VQA: Visual Question Answering

Stanislaw Antol*[1] Aishwarya Agrawal*[1] Jiasen Lu[1] Margaret Mitchell[2]
Dhruv Batra[1] C. Lawrence Zitnick[2] Devi Parikh[1]
[1]Virginia Tech [2]Microsoft Research
[1]{santol, aish, jiasenlu, dbatra, parikh}@vt.edu [2]{memitc, larryz}@microsoft.com

Abstract

We propose the task of free-form *and* open-ended *Visual Question Answering (VQA). Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. Mirroring real-world scenarios, such as helping the visually impaired, both the questions and answers are open-ended. Visual questions selectively target different areas of an image, including background details and underlying context. As a result, a system that succeeds at VQA typically needs a more detailed understanding of the image and complex reasoning than a system producing generic image captions. Moreover, VQA is amenable to automatic evaluation, since many open-ended answers contain only a few words or a closed set of answers that can be provided in a multiple-choice format. We provide a dataset containing ∼0.25M images, ∼0.76M questions, and ∼10M answers (www.visualqa.org), and discuss the information it provides. Numerous baselines for VQA are provided and compared with human performance.*

1. Introduction

We are witnessing a renewed excitement in multi-discipline Artificial Intelligence (AI) research problems. In particular, research in image and video captioning that combines Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation & Reasoning (KR) has dramatically increased in the past year [13, 7, 9, 32, 21, 19, 45]. Part of this excitement stems from a belief that multi-discipline tasks like image captioning are a step towards solving AI. However, the current state of the art demonstrates that a coarse scene-level understanding of an image paired with word n-gram statistics suffices to generate reasonable image captions, which suggests image captioning may not be as "AI-complete" as desired.

What makes for a compelling "AI-complete" task? We believe that in order to spawn the next generation of AI algorithms, an ideal task should (i) require *multi-modal knowledge* beyond a single sub-domain (such as CV) and (ii) have a well-defined *quantitative evaluation metric* to track progress. For some tasks, such as image captioning, automatic evaluation is still a difficult and open research problem [43, 10, 18].

Figure 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions.

In this paper, we introduce the task of *free-form* and *open-ended* Visual Question Answering (VQA). A VQA system takes as input an image and a free-form, open-ended, natural-language question about the image and produces a natural-language answer as the output. This goal-driven task is applicable to scenarios encountered when visually-impaired users [2] or intelligence analysts actively elicit visual information. Example questions are shown in Fig. 1.

Open-ended questions require a potentially vast set of AI capabilities to answer – fine-grained recognition (*e.g.*, "What kind of cheese is on the pizza?"), object detection (*e.g.*, "How many bikes are there?"), activity recognition (*e.g.*, "Is this man crying?"), knowledge base reasoning (*e.g.*, "Is this a vegetarian pizza?"), and commonsense reasoning (*e.g.*, "Does this person have 20/20 vision?", "Is this person expecting company?").

VQA [16, 30, 42, 2] is also amenable to automatic quantitative evaluation, making it possible to effectively track

* The first two authors contributed equally.

progress on this task. While the answer to many questions is simply "yes" or "no", the process for determining a correct answer is typically far from trivial (e.g. in Fig. 1, "Does this person have 20/20 vision?"). Moreover, since questions about images often tend to seek specific information, simple one-to-three word answers are sufficient for many questions. In such scenarios, we can easily evaluate a proposed algorithm by the number of questions it answers correctly. In this paper, we present both an open-ended answering task and a multiple-choice task [38, 27]. Unlike the open-answer task that requires a free-form response, the multiple-choice task only requires an algorithm to pick from a predefined list of possible answers.

We present a large dataset that contains 204,721 images from the MS COCO dataset [26] and a newly created abstract scene dataset [48, 1] that contains 50,000 scenes. The MS COCO dataset has images depicting diverse and complex scenes that are effective at eliciting compelling and diverse questions. We collected a new dataset of "realistic" abstract scenes to enable research focused only on the high-level reasoning required for VQA by removing the need to parse real images. Three questions were collected for each image or scene. Each question was answered by ten subjects along with their confidence. The dataset contains over 760K questions with around 10M answers.

While the use of open-ended questions offers many benefits, it is still useful to understand the types of questions that are being asked and which types various algorithms may be good at answering. To this end, we analyze the types of questions asked and the types of answers provided. Through several visualizations, we demonstrate the astonishing diversity of the questions asked. We also explore how the information content of questions and their answers differs from image captions. For baselines, we offer several approaches that use a combination of both text and state-of-the-art visual features [23]. As part of the VQA initiative, we will organize an annual challenge and associated workshop to discuss state-of-the-art methods and best practices.

VQA poses a rich set of challenges, many of which have been viewed as the holy grail of automatic image understanding and AI in general. However, it includes as building blocks several components that the CV, NLP, and KR [4, 6, 25, 29, 3] communities have made significant progress on during the past few decades. VQA provides an attractive balance between pushing the state of the art, while being accessible enough for the communities to start making progress on the task.

