Visual7W: Grounded Question Answering in Images

Yuke Zhu[†]   Oliver Groth[‡]   Michael Bernstein[†]   Li Fei-Fei[†]

[†]Computer Science Department, Stanford University

[‡]Computer Science Department, Dresden University of Technology

Abstract

We have seen great progress in basic perceptual tasks such as object recognition and detection. However, AI models still fail to match humans in high-level vision tasks due to the lack of capacities for deeper reasoning. Recently the new task of visual question answering (QA) has been proposed to evaluate a model's capacity for deep image understanding. Previous works have established a loose, global association between QA sentences and images. However, many questions and answers, in practice, relate to local regions in the images. We establish a semantic link between textual descriptions and image regions by object-level grounding. It enables a new type of QA with visual answers, in addition to textual answers used in previous work. We study the visual QA tasks in a grounded setting with a large collection of 7W multiple-choice QA pairs. Furthermore, we evaluate human performance and several baseline models on the QA tasks. Finally, we propose a novel LSTM model with spatial attention to tackle the 7W QA tasks.

1. Introduction

The recent development of deep learning technologies has achieved successes in many perceptual visual tasks such as object recognition, image classification and pose estimation [15, 21, 25, 38, 39, 42, 43]. Yet the status quo of computer vision is still far from matching human capabilities, especially when it comes to understanding an image in all its details. Recently, visual question answering (QA) has been proposed as a proxy task for evaluating a vision system's capacity for deeper image understanding. Several QA datasets [1, 7, 27, 36, 49] have been released since last year. They contributed valuable data for training visual QA systems and introduced various tasks, from picking correct multiple-choice answers [1] to filling in blanks [49].

Pioneer work in image captioning [4, 5, 13, 45, 47], sentence-based image retrieval [14, 40] and visual QA [1, 7, 36] shows promising results. These works aimed at establishing a global association between sentences and images. However, as Flickr30K [34, 48] and Visual Madlibs [49]
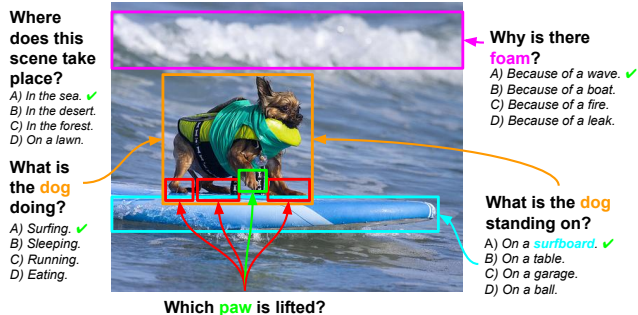
Figure 1: Deep image understanding relies on detailed knowledge about different image parts. We employ diverse questions to acquire detailed information on images, ground objects mentioned in text with their visual appearances, and provide a multiple-choice setting for evaluating a visual question answering task with both textual and visual answers.

suggest, a tighter semantic link between textual descriptions and corresponding visual regions is a key ingredient for better models. As Fig. 1 shows, the localization of objects can be a critical step to understand images better and solve image-related questions. Providing these image-text correspondences is called *grounding*. Inspired by Geman et al.'s prototype of a visual Turing test based on image regions [8] and the comprehensive data collection of QA pairs on COCO images [25] such as VQA [1] and Baidu [7], we fuse visual QA and grounding in order to create a new QA dataset with dense annotations and a more flexible evaluation environment. Object-level grounding provides a stronger link between QA pairs and images than global image-level associations. Furthermore, it allows us to resolve coreference ambiguity [19, 35] and to understand object distributions in QA, and enables visually grounded answers that consist of object bounding boxes.

Motivated by the goal of developing a model for visual QA based on grounded regions, our paper introduces a dataset that extends previous approaches [1, 7, 36] and proposes an attention-based model to perform this task. We collected 327,939 QA pairs on 47,300 COCO images [25], together with 1,311,756 human-generated multiple-choices and 561,459 object groundings from 36,579 categories. Our data collection was inspired by the age-old idea of the *W*

1

questions in journalism to describe a complete story [22]. The *7W* questions roughly correspond to an array of standard vision tasks: *what* [9, 15, 39], *where* [24, 50], *when* [30, 32], *who* [35, 42], *why* [33], *how* [23, 31] and *which* [16, 17]. The Visual7W dataset features richer questions and longer answers than VQA [1]. In addition, we provide complete grounding annotations that link the object mentions in the QA sentences to their bounding boxes in the images and therefore introduce a new QA type with image regions as the visually grounded answers. We refer to questions with textual answers as *telling* questions (*what*, *where*, *when*, *who*, *why* and *how*) and to such with visual answers as *pointing* questions (*which*). We provide a detailed comparison and data analysis in Sec. 4.

A salient property of our dataset is the notable gap between human performance (96.6%) and state-of-the-art LSTM models [28] (52.1%) on the visual QA tasks. We add a new spatial attention mechanism to an LSTM architecture for tackling the visually grounded QA tasks with both textual and visual answers (see Sec. 5). The model aims to capture the intuition that answers to image-related questions usually correspond with specific image regions. It learns to attend to the pertinent regions as it reads the question tokens in a sequence. We achieve state-of-the-art performance with 55.6%, and find correlations between the model's attention heat maps and the object groundings (see Sec. 6). Due to the large performance gap between human and machine, we envision our dataset and visually grounded QA tasks to contribute to a long-term joint effort from several communities such as vision, natural language processing and knowledge to close the gap together.

