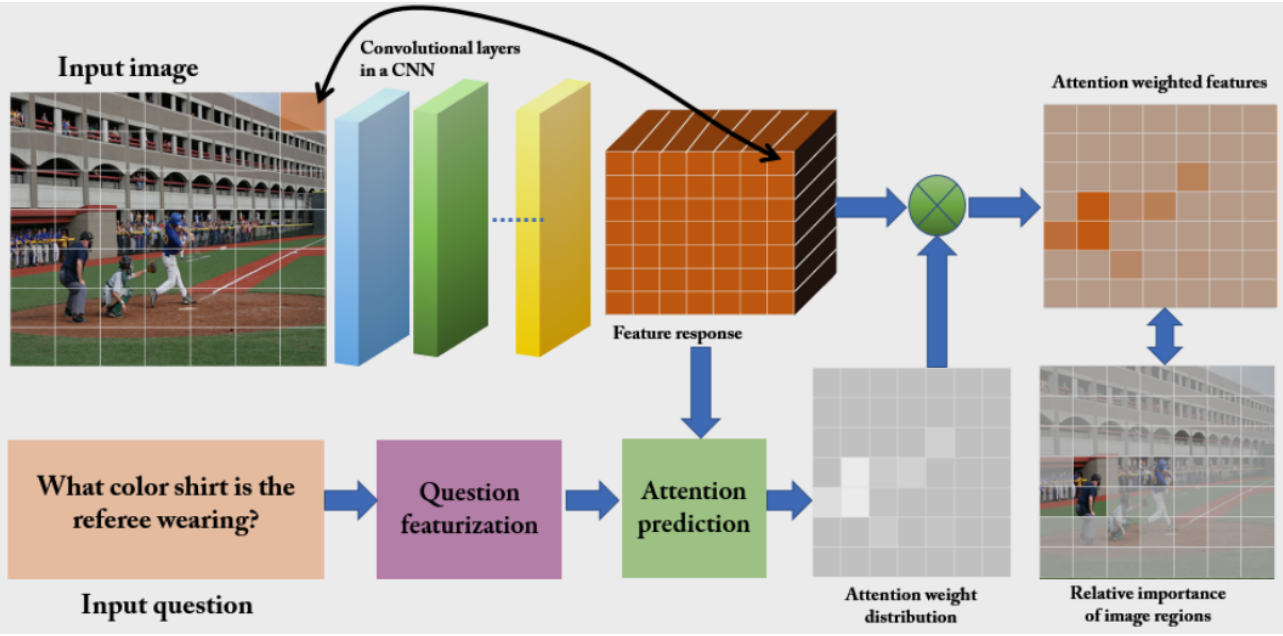
5.2 Bayesian and Question-Aware Models

VQA requires drawing inferences and modeling relationships between the question and the image.

Once the questions and images are featurized, modeling co-occurrence statistics of the question and image features can be helpful for drawing inferences about the correct answers.

Two major Bayesian VQA frameworks have explored modeling these relationships.	
[30]	the first Bayesian framework for VQA used semantic segmentation to identify the objects in an image and their positions. a Bayesian algorithm trained to model the spatial relationships of the objects, used to compute each answer's probability. the earliest known algorithm for VQA surpassed by simple baseline models dependent on the results of the semantic segmentation, which was imperfect.
[36]	exploited the fact that the type of answer can be predicted using solely the question used a variant of quadratic discriminant analysis, which modeled the probability of image features given the question features and the answer type. ResNet-152 was used for the image features skip-thought vectors were used to represent the question.

5.3 Attention Based Models



a common way to incorporate attention into a VQA system

Using.
Attentive models attempt to overcome the limitation:

global features alone may obscure task-relevant regions of the input space

learn to ‘attend’ to the most relevant regions of the input space.
Attention models have shown great successes in other vision and NLP tasks:
object recognition [65]
captioning [20]
machine translation [66, 67]

used spatial attention to create region-specific CNN features, rather than using global features from the entire image.
incorporating attention into the text representation.
The basic idea:
1) certain visual regions in an image and certain words in a question are more informative than others for answering a given question.
2) Global image features may not be granular enough to address region specific questions.

represent the visual features at all spatial regions, instead of solely at the global level.
local features from relevant regions can be given higher prominence based on the question asked.

two common ways to achieve local feature encoding:
1) impose a uniform grid over all image locations, with the local image features present at each grid location.
often done by operating on the last CNN layer prior to the final spatial pooling that flattens the features.
The relevance of each grid location is then determined by the question.
2) generate region proposals (bounding boxes) for an image
encode each of these boxes using a CNN
determine the relevance of each box’s features using the question.
using spatial visual attention for VQA [63, 49, 52, 48, 54, 51, 46, 55], significant differences among these methods:
The Focus Regions for VQA [63] and Focused Dynamic Attention (FDA) models [51]:
both used Edge Boxes [68] to generate bounding box region proposals for images.

