VQA: Visual Question Answering

www.visualqa.org

Aishwarya Agrawal*, Jiasen Lu*, Stanislaw Antol*,
Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh

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 and methods for VQA are provided and compared with human performance. Our VQA demo is available on CloudCV (http://cloudcv.org/vqa).

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 [16], [9], [12], [38], [26], [24], [53]. 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 [51], [13], [22].

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 [3] 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 [19], [36], [50], [3] is also amenable to automatic quantitative evaluation, making it possible to effectively track 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 [45], [33]. Unlike the open-ended task that requires a free-form response, the multiple-choice task only requires an

Fig. 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.

- *The first three authors contributed equally.
- A. Agrawal, J. Lu and S. Antol are with Virginia Tech.
- M. Mitchell is with Microsoft Research, Redmond.
- C. L. Zitnick is with Facebook AI Research.
- D. Batra and D. Parikh are with Georgia Institute of Technology.

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 [32] and a newly created abstract scene dataset [57], [2] 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 [29]. 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 [5], [8], [31], [35], [4] 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 [19], [36], [50], [3]. However, unlike our work, these are fairly restricted (sometimes synthetic) settings with small datasets. For instance, [36] only considers questions whose answers come from a predefined closed world of 16 basic colors or 894 object categories. [19] 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 [19], [36] (>250,000 *vs*. 2,591 and 1,449 images respectively). The proposed VQA task has connections to other related work: [50] has studied joint parsing of videos and corresponding text to answer queries on two datasets containing 15 video clips each. [3] uses crowdsourced workers to answer questions about visual content asked by visually-impaired users. In concurrent work, [37] proposed combining an LSTM for the question with a CNN for the image to generate an answer. In their model, the LSTM question representation is conditioned on the CNN image features at each time step, and the final LSTM hidden state is used to sequentially decode the answer phrase. In contrast, the model developed in this paper explores "late fusion" – *i.e*., the LSTM question representation and the CNN image features are computed independently, *fused* via an element-wise multiplication, and then passed through fully-connected layers to generate a softmax distribution over output answer classes. [34] generates abstract scenes to capture visual common sense relevant to answering (purely textual) fill-in-the-blank and visual paraphrasing questions. [47] and [52] use visual information to assess the plausibility of common sense assertions. [55] 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, [18] collected questions & answers in Chinese (later translated to English by humans) for COCO images. [44] 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 [15], [14], [54], [45]). Other related textual tasks include sentence completion (*e.g*., [45] with multiple-choice answers). These approaches provide inspiration for VQA techniques. One key concern in text is the *grounding* of questions. For instance, [54] 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 [11], [29], image captioning [30], [17], [40], [9], [16], [53], [12], [24], [38], [26] and video captioning [46], [21], 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 [53]). The questions in VQA require detailed specific information about the image for which generic image captions are of little use [3].

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 [28], [43] or generating referring expressions [25], [42] for a particular object in an image that 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

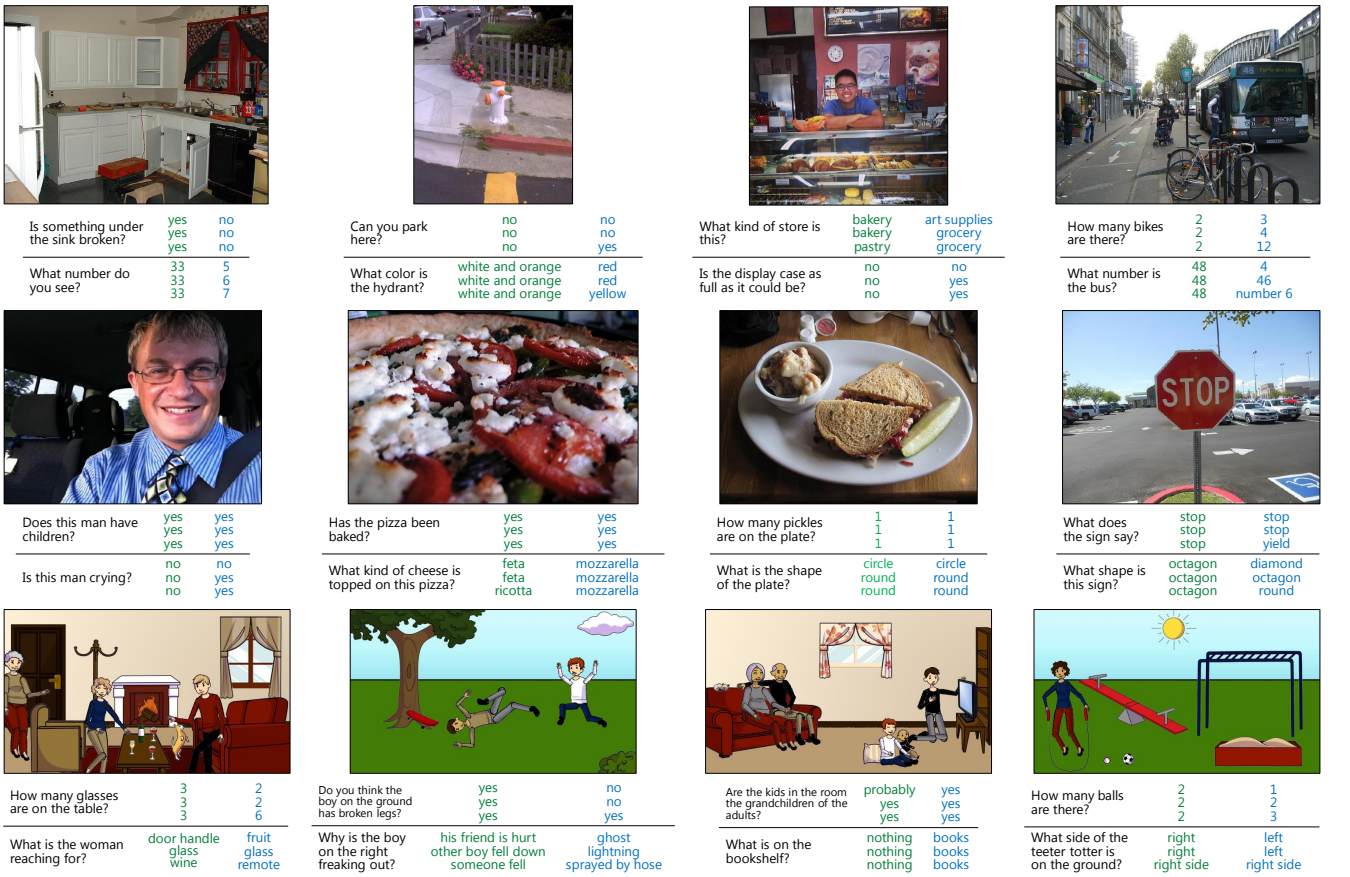


Fig. 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset. See the appendix for more examples.

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 baselines' & methods' 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) [32] 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 [2], [57], [58], [59] containing 50K scenes. The dataset contains 20 "paperdoll" human models [2] 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 [57], [58], [59]. See the appendix 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 [32] (including test-dev, test-standard, test-challenge, test-reserve). For the VQA challenge (see section 6), test-dev is used for debugging and validation experiments and allows for unlimited submission to the evaluation server. Test-standard is the 'default' test data for the VQA competition. When comparing to the state of the art (e.g., in papers), results should be reported on test-standard. Test-standard is also used to maintain a public leaderboard that is updated upon submission. Test-reserve is used to protect against possible overfitting. If there are substantial differences between a method's scores on test-standard and test-reserve, this raises a red-flag and prompts further investigation. Results on test-reserve are not publicly revealed. Finally, test-challenge is used to determine the winners of the challenge.

For abstract scenes, we created splits for standardization, separating the scenes into 20K/10K/20K for train/val/test splits, respectively. There are no subsplits (test-dev, test-standard, test-challenge, test-reserve) for abstract scenes.

Captions. The MS COCO dataset [32], [7] 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 questions is a significant challenge. Many simple questions

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

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 appendix, 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!*
> *Ask a question about this scene that this smart robot probably can not answer, but any human can easily answer while looking at the scene in the image.*"

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 ∼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 the appendix for more details about the user interface to collect answers. See Section 4 for an analysis of the answers provided.

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

For the open-ended task, the generated answers are evaluated using the following accuracy metric:

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

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 [39], since they often group together words that we wish to distinguish, such as "left" and "right". We also avoid using evaluation metrics from machine translation such as BLEU and ROUGE because such metrics are typically applicable and reliable for sentences containing multiple words. In VQA, most answers (89.32%) are single word; thus there no high-order n-gram matches between predicted answers and ground-truth answers, and low-order n-gram matches degenerate to exact-string matching. Moreover, these automatic metrics such as BLEU and ROUGE have been found to poorly correlate with human judgement for tasks such as image caption evaluation [6].

For multiple-choice task, 18 candidate answers are created for each question. As with the open-ended task, the accuracy of a chosen option is computed based on the number of human subjects who provided that answer (divided 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. See the appendix for more details about the user interface to collect these answers. 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 appendix.

Note that all 18 candidate answers are unique. But since 10 different subjects answered every question, it is possible that more than one of those 10 answers be present in the 18 choices. In such cases, according to the accuracy metric, multiple options could have a non-zero accuracy.

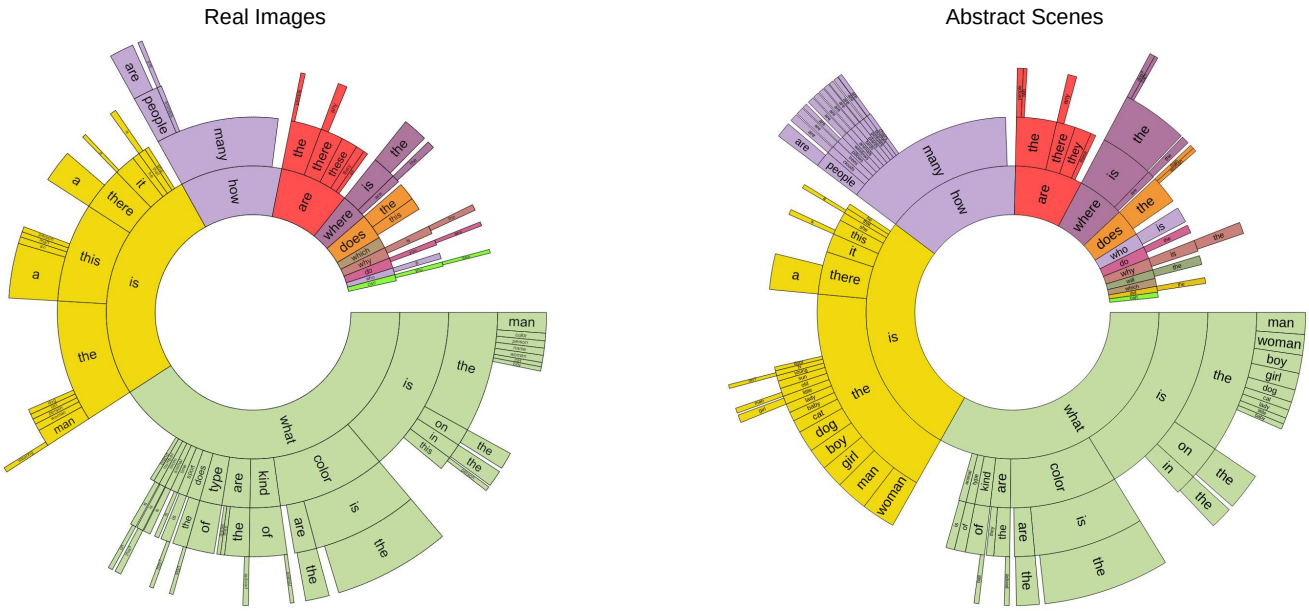


Fig. 3: Distribution of questions by their first four words for a random sample of 60K questions for real images (left) and all questions for abstract scenes (right). The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing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

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 [32] 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 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 appendix for visualizations for "What is..." questions.

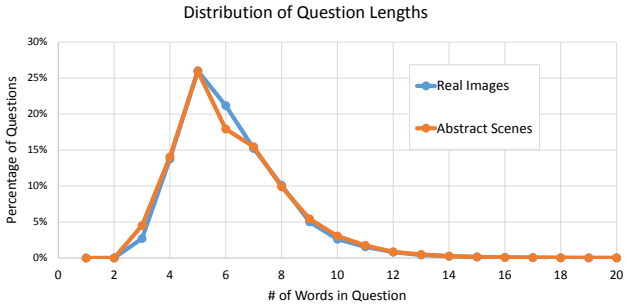


Fig. 4: Percentage of questions with different word lengths for real images and abstract scenes.