The Visual7W dataset constitutes a part of the Visual Genome project [20]. Visual Genome contains 1.7 million QA pairs of the 7W question types, which offers the largest visual QA collection to date for training models. The QA pairs in Visual7W are a subset of the 1.7 million QA pairs from Visual Genome. Moreover, Visual7W includes extra annotations such as object groundings, multiple choices and human experiments, making it a clean and complete benchmark for evaluation and analysis.

2. Related Work

Vision + Language. There have been years of effort in connecting the visual and textual information for joint learning [2, 19, 33, 35, 40, 52]. Image and video captioning has become a popular task in the past year [4, 5, 13, 37, 45, 47]. The goal is to generate text snippets to describe the images and regions instead of just predicting a few labels. Visual question answering is a natural extension to the captioning tasks, but is more interactive and has a stronger connection to real-world applications [3].

Text-based question answering. Question answering in NLP has been a well-established problem. Successful applications can be seen in voice assistants in mobile devices, search engines and game shows (e.g., IBM Waston). Traditional question answering system relies on an elaborate pipeline of models involving natural language parsing, knowledge base querying, and answer generation [6]. Recent neural network models attempt to learn end-to-end directly from questions and answers [12, 46].
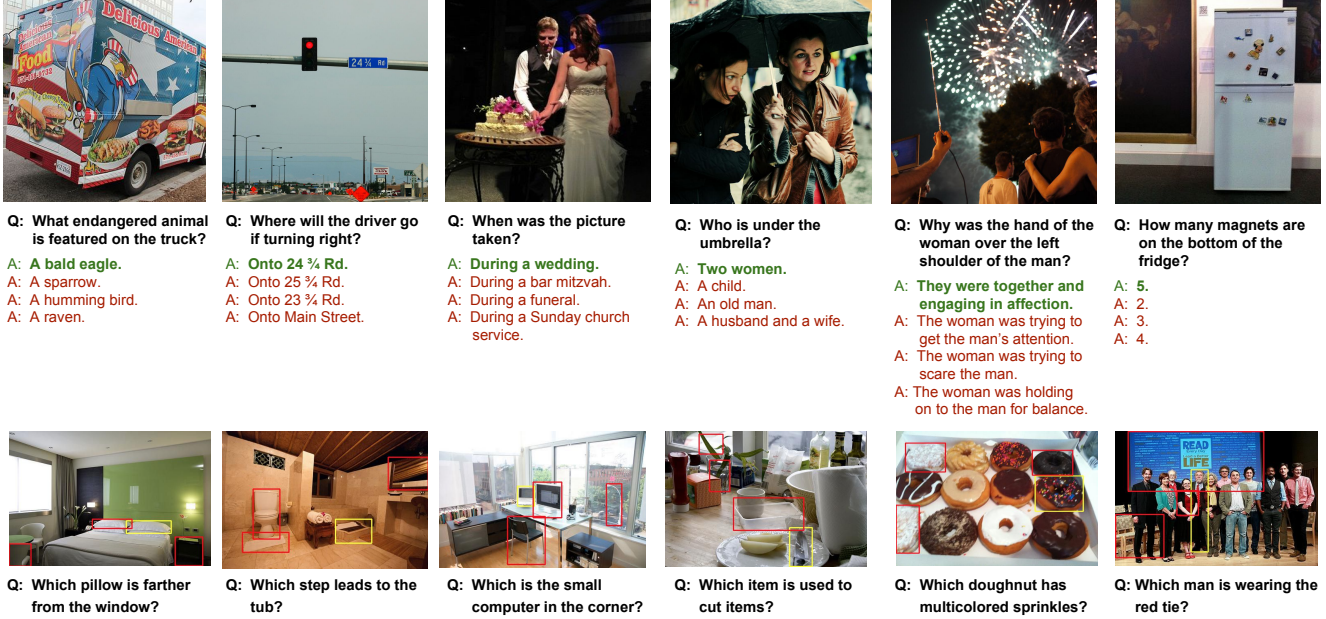
Visual question answering. Geman et al. [8] proposed a restricted visual Turing test to evaluate visual understanding. The DAQUAR dataset is the first toy-sized QA benchmark built upon indoor scene RGB-D images. Most of the other datasets [1, 7, 36, 49] collected QA pairs on Microsoft COCO images [25], either generated automatically by NLP tools [36] or written by human workers [1, 7, 49]. Following these datasets, an array of models has been proposed to tackle the visual QA tasks. The proposed models range from probabilistic inference [27, 44, 51] and recurrent neural networks [1, 7, 28, 36] to convolutional networks [26]. Previous visual QA datasets evaluate textual answers on images while omitting the links between the object mentions and their visual appearances. Inspired by Geman et al. [8], we establish the link by grounding objects in the images and perform experiments in the grounded QA setting.

3. Creating the Visual7W Dataset

We elaborate on the details of the data collection we conducted upon 47,300 images from COCO [25] (a subset of images from Visual Genome [20]). We leverage the six W questions (*what*, *where*, *when*, *who*, *why*, and *how*) to systematically examine a model's capability for visual understanding, and append a 7th *which* question category. This extends existing visual QA setups [1, 7, 36] to accommodate visual answers. We standardize the visual QA tasks with multi-modal answers in a multiple-choice format. Each question comes with four answer candidates, with one being the correct answer. In addition, we ground all the objects mentioned in the QA pairs to their corresponding bounding boxes in the images. The object-level groundings enable examining the object distributions and resolve the coreference ambiguity [19, 35].