[63]	a CNN used to extract features from each of these boxes. The input to their VQA system consisted of these CNN features question features one of the multiple choice answers trained to produced a score for each multiple-choice answer the highest scoring answer was selected. The score is calculated using a weighted average of scores from each of the regions where the weights are simply learned by passing the dot product of regional CNN feature and question embedding to a fully connected layer.
FDA [51]	only use the region proposals that have the objects mentioned in the question. Their VQA algorithm requires as input a list of bounding boxes with their corresponding object label. During training, the object labels and bounding boxes are obtained from COCO annotations. During test, the labels are obtained by classifying each bounding box using ResNet [2]. Subsequently, word2vec [69] was used to compute the similarity

	between words in the question and the object labels assigned to each of the bounding boxes. Any box with a score greater than 0.5 is successively fed into an LSTM network. At the last time-step, global CNN features from the entire image are also fed into the network, giving it access to both global and local features. A separate LSTM was also used as the question representation. The output from these two LSTMs are then fed into a fully connected layer that is fed to a softmax classifier to produce the answer predictions.
Stacked Attention Network (SAN) [49] Dynamic Memory Network (DMN) [52]	used visual features from the spatial grid of a CNN's feature maps used the last convolutional layer from VGG-19 with 448 × 448 images to produce a 14 × 14 filter response map with 512 dimensional features at each grid location. SAN [49] an attention layer is specified by a single layer of weights that uses the question and the CNN feature map with a softmax activation function to compute the attention distribution across image locations This distribution is then applied to the CNN feature map to pool across spatial feature locations using a weighted sum, which generates a global image representation that emphasizes certain spatial regions more than others. This feature vector is then combined with a vector of question features to create a representation that can be used with a softmax layer to predict the answer. this approach is generalized to handle multiple (stacked) attention layers, enabling the system to model complex relationships among multiple objects in an image.
Spatial Memory Network [48]	spatial attention is produced by estimating the correlation of image patches with individual words in the question. This word-guided attention is used to predict an attention distribution, which is then used to compute the weighted sum of the visual features embedding across image regions. Two different models were then explored: 1) In the one-hop model, the features encoding the entire question are combined with the weighted visual features to predict the answer. 2) In the two-hop model, the combination of the visual and question features is looped back into the attentive mechanism for refining the attention distribution.
[52]	used a modified Dynamic Memory Network (DMN) [70]. A DMN consists of an input module, an episodic memory module, and an answering module. DMNs have been used for text based QA, where each word in a sentence is fed into a recurrent neural network and the output of the network is used to extract 'facts.' Then, the episodic memory module makes multiple passes over a subset of these facts. With each pass, the internal memory representation of the network is updated. An answering module uses the final state of the memory representation and the input question to predict an answer. To use a DMN for VQA, they used visual facts in addition to text. To generate visual facts, the CNN features at each spatial grid location are treated as words in a sentence that are sequentially fed into a recurrent neural network.

	The episodic memory module then makes passes through both text and visual facts to update its memory. The answering module remains unchanged.
Hierarchical Co-Attention model [54]	applies attention to image and question visual attention similar to Spatial Memory Network [48]. uses a hierarchical encoding of the question the word level (using a one-hot encoding) the phrase level (using bi- or tri-gram window size) the question level (using the final time-step of an LSTM network) use two different attentive mechanisms: The parallel co-attention approach simultaneously attended to both the question and image The alternative co-attention approach alternated between attending to the question or the image allowed the relevance of words in the question and the relevance of specific image regions to be determined by each other The answer prediction is made by recursively combining the co-attended features from all three levels of the question hierarchy.
[56]	Using joint attention for image and question features allow image and question attention to guide each other directing attention to relevant words and visual regions simultaneously visual and question input are jointly represented by a memory vector that is used to simultaneously predict attention for both question and image features The attentive mechanism computes updated image and question representations, which are then used to recursively update the memory vector. This recursive memory update mechanism can be repeated K times to refine the attention in multiple steps. a value of K = 2 worked best for COCO-VQA.

5.4 Bilinear Pooling Methods

VQA relies on jointly analyzing the image and the question.

Early models:
combining their respective features using simple methods
concatenation
element-wise product between the question and image features
outer-product between these two streams of information.
Similar ideas were shown to work well for improving fine-grained image recognition [71].

two most prominent VQA methods using bilinear pooling [46, 57].	
Multimodal Compact Bilinear (MCB) pooling[46]	combining image and text features in VQA approximate the outer-product between the image and text features, allowing a deeper interaction between the two modalities does the outer-product in a lower dimensional space. then used to predict which spatial features are relevant to the question. In a variation of this model, a soft-attention mechanism, similar to the method in [49], was also used, with the only major change being the use of MCB for combining text and question features instead of element-wise multiplication in [49].

	very good results on COCO-VQA the winner of the 2016 VQA Challenge workshop
[57]	MCB is too computationally expensive, despite using an approximate outer-product. use a multi-modal low-rank bi-linear pooling (MLB) scheme that uses the Hadamard product and a linear mapping to achieve approximate bilinear pooling. When used with a spatial visual attention mechanism, MLB rivaled MCB at VQA, but with reduced computational complexity and using a neural network with fewer parameters.