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 appendix 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

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

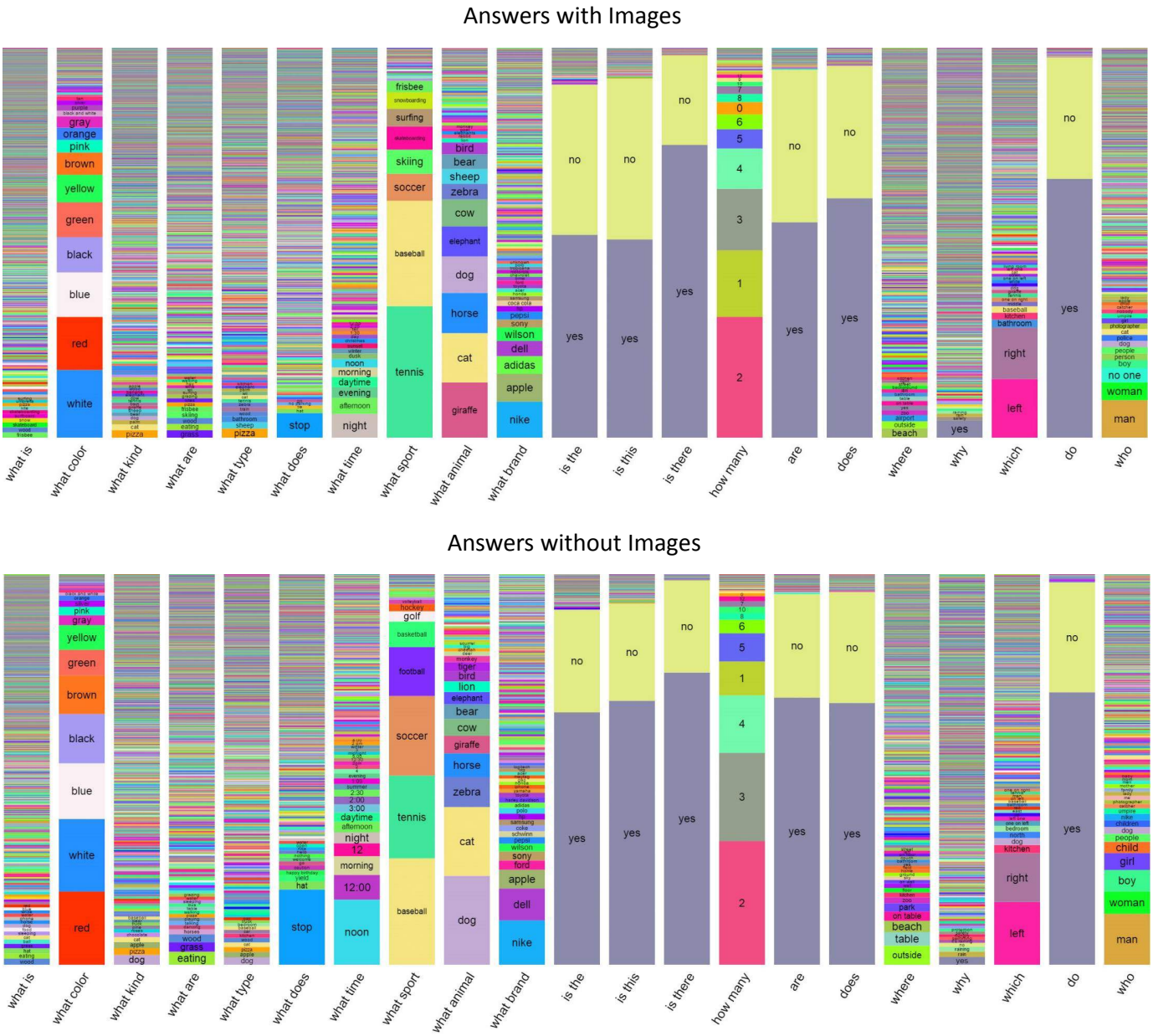


Fig. 5: Distribution of answers per question type for a random sample of 60K questions for real images when subjects provide answers when given the image (top) and when not given the image (bottom).

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 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

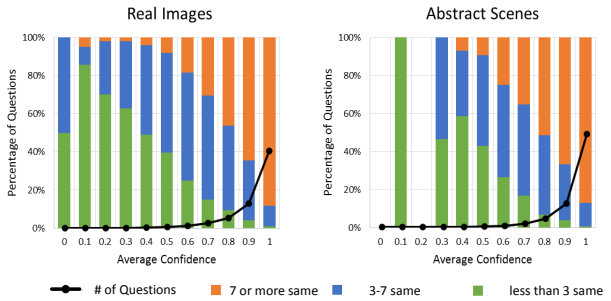


Fig. 6: Number of questions per average confidence score (0 = not confident, 1 = confident) for real images and abstract scenes (black lines). Percentage of questions where 7 or more answers are same, 3-7 are same, less than 3 are same (color bars).

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 perform exact string matching and do not account for synonyms, plurality, *etc*. Note that the automatic determination of synonyms is a 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 –

1) Whether or not they believed a question required commonsense to answer the question, and

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).

2) The youngest age group that they believe a person must be in order to be able to correctly 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 questions 3 or more subjects voted 'yes' to commonsense, (18.14%: 6 or more). In the 'perceived 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%. In Figure 7 we show several questions for which a majority of subjects picked the specified age range. Surprisingly the perceived age needed to answer the questions is fairly well distributed across the different age ranges. As expected the questions that were judged answerable by an adult (18+) generally need specialized knowledge, whereas those answerable by a toddler (3-4) are more generic.

We measure the degree of commonsense required to answer a question as the percentage of subjects (out of 10) who voted "yes" in our "whether or not a question requires commonsense" study. A fine-grained breakdown of average age and average degree of common sense (on a scale of $0 - 100$) required to answer a question is shown in Table 3. The average age and the average degree of commonsense across all questions is 8.92 and 31.01% respectively.

It is important to distinguish between:

1) How old someone needs to be to be able to answer a question correctly, and
2) How old people *think* someone needs to be to be able to answer a question correctly.

Our age annotations capture the latter – perceptions of MTurk workers in an uncontrolled environment. As such, the relative ordering of question types in Table 3 is more important than absolute age numbers. 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

3-4 (15.3%)	**5-8** (39.7%)	**9-12** (28.4%)	**13-17** (11.2%)	**18+** (5.5%)
Is that a bird in the sky?	How many pizzas are shown?	Where was this picture taken?	Is he likely to get mugged if he walked down a dark alleyway like this?	What type of architecture is this?
What color is the shoe?	What are the sheep eating?	What ceremony does the cake commemorate?	Is this a vegetarian meal?	Is this a Flemish bricklaying pattern?
How many zebras are there?	What color is his hair?	Are these boats too tall to fit under the bridge?	What type of beverage is in the glass?	How many calories are in this pizza?
Is there food on the table?	What sport is being played?	What is the name of the white shape under the batter?	Can you name the performer in the purple costume?	What government document is needed to partake in this activity?
Is this man wearing shoes?	Name one ingredient in the skillet.	Is this at the stadium?	Besides these humans, what other animals eat here?	What is the make and model of this vehicle?

Fig. 7: Example questions judged by Mturk workers to be answerable by different age groups. The percentage of questions falling into each age group is shown in parentheses.

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 the appendix for details.

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. 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).

5.1 Baselines

We implemented the following baselines:

1) **random:** We randomly choose an answer from the top 1K answers of the VQA train/val dataset.
2) **prior ("yes"):** We always select the most popular answer ("yes") for both the open-ended and multiple-choice tasks. Note that "yes" is always one of the choices for the multiple-choice questions.
3) **per Q-type prior:** For the open-ended task, we pick the most popular answer per question type (see the appendix for details). For the multiple-choice task, we pick the answer (from the provided choices) that is most similar to the picked answer for the open-ended task using cosine similarity in Word2Vec[39] feature space.
4) **nearest neighbor:** Given a test image, question pair, we first find the K nearest neighbor questions and associated images from the training set. See appendix for details on how neighbors are found. Next, for the open-ended task, we pick the most frequent ground truth answer from this set of nearest neighbor question, image pairs. Similar to the "per Q-type prior" baseline, for the multiple-choice task, we pick the answer (from the provided choices) that is most similar to the picked answer for the open-ended task using cosine similarity in Word2Vec[39] feature space.

5.2 Methods

For our methods, we develop a 2-channel vision (image) + language (question) model that culminates with a softmax over K possible outputs. 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 describe the different components of our model below:

Image Channel: This channel provides an embedding for the image. We experiment with two embeddings –

1) **I:** The activations from the last hidden layer of VGGNet [48] are used as 4096-dim image embedding.
2) **norm I:** These are ℓ_2 normalized activations from the last hidden layer of VGGNet [48].

Question Channel: This channel provides an embedding for the question. We experiment with three embeddings –

1) **Bag-of-Words Question (BoW Q)**: 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-dim embedding for the question.
2) **LSTM Q:** An LSTM with one hidden layer is used to obtain 1024-dim embedding for the question. The embedding obtained from the LSTM is a concatenation of last cell state and last hidden state representations (each being 512-dim) from the hidden layer of the LSTM. Each question word is encoded with 300-dim embedding by a fully-connected layer + tanh non-linearity which is then fed to the LSTM. The input vocabulary to the embedding layer consists of all the question words seen in the training dataset.
3) **deeper LSTM Q:** An LSTM with two hidden layers is used to obtain 2048-dim embedding for the question. The embedding obtained from the LSTM is a concatenation of last cell state and last hidden state representations (each

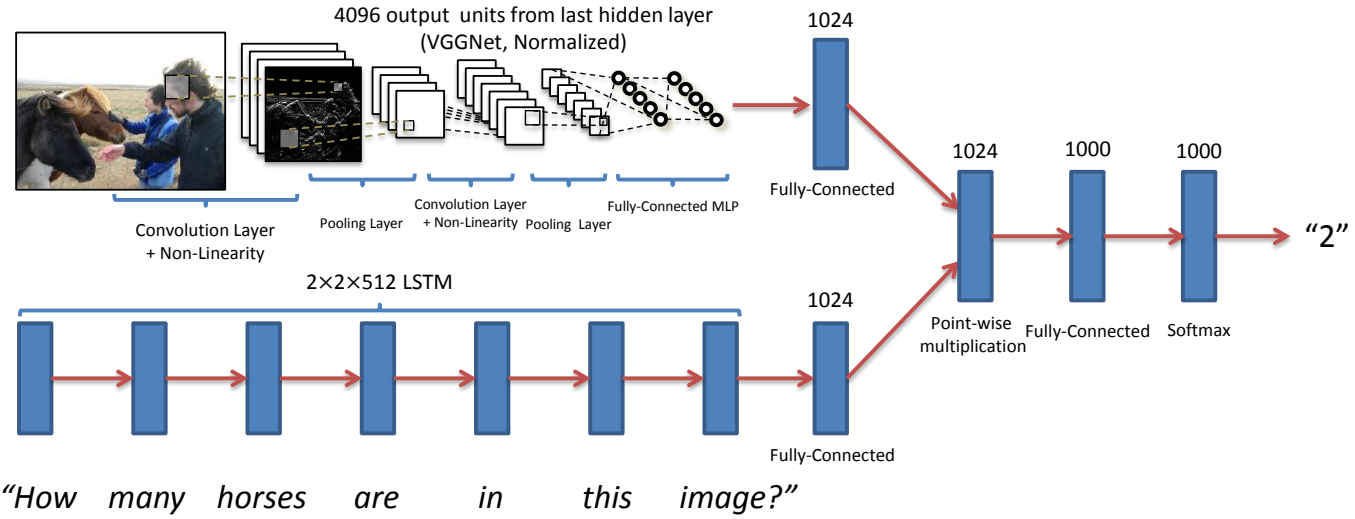


Fig. 8: Our best performing model (deeper LSTM Q + norm I). This model uses a two layer LSTM to encode the questions and the last hidden layer of VGGNet [48] to encode the images. The image features are then ℓ_2 normalized. Both the question and image features are transformed to a common space and fused via element-wise multiplication, which is then passed through a fully connected layer followed by a softmax layer to obtain a distribution over answers.

being 512-dim) from each of the two hidden layers of the LSTM. Hence 2 (hidden layers) x 2 (cell state and hidden state) x 512 (dimensionality of each of the cell states, as well as hidden states) in Fig. 8. This is followed by a fully-connected layer + tanh non-linearity to transform 2048-dim embedding to 1024-dim. The question words are encoded in the same way as in LSTM Q.

Multi-Layer Perceptron (MLP): The image and question embeddings are combined to obtain a single embedding.

1) For **BoW Q + I** method, we simply concatenate the BoW Q and I embeddings.
2) For **LSTM Q + I**, and **deeper LSTM Q + norm I** (Fig. 8) methods, the image embedding is first transformed to 1024-dim by a fully-connected layer + tanh non-linearity to match the LSTM embedding of the question. The transformed image and LSTM embeddings (being in a common space) are then fused via element-wise multiplication.

This combined image + question embedding is then passed to an MLP – a fully connected neural network classifier with 2 hidden layers and 1000 hidden units (dropout 0.5) in each layer with tanh non-linearity, followed by a softmax layer to obtain a distribution over K answers. The entire model is learned end-to-end with a cross-entropy loss. VGGNet parameters are frozen to those learned for ImageNet classification and not fine-tuned in the image channel.

We also experimented with providing captions as input to our model. 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 caption embedding (**Caption**). For **BoW Question + Caption (BoW Q + C)** method, we simply concatenate the BoW Q and C embeddings.

For testing, we report the result on two different tasks: open-ended 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.

	Open-Ended				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
prior ("yes")	29.66	70.81	00.39	01.15	29.66	70.81	00.39	01.15
per Q-type prior	37.54	71.03	35.77	09.38	39.45	71.02	35.86	13.34
nearest neighbor	42.70	71.89	24.36	21.94	48.49	71.94	26.00	33.56
BoW Q	48.09	75.66	36.70	27.14	53.68	75.71	37.05	38.64
I	28.13	64.01	00.42	03.77	30.53	69.87	00.45	03.76
BoW 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
deeper LSTM Q	50.39	78.41	34.68	30.03	55.88	78.45	35.91	41.13
deeper LSTM Q + norm I	**57.75**	**80.50**	**36.77**	**43.08**	**62.70**	**80.52**	**38.22**	**53.01**
Caption	26.70	65.50	02.03	03.86	28.29	69.79	02.06	03.82
BoW 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-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image, C = Caption. (Caption and BoW Q + C results are on val). See text for details.