3.1. Collecting the 7W Questions

The data collection tasks are conducted on Amazon Mechanical Turk (AMT), an online crowdsourcing platform. The online workers are asked to write pairs of question and answer based on image content. We instruct the workers to be concise and unambiguous to avoid wordy or speculative questions. To obtain a clean set of high-quality QA pairs, we ask three AMT workers to label each pair as *good* or *bad* independently. The workers judge each pair by whether an average person is able to tell the answer when seeing the

Figure 2 (first row):

Q: What endangered animal is featured on the truck?
A: **A bald eagle.**
A: A sparrow.
A: A humming bird.
A: A raven.

Q: Where will the driver go if turning right?
A: **Onto 24 ¾ Rd.**
A: Onto 25 ¾ Rd.
A: Onto 23 ¾ Rd.
A: Onto Main Street.

Q: When was the picture taken?
A: **During a wedding.**
A: During a bar mitzvah.
A: During a funeral.
A: During a Sunday church service.

Q: Who is under the umbrella?
A: **Two women.**
A: A child.
A: An old man.
A: A husband and a wife.

Q: Why was the hand of the woman over the left shoulder of the man?
A: **They were together and engaging in affection.**
A: The woman was trying to get the man's attention.
A: The woman was trying to scare the man.
A: The woman was holding on to the man for balance.

Q: How many magnets are on the bottom of the fridge?
A: **5.**
A: 2.
A: 3.
A: 4.

Figure 2 (second row):

Q: Which pillow is farther from the window?

Q: Which step leads to the tub?

Q: Which is the small computer in the corner?

Q: Which item is used to cut items?

Q: Which doughnut has multicolored sprinkles?

Q: Which man is wearing the red tie?

Figure 2: Examples of multiple-choice QA from the 7W question categories. The first row shows *telling* questions where the green answer is the ground-truth, and the red ones are human-generated wrong answers. The *what*, *who* and *how* questions often pertain to recognition tasks with spatial reasoning. The *where*, *when* and *why* questions usually involve high-level common sense reasoning. The second row depicts *pointing* (*which*) questions where the yellow box is the correct answer and the red boxes are human-generated wrong answers. These four answers form a multiple-choice test for each question.

image. We accept the QA pairs with at least two positive votes. We notice varying acceptance rates between categories, ranging from 92% for *what* to 63% for *why*. The overall acceptance rate is 85.8%.

VQA [1] relied on both human workers and automatic methods to generate a pool of candidate answers. We find that human-generated answers produce the best quality; on the contrary, automatic methods are prone to introducing candidate answers paraphrasing the ground-truth answers. For the *telling* questions, the human workers write three plausible answers to each question without seeing the image. To ensure the uniqueness of correct answers, we provide the ground-truth answers to the workers, and instruct them to write answers of different meanings. For the *pointing* questions, the workers draw three bounding boxes of other objects in the image, ensuring that these boxes cannot be taken as the correct answer. We provide examples from the 7W categories in Fig. 2.

3.2. Collecting Object-level Groundings

We collect object-level groundings by linking the object mentions in the QA pairs to their bounding boxes in the images. We ask the AMT workers to extract the object mentions from the QA pairs and draw boxes on the images. We collect additional groundings for the multiple choices of the *pointing* questions. Duplicate boxes are removed based on the object names with an Intersection-over-Union threshold

question ambiguity

What is the **man** wearing?
A black shirt.

answer ambiguity

What is the vehicle with ads on it?
A **red bus**.

Figure 3: Coreference ambiguity arises when an object mention has multiple correspondences in an image, and the textual context is insufficient to tell it apart. The answer to the left question can be either *gray*, *yellow* or *black*, depending on which man is meant. In the right example, the generic phrase *red bus* can refer to both buses in the image. Thus an algorithm might answer correctly even if referring to the wrong bus.

of 0.5. In total, we have collected 561,459 object bounding boxes, on average 12 boxes per image.

The benefits of object-level groundings are three-fold: 1) it resolves the coreference ambiguity problem between QA sentences and images; 2) it extends the existing visual QA setups to accommodate visual answers; and 3) it offers a means to understand the distribution of objects, shedding light on the essential knowledge to be acquired for tackling the QA tasks (see Sec. 4).

We illustrate examples of coreference ambiguity in Fig. 3. Ambiguity might cause a question to have more than

Table 1: Comparisons on Existing Visual Question Answering Datasets

	# QA	# Images	AvgQLen	AvgALen	LongAns	TopAns	HumanPerf	COCO	MC	Grounding	VisualAns
DAQUAR [27]	12,468	1,447	11.5 ± 2.4	1.2 ± 0.5	3.4%	96.4%					
Visual Madlibs [49]	56,468	9,688	4.9 ± 2.4	2.8 ± 2.0	47.4%	57.9%			✓		
COCO-QA [36]	117,684	69,172	8.7 ± 2.7	1.0 ± 0	0.0%	100%		✓			
Baidu [7]	316,193	316,193	-	-	-	-		✓			
VQA [1]	614,163	204,721	6.2 ± 2.0	1.1 ± 0.4	3.8%	82.7%	✓	✓	✓		
Visual7W (Ours)	327,939	47,300	6.9 ± 2.4	2.0 ± 1.4	27.6%	63.5%	✓	✓	✓	✓	✓

one plausible answers at test time, thus complicating evaluation. Our online study shows that, such ambiguity occurs in 1% of the accepted questions and 7% of the accepted answers. This illustrates a drawback of existing visual QA setups [1, 7, 27, 36, 49], where in the absence of object-level groundings the textual questions and answers are only loosely coupled to the images.