2. Related Work

VQA Efforts. Several recent papers have begun to study visual question answering [16, 30, 42, 2]. However, unlike our work, these are fairly restricted (sometimes synthetic) settings with small datasets. For instance, [30] only considers questions whose answers come from a predefined closed world of 16 basic colors or 894 object categories. [16] also considers questions generated from templates from a fixed vocabulary of objects, attributes, relationships between objects, *etc*. In contrast, our proposed task involves *open-ended*, *free-form* questions and answers provided by humans. Our goal is to increase the diversity of knowledge and kinds of reasoning needed to provide correct answers. Critical to achieving success on this more difficult and unconstrained task, our VQA dataset is *two orders of magnitude* larger than [16, 30] (>250,000 *vs*. 2,591 and 1,449 images respectively). The proposed VQA task has connections to other related work: [42] has studied joint parsing of videos and corresponding text to answer queries on two datasets containing 15 video clips each. [2] uses crowdsourced workers to answer questions about visual content asked by visually-impaired users. In concurrent work, [31] proposed combining an LSTM for the question with a CNN for the image to generate an answer – a similar model is evaluated in this paper. [28] generates abstract scenes to capture visual common sense relevant to answering (purely textual) fill-in-the-blank and visual paraphrasing questions. [40] and [44] use visual information to assess the plausibility of common sense assertions. [47] introduced a dataset of 10k images and prompted captions that describe specific aspects of a scene (*e.g*., individual objects, what will happen next). Concurrent with our work, [15] collected questions & answers in Chinese (later translated to English) for COCO images. [37] automatically generated four types of questions (object, count, color, location) using COCO captions.

Text-based Q&A is a well studied problem in the NLP and text processing communities (recent examples being [12, 11, 46, 38]). Other related textual tasks include sentence completion (*e.g*., [38] with multiple-choice answers). These approaches provide inspiration for VQA techniques. One key concern in text is the *grounding* of questions. For instance, [46] synthesized textual descriptions and QA-pairs grounded in a simulation of actors and objects in a fixed set of locations. VQA is naturally grounded in images – requiring the understanding of both text (questions) and vision (images). Our questions are generated by humans, making the need for commonsense knowledge and complex reasoning more essential.

Describing Visual Content. Related to VQA are the tasks of image tagging [8, 23], image captioning [24, 14, 34, 7, 13, 45, 9, 19, 32, 21] and video captioning [39, 17], where words or sentences are generated to describe visual content. While these tasks require both visual and semantic knowledge, captions can often be non-specific (e.g., observed by [45]). The questions in VQA require detailed specific information about the image for which generic image captions are of little use [2].

Other Vision+Language Tasks. Several recent papers have explored tasks at the intersection of vision and language that are easier to evaluate than image captioning, such as coreference resolution [22, 36] or generating referring expressions [20, 35] for a particular object in an image that

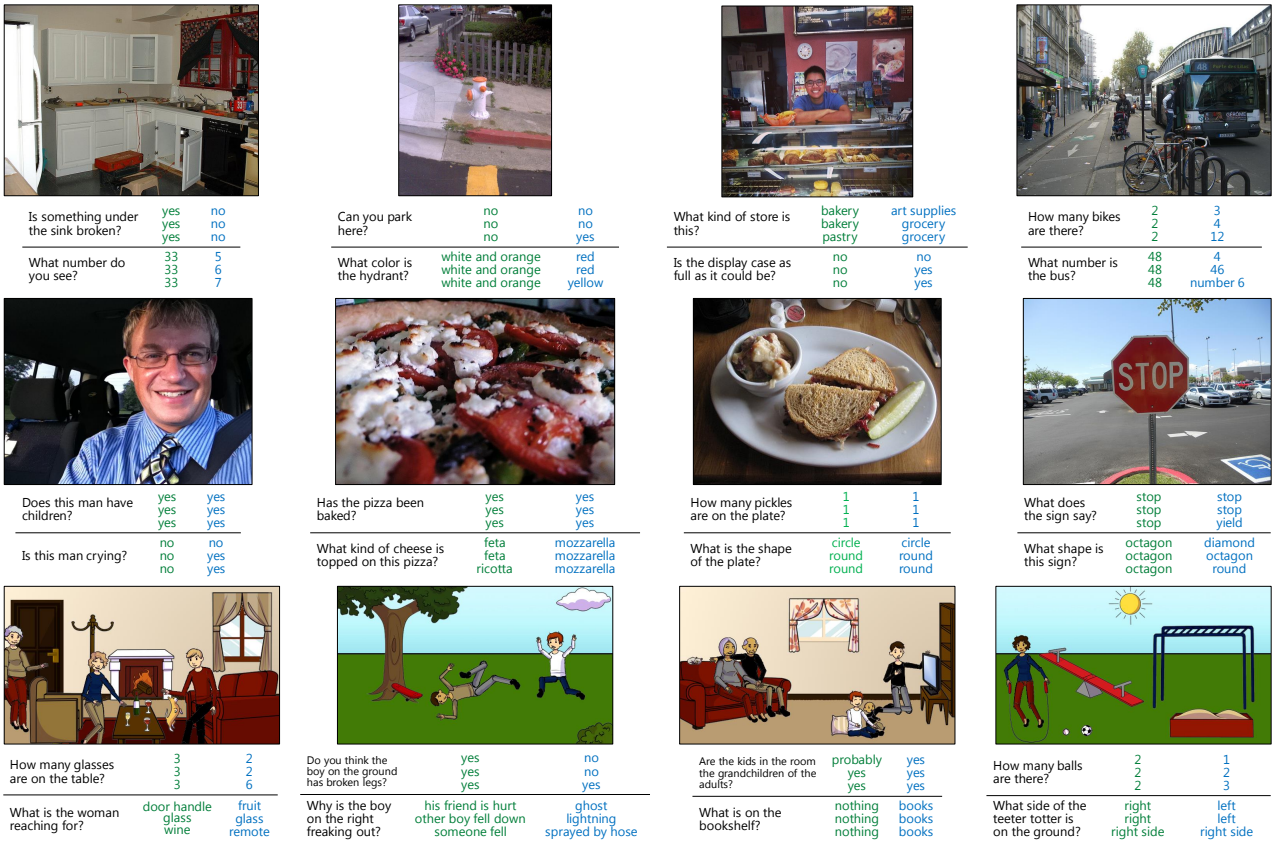

Figure 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the supplementary material for more examples.

would allow a human to identify which object is being referred to (*e.g.*, "the one in a red shirt", "the dog on the left"). While task-driven and concrete, a limited set of visual concepts (*e.g.*, color, location) tend to be captured by referring expressions. As we demonstrate, a richer variety of visual concepts emerge from visual questions and their answers.