5.5 Compositional VQA Models

In VQA, questions often require multiple steps of reasoning to answer properly.

Two compositional frameworks have been proposed for VQA that attempt to tackle solving VQA in a series of sub-steps [44, 50, 55].	
Neural Module Network (NMN) [44, 50] framework	uses external question parsers to find the sub-task in the question whereas Recurrent Answering Units (RAU) [55] is trained end-to-end and sub-tasks can be implicitly learned. NMN is an especially interesting approach to VQA [44, 50]. The NMN framework treats VQA as a sequence of sub-tasks that are carried out by separate neural sub-networks. Each of the sub-network performs a single well-defined task, e.g., the find[X] module produces a heat map for the presence of certain object. Other modules include describe, measure, and transform. These modules then must be assembled into a meaningful layout. Two methods have been explored for inferring the required layout. [44], a natural language parser is used on the input question to both find the sub-tasks in the question and to infer the required layout of the sub-tasks that when executed in sequence would produce an answer to the given question [44]. [50], using algorithms to dynamically select the best layout for the given question from a set of automatically generated layout candidates.
RAU model [55]	implicitly perform compositional reasoning without depending on an external language parser. used multiple self-contained answering units that can solve VQA sub-tasks. These answering units are arranged in recurrent manner. Each answering unit on the chain is equipped with an attentive mechanism derived from [49] and a classifier. The authors' claimed that the inclusion of multiple recurrent answering units allows inferring the answer from a series of sub-tasks solved by each answering unit. However, they did not perform visualization or ablation studies to show how the answer might get refined in each time-step. This makes it difficult to assess whether progressive refinement and reasoning is

	occurring or not, especially considering that the complete image and question information is available to all answering units at every time step and that only the output from the first answering unit is used during the test stage.

5.6 Other Noteworthy Models

Answering questions about images can often require information beyond what can be directly inferred by analyzing the image.
Having knowledge about the uses and typical context for the objects present in an image can be helpful for VQA.

could use knowledge bank about particular animals to answer questions	
[37]	the knowledge bank improved performance The external knowledge bases were tailored to general information obtained from DBpedia [72], it is possible that using a source tailored to VQA could yield greater improvement.
[47]	incorporated a Dynamic Parameter Prediction layer into the fully connected layers of a CNN. The parameters of this layer are predicted from the question by using a recurrent neural network. allows the visual features to be specific to the question before the final classification step. can be seen as a kind of implicit attentive mechanism modifies the visual input based on the question.
[53]	Multimodal Residual Networks (MRN) motivated by the success of the ResNet architecture in image classification. a modification of ResNet [2] to use both visual and question features in the residual mapping. The visual and question embedding are allowed to have their own residual blocks with skip connections. after each residual block the visual data is inter-weaved with the question embedding. explored several alternate arrangement for constructing the residual architecture with multi-modal input and chose the above network based on performance.

5.7 What methods and techniques work better?

Overview of different methods that were evaluated on open-ended COCO-VQA and their design choices. Results are report on the ‘test-dev’ split when ‘test-standard’ results are not available (Denoted by *).

	Accuracy (%) (Acc_{VQA})	CNN Network	Use of Attention	Ext. Data	Compositional
LSTM Q+I [32]	54.1	VGGNet	-	-	-
iBOWIMG [38]	55.9	GoogLeNet	-	-	-
DPPNet [47]	57.4	VGGNet	-	-	-
SMem [48]	58.2	GoogLeNet	y	-	-
SAN [49]	58.9	GoogLeNet	y	-	-

NMN [44]	58.7	VGGNet	y	-	y
D-NMN [50]	59.4	VGGNet	y	-	y
AMA [37]	59.4	VGGNet	-	y	-
FDA [51]	59.5	ResNet	y	-	-
HYBRID [36]	60.1	ResNet	-	-	-
DMN+ [52]	60.4	ResNet	y	-	-
MRN [53]	61.8	ResNet	y	-	-
HieCoAtten-VGG* [54]	60.5	VGGNet	y	-	-
HieCoAtten-ResNet [54]	62.1	ResNet	y	-	-
RAU VGG* [55]	61.3	VGGNet	y	-	-
RAU ResNet [55]	63.2	ResNet	y	-	-
MCB* [46]	61.2	ResNet	-	-	-
MCB-ATT* [46]	64.2	ResNet	y	-	-
DAN-VGG* [56]	62.0	VGGNet	y	-	-
DAN-ResNet [56]	64.3	ResNet	y	-	-
MLB [57]	65.1	ResNet	y	-	-
MLB+VG* [57]	65.8	ResNet	y	y	-
MCB-ensemble [46]	66.5	ResNet	y	y	-

ResNet produces superior performance over VGGNet or GoogLeNet	
[55]	an increase of 2% by using ResNet-101 instead of the VGG-16 CNN for image features.
[54]	an increase of 1.3% when making the same change in their model
[56]	changing VGG-19 to ResNet-152 increased performance by 2.3%

spatial attention can be used to increase performance for a model.
[46][54]
the attentive version performed better.
attention alone does not appear to be sufficient.
Bayesian and compositional architectures do not significantly improve over comparable models

[36]
the model performed competitively only after it was combined with an MLP model.
It is unclear whether the increase was due to model averaging or the proposed Bayesian method.