5.3 Results

Table 2 shows the accuracy of our baselines and methods for both the open-ended and multiple-choice tasks on the VQA test-dev for real images.

As expected, the vision-alone model (I) that completely ignores the question performs rather poorly (open-ended: 28.13% / multiple-choice: 30.53%). In fact, on open-ended task, the vision-alone model (I) performs worse than the prior ("yes") baseline, which ignores both the image *and* question (responding to every question with a "yes").

Interestingly, the language-alone methods (per Q-type prior, BoW Q, LSTM Q) that ignore the image perform surprisingly well, with BoW Q achieving 48.09% on open-ended (53.68% on multiple-choice) and LSTM Q achieving 48.76% on open-ended (54.75% on multiple-choice); both outperforming the nearest neighbor baseline (open-ended: 42.70%, multiple-

choice: 48.49%). Our quantitative results and analyses suggest that this might be due to the language-model exploiting subtle statistical priors about the question types (e.g. "What color is the banana?" can be answered with "yellow" without looking at the image). For a detailed discussion of the subtle biases in the questions, please see [56].

The accuracy of our **best model** (deeper LSTM Q + norm I (Fig. 8), selected using VQA test-dev accuracies) on VQA test-standard is 58.16% (open-ended) / 63.09% (multiple-choice). We can see that our model is able to significantly outperform both the vision-alone and language-alone baselines. As a general trend, results on multiple-choice are better than open-ended. All methods are significantly worse than human performance.

Our VQA demo is available on CloudCV [1] – http://cloudcv.org/vqa. This will be updated with newer models as we develop them.

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.

We also analyzed the accuracies of our best model (deeper LSTM Q + norm I) on a subset of questions with certain specific (ground truth) answers. In Fig. 9, we show the average accuracy of the model on questions with 50 most frequent ground truth answers on the VQA validation set (plot is sorted by accuracy, not frequency). We can see that the model performs well for answers that are common visual objects such as "wii", "tennis", "bathroom" while the performance is somewhat underwhelming for counts (*e.g.*, "2", "1", "3"), and particularly poor for higher counts (*e.g.*, "5", "6", "10", "8", "7").

In Fig. 10, we show the distribution of 50 most frequently predicted answers when the system is correct on the VQA validation set (plot is sorted by prediction frequency, not accuracy). In this analysis, "system is correct" implies that it has VQA accuracy 1.0 (see section 3 for accuracy metric). We can see that the frequent ground truth answers (*e.g.*, "yes", "no", "2", "white", "red", "blue", "1", "green") are more frequently predicted than others when the model is correct.

Finally, evaluating our best model (deeper LSTM Q + norm I) on the validation questions for which we have age annotations (how old a human needs to be to answer the question correctly), we estimate that our model performs as well as a 4.74 year old child! The average age required on the same set of questions is 8.98. Evaluating the same model on the validation questions for which we have commonsense annotations (whether the question requires commonsense to answer it), we estimate that it has degree of commonsense of 17.35%. The average degree of commonsense required on same set of questions is 31.23%. Again, these estimates reflect the age and commonsense perceived by MTurk workers that would be required to answer the question. See the appendix for details.

Question Type	Open-Ended K = 1000 Q	Open-Ended K = 1000 Q + I	Open-Ended K = 1000 Q + C	Open-Ended Human Q	Open-Ended Human Q + I	Human Age To Be Able To Answer	Commonsense To Be Able To Answer (%)
what is (13.84)	23.57	34.28	43.88	16.86	73.68	09.07	27.52
what color (08.98)	33.37	43.53	48.61	28.71	86.06	06.60	13.22
what kind (02.49)	27.78	42.72	43.88	19.10	70.11	10.55	40.34
what are (02.32)	25.47	39.10	47.27	17.72	69.49	09.03	28.72
what type (01.78)	27.68	42.62	44.32	19.53	70.65	11.04	38.92
is the (10.16)	70.76	69.87	70.50	65.24	95.67	08.51	30.30
is this (08.26)	70.34	70.79	71.54	63.35	95.43	10.13	45.32
how many (10.28)	43.78	40.33	47.52	30.45	86.32	07.67	15.93
are (07.57)	73.96	73.58	72.43	67.10	95.24	08.65	30.63
does (02.75)	76.81	75.81	75.88	69.96	95.70	09.29	38.97
where (02.90)	16.21	23.49	29.47	11.09	43.56	09.54	36.51
is there (03.60)	86.50	86.37	85.88	72.48	96.43	08.25	19.88
why (01.20)	16.24	13.94	14.54	11.80	21.50	11.18	73.56
which (01.21)	29.50	34.83	40.84	25.64	67.44	09.27	30.00
do (01.15)	77.73	79.31	74.63	71.33	95.44	09.23	37.68
what does (01.12)	19.58	20.00	23.19	11.12	75.88	10.02	33.27
what time (00.67)	8.35	14.00	18.28	07.64	58.98	09.81	31.83
who (00.77)	19.75	20.43	27.28	14.69	56.93	09.49	43.82
what sport (00.81)	37.96	81.12	93.87	17.86	95.59	08.07	31.87
what animal (00.53)	23.12	59.70	71.02	17.67	92.51	06.75	18.04
what brand (00.36)	40.13	36.84	32.19	25.34	80.95	12.50	41.33

TABLE 3: Open-ended test-dev results for different question types on real images (Q+C is reported on val). Machine performance is reported using the bag-of-words representation for questions. 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 and second last columns respectively show the average human age and average degree of commonsense required to answer the questions (as reported by AMT workers), respectively. See text for details.

We further analyzed the performance of the model for different age groups on the validation questions for which we have age annotations. In Fig. 11, we computed the average accuracy of the predictions made by the model for questions belonging to different age groups. Perhaps as expected, the accuracy of the model decreases as the age of the question increases (from 61.07% at $3-4$ age group to 47.83% at $18+$ age group).

In Fig. 12, we show the distribution of age of questions for different levels of accuracies achieved by our system on the validation questions for which we have age annotations. It is interesting to see that the relative proportions of different age groups is consistent across all accuracy bins with questions belonging to the age group 5-8 comprising the majority of the predictions which is expected because 5-8 is the most common age group in the dataset (see Fig. 7).

Table 4 shows the accuracy of different ablated versions of our best model (deeper LSTM Q + norm I) for both the open-ended and multiple-choice tasks on the VQA test-dev for real images. The different ablated versions are as follows –

1) **Without I Norm:** In this model, the activations from the last hidden layer of VGGNet [48] are not ℓ_2-normalized. Comparing the accuracies in Table 4 and Table 2, we can see that ℓ_2-normalization of image features boosts the performance by 0.16% for open-ended task and by 0.24% for multiple-choice task.
2) **Concatenation:** In this model, the transformed image and LSTM embeddings are concatenated (instead of element-wise multiplied), resulting in doubling the number of

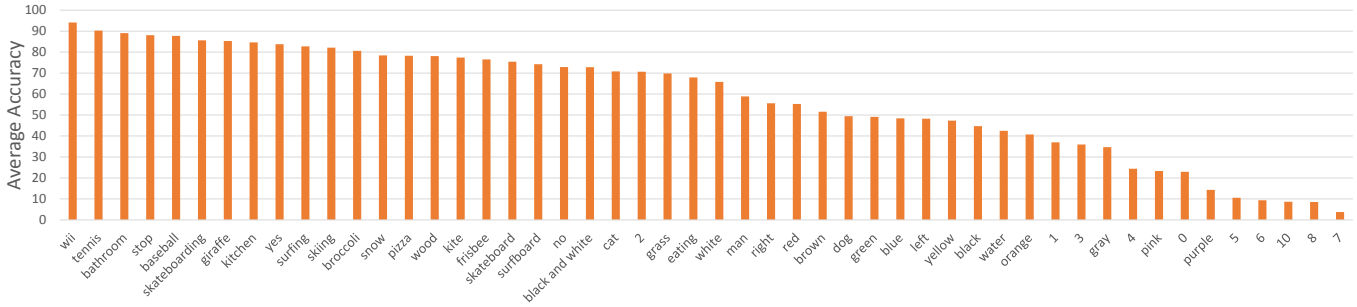


Fig. 9: Pr (system is correct | answer) for 50 most frequent ground truth answers on the VQA validation set (plot is sorted by accuracy, not frequency). System refers to our best model (deeper LSTM Q + norm I).

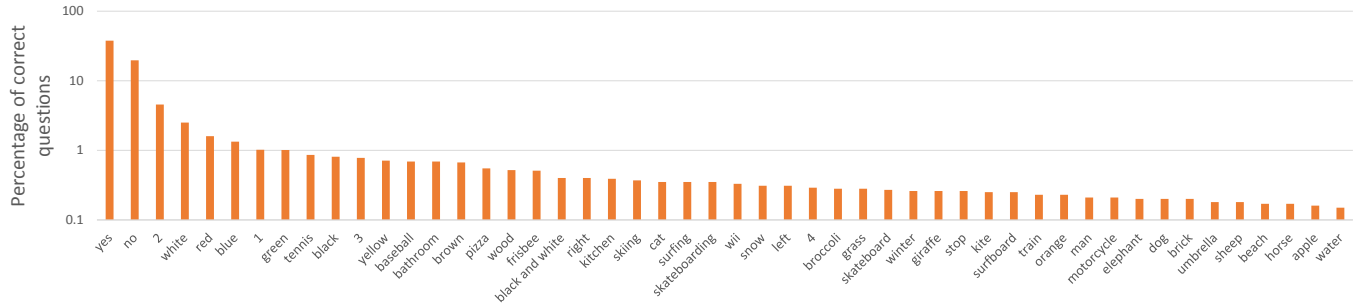


Fig. 10: Pr (answer | system is correct) for 50 most frequently predicted answers on the VQA validation set (plot is sorted by prediction frequency, not accuracy). System refers to our best model (deeper LSTM Q + norm I).

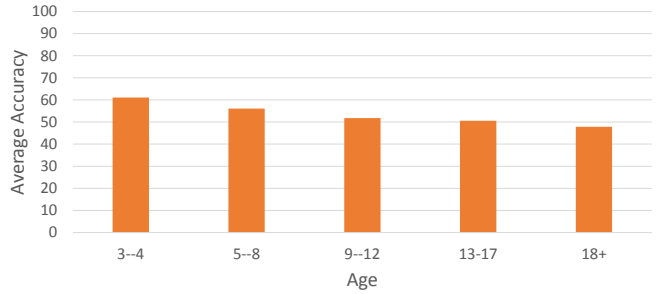


Fig. 11: Pr (system is correct | age of question) on the VQA validation set. System refers to our best model (deeper LSTM Q + norm I).

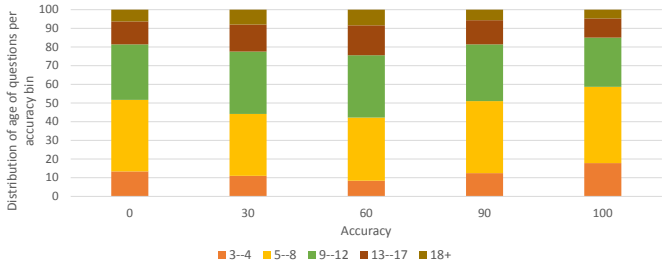


Fig. 12: Pr (age of question | system is correct) on the VQA validation set. System refers to our best model (deeper LSTM Q + norm I).

parameters in the following fully-connected layer. Comparing the accuracies in Table 4 and Table 2, we can see that element-wise fusion performs better by 0.95% for open-ended task and by 1.24% for multiple-choice task.

3) **K = 500:** In this model, we use K = 500 most frequent answers as possible outputs. Comparing the accuracies in Table 4 and Table 2, we can see that K = 1000 performs better than K = 500 by 0.82% for open-ended task and by 1.92% for multiple-choice task.
4) **K = 2000:** In this model, we use K = 2000 most frequent answers as possible outputs. Comparing the accuracies in Table 4 and Table 2, we can see that K = 2000 performs better then K = 1000 by 0.40% for open-ended task and by 1.16% for multiple-choice task.
5) **Truncated Q Vocab @ 5:** In this model, the input vocabulary to the embedding layer (which encodes the question words) consists of only those question words which occur atleast 5 times in the training dataset, thus reducing the vocabulary size from 14770 (when all question words are used) to 5134 (65.24% reduction). Remaining question words are replaced with UNK (unknown) tokens. Comparing the accuracies in Table 4 and Table 2, we can see that truncating the question vocabulary @ 5 performs better than using all questions words by 0.24% for open-ended task and by 0.17% for multiple-choice task.
6) **Truncated Q Vocab @ 11:** In this model, the input vocabulary to the embedding layer (which encodes the question words) consists of only those question words which occur atleast 11 times in the training dataset, thus reducing the vocabulary size from 14770 (when all question words are used) to 3561 (75.89% reduction). Remaining question words are replaced with UNK (unknown) tokens. Comparing the accuracies in Table 4 and Table 2, we can see that truncating the question vocabulary @ 11 performs better than using all questions words by 0.06% for open-ended task and by 0.02% for multiple-choice task.
7) **Filtered Dataset:** We created a filtered version of the VQA train + val dataset in which we only keep the answers with subject confidence "yes". Also, we keep only those questions for which at least 50% (5 out of 10) answers are annotated with subject confidence "yes". The resulting filtered dataset consists of 344600 questions,

	Open-Ended				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
Without I Norm	57.59	80.41	36.63	42.84	62.46	80.43	38.10	52.62
Concatenation	56.80	78.49	35.08	43.19	61.46	78.52	36.43	52.54
K = 500	56.93	80.61	36.24	41.39	60.78	80.64	37.44	49.10
K = 2000	58.15	80.56	37.04	43.79	63.86	80.59	38.97	55.20
Truncated Q Vocab @ 5	57.99	80.67	36.99	43.38	62.87	80.71	38.22	53.20
Truncated Q Vocab @ 11	57.81	80.42	36.97	43.22	62.72	80.45	38.30	53.09
Filtered Dataset	56.62	80.19	37.48	40.95	60.82	80.19	37.48	49.57

TABLE 4: Accuracy of ablated versions of our best model (deeper LSTM Q + norm I) for the open-ended and multiple-choice tasks on the VQA test-dev for real images. Q = Question, I = Image. See text for details.