3. VQA Dataset Collection

We now describe the Visual Question Answering (VQA) dataset. We begin by describing the real images and abstract scenes used to collect the questions. Next, we describe our process of collecting questions and their corresponding answers. Analysis of the questions and answers gathered as well as baseline results are provided in following sections.

Real Images. We use the 123,287 training and validation images and 81,434 test images from the newly-released Microsoft Common Objects in Context (MS COCO) [26] dataset. The MS COCO dataset was gathered to find images containing multiple objects and rich contextual information. Given the visual complexity of these images, they are well-suited for our VQA task. The more diverse our collection of images, the more diverse, comprehensive, and interesting the resultant set of questions and their answers.

Abstract Scenes. The VQA task with real images requires the use of complex and often noisy visual recognizers. To attract researchers interested in exploring the high-level reasoning required for VQA, but not the low-level vision tasks, we create a new abstract scenes dataset [1, 48, 49, 50] containing 50K scenes. The dataset contains 20 "paperdoll" human models [1] spanning genders, races, and ages with 8 different expressions. The limbs are adjustable to allow for continuous pose variations. The clipart may be used to depict both indoor and outdoor scenes. The set contains over 100 objects and 31 animals in various poses. The use of this clipart enables the creation of more realistic scenes (see bottom row of Fig. 2) that more closely mirror real images than previous papers [48, 49, 50]. See the supp. material for the user interface, additional details, and examples.

Splits. For real images, we follow the same train/val/test split strategy as the MC COCO dataset [26] (including test-dev, test-standard, test-challenge, test-reserve). For abstract scenes, we create standard splits, separating the scenes into 20K/10K/20K for train/val/test splits, respectively.

Captions. The MS COCO dataset [26, 5] already contains five single-sentence captions for all images. We also collected five single-captions for all abstract scenes using the same user interface[1] for collection.

Questions. Collecting interesting, diverse, and well-posed

[1] https://github.com/tylin/coco-ui

questions is a significant challenge. Many simple questions may only require low-level computer vision knowledge, such as "What color is the cat?" or "How many chairs are present in the scene?". However, we also want questions that require commonsense knowledge about the scene, such as "What sound does the pictured animal make?". Importantly, questions should also *require* the image to correctly answer and not be answerable using just commonsense information, *e.g.*, in Fig. 1, "What is the mustache made of?". By having a wide variety of question types and difficulty, we may be able to measure the continual progress of both visual understanding and commonsense reasoning.

We tested and evaluated a number of user interfaces for collecting such "interesting" questions. Specifically, we ran pilot studies asking human subjects to ask questions about a given image that they believe a "toddler", "alien", or "smart robot" would have trouble answering. We found the "smart robot" interface to elicit the most interesting and diverse questions. As shown in the supplementary material, our final interface stated "*We have built a smart robot. It understands a lot about images. It can recognize and name all the objects, it knows where the objects are, it can recognize the scene (*e.g.*, kitchen, beach), people's expressions and poses, and properties of objects (*e.g.*, color of objects, their texture). Your task is to stump this smart robot!*". To bias against generic image-independent questions, subjects were instructed to ask questions that *require* the image to answer.

The same user interface was used for both the real images and abstract scenes. In total, three questions from unique workers were gathered for each image/scene. When writing a question, the subjects were shown the previous questions already asked for that image to increase the question diversity. In total, the dataset contains over $\sim$0.76M questions.

Answers. Open-ended questions result in a diverse set of possible answers. For many questions, a simple "yes" or "no" response is sufficient. However, other questions may require a short phrase. Multiple different answers may also be correct. For instance, the answers "white", "tan", or "off-white" may all be correct answers to the same question. Human subjects may also disagree on the "correct" answer, *e.g.*, some saying "yes" while others say "no". To handle these discrepancies, we gather *10 answers for each question from unique workers*, while also ensuring that the worker answering a question did not ask it. We ask the subjects to provide answers that are "a brief phrase and not a complete sentence. Respond matter-of-factly and avoid using conversational language or inserting your opinion." In addition to answering the questions, the subjects were asked "Do you think you were able to answer the question correctly?" and given the choices of "no", "maybe", and "yes". See Sec. 4 for an analysis of the answers provided.

For testing, we offer two modalities for answering the questions: (i) **open-answer** and (ii) **multiple-choice**.

For the open-answer task, the generated answers are evaluated using the following accuracy metric: $\min(\frac{\text{\# humans that provided that answer}}{3}, 1)$, *i.e.*, an answer is deemed 100% accurate if at least 3 workers provided that exact answer.[2] Before comparison, all responses are made lowercase, numbers converted to digits, and punctuation & articles removed. We avoid using soft metrics such as Word2Vec [33], since they often group together words that we wish to distinguish, such as "left" and "right".