4. Comparison and Analysis

In this section, we analyze our Visual7W dataset collected on COCO images (cf. Table 1, *COCO*), present its key features, and provide comparisons of our dataset with previous work. We summarize important metrics of existing visual QA datasets in Table 1.[1]

Advantages of Grounding The unique feature of our Visual7W dataset is the grounding annotations of all textually mentioned objects (cf. Table 1, *Grounding*). In total we have collected 561,459 object groundings, which enables the new type of visual answers in the form of bounding boxes (cf. Table 1, *VisualAns*). Examining the object distribution in the QA pairs sheds light on the focus of the questions and the essential knowledge to be acquired for answering them. Our object groundings spread across 36,579 categories (distinct object names), thereby exhibiting a long tail pattern where 85% of the categories have fewer than 5 instances (see Fig. 4). The open-vocabulary annotations of objects, in contrast with traditional image datasets focusing on predefined categories and salient objects [25, 38], provide a broad coverage of objects in the images.

Human-Machine Performance Gap We expect that a good QA benchmark should exhibit a sufficient performance gap between humans and state-of-the-art models, leaving room for future research to explore. Additionally a nearly perfect human performance is desired to certify the quality of its questions. On Visual7W, we conducted two experiments to measure human performance (cf. Table 1, *HumanPerf*), as well as examining the percentage of questions that can be answered without images. Our results show both strong human performance and a strong interdependency between images and QA pairs. We provide the
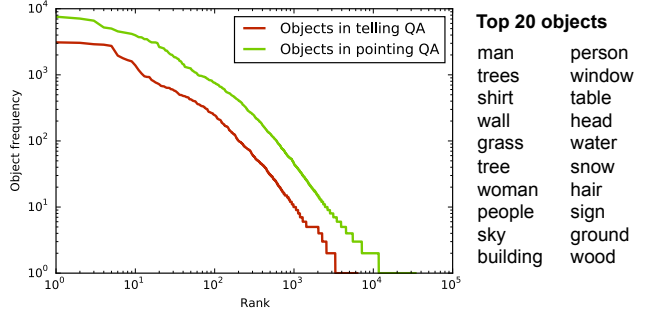
Figure 4: Object distribution in *telling* and *pointing* QA. The rank of an object category is based on its frequency with rank #1 referring to the most frequent one. The *pointing* QA pairs cover an order of magnitude more objects than the *telling* QA pairs. The top 20 object categories indicate that the object distribution's bias towards persons, daily-life objects and natural entities.

detailed analysis and comparisons with the state-of-the-art automatic models in Sec. 6.

Table 2: Model and Human Performances on QA Datasets

	Model	Human	Δ
VQA (open-ended) [1]	0.54	0.83	0.29
VQA (multiple-choice) [1]	0.57	0.92	0.35
Facebook bAbI [46]	0.92	~1.0	0.08
Ours (*telling* QA)	0.54	0.96	**0.42**
Ours (*pointing* QA)	0.56	0.97	0.41

Table 2 compares Visual7W with VQA [1] and Facebook bAbI [46], which have reported model and human performances. Facebook bAbI [46] is a textual QA dataset claiming that humans can potentially achieve 100% accuracy yet without explicit experimental proof. For VQA [1], numbers are reported for both multiple-choice and open-ended evaluation setups. Visual7W features the largest performance gap (Δ), a desirable property for a challenging and long-lasting evaluation task. At the same time, the nearly perfect human performance proves high quality of the 7W questions.

QA Diversity The diversity of QA pairs is an important feature of a good QA dataset as it reflects a broad coverage of image details, introduces complexity and potentially requires a broad range of skills for solving the questions. To obtain diverse QA pairs, we decided to rule out binary questions, contrasting Geman et al.'s proposal [8] and VQA's approach [1]. We hypothesize that this encourages workers to write more complex questions and also prevents inflating answer baselines with simple yes/no answers.

---

[1] We report the statistics of VQA dataset [1] with its real images and Visual Madlibs [49] with its filtered hard tasks. The fill-in-the-blank tasks in Visual Madlibs [49], where the answers are sentence fragments, differ from other QA tasks, resulting in distinct statistics. We omit some statistics for Baidu [7] due to its partial release.
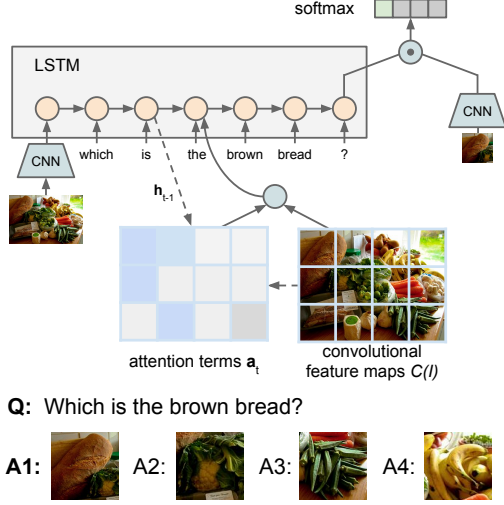
Figure 5: Diagram of the recurrent neural network model for *pointing* QA. At the encoding stage, the model reads the image and the question tokens word by word. At each word, it computes attention terms based on the previous hidden state and the convolutional feature map, deciding which regions to focus on. At the decoding stage, it computes the log-likelihood of an answer by a dot product between its transformed visual feature (fc7) and the last LSTM hidden state.