	Open-Ended				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
snubi-naverlabs	60.60	82.23	38.22	46.99	64.95	82.25	39.56	55.68
MM_PaloAlto	60.36	80.43	36.82	48.33	–	–	–	–
LV-NUS	59.54	81.34	35.67	46.10	64.18	81.25	38.30	55.20
ACVT_Adelaide	59.44	81.07	37.12	45.83	–	–	–	–
global_vision	58.43	78.24	36.27	46.32	–	–	–	–
deeper LSTM Q + norm I	58.16	80.56	36.53	43.73	63.09	80.59	37.70	53.64
iBOWIMG	–	–	–	–	61.97	76.86	37.30	54.60

TABLE 5: Test-standard accuracy of our best model (deeper LSTM Q + norm I) compared to test-standard accuracies of other entries for the open-ended and multiple-choice tasks in the respective VQA Real Image Challenge leaderboards (as of October 28, 2016).

compared to 369861 questions in the original dataset, thus leading to only 6.83% reduction in the size of the dataset. The filtered dataset has 8.77 answers per question on average. We did not filter the test set so that accuracies of the model trained on the filtered dataset can be compared with that of the model trained on the original dataset. The row "Filtered Dataset" in Table 4 shows the performance of the deeper LSTM Q + norm I model when trained on the filtered dataset. Comparing these accuracies with the corresponding accuracies in Table 2, we can see that the model trained on filtered version performs worse by 1.13% for open-ended task and by 1.88% for multiple-choice task.

6 VQA Challenge and Workshop

We have set up an evaluation server[3] where results may be uploaded for the test set and it returns an accuracy breakdown. We are organizing an annual challenge and workshop to facilitate systematic progress in this area; the first instance of the workshop will be held at CVPR 2016[4]. We suggest that papers reporting results on the VQA dataset –

1) Report test-standard accuracies, which can be calculated using either of the non-test-dev phases, i.e., "test2015" or "Challenge test2015" on the following links: [oe-real | oe-abstract | mc-real | mc-abstract].
2) Compare their test-standard accuracies with those on the corresponding test2015 leaderboards [oe-real-leaderboard | oe-abstract-leaderboard | mc-real-leaderboard | mc-abstract-leaderboard].

For more details, please see the challenge page[5]. Screenshots of leaderboards for open-ended-real and multiple-choice-real are shown in Fig. 13. We also compare the test-standard accuracies of our best model (deeper LSTM Q + norm I) for both open-ended and multiple-choice tasks (real images) with other entries (as of October 28, 2016) on the corresponding leaderboards in Table 5.

7 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 containing over 250K images, 760K questions, and around 10M answers. 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 [3], or the questions could focused on one specific domain (say sports). Bigham *et al.* [3] 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.

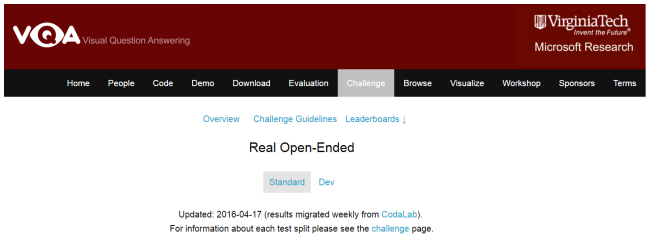
Acknowledgements. We would like to acknowledge the countless hours of effort provided by the workers on Amazon Mechanical Turk. This work was supported in part by the The Paul G. Allen Family Foundation via an award to D.P., ICTAS at Virginia Tech via awards to D.B. and D.P., Google Faculty Research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

Appendix Overview

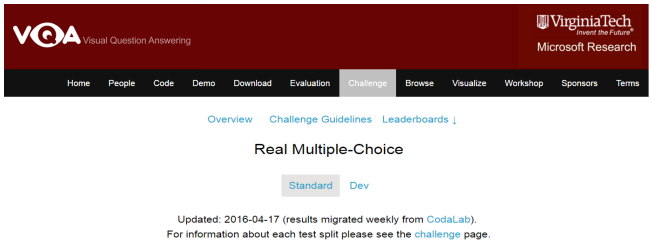
In the appendix, we provide:

I - Additional analysis comparing captions and Q&A data
II - Qualitative visualizations for "What is" questions
III - Human accuracy on multiple-choice questions
IV - Details on VQA baselines
V - "Age" and "Commonsense" of our model
VI - Details on the abstract scene dataset
VII - User interfaces used to collect the dataset
VIII - List of the top answers in the dataset

3. http://visualqa.org/challenge.html
4. http://www.visualqa.org/workshop.html
5. http://visualqa.org/challenge.html



	By Answer Type			Overall
	Yes/No	Number	Other	
snubi-naverlabs[5]	82.23	38.22	46.99	60.6
MM_PaloAlto[3]	80.43	36.82	48.33	60.36
LV-NUS[2]	81.34	35.67	46.1	59.54
ACVT_Adelaide[1]	81.07	37.12	45.83	59.44
global_vision[4]	78.24	36.27	46.32	58.43
vqateam-deeperLSTM_NormlizeCNN[7]	80.56	36.53	43.73	58.16
vqateam-lstm_cnn[8]	79.01	35.55	36.8	54.06
vqateam-q_lstm_alone[11]	78.12	34.94	26.99	48.89
vqateam-nearest_neighbor[9]	71.73	24.31	22	42.73
vqateam-prior_per_qtype[10]	71.17	35.63	9.32	37.55
vqateam-all_yes[6]	70.53	0.43	1.26	29.72



	By Answer Type			Overall
	Yes/No	Number	Other	
snubi-naverlabs[3]	82.25	39.56	55.68	64.95
LV-NUS[1]	81.25	38.3	55.2	64.18
vqateam-deeperLSTM_NormlizeCNN[5]	80.59	37.7	53.64	63.09
iBOWIMG[2]	76.86	37.3	54.6	61.97
vqateam-lstm_cnn[6]	79.02	36.1	43.93	57.57
vqateam-q_lstm_alone[9]	78.12	35.86	39.44	55.01
vqateam-nearest_neighbor[7]	71.75	25.81	34.09	48.75
vqateam-prior_per_qtype[8]	71.15	35.7	13.1	39.38
vqateam-all_yes[4]	70.53	0.43	1.26	29.72

Fig. 13: Leaderboard showing test-standard accuracies for VQA Real Image Challenge (Open-Ended) on left and leaderboard showing test-standard accuracies for VQA Real Image Challenge (Multiple-Choice) on right (snapshot from October 28, 2016).

IX - Additional examples from the VQA dataset

APPENDIX I: CAPTIONS *vs.* QUESTIONS

Do questions and answers provide further information about the visual world beyond that captured by captions? One method for determining whether the information captured by questions & answers is different from the information captured by captions is to measure some of the differences in the word distributions from the two datasets. We cast this comparison in terms of nouns, verbs, and adjectives by extracting all words from the caption data (MS COCO captions for real images and captions collected by us for abstract scenes) using the Stanford part-of-speech (POS)[6] tagger [49]. We normalize the word frequencies from captions, questions, and answers per image, and compare captions *vs.* questions and answers combined. Using a Kolmogorov-Smirnov test to determine whether the underlying distributions of the two datasets differ, we find a significant difference for all three parts of speech ($p < .001$) for both real images and abstract scenes. This helps motivate the VQA task as a way to learn information about visual scenes; although both captions and questions & answers provide information about the visual world, they do it from different perspectives, with different underlying biases [20], and can function as complementary to one another.

We illustrate the similarities and differences between the word distributions in captions vs. questions & answers as Venn-style word clouds [10] with size indicating the normalized count – Fig. 15 (nouns), Fig. 16 (verbs), and Fig. 17 (adjectives) for real images and Fig. 18 (nouns), Fig. 19 (verbs), and Fig. 20 (adjectives) for abstract scenes.[7] The left side shows the top words in questions & answers, the right the top words in captions, and the center the words common to both, with size indicating the harmonic mean of the counts.

We see that adjectives in captions capture some clearly visual properties discussed in previous work on vision to language [41], such as material and pattern, while the questions & answers have more adjectives that capture what is usual (*e.g.*, "dominant", "approximate", "higher") and other kinds of commonsense properties (*e.g.*, "edible", "possible", "unsafe", "acceptable"). Interestingly, we see that question & answer nouns capture information about "ethnicity" and "hairstyle", while caption nouns capture information about pluralized visible objects (*e.g.*, "cellphones", "daughters") and groups (*e.g.*, "trio", "some"), among other differences. "Man" and "people" are common in both captions and questions & answers.

One key piece to understanding the visual world is understanding spatial relationships, and so we additionally extract spatial prepositions and plot their proportions in the captions *vs.* the questions & answers data in Fig. 14 (left) for real images and Fig. 14 (right) for abstract scenes. We see that questions &

6. Noun tags begin with NN, verb tags begin with VB, adjective tags begin with JJ, and prepositions are tagged as IN.

7. Visualization created using http://worditout.com/.

answers have a higher proportion of specific spatial relations (*i.e.*, “in”, “on”) compared to captions, which have a higher proportion of general spatial relations (*i.e.*, “with”, “near”).

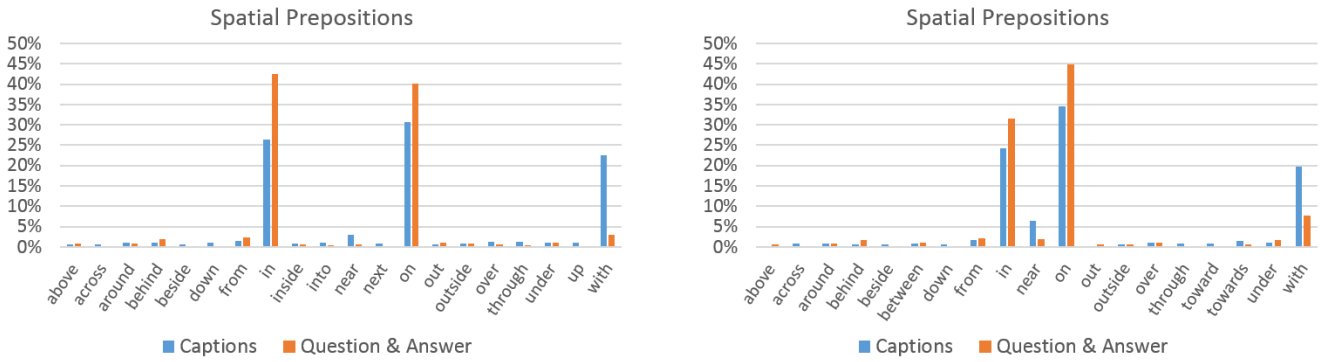


Fig. 14: Proportions of spatial prepositions in the captions and question & answers for real images (left) and abstract scenes (right).

Fig. 15: Venn-style word clouds [10] for nouns with size indicating the normalized count for real images.

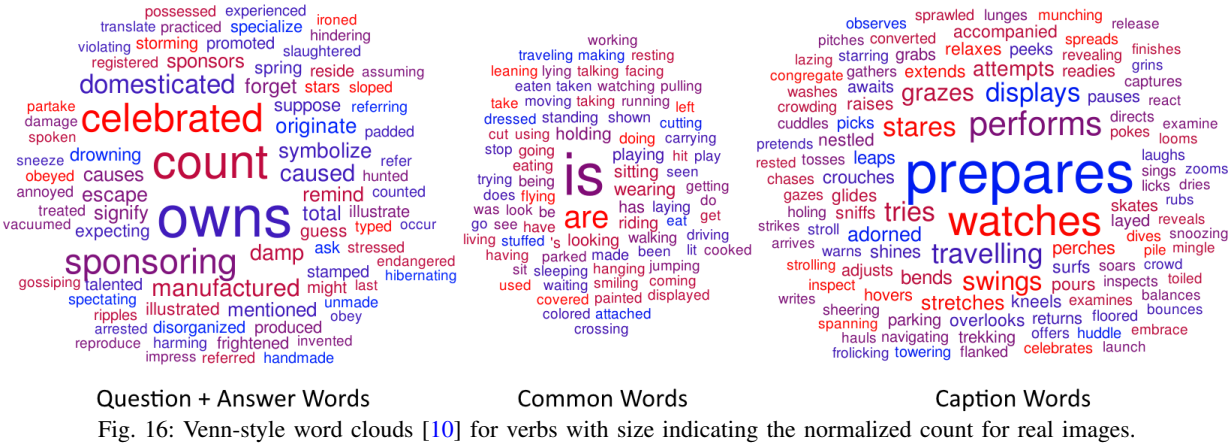


Fig. 16: Venn-style word clouds [10] for verbs with size indicating the normalized count for real images.