For multiple-choice task, 18 candidate answers are created for each question. As with the open-answer task, the accuracy of a chosen option is computed based on the number of human subjects who provided that answer (scaled by 3 and clipped at 1). We generate a candidate set of correct and incorrect answers from four sets of answers: **Correct:** The most common (out of ten) correct answer. **Plausible:** To generate incorrect, but still plausible answers we ask three subjects to answer the questions without seeing the image. If three unique answers are not found, we gather additional answers from nearest neighbor questions using a bag-of-words model. The use of these answers helps ensure the image, and not just commonsense knowledge, is necessary to answer the question. **Popular:** These are the 10 most popular answers. For instance, these are "yes", "no", "2", "1", "white", "3", "red", "blue", "4", "green" for real images. The inclusion of the most popular answers makes it more difficult for algorithms to infer the type of question from the set of answers provided, *i.e.*, learning that it is a "yes or no" question just because "yes" and "no" are present in the answers. **Random:** Correct answers from random questions in the dataset. To generate a total of 18 candidate answers, we first find the union of the correct, plausible, and popular answers. We include random answers until 18 unique answers are found. The order of the answers is randomized. Example multiple choice questions are in the supplement.

4. VQA Dataset Analysis

In this section, we provide an analysis of the questions and answers in the VQA train dataset. To gain an understanding of the types of questions asked and answers provided, we visualize the distribution of question types and answers. We also explore how often the questions may be answered without the image using just commonsense information. Finally, we analyze whether the information contained in an image caption is sufficient to answer the questions.

The dataset includes 614,163 questions and 7,984,119 answers (including answers provided by workers with and without looking at the image) for 204,721 images from the MS COCO dataset [26] and 150,000 questions with 1,950,000 answers for $50,000$ abstract scenes.

4.1. Questions

Types of Question. Given the structure of questions generated in the English language, we can cluster questions into

[2] In order to be consistent with 'human accuracies' reported in Sec. 4, machine accuracies are averaged over all $\binom{10}{9}$ sets of human annotators

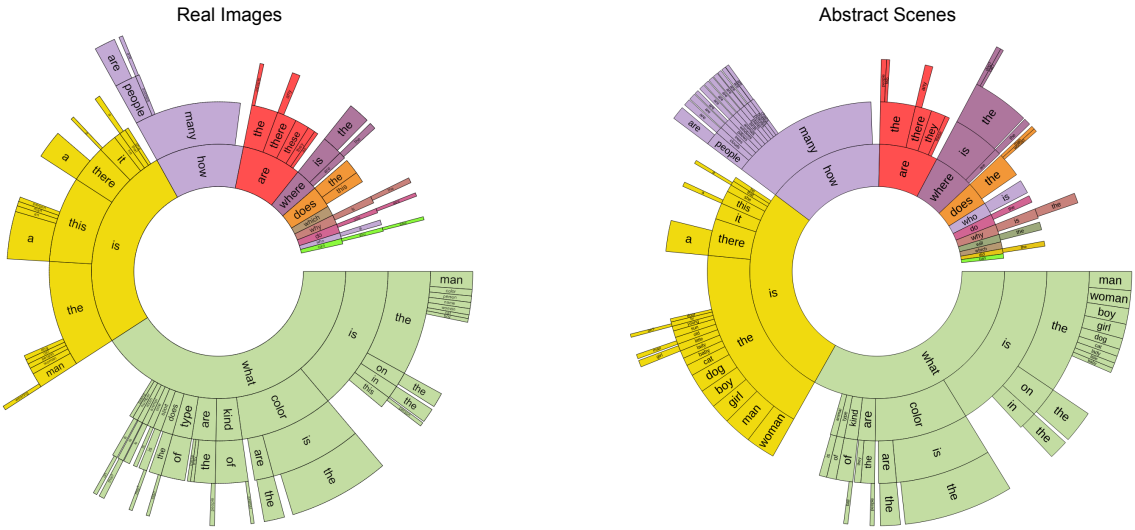

Figure 3: Distribution of questions by their first four words for a random sample of 60K questions for real images (left) and all questions for abstract scenes (right). The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show.

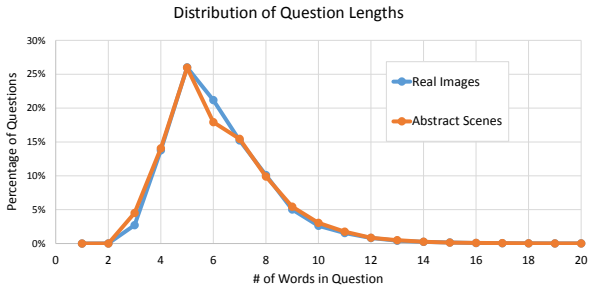

Figure 4: Percentage of questions with different word lengths for real images and abstract scenes.

different types based on the words that start the question. Fig. 3 shows the distribution of questions based on the first four words of the questions for both the real images (left) and abstract scenes (right). Interestingly, the distribution of questions is quite similar for both real images and abstract scenes. This helps demonstrate that the type of questions elicited by the abstract scenes is similar to those elicited by the real images. There exists a surprising variety of question types, including "What is...", "Is there...", "How many...", and "Does the...". Quantitatively, the percentage of questions for different types is shown in Table 3. Several example questions and answers are shown in Fig. 2. A particularly interesting type of question is the "What is..." questions, since they have a diverse set of possible answers. See the supp. for visualizations for "What is..." questions.

Lengths. Fig. 4 shows the distribution of question lengths. We see that most questions range from four to ten words.

4.2. Answers

Typical Answers. Fig. 5 (top) shows the distribution of answers for several question types. We can see that a number of question types, such as "Is the...", "Are...", and "Does..." are typically answered using "yes" and "no" as answers. Other questions such as "What is..." and "What type..." have a rich diversity of responses. Other question types such as "What color..." or "Which..." have more specialized responses, such as colors, or "left" and "right". See the supplement for a list of the most popular answers.