When examining the richness of QA pairs, the length of questions and answers (cf. Table 1, *AvgQLen*, *AvgALen*) is a rough indicator for the amount of information and complexity they contain. The overall average question and answer lengths are 6.9 and 2.0 words respectively. The *pointing* questions have the longest average question length. The *telling* questions exhibit a long-tail distribution where 51.2%, 21.2%, and 16.6% of their answers have one, two or three words respectively. Many answers to *where* and *why* questions are phrases and sentences, with an average of 3 words. In general, our dataset features long answers where 27.6% of the questions have answers of more than two words (cf. Table 1, *LongAns*). In contrast, 89% of answers in VQA [1], 90% of answers in DAQUAR [27] and all answers in COCO-QA [36] are a single word. We also capture more long-tail answers as our 1,000 most frequent answers only account for 63.5% of all our answers (cf. Table 1, *TopAns*). Finally we provide human created multiple-choices for evaluation (cf. Table 1, *MC*).

5. Attention-based Model for Grounded QA

The visual QA tasks are visually grounded, as local image regions are pertinent to answering questions in many cases. For instance, in the first *pointing* QA example of Fig. 2 the regions of the window and the pillows reveal the answer, while other regions are irrelevant to the question. We capture this intuition by introducing a spatial attention mechanism similar to the model for image captioning [47].

5.1. Recurrent QA Models with Spatial Attention

LSTM models [11] have achieved state-of-the-art results in several sequence processing tasks [5, 13, 41]. They have also been used to tackle visual QA tasks [1, 7, 28]. These models represent images by their global features, lacking a mechanism to understand local image regions. We add spatial attention [10, 47] to the standard LSTM model for visual QA, illustrated in Fig. 5. We consider QA as a two-stage process [7, 28]. At the encoding stage, the model memorizes the image and the question into a hidden state vector (the gray box in Fig. 5). At the decoding stage, the model selects an answer from the multiple choices based on its memory (the *softmax* layer in Fig. 5). We use the same encoder structure for all visual QA tasks but different decoders for the *telling* and *pointing* QA tasks. Given an image I and a question $Q = (q_1, q_2, \ldots, q_m)$, we learn the embeddings of the image and the word tokens as follow:

$$v_0 = W_i[F(I)] + b_i \tag{1}$$
$$v_i = W_w[OH(t_i)], i = 1, \ldots, m \tag{2}$$

where $F(\cdot)$ transforms an image I from pixel space to a 4096-dimensional feature representation. We extract the activations from the last fully connected layer (fc7) of a pre-trained CNN model VGG-16 [39]. $OH(\cdot)$ transforms a word token to its one-hot representation, an indicator column vector where there is a single one at the index of the token in the word vocabulary. The W_i matrix transforms the 4096-dimensional image features into the d_i-dimensional embedding space v_0, and the W_w transforms the one-hot vectors into the d_w-dimensional embedding space v_i. We set d_i and d_w to the same value of 512. Thus, we take the image as the first input token. These embedding vectors $v_{0,1,\ldots,m}$ are fed into the LSTM model one by one. The update rules of our LSTM model can be defined as follow:

$$\mathbf{i}_t = \sigma(W_{vi}v_t + W_{hi}\mathbf{h}_{t-1} + W_{ri}\mathbf{r}_t + b_i) \tag{3}$$
$$\mathbf{f}_t = \sigma(W_{vf}v_t + W_{hf}\mathbf{h}_{t-1} + W_{rf}\mathbf{r}_t + b_f) \tag{4}$$
$$\mathbf{o}_t = \sigma(W_{vo}v_t + W_{ho}\mathbf{h}_{t-1} + W_{ro}\mathbf{r}_t + b_o) \tag{5}$$
$$\mathbf{g}_t = \phi(W_{vg}v_t + W_{hg}\mathbf{h}_{t-1} + W_{rg}\mathbf{r}_t + b_g) \tag{6}$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \tag{7}$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \phi(\mathbf{c}_t) \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function, $\phi(\cdot)$ is the tanh function, and $\odot$ is the element-wise multiplication operator. The attention mechanism is introduced by the term $\mathbf{r}_t$, which is a weighted average of convolutional features that depends upon the previous hidden state and the convolutional features. The exact formulation is as follows:

$$\mathbf{e}_t = w_a^T \tanh(W_{he}\mathbf{h}_{t-1} + W_{ce}C(I)) + b_a \tag{9}$$
$$\mathbf{a}_t = \text{softmax}(\mathbf{e}_t) \tag{10}$$
$$\mathbf{r}_t = \mathbf{a}_t^T C(I) \tag{11}$$
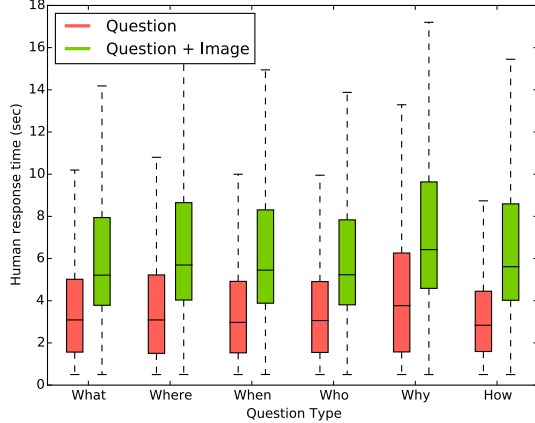
Figure 6: Response time of human subjects on the *telling* QA tasks. The boxes go from the first quartile to the third quartile of the response time values. The bars in the centers of the boxes indicate the median response time of each category.