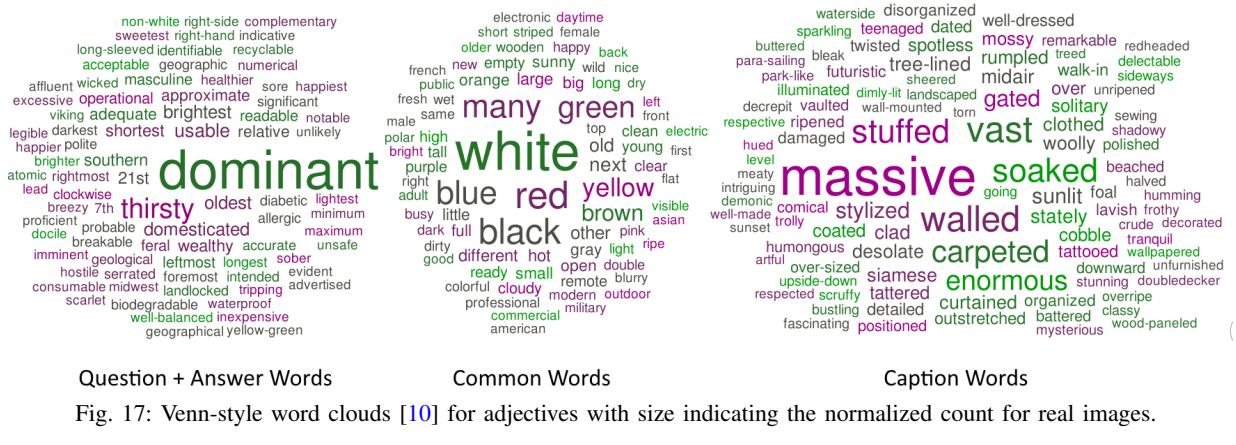


Fig. 17: Venn-style word clouds [10] for adjectives with size indicating the normalized count for real images.

Fig. 18: Venn-style word clouds [10] for nouns with size indicating the normalized count for abstract scenes.

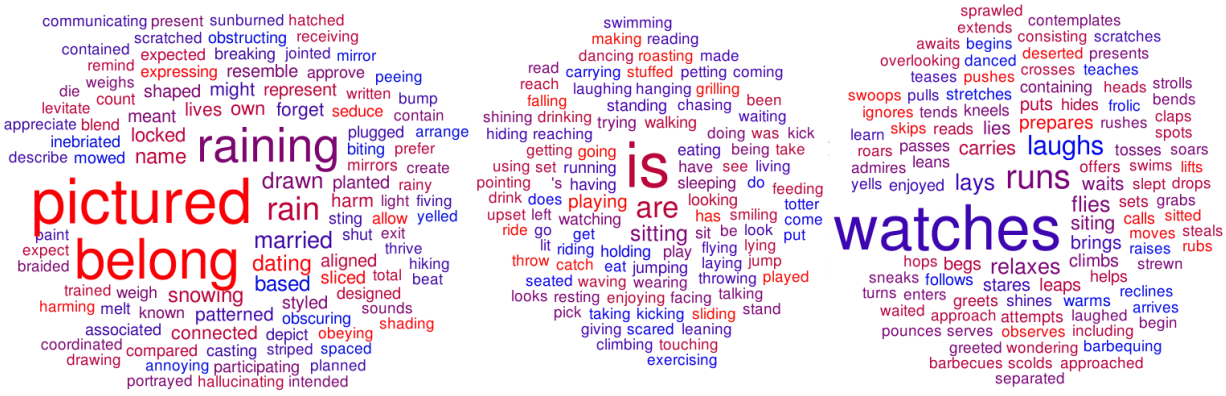


Fig. 19: Venn-style word clouds [10] for verbs with size indicating the normalized count for abstract scenes.

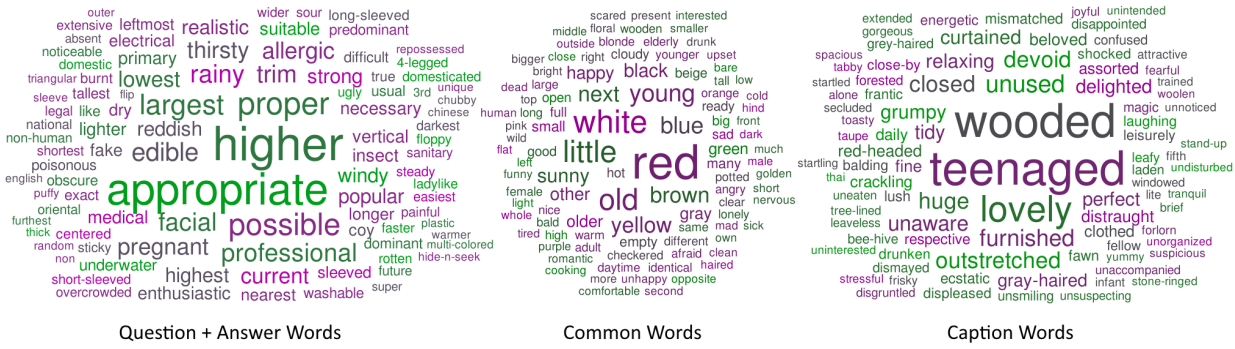


Fig. 20: Venn-style word clouds [10] for adjectives with size indicating the normalized count for abstract scenes.

Dataset	Accuracy Metric	All	Yes/No	Number	Other
Real	MC majority vote	91.54	97.40	86.97	87.91
	MC average	88.53	94.40	84.99	84.64
	Open-Ended	80.62	94.78	78.46	69.69
Abstract	MC majority vote	93.57	97.78	96.71	88.73
	MC average	90.40	94.59	94.36	85.32
	Open-Ended	85.66	95.32	94.17	74.12

TABLE 6: For each of the two datasets, real and abstract, first two rows are the human accuracies for multiple-choice questions when subjects were shown both the image and the question. Majority vote means we consider the answer picked by majority of the three subjects to be the predicted answer by humans and compute accuracy of that answer for each question. Average means we compute the accuracy of each of the answers picked by the subjects and record their average for each question. The last row is the inter-human agreement for open-ended answers task when subjects were shown both the image and the question. All accuracies are evaluated on a random subset of 3000 questions.

Appendix II: "What is" Analysis

In Fig. 21, we show the distribution of questions starting with "What is" by their first five words for both real images and abstract scenes. Note the diversity of objects referenced in the questions, as well as, the relations between objects, such as "holding" and "sitting on". In Fig. 22, we show the distribution of answers for "What is" questions ending in different words. For instance, questions ending in "eating" have answers such as "pizza", "watermelon" and "hot dog". Notice the diversity in answers for some questions, such as those that end with "for?" or "picture?". Other questions result in intuitive responses, such as "holding?" and the response "umbrella".

Appendix III: Multiple-Choice Human Accuracy

To compute human accuracy for multiple-choice questions, we collected three human answers per question on a random subset of 3,000 questions for both real images and abstract scenes. In Table 6, we show the human accuracies for multiple choice questions. Table 6 also shows the inter-human agreement for open-ended answer task. In comparison to open-ended answer, the multiple-choice accuracies are more or less same for "yes/no" questions and significantly better ($\approx 15\%$ increase for real images and $\approx 11\%$ increase for abstract scenes) for "other" questions. Since "other" questions may be ambiguous, the increase in accuracy using multiple choice is not surprising.

Appendix IV: Details on VQA baselines

"per Q-type prior" baseline. We decide on different question types based on first few words of questions in the real images training set and ensure that each question type has at least 30 questions in the training dataset. The most popular answer for each question type is also computed on real images training set.

"nearest neighbor" baseline. For every question in the VQA test-standard set, we find its k nearest neighbor questions in the training set using cosine similarity in Skip-Thought [27] feature space. We also experimented with bag of words and Word2Vec [39] feature spaces but we obtained the best performance with Skip-Thought. In this set of k questions and their associated images, we find the image which is most similar to the query image using cosine similarity in fc7 feature space. We use the fc7 features from the caffenet model in BVLC Caffe [23]. The most common ground truth answer of this most similar image and question pair is the predicted answer for the query image and question pair. We pick $k = 4$ on the test-dev set.

Appendix V: "Age" and "Commonsense" of our model

We estimate the age and degree of commonsense of our **best model** (deeper LSTM Q + norm I), selected using VQA test-dev accuracies). To estimate the age, we compute a weighted average of the average age per question, weighted by the accuracy of the model's predicted answer for that question, on the subset of questions in the VQA validation set for which we have age annotations (how old a human needs to be to answer the question correctly). To estimate the degree of commonsense, we compute a weighted average of the average degree of commonsense per question, weighted by the accuracy of the model's predicted answer for that question, on the subset of questions in the VQA validation set for which we have commonsense annotations (whether the question requires commonsense to answer it).

Appendix VI: Abstract Scenes Dataset

In Fig. 23 (left), we show a subset of the objects that are present in the abstract scenes dataset. For more examples of the scenes generated, please see Fig. 28. The user interface used to create the scenes is shown in Fig. 23 (right). Subjects used a drag-and-drop interface to create the scenes. Each object could be flipped horizontally and scaled. The scale of the object determined the rendering order of the objects. Many objects have different attributes corresponding to different poses or types. Most animals have five different discrete poses. Humans have eight discrete expressions and their poses may be continuously adjusted using a "paperdoll" model [2].

Appendix VII: User Interfaces

In Fig. 24, we show the AMT interface that we used to collect questions for images. Note that we tell the workers that the robot already knows the answer to the previously asked question(s), inspiring them to ask different kinds of questions, thereby increasing the diversity of our dataset.

Fig. 25 shows the AMT interface used for collecting answers to the previously collected questions when subjects were shown the corresponding images. Fig. 26 shows the interface that was used to collect answers to questions when subjects were not shown the corresponding image (*i.e.*, to help in gathering incorrect, but plausible, answers for the multiple-choice task and to assess how accurately the questions can be answered using common sense knowledge alone).

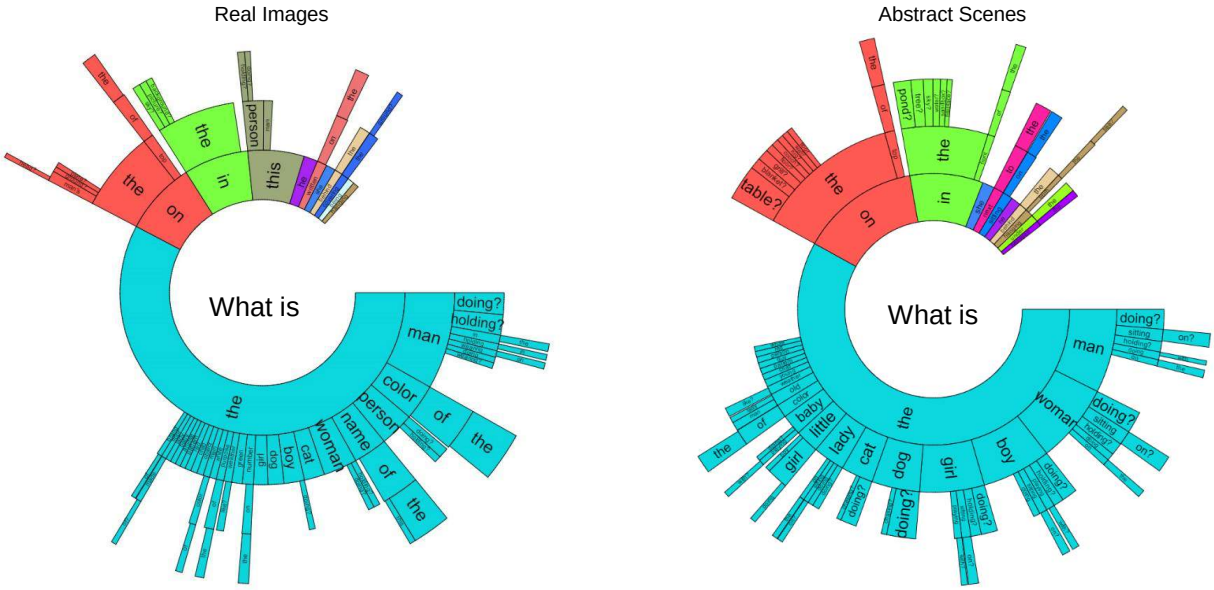


Fig. 21: Distribution of questions starting with "What is" by their first five words for a random sample of 60K questions for real images (left) and all questions for abstract scenes (right). The ordering of the words starts towards the center and radiates outwards. The arc length is proportional to the number of questions containing the word. White areas are words with contributions too small to show.