[44][50]
the NMN models do not outperform comparable non-compositional models, e.g., [49].

It is possible that both of these methods perform well on specific VQA sub-tasks, e.g., NMN was shown to be specially helpful for positional reasoning questions in the SHAPES dataset.
However, since major datasets do not provide a detailed breakdown of question types, it is not possible to quantify how systems perform on specific question types.
any improvements on rare question types will have negligible impact on the overall performance score, making it difficult to properly evaluate the benefits of these methods.

6 Discussion

there is still a significant gap between the best methods and humans.
remains unclear whether the improvements in performance come from the mechanisms incorporated into later systems, e.g., attention, or if it is due to other factors.
difficult to decouple the contributions of text and image data in isolation.
numerous challenges to comparing algorithms due to the variations in how they are evaluated.

6.1 Vision vs. Language in VQA

VQA consists of two distinct data streams that need to be correctly used to ensure robust performance: images and questions.

Ablation studies [36, 32] have routinely shown that question only models perform drastically better than image only models, especially on open-ended COCO-VQA.
On COCO-QA, simple image-blind models that use only the question can achieve 50% accuracy with the gain from using the image being comparatively modest [36].
[36] shown that for DAQUAR-37, using a better language embedding with an image-blind model produced results superior to earlier works employing both images and questions

This is primarily due to two factors.
1) the question severely constrains the kinds of answers expected in many cases, essentially turning an open-ended question into a multiple-choice one
2) the datasets tend to have strong bias.

These two factors make language a much stronger prior than the image features alone.
The predictive power of language over images have been corroborated by ablation studies.
[73]
studied a model that had been trained using both image and question features.
then studied how the predictions of the model differed when it was given only the image or only the question, compared to when it was given both.
found that the image-only model’s predictions differed from the combined model 40% more often than the question only model.
also showed that the way the question is phrased strongly biases the answer.
When training a neural network, these regularities will be incorporated into the model.
While this produces increased performance on the dataset, it is potentially detrimental to creating a general VQA system.

[42]
bias in VQA was studied using synthetic cartoon images.
created a dataset with solely binary questions, in which the same question could be asked about two images that were mostly identical, except for a minor change that caused the correct answer to be different.
found that a model trained on an unbalanced version of this dataset performed 11% worse (absolute difference) on a balanced test dataset compared to a model trained on a balanced version of the dataset.

We conducted two experiments to assess the effect of language bias in VQA.
1) used the model 3 from [38].
trained on COCO-VQA
allows the contribution of the question and image features to be assessed independently by splitting the weights of the softmax output layer into image and question components.
asked simple binary questions with a relatively equal prior for both choices so that the image must be analyzed to answer the question.
the system performs poorly, especially when considering that the baseline accuracy for yes/no questions for COCO-VQA is about 80%.

2) studied how language bias affected the more complex MCB-ensemble model [46] that was trained on COCO-VQA.
the winner of the 2016 VQA Challenge workshop.
created a small dataset consisting only of yes/no questions.
used annotations from the validation split of the COCO dataset to determine whether an image contained a person, and then asked an equal number of 'yes' and 'no' questions about whether there are any people present.
We used the questions 'Are there any people in the photo?', 'Is there a person in the picture?', and 'Is there a person in the photo?'
For each variation, there were 38,514 yes/no questions (115,542 total). The accuracy of MCB-ensemble on this dataset was worse than chance (47%), which starkly contrasts with its results on COCO-VQA (i.e., 83% on COCO-VQA yes/no questions).
This is likely due to severe bias in the training dataset, and not due to an inability for MCB to learn the task.

VQA systems are sensitive to the way a question is phrased.
observed similar results when using the system in [32].
created another toy dataset from the validation split of the COCO dataset and used it to evaluate the MCB-ensemble model that was trained on COCO-VQA.
In this toy dataset, the task is to identify which sport was being played. We asked three variations of the same question: 1) 'What are they doing?', 2) 'What are they playing?', and 3) 'What sport are they playing?' Each variation contains 5,237 questions about seven common sports (15,711 questions total).
MCB-ensemble achieved 33.6% for variation 1, 78% for variation 2, and 86.4% for variation 3. The dramatic increase in performance from variation 1 to 2 is caused by the inclusion of keyword 'playing' instead of the generic verb 'doing.' The increment from variation 2 to 3 is caused by explicitly including the keyword 'sport' in the question.
VQA systems are over-dependent on language 'clues' that annotators often include.
language bias is an issue that critically affects the performance of current VQA systems.

current VQA systems are more dependent on the question than the image content.
Language bias in datasets critically affects the performance of the current VQA systems, which limits their deployment.