where $C(I)$ returns the 14×14 512-dimensional convolutional feature maps of image I from the fourth convolutional layer from the same VGG-16 model [39]. The attention term $\mathbf{a}_t$ is a 196-dimensional unit vector, deciding the contribution of each convolutional feature at the t-th step. The standard LSTM model can be considered as a special case with each element in $\mathbf{a}_t$ set uniformly. W_i, b_i, W_w and all the Ws and bs in the LSTM model and attention terms are learnable parameters.

5.2. Learning and Inference

The model first reads the image v_0 and all the question tokens v_{q_1}, $v_{q_2}, \ldots, v_{q_m}$ until reaching the question mark (i.e., end token of the question sequence). When training for *telling* QA, we continue to feed the ground-truth answer tokens $v_{a_1}, v_{a_2}, \ldots, v_{a_n}$ into the model. For *pointing* QA, we compute the log-likelihood of an candidate region by a dot product between its transformed visual feature (fc7) and the last LSTM hidden state (see Fig. 5). We use cross-entropy loss to train the model parameters with backpropagation. During testing, we select the candidate answer with the largest log-likelihood. We set the hyperparameters using the validation set. The dimensions of the LSTM gates and memory cells are 512 in all the experiments. The model is trained with Adam update rule [18], mini-batch size 128, and a global learning rate of 10^{-4}.

6. Experiments

We evaluate the human and model performances on the QA tasks. We report a reasonably challenging performance delta leaving sufficient room for future research to explore.

6.1. Experiment Setups

As the 7W QA tasks have been formulated in a multiple-choice format, we use the same procedure to evaluate hu-

man and model performances. At test time, the input is an image and a natural language question, followed by four multiple choices. In *telling* QA, the multiple choices are written in natural language; whereas, in *pointing* QA, each multiple choice corresponds to an image region. We say the model is correct on a question if it picks the correct answer among the candidates. Accuracy is used to measure the performance. An alternative method to evaluate *telling* QA is to let the model predict open-ended text outputs [1]. This approach works well on short answers; however, it performs poorly on long answers, where there are many ways of paraphrasing the same meaning. We make the training, validation and test splits, each with 50%, 20%, 30% of the pairs respectively. The numbers are reported on the test set.

6.2. 7W QA Experiments

6.2.1 Human Experiments on 7W QA

We evaluate human performances on the multiple-choice 7W QA. We want to measure in these experiments 1) how well humans can perform in the visual QA task and 2) whether humans can use common sense to answer questions without seeing the images.

We conduct two sets of human experiments. In the first experiment (Question), a group of five AMT workers are asked to guess the best possible answers from the multiple choices without seeing the images. In the second experiment (Question + Image), we have a different group of five workers to answer the same questions given the images. The first block in Table 3 reports the human performances on these experiments. We measure the mean accuracy over the QA pairs where we take the majority votes among the five human responses. Even without the images, humans manage to guess the most plausible answers in some cases. Human subjects achieve 35.3% accuracy, 10% higher than chance. The human performance without images is remarkably high (43.9%) for the *why* questions, indicating that many *why* questions encode a fair amount of common sense that humans are able to infer without visual cue. However, images are important in the majority of the questions. Human performance is significantly improved when the images are provided. Overall, humans achieve a high accuracy of 96.6% on the 7W QA tasks.

Fig. 6 shows the box plots of response time of the human subjects for *telling* QA. Human subjects spend double the time to respond when the images are displayed. In addition, *why* questions take a longer average response time compared to the other five question types. Human subjects spend an average of 9.3 seconds on *pointing* questions. However, that experiment was conducted in a different user interface, where workers click on the answer boxes in the image. Thus, the response time is not comparable with the *telling* QA tasks. Interestingly, longer response time does not imply higher performance. Human subjects spend more

Table 3: Human and model performances in the multiple-choice 7W QA tasks (in accuracy)

Method	Telling						Pointing	Overall
	What	Where	When	Who	Why	How	Which	
Human (Question)	0.356	0.322	0.393	0.342	0.439	0.337	-	0.353
Human (Question + Image)	0.965	0.957	0.944	0.965	0.927	0.942	0.973	0.966
Logistic Regression (Question)	0.420	0.375	0.666	0.510	0.354	0.458	0.354	0.383
Logistic Regression (Image)	0.408	0.426	0.438	0.415	0.337	0.303	0.256	0.305
Logistic Regression (Question + Image)	0.429	0.454	0.621	0.501	0.343	0.356	0.307	0.352
LSTM (Question)	0.430	0.414	0.693	0.538	0.491	0.484	-	0.462
LSTM (Image)	0.422	0.497	0.660	0.523	0.475	0.468	0.299	0.359
LSTM (Question + Image) [28]	0.489	0.544	0.713	0.581	0.513	**0.503**	0.521	0.521
Ours, LSTM-Att (Question + Image)	**0.515**	**0.570**	**0.750**	**0.595**	**0.555**	0.498	**0.561**	**0.556**

Figure 7: Qualitative results of human subjects and the state-of-the-art model (LSTM-Att) on multiple-choice QAs. We illustrate the prediction results of six multiple-choice QAs, with and without images. The green answer corresponds to the correct answer to each question, and the rest three are wrong answer candidates. We take the majority votes of five human subjects as the human predictions (H) and the top predictions from the model (M). The correct predictions are indicated by check marks.

time on questions with lower accuracy. The Pearson correlation coefficient between the average response time and the average accuracy is -0.135, indicating a weak negative correlation between the response time and human accuracy.