Lengths. Most answers consist of a single word, with the distribution of answers containing one, two, or three words, respectively being 89.32%, 6.91%, and 2.74% for real images and 90.51%, 5.89%, and 2.49% for abstract scenes. The brevity of answers is not surprising, since the questions tend to elicit specific information from the images. This is in contrast with image captions that generically describe the entire image and hence tend to be longer. The brevity of our answers makes automatic evaluation feasible. While it may be tempting to believe the brevity of the answers makes the problem easier, recall that they are human-provided open-ended answers to open-ended questions. The questions typically require complex reasoning to arrive at these deceptively simple answers (see Fig. 2). There are currently 23,234 unique one-word answers in our dataset for real images and 3,770 for abstract scenes.

'Yes/No' and 'Number' Answers. Many questions are answered using either "yes" or "no" (or sometimes "maybe") – 38.37% and 40.66% of the questions on real images and abstract scenes respectively. Among these 'yes/no' questions, there is a bias towards "yes" – 58.83% and 55.86% of 'yes/no' answers are "yes" for real images and abstract scenes. Question types such as "How many..." are answered using numbers – 12.31% and 14.48% of the questions on real images and abstract scenes are 'number' questions. "2" is the most popular answer among the 'number' questions, making up 26.04% of the 'number' answers for

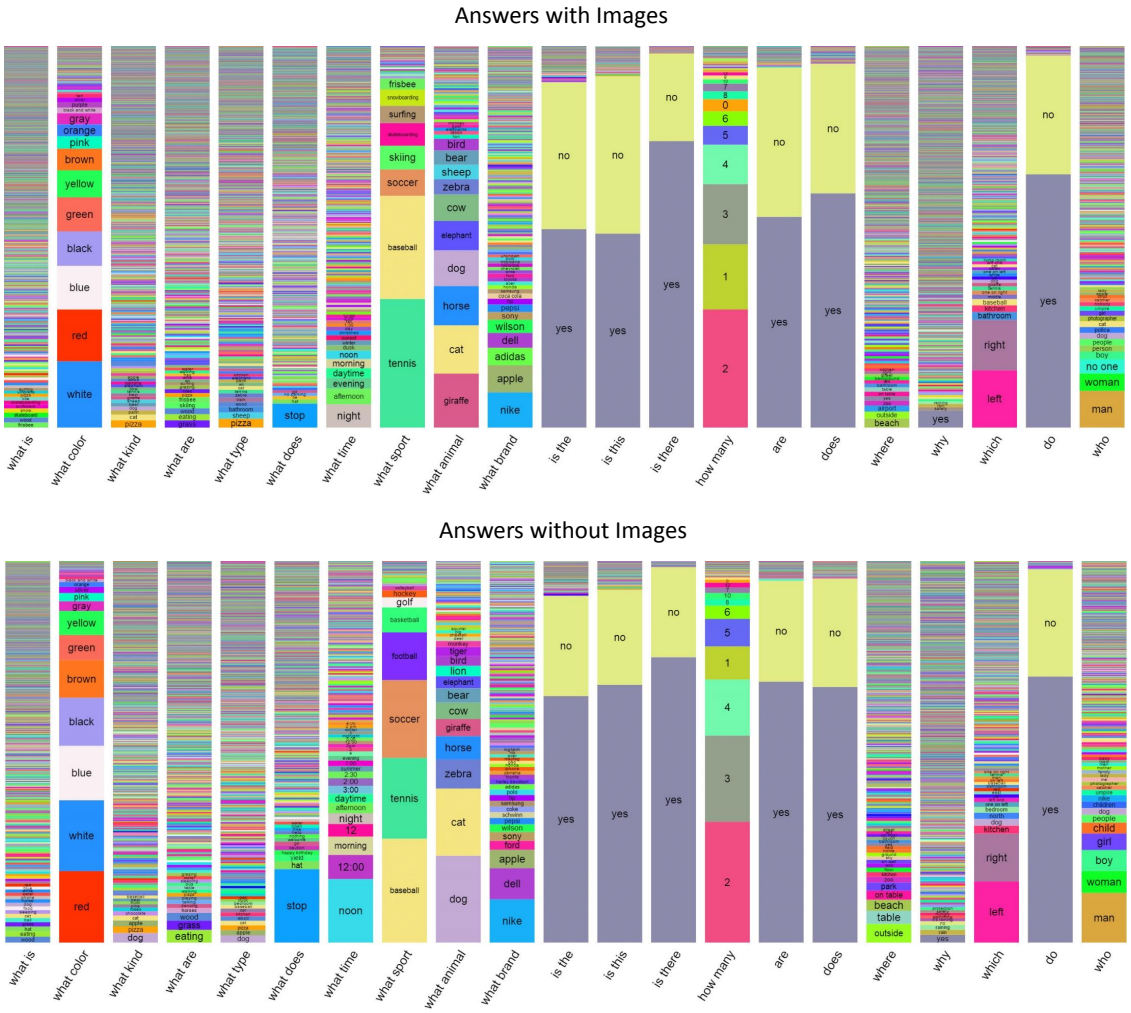

Figure 5: Distribution of answers per question type for a random sample of 60K questions for real images when subjects provide answers when given the image (top) and when not given the image (bottom).

real images and 39.85% for abstract scenes.

Subject Confidence. When the subjects answered the questions, we asked "Do you think you were able to answer the question correctly?". Fig. 6 shows the distribution of responses. A majority of the answers were labeled as confident for both real images and abstract scenes.

Inter-human Agreement. Does the self-judgment of confidence correspond to the answer agreement between subjects? Fig. 6 shows the percentage of questions in which (i) 7 or more, (ii) $3-7$, or (iii) less than 3 subjects agree on the answers given their average confidence score (0 = not confident, 1 = confident). As expected, the agreement between subjects increases with confidence. However, even if all of the subjects are confident the answers may still vary. This is not surprising since some answers may vary, yet have very similar meaning, such as "happy" and "joyful".