New VQA datasets must endeavor to compensate for this issue, by either having questions that force analysis of image content and/or by making datasets less biased.

6.2 How useful is attention for VQA?

difficult to determine how much attention helps VQA algorithms.
In ablation studies, when attentive mechanisms are removed from models it impairs their performance [46, 54].

Currently, the best model for COCO-VQA does employ spatial visual attention [46], but simple models that do not use attention have been shown to exceed earlier models that used complex attentive mechanisms.

[62]
an attention-free model that used multiple global image feature representations (VGG-19, ResNet-101, and ResNet-152), instead of a single CNN, performed very well compared some attentive models.
They combined image and question features using both element-wise multiplication and addition, instead of solely concatenating them.
Combined with ensembling, this yielded results significantly higher than the complex attention-based models used in [49] and [52].
Similar results have been obtained by other systems that do not employ spatial attention[36, 64, 53].

Attention alone does not ensure good VQA performance,
but incorporating attention into a VQA model appears to improve performance over the same model when attention is not used.

[74]
methods commonly used to incorporate spatial attention to specific image features do not cause models to attend to the same regions as humans tasked with VQA.
They made this observation using both the attentive mechanisms used in [49] and [54].
This may be because the regions the model learns to attend to are discriminative due to biases in the dataset and not due to where the algorithm should attend.
attentive mechanisms may not be correctly deployed due to biases.

6.3 Bias Impairs Method Evaluation

Dataset bias significantly impairs the ability to evaluate VQA algorithms.
Questions that require the use of the image content are often relatively easy to answer.
Many are about the presence of objects or scene attributes.
These questions tend to be handled well by CNNs and also have strong language biases.
Harder questions, such as those beginning with ‘Why’ are comparatively rare.
This has serious implications for evaluating performance.
For COCO-VQA (train and validation partitions), a system that improves accuracy on questions beginning with ‘Is’ and ’Are’ by 15% will increase overall accuracy by 5%.
However, the same increase in both ‘Why’ and ‘Where’ questions will only increase accuracy by 0.6%.
In fact, even if all ‘Why’ and ‘Where’ questions are answered correctly, the overall increase in accuracy will only be 4.1%.
On the other hand, answering ‘yes’ to all questions beginning with ‘Is there’ will yield an accuracy of 85.2% on those questions.
These problems could be overcome if each type of question was evaluated in isolation, and then the mean accuracy across question types was used instead of overall accuracy for benchmarking the algorithms.
This approach is similar to the mean per-class accuracy metric used for evaluating object classification algorithms, which was adopted due to bias in the amount of test data available for different object categories.

6.4 Are Binary Questions Sufficient?

Using binary (yes/no or true/false) questions to evaluate algorithms has attracted significant discussion in the VQA community.

The main argument against using binary questions is the lack of complex questions and the relative ease in answering the questions that are typically generated by human annotators.
Visual Genome and Visual7W exclude binary questions altogether. The authors argued that this choice would encourage more complex questions from the annotators.

On the other hand, binary questions are easy to evaluate and these questions can, in theory, encompass an enormous variety of tasks.

The SHAPES dataset [44] uses binary questions exclusively but contains complex questions involving spatial reasoning, counting, and drawing inferences (see Figure 6). Using cartoon images, [42] also showed that these questions can be especially difficult for VQA algorithms when the dataset is balanced.

there are challenges to creating balanced binary questions for real world imagery.
In COCO-VQA, ‘yes’ is a much more common answer than ‘no,’ simply because people tend to ask questions biased toward ‘yes’ as an answer.
As long as bias is controlled, yes/no questions can play an important role in future VQA benchmarks, but a VQA system should be capable of more than solely binary questions so that its abilities can be fully assessed.
All real-world applications for VQA, such as enabling the blind to ask questions about visual content, require the output of the VQA system to be open-ended.
A system that can solely handle binary questions will have limited real-world utility.

6.5 Open Ended vs. Multiple Choice

Because it is challenging to evaluate open-ended multi-word answers, multiple-choice has been proposed as a way to evaluate VQA algorithms.
As long as the alternatives are sufficiently difficult, a system could be evaluated in this manner but then be deployed to answer open-ended questions.
For these reasons, multiple choice is used to evaluate Visual7W, Visual Genome, and a variant of The VQA Dataset.

In this framework:
an algorithm has access to a number of possible answers (e.g., 18 for COCO-VQA), along with the question and image.
It must then select among possible choices.

A major problem with multiple-choice evaluation：
the problem can be reduced to determining which of the answers is correct instead of actually answering the question.