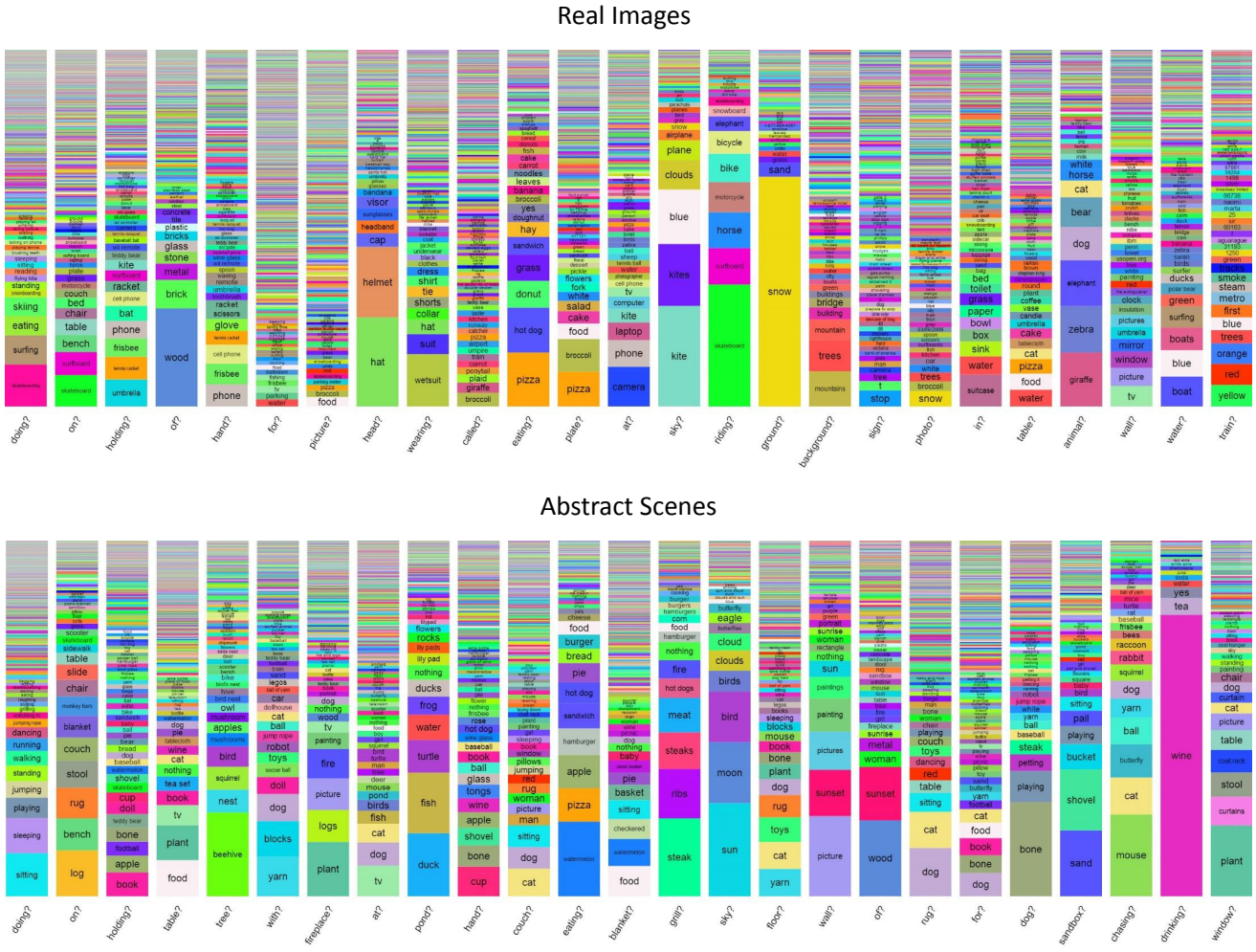


Fig. 22: Distribution of answers for questions starting with "What is" for a random sample of 60K questions for real images (top) and all questions for abstract scenes (bottom). Each column corresponds to questions ending in different words, such as "doing?", "on?", *etc*.

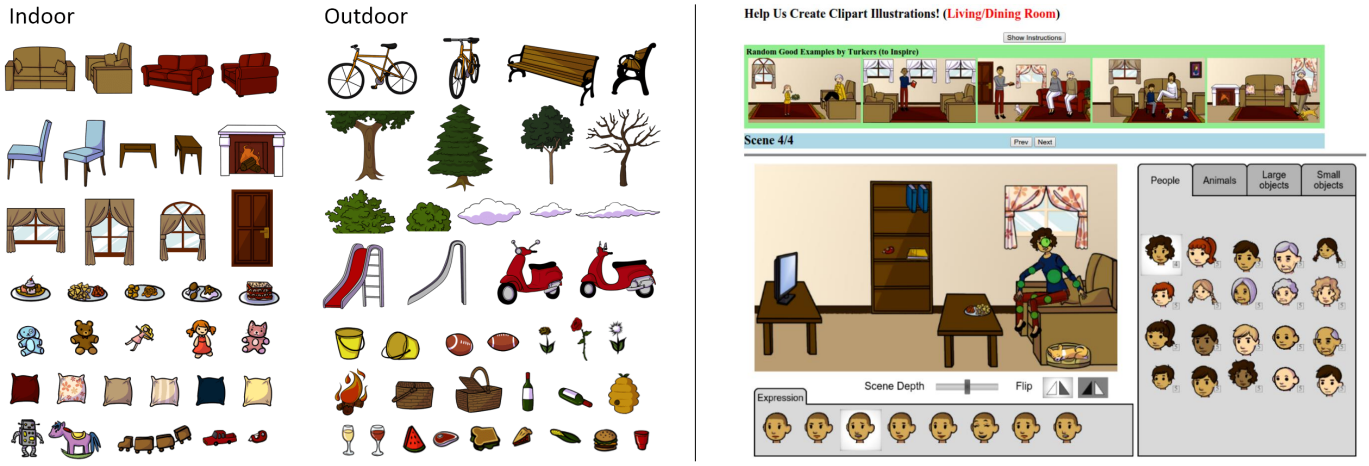


Fig. 23: Left: A small subset of the objects present in the abstract scene dataset. Right: The AMT interface for collecting abstract scenes. The light green circles indicate where users can select to manipulate a person's pose. Different objects may be added to the scene using the folders to the right.

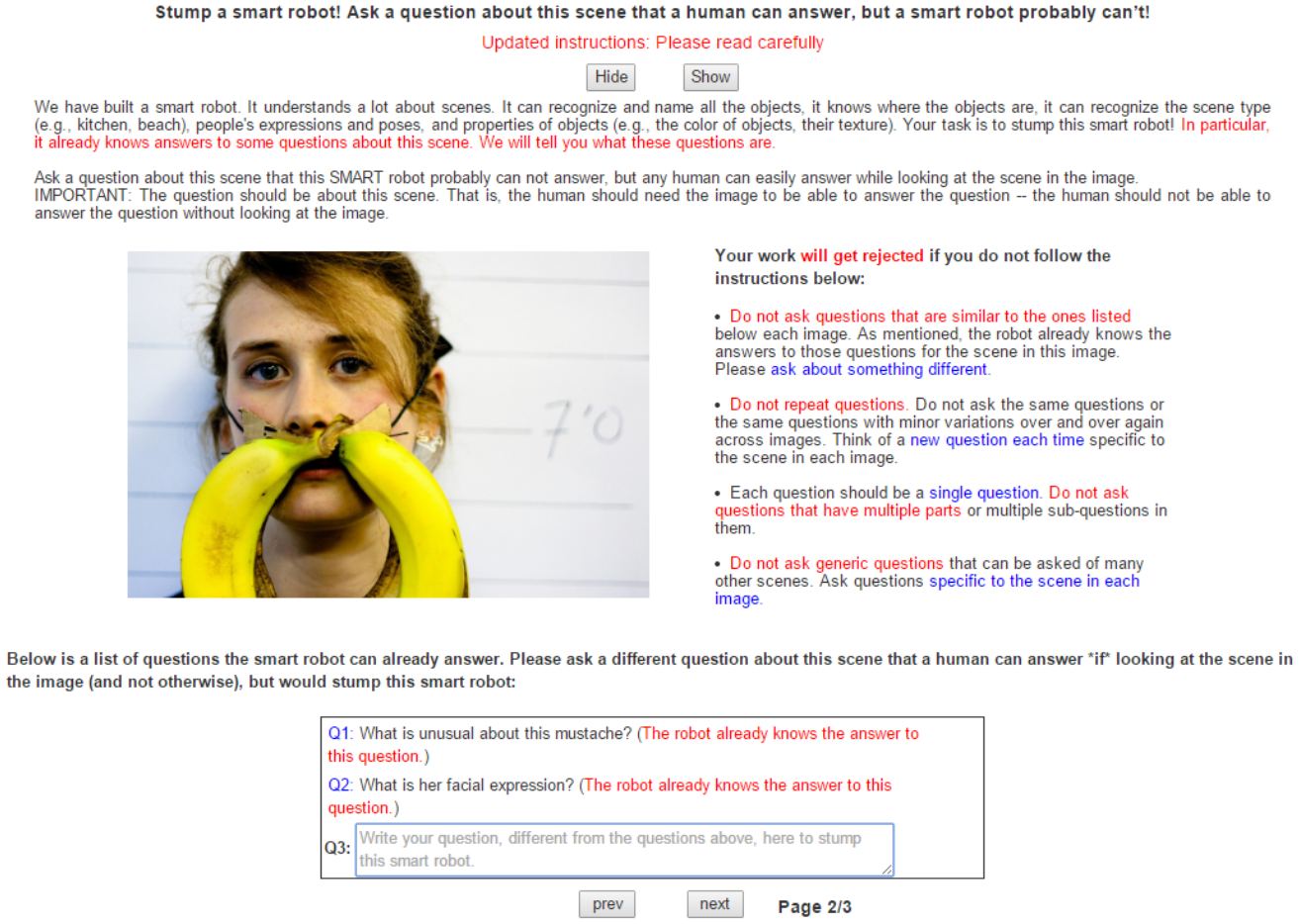


Fig. 24: Our AMT interface for collecting the third question for an image, when subjects were shown previous questions that were collected and were asked to ask a question different from previous questions.

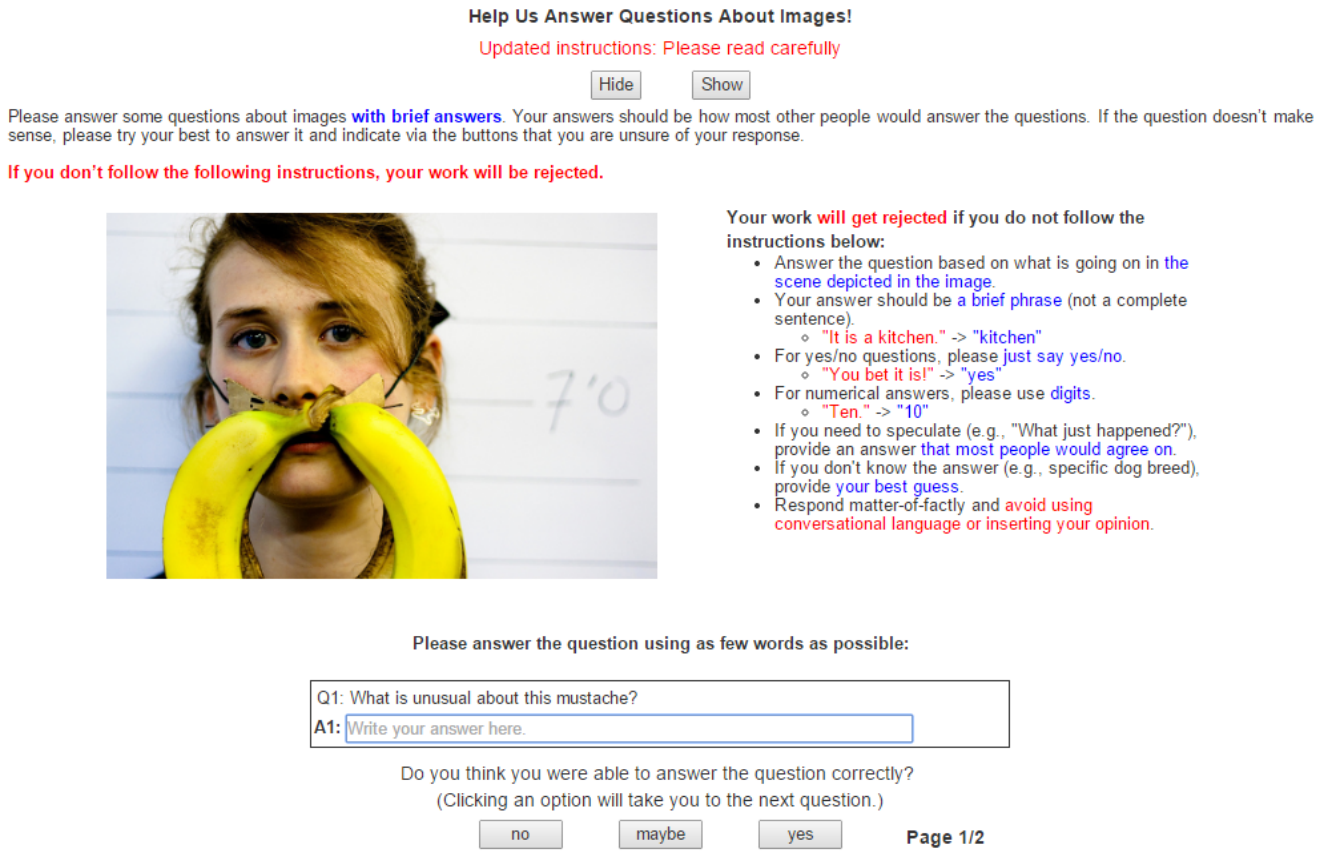


Fig. 25: The AMT interface used to collect answers to a question when subjects were shown the image while answering the question.