As shown in Table 1 (Question + Image), there is significant inter-human agreement in the answers for both real images (83.30%) and abstract scenes (87.49%). Note that on average each question has 2.70 unique answers for real images and 2.39 for abstract scenes. The agreement is significantly higher ($> 95\%$) for "yes/no" questions and lower for other questions ($< 76\%$), possibly due to the fact that we do exact string matching and do not account for synonyms, plurality, *etc*. Note that the automatic determination of synonyms is a

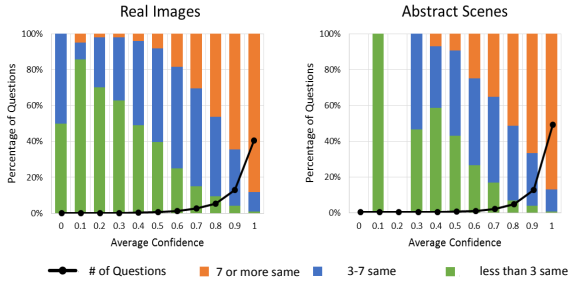

Figure 6: Number of questions per average confidence score (0 = not confident, 1 = confident) for real images and abstract scenes (black lines). Percentage of questions where 7 or more answers are same, 3-7 are same, less than 3 are same (color bars).

difficult problem, since the level of answer granularity can vary across questions.

4.3. Commonsense Knowledge

Is the Image Necessary? Clearly, some questions can sometimes be answered correctly using commonsense knowledge alone without the need for an image, *e.g.*, "What is the color of the fire hydrant?". We explore this issue by asking three subjects to answer the questions *without seeing the image* (see the examples in blue in Fig. 2). In Table 1 (Question), we show the percentage of questions for which the correct answer is provided over all questions, "yes/no" questions, and the other questions that are not "yes/no". For "yes/no" questions, the human subjects respond better than chance. For other questions, humans are only correct about 21% of the time. This demonstrates that understanding the visual information is critical to VQA and that commonsense information alone is not sufficient.

To show the qualitative difference in answers provided with and without images, we show the distribution of answers for various question types in Fig. 5 (bottom). The distribution of colors, numbers, and even "yes/no" responses is surprisingly different for answers with and without images.

Which Questions Require Common Sense? In order to identify questions that require commonsense reasoning to answer, we conducted two AMT studies (on a subset 10K questions from the real images of VQA trainval) asking subjects – (i) whether or not a question required knowledge external to the image, and (ii) the youngest age group that could answer the question – toddler (3-4), younger child (5-8), older child (9-12), teenager (13-17), adult (18+). Each question was shown to 10 subjects. We found that for 47.43% of question 3 or more subjects voted 'yes' to commonsense, (18.14%: 6 or more). In the 'human age required to answer question' study, we found the following distribution of responses: toddler: 15.3%, younger child: 39.7%, older child: 28.4%, teenager: 11.2%, adult: 5.5%. A fine-grained breakdown of average age required to answer a question is shown in Table 3. The two rankings of questions in terms of common sense required according to the two studies were largely correlated (Pearson's rank correlation: 0.58).

4.4. Captions *vs*. Questions

Do generic image captions provide enough information to answer the questions? Table 1 (Question + Caption) shows the percentage of questions answered correctly when human subjects are given the question and a human-provided caption describing the image, but not the image. As expected, the results are better than when humans are shown the questions alone. However, the accuracies are significantly lower than when subjects are shown the actual image. This demonstrates that in order to answer the questions correctly, deeper image understanding (beyond what image captions typically capture) is necessary. In fact, we find that the distributions of nouns, verbs, and adjectives mentioned in captions is statistically significantly different from those mentioned in our questions + answers (Kolmogorov-Smirnov test, $p < .001$) for both real images and abstract scenes. See supplementary material for details.

Dataset	Input	All	Yes/No	Number	Other
Real	Question	40.81	67.60	25.77	21.22
	Question + Caption*	57.47	78.97	39.68	44.41
	Question + Image	83.30	95.77	83.39	72.67
Abstract	Question	43.27	66.65	28.52	23.66
	Question + Caption*	54.34	74.70	41.19	40.18
	Question + Image	87.49	95.96	95.04	75.33

Table 1: Test-standard accuracy of human subjects when asked to answer the question without seeing the image (Question), seeing just a caption of the image and not the image itself (Question + Caption), and seeing the image (Question + Image). Results are shown for all questions, "yes/no" & "number" questions, and other questions that are neither answered "yes/no" nor number. All answers are free-form and not multiple-choice. *These accuracies are evaluated on a subset of 3K train questions (1K images).

5. VQA Baselines and Methods

In this section, we explore the difficulty of the VQA dataset for the MS COCO images using several baselines and novel methods. For reference, if we randomly choose an answer from the top 1K answers of the VQA train/val dataset, the test-standard accuracy is 0.12%. If we always select the most popular answer ("yes"), the accuracy is 29.72%. Picking the most popular answer per question type does 36.18% and a nearest neighbor approach does 40.61% on val (see the supplement for details).

We train on VQA train+val. Unless stated otherwise, all human accuracies are on test-standard, machine accuracies are on test-dev, and results involving human captions (in gray font) are trained on train and tested on val (because captions are not available for test).