[64]
formulated VQA as an answer scoring task
the system was trained to produce a score based on the image, question, and potential answers.
The answers themselves were fed into the system as features.
It achieved state-of-the-art results on Visual7W and rivals the best methods on COCO-VQA, with their method performing better than many complex systems that use attention.
To a large extent, we believe their system performed well because it learned to better exploit biases in the answers instead of reasoning about images.
On Visual7W, they showed that a variant of their system that used solely the answers and was both image- and question-blind rivaled baselines using the question and image.

We argue that any VQA system should be able to operate without being given answers as inputs.

Multiple-choice can be an important ingredient for evaluating multi-word answers, but it alone is not sufficient.
When multiple-choice is used, the choices must be selected carefully to ensure that a question is hard and not deducible from the provided answers alone.
A system that is solely capable of operating with answers provided is not really solving VQA, because these are not available when a system is deployed.

7 Recommendations for Future VQA Datasets

Existing VQA benchmarks are not sufficient to evaluate whether an algorithm has ‘solved’ VQA.

Future datasets：
1) larger.
algorithms do not have enough data for training and evaluation.
We did a small experiment where we trained a simple MLP baseline model for VQA using ResNet-152 image features and skip-thought features for the questions, and we assessed performance as a function of the amount of training data available on COCO-VQA.
even on datasets that are biased, increasing the size of the dataset could significantly improve accuracy.

2) less biased.
For real-world open-ended VQA, this will be difficult to achieve without carefully instructing the humans that generate the questions.
for VQA this problem is compounded by bias in the questions as well.

3) more nuanced analysis
All of the publicly released datasets use evaluation metrics that treat every question with equal weight,
but some kinds of questions are far easier, either because of bias or because existing algorithms excel at answering that kind of question, e.g., object recognition questions.
Some datasets such as COCO-QA have divided VQA questions into distinct categories, e.g., for COCO-QA these are color, counting (number), location, and object.
We believe that mean per-question type performance should replace standard accuracy, so each question would not have equal weight in evaluating performance.
This would go a long way towards making a VQA algorithm have to perform well at a wide variety of question types to perform well overall, otherwise a system that excelled at answering ‘Why’ questions but was slightly worse than another model at more common questions would not be fairly evaluated.
To do this, each question would need to be assigned a category.
We believe this effort would make benchmark results significantly more meaningful.
The scores on each question type could also be used to compare algorithms to see which kind of questions they excel at.

8 Conclusions

VQA is an important basic research problem in computer vision and natural language processing that requires a system to do much more than task specific algorithms, such as object recognition and object detection.
An algorithm that can answer arbitrary questions about images would be a milestone in artificial intelligence.
We believe that VQA should be a necessary part of any visual Turing test.

the field needs a dataset that evaluates the important characteristics of a VQA algorithm, so that if an algorithm performs well on that dataset then it means it is doing well on VQA in general.

Future work on VQA：
the creation of larger and far more varied datasets.
Bias in these datasets will be difficult to overcome, but evaluating different kinds of questions individually in a nuanced manner, rather than using naive accuracy alone, will help significantly develop VQA algorithms that can reason about image content, but these algorithms may lead to significant new areas of research.

References

[1]	K. Simonyan A. Zisserman	Very deep convolutional networks for large-scale image recognition	ICLR	2015
[2]	K. He X. Zhang S. Ren J. Sun	Deep residual learning for image recognition	CVPR	2016
[3]	J. Redmon S. Divvala R. Girshick A. Farhadi	You only look once: Unified, real-time object detection	CVPR	2016
[4]	S. Ren K. He R. Girshick J. Sun	Faster r-cnn: Towards real-time object detection with region proposal networks	NIPS	2015
[5]	J. Donahue L. A. Hendricks S. Guadarrama M. Rohrbach S. Venugopalan K. Saenko T. Darrell	Long-term recurrent convolutional networks for visual recognition and description	CVPR	2015
[6]	A. Karpathy G. Toderici S. Shetty T. Leung R. Sukthankar L. Fei-Fei	Large-scale video classification with convolutional neural networks	CVPR	2014
[7]	K. Simonyan A. Zisserman	Two-stream convolutional networks for action recognition in videos	NIPS	2014
[8]	D. Geman S. Geman N. Hallonquist L. Younes	Visual turing test for computer vision systems	PNAS	2015
[9]	M. Malinowski M. Fritz	Towards a visual turing challenge	arXiv	2014