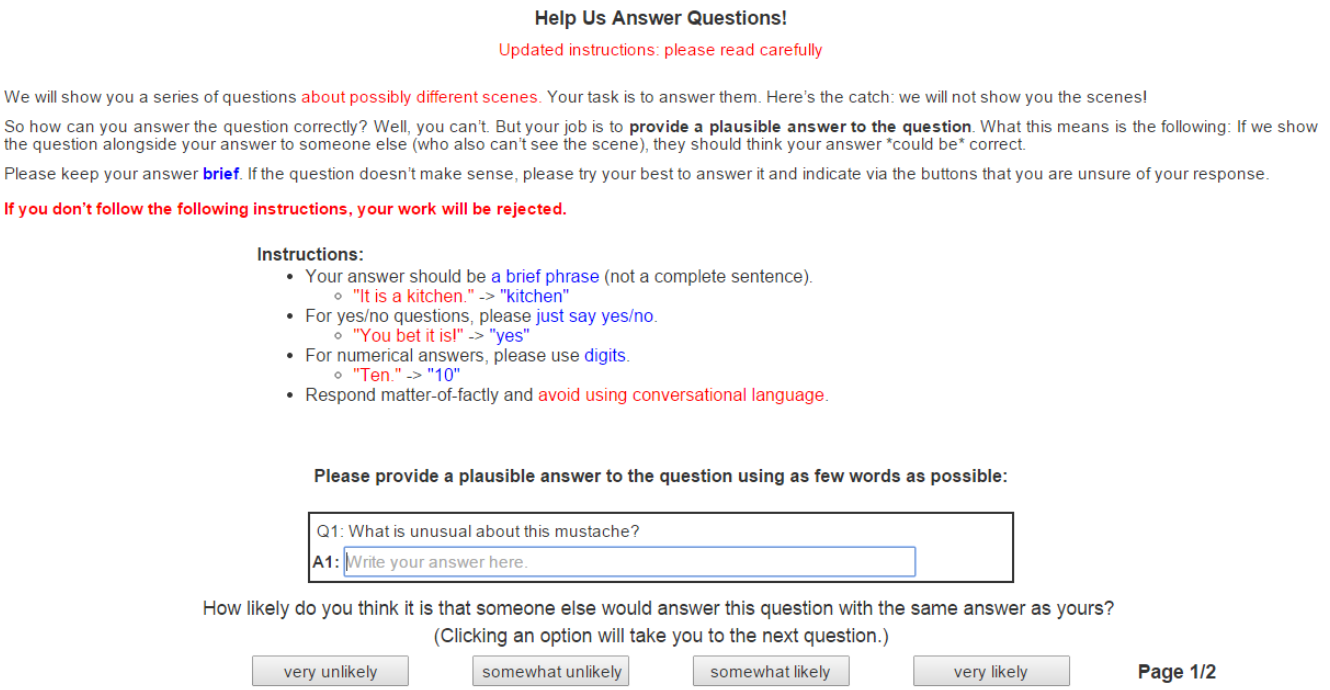


Fig. 26: The AMT interface used to collect answers to a question when subjects were not shown the image while answering the question using only commonsense to collect the plausible, but incorrect, multiple-choice answers.

APPENDIX VIII: ANSWER DISTRIBUTION

The top 250 answers in our real images dataset along with their counts and percentage counts are given below. The answers have been presented in different colors to show the different Part-of-Speech (POS) tagging of the answers with the following color code: yes/no, noun, verb, adjective, adverb, and numeral.

"yes" (566613, 22.82%), "no" (381307, 15.35%), "2" (80031, 3.22%), "1" (46537, 1.87%), "white" (41753, 1.68%), "3" (41334, 1.66%), "red" (33834, 1.36%), "blue" (28881, 1.16%), "4" (27174, 1.09%), "green" (22453, 0.9%), "black" (21852, 0.88%), "yellow" (17312, 0.7%), "brown" (14488, 0.58%), "5" (14373, 0.58%), "tennis" (10941, 0.44%),"baseball" (10299, 0.41%), "6" (10103, 0.41%), "orange" (9136, 0.37%), "0" (8812, 0.35%), "bathroom" (8473, 0.34%), "wood" (8219, 0.33%), "right" (8209, 0.33%), "left" (8058, 0.32%), "frisbee" (7671, 0.31%), "pink" (7519, 0.3%), "gray" (7385, 0.3%), "pizza" (6892, 0.28%), "7" (6005, 0.24%), "kitchen" (5926, 0.24%), "8" (5592, 0.23%), "cat" (5514, 0.22%), "skiing" (5189, 0.21%), "skateboarding" (5122, 0.21%), "dog" (5092, 0.21%), "snow" (4867, 0.2%), "black and white" (4852, 0.2%), "skateboard" (4697, 0.19%), "surfing" (4544, 0.18%), "water" (4513, 0.18%), "giraffe" (4027, 0.16%), "grass" (3979, 0.16%), "surfboard" (3934, 0.16%), "wii" (3898, 0.16%), "kite" (3852, 0.16%), "10" (3756, 0.15%), "purple" (3722, 0.15%), "elephant" (3646, 0.15%), "broccoli" (3604, 0.15%), "man" (3590, 0.14%), "winter" (3490, 0.14%), "stop" (3413, 0.14%), "train" (3226, 0.13%), "9" (3217, 0.13%), "apple" (3189, 0.13%), "silver" (3186, 0.13%), "horse" (3159, 0.13%), "banana" (3151, 0.13%), "umbrella" (3139, 0.13%), "eating" (3117, 0.13%), "sheep" (2927, 0.12%), "bear" (2803, 0.11%), "phone" (2772, 0.11%), "12" (2633, 0.11%), "motorcycle" (2608, 0.11%), "cake" (2602, 0.1%), "wine" (2574, 0.1%), "beach" (2536, 0.1%), "soccer" (2504, 0.1%), "sunny" (2475, 0.1%), "zebra" (2403, 0.1%), "tan" (2402, 0.1%), "brick" (2395, 0.1%), "female" (2372, 0.1%), "bananas" (2350, 0.09%), "table" (2331, 0.09%), "laptop" (2316, 0.09%), "hat" (2277, 0.09%), "bench" (2259, 0.09%), "flowers" (2219, 0.09%), "woman" (2197, 0.09%), "male" (2170, 0.09%), "cow" (2084, 0.08%), "food" (2083, 0.08%), "living room" (2022, 0.08%), "bus" (2011, 0.08%), "snowboarding" (1990, 0.08%), "kites" (1979, 0.08%), "cell phone" (1943, 0.08%), "helmet" (1885, 0.08%), "maybe" (1853, 0.07%), "outside" (1846, 0.07%), "hot dog" (1809, 0.07%), "night" (1805, 0.07%), "trees" (1785, 0.07%), "11" (1753, 0.07%), "bird" (1739, 0.07%), "down" (1732, 0.07%), "bed" (1587, 0.06%), "camera" (1560, 0.06%), "tree" (1547, 0.06%), "christmas" (1544, 0.06%), "fence" (1543, 0.06%), "nothing" (1538, 0.06%), "unknown" (1532, 0.06%), "tennis racket" (1525, 0.06%), "red and white" (1518, 0.06%), "bedroom" (1500, 0.06%), "bat" (1494, 0.06%), "glasses" (1491, 0.06%), "tile" (1487, 0.06%), "metal" (1470, 0.06%), "blue and white" (1440, 0.06%), "fork" (1439, 0.06%), "plane" (1439, 0.06%), "airport" (1422, 0.06%), "cloudy" (1413, 0.06%), "15" (1407, 0.06%), "up" (1399, 0.06%), "blonde" (1398, 0.06%), "day" (1396, 0.06%), "teddy bear" (1386, 0.06%), "glass" (1379, 0.06%), "20" (1365, 0.05%), "beer" (1345, 0.05%), "car" (1331, 0.05%), "sitting" (1328, 0.05%), "boat" (1326, 0.05%), "standing" (1326, 0.05%), "clear" (1318, 0.05%), "13" (1318, 0.05%), "nike" (1293, 0.05%), "sand" (1282, 0.05%), "open" (1279, 0.05%), "cows" (1271, 0.05%), "bike" (1267, 0.05%), "chocolate" (1266, 0.05%), "donut" (1263, 0.05%), "airplane" (1247, 0.05%), "birthday" (1241, 0.05%), "carrots" (1239, 0.05%), "skis" (1220, 0.05%), "girl" (1220, 0.05%), "many" (1211, 0.05%), "zoo" (1204, 0.05%), "suitcase" (1199, 0.05%), "old" (1180, 0.05%), "chair" (1174, 0.05%), "beige" (1170, 0.05%), "ball" (1169, 0.05%), "ocean" (1168, 0.05%), "sandwich" (1168, 0.05%), "tie" (1166, 0.05%), "horses" (1163, 0.05%), "palm" (1163, 0.05%), "stripes" (1155, 0.05%), "fall" (1146, 0.05%), "cheese" (1142, 0.05%), "scissors" (1134, 0.05%), "round" (1125, 0.05%), "chinese" (1123, 0.05%), "knife" (1120, 0.05%), "14" (1110, 0.04%), "toilet" (1099, 0.04%), "don't know" (1085, 0.04%), "snowboard" (1083, 0.04%), "truck" (1076, 0.04%), "boy" (1070, 0.04%), "coffee" (1070, 0.04%), "cold" (1064, 0.04%), "fruit" (1064, 0.04%), "walking" (1053, 0.04%), "wedding" (1051, 0.04%), "lot" (1050, 0.04%), "sunglasses" (1047, 0.04%), "mountains" (1030, 0.04%), "wall" (1009, 0.04%), "elephants" (1006, 0.04%), "wetsuit" (998, 0.04%), "square" (994, 0.04%), "toothbrush" (989, 0.04%), "sleeping" (986, 0.04%), "fire hydrant" (977, 0.04%), "bicycle" (973, 0.04%), "overcast" (968, 0.04%), "donuts" (961, 0.04%), "plastic" (961, 0.04%), "breakfast" (955, 0.04%), "tv" (953, 0.04%), "paper" (952, 0.04%), "ground" (949, 0.04%), "asian" (938, 0.04%), "plaid" (936, 0.04%), "dirt" (933, 0.04%), "mirror" (928, 0.04%), "usa" (928, 0.04%), "chicken" (925, 0.04%), "plate" (920, 0.04%), "clock" (912, 0.04%), "luggage" (908, 0.04%), "none" (908, 0.04%), "street" (905, 0.04%), "on table" (904, 0.04%), "spoon" (899, 0.04%), "cooking" (898, 0.04%), "daytime" (896, 0.04%), "16" (893, 0.04%), "africa" (890, 0.04%), "stone" (884, 0.04%), "not sure" (873, 0.04%), "window" (868, 0.03%), "sun" (865, 0.03%), "gold" (860, 0.03%), "people" (856, 0.03%), "racket" (847, 0.03%), "zebras" (845, 0.03%), "carrot" (841, 0.03%), "person" (835, 0.03%), "fish" (835, 0.03%), "happy" (824, 0.03%), "circle" (822, 0.03%), "oranges" (817, 0.03%), "backpack" (812, 0.03%), "25" (810, 0.03%), "leaves" (809, 0.03%), "watch" (804, 0.03%), "mountain" (800, 0.03%), "no one" (798, 0.03%), "ski poles" (792, 0.03%), "city" (791, 0.03%), "couch" (790, 0.03%), "afternoon" (782, 0.03%), "jeans" (781, 0.03%), "brown and white" (779, 0.03%), "summer" (774, 0.03%), "giraffes" (772, 0.03%), "computer" (771, 0.03%), "refrigerator" (768, 0.03%), "birds" (762, 0.03%), "child" (761, 0.03%), "park" (759, 0.03%), "flying kite" (756, 0.03%), "restaurant" (747, 0.03%), "evening" (738, 0.03%), "graffiti" (736, 0.03%), "30" (730, 0.03%), "grazing" (727, 0.03%), "flower" (723, 0.03%), "remote" (720, 0.03%), "hay" (719, 0.03%), "50" (716, 0.03%).

APPENDIX IX: ADDITIONAL EXAMPLES

To provide insight into the dataset, we provide additional examples. In Fig. 27, Fig. 28, and Fig. 29, we show a random selection of the VQA dataset for the MS COCO [32] images, abstract scenes, and multiple-choice questions, respectively.