6.2.2 Model Experiments on 7W QA

Having examined human performance, our next question is how well the state-of-the-art models can perform in the 7W QA task. We evaluate automatic models on the 7W QA tasks in three sets of experiments: without images (Question), without questions (Image) and with images (Question + Image). In the experiments without images (questions), we zero out the image (questions) features. We briefly describe the three models we used in the experiments:

Logistic Regression A logistic regression model that predicts the answer from a concatenation of image fc7 feature and question feature. The questions are represented by 200-dimensional averaged word embeddings from a pre-trained model [29]. For *telling* QA, we take the top-5000 most frequent answers (79.2% of the training set answers) as the
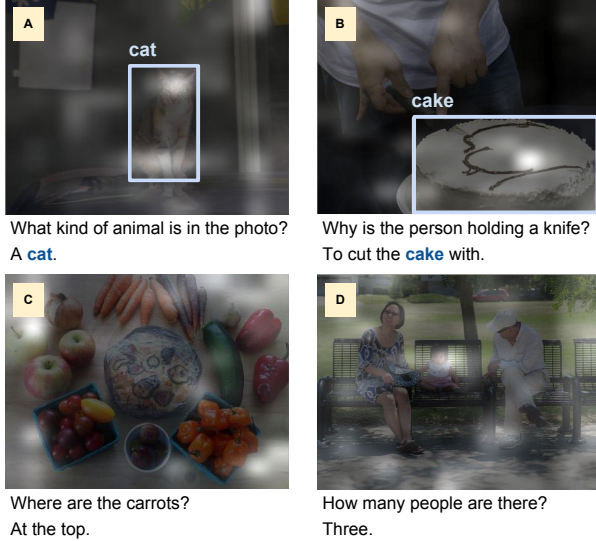
class labels. At test time, we select the top-scoring answer candidate. For *pointing* QA, we perform k-means to cluster training set regions by fc7 features into 5000 clusters, used as class labels. At test time, we select the answer candidate closest to the centroid of the predicted cluster.

LSTM The LSTM model in Malinowski and Fritz [28] for visual QA with no attention modeling, which can be considered as a simplified version of our full model with the attention terms set to be uniform.

LSTM-Att Our LSTM model with spatial attention introduced in Sec. 5, where the attention terms in Eq. (10) determines which region to focus on at each step.

We report the results in Table 3. All the baseline models surpass the chance performance (25%). The logistic regression baseline yields the best performance when only the question features are provided. Having the global image features hurts its performance, indicating the importance of understanding local image regions rather than a holistic representation. Interestingly, the LSTM performance (46.2%) significantly outperforms human perfor-

What kind of animal is in the photo?
A **cat**.

Why is the person holding a knife?
To cut the **cake** with.

Where are the carrots?
At the top.
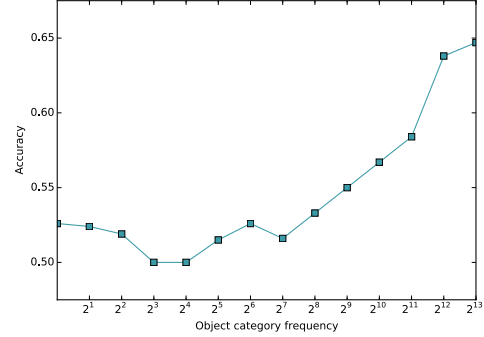
How many people are there?
Three.

Figure 8: Object groundings and attention heat maps. We visualize the attention heat maps (with Gaussian blur) on the images. The brighter regions indicate larger attention terms, i.e., where the model focuses. The bounding boxes show the object-level groundings of the objects mentioned in the answers.

Figure 9: Impact of object category frequency on the model accuracy in the *pointing* QA task. The x-axis shows the upper bound object category frequency of each bin. The y-axis shows the mean accuracy within each bin. The accuracy increases gradually as the model sees more instances from the same category. Meanwhile, the model manages to handle infrequent categories by transferring knowledge from larger categories.

mance (35.3%) when the images are not present. Human subjects are not *trained* before answering the questions; however, the LSTM model manages to learn the priors of answers from the training set. In addition, both the questions and image content contribute to better results. The Question + Image baseline shows large improvement on overall accuracy (52.1%) than the ones when either the question or the image is absent. Finally, our attention-based LSTM model (LSTM-Att) outperforms other baselines on all question types, except the *how* category, achieving the best model performance of 55.6%.

We show qualitative results of human experiments and the LSTM models on the *telling* QA task in Fig. 7. Human subjects fail to tell a sheep apart from a goat in the last example, whereas the LSTM model gives the correct answer. Yet humans successfully answer the fourth *why* question when seeing the image, where the LSTM model fails in both cases.