[10]	T.-Y. Lin M. Maire S. Belongie J. Hays P. Perona D. Ramanan P. Dollár C. L. Zitnick,	Microsoft coco: Common objects in context	ECCV	2014
[11]	C. Szegedy A. Toshev D. Erhan	Deep neural networks for object detection	NIPS	2013
[12]	J. Long E. Shelhamer T. Darrell	Fully convolutional networks for semantic segmentation	CVPR	2015
[13]	H. Noh S. Hong B. Han	Learning deconvolution network for semantic segmentation	CVPR	2015
[14]	N. Silberman D. Sontag R. Fergus	Instance segmentation of indoor scenes using a coverage loss	ECCV	2014
[15]	Z. Zhang A. G. Schwing S. Fidler R. Urtasun	Monocular object instance segmentation and depth ordering with CNNs	CVPR	2015
[16]	Z. Zhang S. Fidler R. Urtasun	Instance-level segmentation with deep densely connected MRFs	CVPR	2016
[17]	A. Karpathy L. Fei-Fei,	Deep visual-semantic alignments for generating image descriptions	CVPR	2015
[18]	J. Mao W. Xu Y. Yang J. Wang Z. Huang A. Yuille	Deep captioning with multi-modal recurrent neural networks (m-rnn)	ICLR	2015
[19]	O. Vinyals A. Toshev S. Bengio D. Erhan	Show and tell: A neural image caption generator	CVPR	2015
[20]	K. Xu J. Ba R. Kiros A. Courville R. Salakhutdinov R. Zemel Y. Bengio	Show, attend and tell: Neural image caption generation with visual attention	ICML	2015
[21]	K. Papineni	BLEU: a method for automatic evaluation of machine	ACL	2002

	S. Roukos T. Ward W.-J. Zhu	translation		
[22]	C.-Y. Lin	Rouge: A package for automatic evaluation of summaries	ACL .	2004
[23]	S. Banerjee A. Lavie	Meteor: An automatic metric for mt evaluation with improved correlation with human judgments	ACL	2005
[24]	R. Vedantam, C. Lawrence Zitnick D. Parikh	Cider: Consensus-based image description evaluation	CVPR	2015
[25]	C. Callison-Burch M. Osborne P. Koehn	Re-evaluation the role of BLEU in machine translation research		2006
[26]	H. Fang S. Gupta F. Iandola R. Srivastava L. Deng P. Dollár J. Gao X. He M. Mitchell J. Platt, et al.	From captions to visual concepts and back	CVPR	2015
[27]	R. Bernardi R. Cakici D. Elliott A. Erdem E. Erdem N. Ikizler-Cinbis F. Keller A. Muscat B. Plank	Automatic description generation from images: A survey of models, datasets, and evaluation measures	JAIR	2016
[28]	J. Devlin S. Gupta R. Girshick M. Mitchell C. L. Zitnick	Exploring nearest neighbor approaches for image captioning	arXiv	2015
[29]	J. Johnson A. Karpathy L. Fei-Fei	Densecap: Fully convolutional localization networks for dense captioning	CVPR	2016
[30]	M. Malinowski M. Fritz	A multi-world approach to question answering about real-world scenes based on uncertain input	NIPS	2014
[31]	M. Ren R. Kiros R. Zemel	Exploring models and data for image question answering	NIPS	2015

[32]	S. Antol A. Agrawal J. Lu M. Mitchell D. Batra C. L. Zitnick D. Parikh	VQA: Visual question answering	ICCV	2015
[33]	H. Gao J. Mao J. Zhou Z. Huang L. Wang W. Xu	Are you talking to a machine? Dataset and methods for multilingual image question answering	NIPS	2015
[34]	Y. Zhu O. Groth M. Bernstein L. Fei-Fei	Visual7w: Grounded question answering in images	CVPR	2016
[35]	R. Krishna Y. Zhu O. Groth J. Johnson K. Hata J. Kravitz S. Chen Y. Kalantidis L.-J. Li D. A. Shamma et al.	Visual genome: Connecting language and vision using crowd-sourced dense image annotations	IJCV	2017
[36]	K. Kafle C. Kanan	Answer-type prediction for visual question answering	CVPR	2016
[37]	Q. Wu P. Wang C. Shen A. van den Hengel A. R. Dick	Ask me anything: Free-form visual question answering based on knowledge from external sources	CVPR	2016
[38]	B. Zhou Y. Tian S. Sukhbaatar A. Szlam R. Fergus	Simple baseline for visual question answering	arXiv	2015
[39]	N. Silberman D. Hoiem P. Kohli R. Fergus	Indoor segmentation and support inference from rgbd images	ECCV	2012 .
[40]	M. Malinowski, M. Rohrbach, and M. Fritz	Ask your neurons: A neural-based approach to answering questions about images	ICCV	2015
[41]	S. Antol	Zero-shot learning via visual abstraction	ECCV	2014