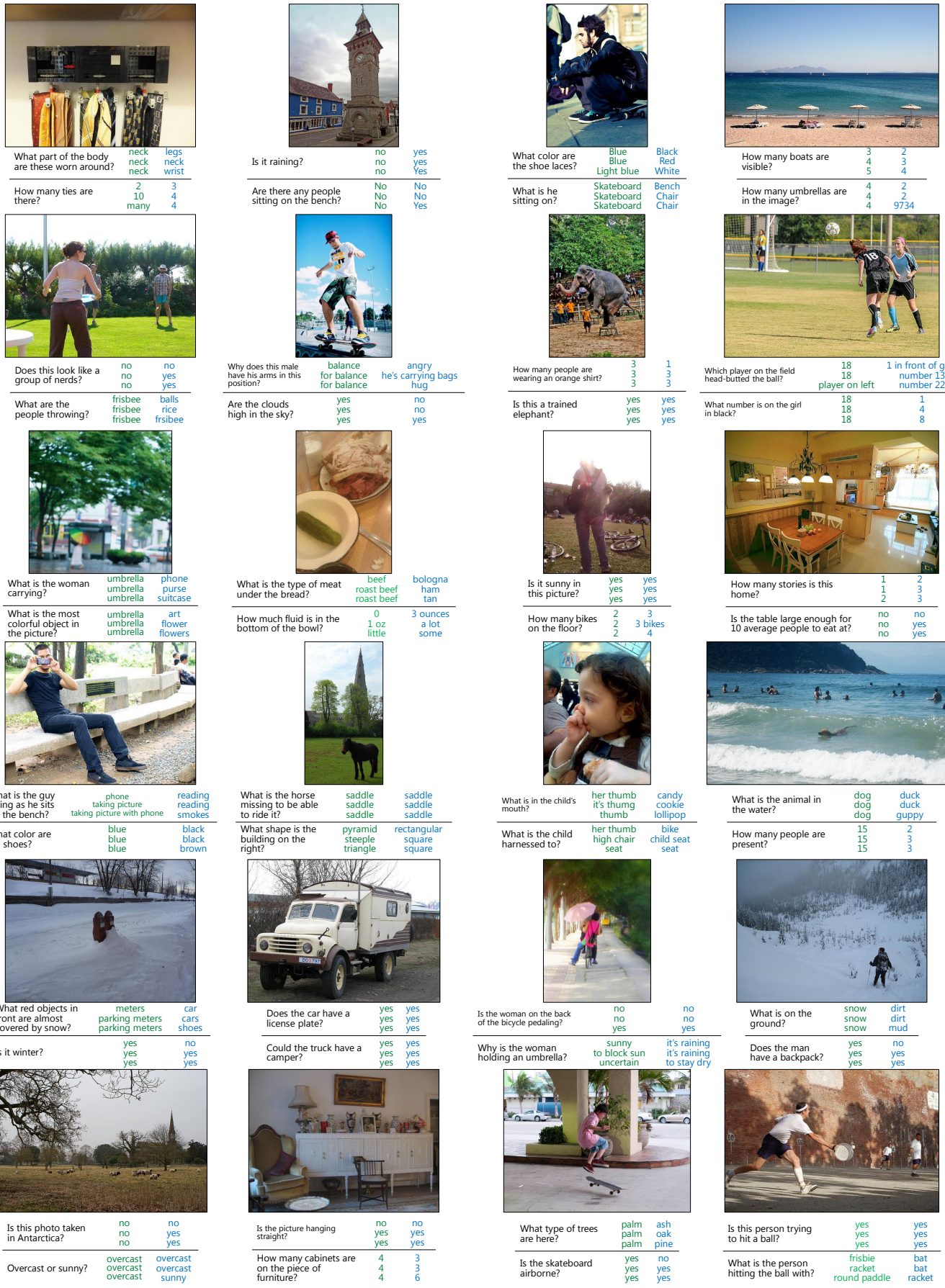


Fig. 27: Random examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the real image dataset.

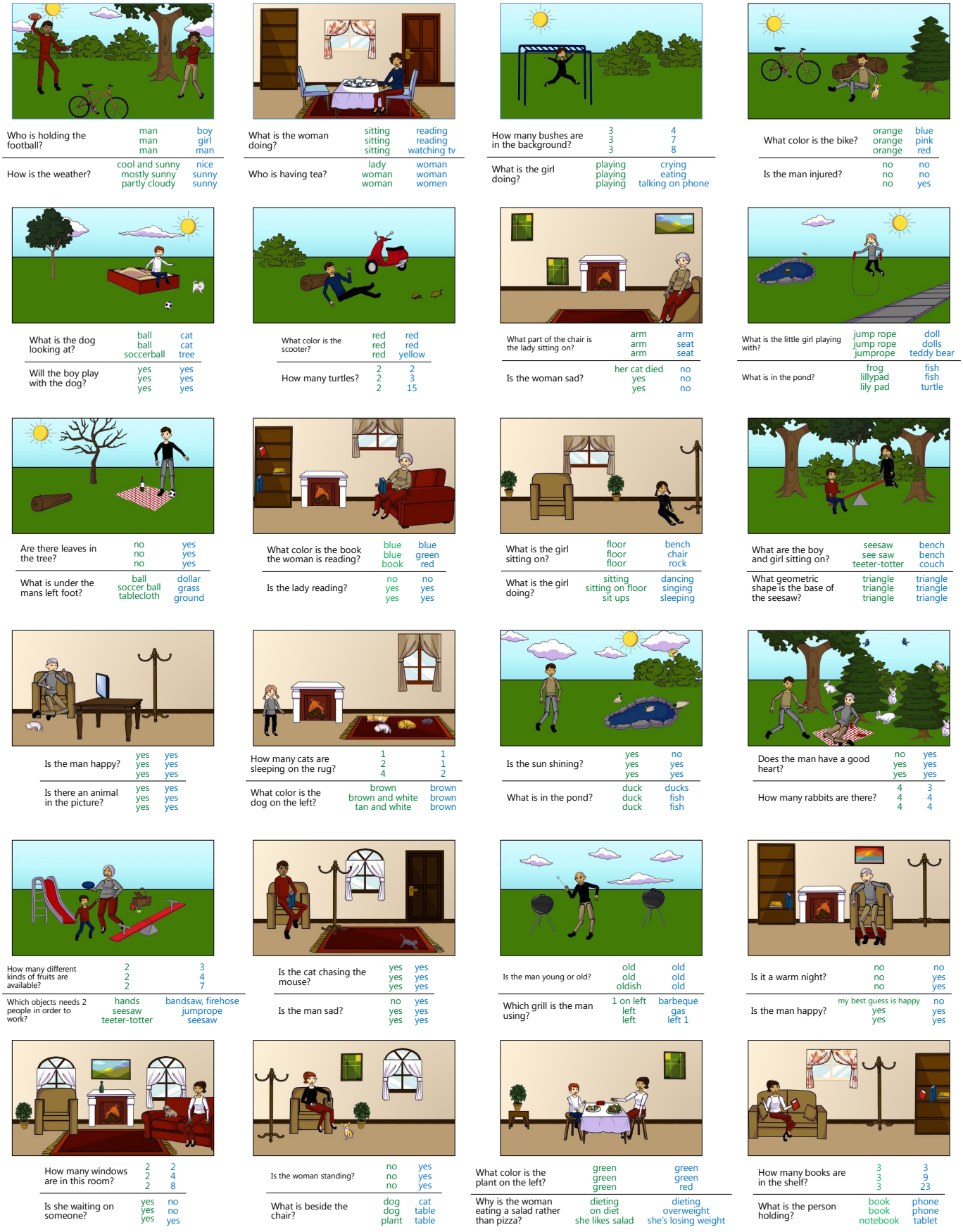


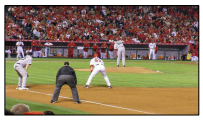
Fig. 28: Random examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the abstract scene dataset.

Q: Where is the kid pointing?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) park	**(l)** up	**(m)** floor mat	**(n)** so people don't get wet
(o) down	**(p)** mom	**(q)** pharos	**(r)** ketchup pickle relish mustard

Q: How many people are in the picture on side of refrigerator?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) 108 mph	**(l)** banana, apple	**(m)** 7	**(n)** 10 many
(o) fruit salad	**(p)** full swing	**(q)** 5	**(r)** vattenfall strom fur gewinner

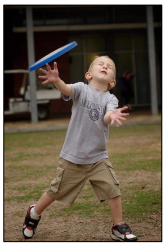


Q: What sport are they playing?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) tennis	**(l)** bodily functions	**(m)** scissors	**(n)** mississippi and meade
(o) baseball	**(p)** frisbee	**(q)** soccer	**(r)** its advertising object

Q: What is the man in gray pant's job?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) cop	**(l)** umpire	**(m)** snowflake	**(n)** banker
(o) chef	**(p)** speedboat	**(q)** 10: 32	**(r)** males



Q: What is the color of freebee?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) brick	**(l)** peach	**(m)** hill	**(n)** vitamin c
(o) brown	**(p)** christleton	**(q)** bonsai tree	**(r)** black

Q: How old is the child?

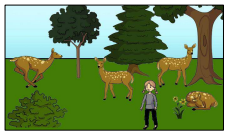
(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) 6	**(l)** 12	**(m)** 10	**(n)** mechanics
(o) 5	**(p)** wait here	**(q)** mad	**(r)** recording studio

Q: Is this person's face painted?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) 4498	**(l)** not	**(m)** camera film	**(n)** keyboard, mouse, booklet
(o) stairs	**(p)** n200	**(q)** public storage	**(r)** pasta, sauce, meat

Q: How many umbrellas are in the photo?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** green
(k) 20	**(l)** 54	**(m)** max payne	**(n)** 62
(o) 12	**(p)** dresses	**(q)** 3 to 5	**(r)** two way traffic



Q: How many of the deer are sleeping?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) 5	**(l)** left of pond	**(m)** 13	**(n)** plants and cat
(o) tree base	**(p)** cement	**(q)** 0	**(r)** green, blue and yellow

Q: What type of wildlife is this park overrun with?

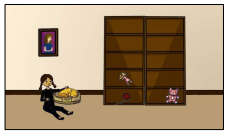
(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) eating	**(l)** deer	**(m)** mosquitoes	**(n)** soup
(o) birds	**(p)** ants	**(q)** girl's	**(r)** woman on right

Q: Where is the blanket?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) fat	**(l)** lying down	**(m)** bed	**(n)** utensils
(o) on bed	**(p)** grass	**(q)** ground	**(r)** watching child

Q: What is for dessert?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) cake	**(l)** pie	**(m)** a	**(n)** doll and dollhouse
(o) ice cream	**(p)** yellow book	**(q)** cheesecake	**(r)** there are no fish

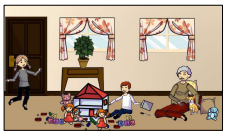


Q: Is the girl standing?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) yes!	**(l)** standing	**(m)** hiding	**(n)** sitting
(o) to sleep	**(p)** bird nest	**(q)** slide	**(r)** park ranger

Q: Does the girl have a lot of toys?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) fork	**(l)** deer	**(m)** rock	**(n)** y
(o) slide	**(p)** yes 3 of them	**(q)** no image	**(r)** children and toys



Q: Why does the little girl not look happy?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) indian	**(l)** upset	**(m)** dog left	**(n)** smiling at it
(o) corner	**(p)** to be pet	**(q)** she fell	**(r)** boy is playing with her toys

Q: Why is the boy playing with his sister's toys?

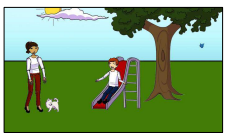
(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) he likes them	**(l)** parking it	**(m)** dogs	**(n)** shelf
(o) he feeds them	**(p)** lonely	**(q)** bored	**(r)** likes them

Q: Why are they standing?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) playing game	**(l)** sheepskin	**(m)** waiting	**(n)** no where to sit
(o) firestone	**(p)** rugby	**(q)** forks	**(r)** waiting for train

Q: Is the TV on?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) shag	**(l)** jeopardy	**(m)** sports	**(n)** between big elephants
(o) edinburgh	**(p)** strawberries	**(q)** tv show	**(r)** white streak on face



Q: How many legs does the dog have?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) outdoors	**(l)** hiding	**(m)** 45	**(n)** sitting in grass
(o) owls	**(p)** 8	**(q)** 12	**(r)** arm of sofa

Q: Is the boy at the top of the ladder?

(a) yes	**(b)** no		
(c) 1	**(d)** 2	**(e)** 3	**(f)** 4
(g) white	**(h)** red	**(i)** blue	**(j)** yellow
(k) not sure	**(l)** yellow dog	**(m)** bottom	**(n)** behind trees
(o) a	**(p)** girl on right	**(q)** top	**(r)** she's in middle

Fig. 29: Random examples of multiple-choice questions for numerous representative examples of the real and abstract scene dataset.

REFERENCES

[1] H. Agrawal, C. S. Mathialagan, Y. Goyal, N. Chavali, P. Banik, A. Mohapatra, A. Osman, and D. Batra. Cloudcv: Large-scale distributed computer vision as a cloud service. In *Mobile Cloud Visual Media Computing*, pages 265–290. Springer International Publishing, 2015. 10

[2] S. Antol, C. L. Zitnick, and D. Parikh. Zero-Shot Learning via Visual Abstraction. In *ECCV*, 2014. 2, 3, 17

[3] 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, 12

[4] 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

[5] 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

[6] X. Chen, H. Fang, T. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015. 4

[7] 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

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

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

[10] G. Coppersmith and E. Kelly. Dynamic wordclouds and vennclouds for exploratory data analysis. In *ACL Workshop on Interactive Language Learning and Visualization*, 2014. 13, 15, 16

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

[12] 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

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

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

[15] 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

[16] 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

[17] 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

[18] 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

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

[20] J. Gordon and B. V. Durme. Reporting bias and knowledge extraction. In *Proceedings of the 3rd Workshop on Knowledge Extraction, at CIKM 2013*, 2013. 13

[21] 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

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

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 17

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

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

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

[27] R. Kiros, Y. Zhu, R. Salakhutdinov, R. S. Zemel, A. Torralba, R. Urtasun, and S. Fidler. Skip-thought vectors. *arXiv preprint arXiv:1506.06726*, 2015. 17

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

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

[30] 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

[31] 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

[32] 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, 5, 21

[33] X. Lin and D. Parikh. Don't Just Listen, Use Your Imagination: Leveraging Visual Common Sense for Non-Visual Tasks. In *CVPR*, 2015. 1

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

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

[36] 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

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

[38] 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

[39] 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, 8, 17

[40] 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

[41] M. Mitchell, K. van Deemter, and E. Reiter. Attributes in visual reference. In *PRE-CogSci*, 2013. 13

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

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

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

[45] M. Richardson, C. J. Burges, and E. Renshaw. MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In *EMNLP*, 2013. 1, 2

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

[47] 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

[48] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 8, 9, 10

[49] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *ACL*, 2003. 13

[50] 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

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

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

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

[54] 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

[55] 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

[56] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh. Yin and yang: Balancing and answering binary visual questions. *CoRR*, abs/1511.05099, 2015. 10

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

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

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