For our baselines, we choose the top $K = 1000$ most frequent answers as possible outputs. This set of answers covers 82.67% of the train+val answers. We experiment with two models: (i) a multi-layer perceptron (MLP) neural network classifier with 2 hidden layers and 1000 hidden units (dropout 0.5) in each layer with tanh non-linearity, and (ii) an LSTM model followed by a softmax layer to generate the answer. We experimented with six inputs for the MLP model. **Question:** The top 1,000 words in the questions are used to create a bag-of-words representation. Since there is a strong correlation between the words that start a question and the answer (see Fig. 5), we find the top 10 first, second, and third words of the questions and create a 30 dimensional bag-of-words representation. These features are concatenated to get a 1,030 dimensional input representation. **Caption:** Similar to Table 1, we assume that a human-generated caption is given as input. We use a bag-of-words representation containing the 1,000 most popular words in the captions as the input feature. **Image:** We use

	Open-Answer				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
Question	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
Image	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
Q+I	52.64	75.55	33.67	37.37	58.97	75.59	34.35	50.33
LSTM Q	48.76	78.20	35.68	26.59	54.75	78.22	36.82	38.78
LSTM Q+I	53.74	78.94	35.24	36.42	57.17	78.95	35.80	43.41
Caption	26.70	65.50	02.03	03.86	28.29	69.79	02.06	03.82
Q+C	54.70	75.82	40.12	42.56	59.85	75.89	41.16	52.53

Table 2: Accuracy of our methods for the open-answer and multiple-choice tasks on the VQA test-dev for real images. (Caption and Q+C results are on val). See text for details.

the last hidden layer of VGGNet [41] as our 4096-dim feature. We also report the learned baseline results on **Question+Image (Q+I)**, **Question+Caption (Q+C)**, and **Question+Image+Caption (Q+I+C)** by simply concatenating the first hidden layer representations of networks trained on each feature individually. The LSTM model uses a one-hot encoding for the question words, and the same image features as above followed by a linear transformation to transform the image features to 1024 dimensions to match the LSTM encoding of the question. The question and image encodings are fused via element-wise multiplication.

For testing, we report the result on two different tasks: open-answer selects the answer with highest activation from all possible K answers and multiple-choice picks the answer that has the highest activation from the potential answers. As shown in Table 2, the accuracy using only the question is $\sim$48%, which demonstrates that the type of question is informative of the answer. As expected, results on multiple-choice are better than open-answer. All methods are significantly worse than human performance.

To gain further insights into these results, we computed accuracies by question type in Table 3. Interestingly, for question types that require more reasoning, such as "Is the" or "How many", the scene-level image features do not provide any additional information. However, for questions that can be answered using scene-level information, such as "What sport," we do see an improvement. Similarly, for questions whose answer may be contained in a generic caption we see improvement, such as "What animal". For all question types, the results are worse than human accuracies.

The accuracy of our **best model** (LSTM Q+I, selected using VQA test-dev accuracies) on VQA test-standard is **54.06%**. Finally, evaluating our model on the questions for which we have annotations for how old a human needs to be to answer the question correctly, we estimate that our model performs as well as a 4.45 year old child! See the supp. for details.

6. Conclusion and Discussion

In conclusion, we introduce the task of Visual Question Answering (VQA). Given an image and an open-ended, natural language question about the image, the task is to provide an accurate natural language answer. We provide a dataset

Question Type	Open-Answer K = 1000 Q	Q + I	Q + C	Open-Answer Human Q	Q + I	Human Age To Be Able To Answer
what is (13.84)	23.57	34.28	43.88	16.86	73.68	09.07
what color (08.98)	33.37	43.53	48.61	28.71	86.06	06.60
what kind (02.49)	27.78	42.72	43.88	19.10	70.11	10.55
what are (02.32)	25.47	39.10	47.27	17.72	69.49	09.03
what type (01.78)	27.68	42.62	44.32	19.53	70.65	11.04
is the (10.16)	70.76	69.87	70.50	65.24	95.67	08.51
is this (08.26)	70.34	70.79	71.54	63.35	95.43	10.13
how many (10.28)	43.78	40.33	47.52	30.45	86.32	07.67
are (07.57)	73.96	73.58	72.43	67.10	95.24	08.65
does (02.75)	76.81	75.81	75.88	69.96	95.70	09.29
where (02.90)	16.21	23.49	29.47	11.09	43.56	09.54
is there (03.60)	86.50	86.37	85.88	72.48	96.43	08.25
why (01.20)	16.24	13.94	14.54	11.80	21.50	11.18
which (01.21)	29.50	34.83	40.84	25.64	67.44	09.27
do (01.15)	77.73	79.31	74.63	71.33	95.44	09.23
what does (01.12)	19.58	20.00	23.19	11.12	75.88	10.02
what time (00.67)	8.35	14.00	18.28	07.64	58.98	09.81
who (00.77)	19.75	20.43	27.28	14.69	56.93	09.49
what sport (00.81)	37.96	81.12	93.87	17.86	95.59	08.07
what animal (00.53)	23.12	59.70	71.02	17.67	92.51	06.75
what brand (00.36)	40.13	36.84	32.19	25.34	80.95	12.50

Table 3: Open-answer test-dev results for different question types on real images (Q+C is reported on val). Questions types are determined by the one or two words that start the question. The percentage of questions for each type is shown in parentheses. Last column shows the human age required to answer the questions (as reported by AMT workers). See text for details.

containing over 250K images, 760K questions, and around 10M answers. We will set up an evaluation server and organize an annual challenge and an associated workshop to facilitate systematic progress. We demonstrate the wide variety of questions and answers in our dataset, as well as the diverse set of AI capabilities in computer vision, natural language processing, and commonsense reasoning required to answer these questions accurately.

The questions we solicited from our human subjects were open-ended and not task-specific. For some application domains, it would be useful to collect task-specific questions. For instance, questions may be gathered from subjects who are visually impaired [2], or the questions could focused on one specific domain (say sports). Bigham *et al*. [2] created an application that allows the visually impaired to capture images and ask open-ended questions that are answered by human subjects. Interestingly, these questions can rarely be answered using generic captions. Training on task-specific datasets may help enable practical VQA applications.