	C. L. Zitnick D. Parikh			
[42]	P. Zhang Y. Goyal D. Summers-Stay D. Batra D. Parikh	Yin and yang: Balancing and answering binary visual questions	CVPR	2016
[43]	B. Thomee D. A. Shamma G. Friedland B. Elizalde K. Ni D. Poland D. Borth L.-J. Li	Yfcc100m: The new data in multimedia research	Communications of the ACM	2016
[44]	J. Andreas M. Rohrbach T. Darrell D. Klein	Deep compositional question answering with neural module networks	CVPR	2016
[45]	Z. Wu M. Palmer	Verbs semantics and lexical selection	ACL	1994
[46]	A. Fukui D. H. Park D. Yang A. Rohrbach T. Darrell M. Rohrbach	Multi-modal compact bilinear pooling for visual question answering and visual grounding	EMNLP	2016
[47]	H. Noh P. H. Seo B. Han	"Image question answering using convolutional neural network with dynamic parameter prediction,"	CVPR	2016
[48]	H. Xu K. Saenko	Ask, attend and answer: Exploring question-guided spatial attention for visual question answering	ECCV	2016
[49]	Z. Yang X. He J. Gao L. Deng A. J. Smola	Stacked attention networks for image question answering	CVPR	2016
[50]	J. Andreas M. Rohrbach T. Darrell D. Klein	Learning to compose neural networks for question answering	NAACL	2016
[51]	I. Ilievski S. Yan J. Feng	A focused dynamic attention model for visual question answering	arXiv	2016
[52]	C. Xiong S. Merity R. Socher	Dynamic memory networks for visual and textual question answering	ICML	2016

[53]	J.-H. Kim S.-W. Lee D.-H. Kwak M.-O. Heo J. Kim J.-W. Ha B.-T. Zhang	Multi-modal residual learning for visual qa	NIPS	2016
[54]	J. Lu J. Yang D. Batra D. Parikh	Hierarchical question-image co-attention for visual question answering	NIPS	2016
[55]	H. Noh B. Han	Training recurrent answering units with joint loss minimization for VQA	arXiv	2016
[56]	H. Nam J.-W. Ha J. Kim	Dual attention networks for multi-modal reasoning and matching	arXiv	2016
[57]	J.-H. Kim K.-W. On J. Kim J.-W. Ha B.-T. Zhang	Hadamard product for low-rank bilinear pooling	arXiv	2016
[58]	C. Szegedy W. Liu Y. Jia P. Sermanet S. Reed D. Anguelov D. Erhan V. Van-houcke A. Rabinovich	Going deeper with convolutions	CVPR	2015
[59]	S. Hochreiter J. Schmidhuber	Long short-term memory	Neural computation	1997
[60]	K. Cho B. Van Merriënboer C. Gulcehre D. Bahdanau F. Bougares H. Schwenk Y. Bengio	Learning phrase representations using rnn encoder-decoder for statistical machine translation	EMNLP	2014
[61]	R. Kiros Y. Zhu R. Salakhutdinov R. S. Zemel A. Torralba R. Urtasun S. Fidler	Skip-thought vectors	NIPS	2015
[62]	K. Saito	Dualnet: Domain-invariant network for visual question	arXiv	2016

	A. Shin Y. Ushiku T. Harada	answering		
[63]	K. J. Shih S. Singh D. Hoiem	Where to look: Focus regions for visual question answering	CVPR	2016
[64]	A. Jabri A. Joulin L. van der Maaten	Revisiting visual question answering baselines	ECCV	2016
[65]	J. Ba V. Mnih K. Kavukcuoglu	Multiple object recognition with visual attention	ICLR	2015
[66]	M.-T. Luong H. Pham C. D. Manning	Effective approaches to attention-based neural machine translation	EMNLP	2015
[67]	D. Bahdanau K. Cho Y. Bengio	Neural machine translation by jointly learning to align and translate	ICLR	2015
[68]	C. L. Zitnick P. Dollár	Edge boxes: Locating object proposals from edges	ECCV	2014
[69]	T. Mikolov J. Dean	Distributed representations of words and phrases and their compositionality	NIPS	2013
[70]	A. Kumar O. Irsoy J. Su J. Bradbury R. English B. Pierce P. Ondruska I. Gulrajani R. Socher	Ask me anything: Dynamic memory networks for natural language processing	ICML	2016
[71]	T.-Y. Lin A. RoyChowdhury S. Maji	Bilinear cnn models for fine-grained visual recognition	ICCV	2015
[72]	J. Lehmann R. Isele, M. Jakob A. Jentzsch D. Kontokostas P. N. Mendes S. Hell-mann M. Morsey P. van Kleef S. Auer et al.	DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia	Semanti c Web	2015
[73]	A. Agrawal, D. Batra, and	Analyzing the behavior of visual question answering models	EMNLP	2016

	D. Parikh			
[74]	A. Das H. Agrawal C. L. Zitnick D. Parikh D. Batra	Human attention in visual question answering: Do humans and deep networks look at the same regions?	EMNLP	2016
[75]	A. Torralba A. Efros	Unbiased look at dataset bias	CVPR	2011

EMNLP： Conference on Empirical Methods on Natural Language Processing
ICCV： The IEEE International Conference on Computer Vision
ICML： International Conference on Machine Learning
NIPS： Advances in Neural Information Processing Systems
ECCV： European Conference on Computer Vision
ICLR： International Conference on Learning Representations
NAACL： Annual Meeting of the North American Chapter of the Association for Computational Linguistics
ACL： Annual Meeting of the Association for Computational Linguistics
PNAS： Proceedings of the National Academy of Sciences
IJCV： International Journal of Computer Vision
JAIR： Journal of Artificial Intelligence Research