The object groundings help us analyzing the behavior of the attention-based model. First, we examine where the model focuses by visualizing the attention terms of Eq. (10). The attention terms vary as the model reads the QA words one by one. We perform max pooling along time to find the maximum attention weight on each of the 14×14 image grid, producing an attention heat map. We see if the model attends to the mentioned objects. The answer object boxes occupy an average of 12% of image area; while the peak of the attention heat map resides in answer object boxes 24% of the time. That indicates a tendency for the model to attend to the answer-related regions. We visualize the atten-

tion heat maps on some example QA pairs in Fig. 8. The top two examples show QA pairs with answers containing an object. The peaks of the attention heat maps reside in the bounding boxes of the target objects. The bottom two examples show QA pairs with answers containing no object. The attention heat maps are scattered around the image grid. For instance, the model attends to the four corners and the borders of the image to look for the carrots in Fig. 8(c).

Furthermore, we use object groundings to examine the model's behavior on the *pointing* QA. Fig. 9 shows the impact of object category frequency on the QA accuracy. We divide the object categories into different bins based on their frequencies (by power of 2) in the training set. We compute the mean accuracy over the test set QA pairs within each bin. We observe increased accuracy for categories with more object instances. However, the model is able to transfer knowledge from common categories to rare ones, generating an adequate performance (over 50%) on object categories with only a few instances.

7. Conclusions

In this paper, we propose to leverage the visually grounded 7W questions to facilitate a deeper understanding of images beyond recognizing objects. Previous visual QA works lack a tight semantic link between textual descriptions and image regions. We link the object mentions to their bounding boxes in the images. Object grounding allows us to resolve coreference ambiguity, understand object distributions, and evaluate on a new type of visually grounded QA. We propose an attention-based LSTM model to achieve the state-of-the-art performance on the QA tasks. Future research directions include exploring ways of utilizing common sense knowledge to improve the model's performance on QA tasks that require complex reasoning.

References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. *ICCV*, 2015. 1, 2, 3, 4, 5, 6

[2] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, 2003. 2

[3] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User Interface Software and Technology*, 2010. 2

[4] X. Chen and C. L. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, 2015. 1, 2

[5] J. Donahue, L. Anne 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, 5

[6] D. Ferrucci et al. Building Watson: An overview of the DeepQA project. *AI Magazine*, 2010. 2

[7] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu. Are you talking to a machine? dataset and methods for multilingual image question answering. *NIPS*, 2015. 1, 2, 4, 5

[8] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015. 1, 2, 4

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2

[10] K. Gregor, I. Danihelka, A. Graves, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 5

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[12] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé III. A neural network for factoid question answering over paragraphs. In *Empirical Methods in Natural Language Processing*, 2014. 2

[13] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *CVPR*, 2015. 1, 2, 5

[14] A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *NIPS*, pages 1889–1897, 2014. 1

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 1, 2

[16] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2

[17] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg. Where to buy it: Matching street clothing photos in online shops. *ICCV*, 2015. 2

[18] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[19] C. Kong, D. Lin, M. Bansal, R. Urtasun, and S. Fidler. What are you talking about? text-to-image coreference. In *CVPR*, 2014. 1, 2

[20] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. In *arXiv preprint arxiv:1602.07332*, 2016. 2

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 1

[22] R. Kuhn and E. Neveu. *Political journalism: New challenges, new practices*. Routledge, 2013. 2

[23] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2

[24] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for ground-to-aerial geolocalization. In *CVPR*, 2015. 2

[25] 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. 1, 2, 4

[26] L. Ma, Z. Lu, and H. Li. Learning to answer questions from image using convolutional neural network. *arXiv preprint arXiv:1506.00333*, 2015. 2

[27] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, pages 1682–1690, 2014. 1, 2, 4, 5

[28] M. Malinowski, M. Rohrbach, and M. Fritz. Ask your neurons: A neural-based approach to answering questions about images. *ICCV*, 2015. 2, 5, 7

[29] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013. 7

[30] F. Palermo, J. Hays, and A. A. Efros. Dating historical color images. In *ECCV*. 2012. 2

[31] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012. 2

[32] L. C. Pickup, Z. Pan, D. Wei, Y. Shih, C. Zhang, A. Zisserman, B. Scholkopf, and W. T. Freeman. Seeing the arrow of time. In *CVPR*, 2014. 2

[33] H. Pirsiavash, C. Vondrick, and A. Torralba. Inferring the why in images. *arXiv preprint arXiv:1406.5472*, 2014. 2

[34] B. Plummer, L. Wang, C. Cervantes, J. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *ICCV*, 2015. 1

[35] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people with "their" names using coreference resolution. In *ECCV*, 2014. 1, 2

[36] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. *NIPS*, 2015. 1, 2, 4, 5

[37] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013. 2

[38] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, pages 1–42, April 2015. 1, 4

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2014. 1, 2, 5, 6

[40] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. 1, 2

[41] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014. 5

[42] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014. 1, 2

[43] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 1

[44] 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. In *IEEE MultiMedia*, 2014. 2

[45] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2

[46] J. Weston, A. Bordes, S. Chopra, and T. Mikolov. Towards ai-complete question answering: a set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015. 2, 4

[47] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015. 1, 2, 5

[48] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 1

[49] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual Madlibs: Fill in the blank Image Generation and Question Answering. *ICCV*, 2015. 1, 2, 4

[50] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, 2014. 2

[51] Y. Zhu, A. Fathi, and L. Fei-Fei. Reasoning about object affordances in a knowledge base representation. *ECCV*, 2014. 2

[52] C. L. Zitnick, D. Parikh, and L. Vanderwende. Learning the visual interpretation of sentences. In *ICCV*, 2013. 2