We believe VQA has the distinctive advantage of pushing the frontiers on "AI-complete" problems, while being amenable to automatic evaluation. Given the recent progress in the community, we believe the time is ripe to take on such an endeavor.

Acknowledgements. We would like to acknowledge the countless hours of effort provided by the workers on Amazon Mechanical Turk. This work was supported in part by the The Paul G. Allen Family Foundation via an award to D.P., ICTAS at Virginia Tech via awards to D.B. and D.P., Google Faculty Research Awards to D.P. and D.B., the National Science Foundation CAREER award to D.B., the Army Research Office YIP Award to D.B., and a Office of Naval Research grant to D.B.

References

[1] S. Antol, C. L. Zitnick, and D. Parikh. Zero-Shot Learning via Visual Abstraction. In *ECCV*, 2014. 2, 3

[2] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. In *User Interface Software and Technology*, 2010. 1, 2, 8

[3] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *International Conference on Management of Data*, 2008. 2

[4] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In *AAAI*, 2010. 2

[5] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015. 3

[6] X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting Visual Knowledge from Web Data. In *ICCV*, 2013. 2

[7] X. Chen and C. L. Zitnick. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In *CVPR*, 2015. 1, 2

[8] J. Deng, A. C. Berg, and L. Fei-Fei. Hierarchical Semantic Indexing for Large Scale Image Retrieval. In *CVPR*, 2011. 2

[9] 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. In *CVPR*, 2015. 1, 2

[10] D. Elliott and F. Keller. Comparing Automatic Evaluation Measures for Image Description. In *ACL*, 2014. 1

[11] A. Fader, L. Zettlemoyer, and O. Etzioni. Paraphrase-Driven Learning for Open Question Answering. In *ACL*, 2013. 2

[12] A. Fader, L. Zettlemoyer, and O. Etzioni. Open Question Answering over Curated and Extracted Knowledge Bases. In *International Conference on Knowledge Discovery and Data Mining*, 2014. 2

[13] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From Captions to Visual Concepts and Back. In *CVPR*, 2015. 1, 2

[14] A. Farhadi, M. Hejrati, A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every Picture Tells a Story: Generating Sentences for Images. In *ECCV*, 2010. 2

[15] H. Gao, J. Mao, J. Zhou, Z. Huang, and A. Yuille. Are you talking to a machine? dataset and methods for multilingual image question answering. In *NIPS*, 2015. 2

[16] D. Geman, S. Geman, N. Hallonquist, and L. Younes. A Visual Turing Test for Computer Vision Systems. In *PNAS*, 2014. 1, 2

[17] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition. In *ICCV*, December 2013. 2

[18] M. Hodosh, P. Young, and J. Hockenmaier. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR*, 2013. 1

[19] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*, 2015. 1, 2

[20] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *EMNLP*, 2014. 2

[21] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *TACL*, 2015. 1, 2

[22] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What Are You Talking About? Text-to-Image Coreference. In *CVPR*, 2014. 2

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012. 2

[24] G. Kulkarni, V. Premraj, S. L. Sagnik Dhar and, Y. Choi, A. C. Berg, and T. L. Berg. Baby Talk: Understanding and Generating Simple Image Descriptions. In *CVPR*, 2011. 2

[25] D. B. Lenat and R. V. Guha. *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., 1989. 2

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, 2014. 2, 3, 4

[27] X. Lin and D. Parikh. Don't Just Listen, Use Your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks. In *CVPR*, 2015. 2

[28] X. Lin and D. Parikh. Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *CVPR*, 2015. 2

[29] H. Liu and P. Singh. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 2004. 2

[30] M. Malinowski and M. Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In *NIPS*, 2014. 1, 2

[31] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 2

[32] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain Images with Multimodal Recurrent Neural Networks. *CoRR*, abs/1410.1090, 2014. 1, 2

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS*, 2013. 4

[34] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. L. Berg, and H. Daume III. Midge: Generating Image Descriptions From Computer Vision Detections. In *ACL*, 2012. 2

[35] M. Mitchell, K. Van Deemter, and E. Reiter. Generating Expressions that Refer to Visible Objects. In *HLT-NAACL*, 2013. 2

[36] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking People with "Their" Names using Coreference Resolution. In *ECCV*, 2014. 2

[37] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015. 2

[38] M. Richardson, C. J. Burges, and E. Renshaw. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *EMNLP*, 2013. 2

[39] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating Video Content to Natural Language Descriptions. In *ICCV*, 2013. 2

[40] F. Sadeghi, S. K. Kumar Divvala, and A. Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, 2015. 2

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 8

[42] 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. *IEEE MultiMedia*, 2014. 1, 2

[43] R. Vedantam, C. L. Zitnick, and D. Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, 2015. 1

[44] R. Vendantam, X. Lin, T. Batra, C. L. Zitnick, and D. Parikh. Learning common sense through visual abstraction. In *ICCV*, 2015. 2

[45] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, 2015. 1, 2

[46] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *CoRR*, abs/1502.05698, 2015. 2

[47] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill-in-the-blank description generation and question answering. In *ICCV*, 2015. 2

[48] C. L. Zitnick and D. Parikh. Bringing Semantics Into Focus Using Visual Abstraction. In *CVPR*, 2013. 2, 3

[49] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the Visual Interpretation of Sentences. In *ICCV*, 2013. 3

[50] C. L. Zitnick, R. Vedantam, and D. Parikh. Adopting Abstract Images for Semantic Scene Understanding. *PAMI*, 2015. 3