Visual Question Answering on Image Sets

Ankan Bansal[1(✉)], Yuting Zhang[2], and Rama Chellappa[1]

[1] University of Maryland, College Park, USA
{ankan,rama}@umd.edu
[2] Amazon Web Services (AWS), Beijing, China
yutingzh@amazon.com

Abstract. We introduce the task of Image-Set Visual Question Answering (ISVQA), which generalizes the commonly studied single-image VQA problem to multi-image settings. Taking a natural language question and a set of images as input, it aims to answer the question based on the content of the images. The questions can be about objects and relationships in one or more images or about the entire scene depicted by the image set. To enable research in this new topic, we introduce two ISVQA datasets – indoor and outdoor scenes. They simulate the real-world scenarios of indoor image collections and multiple car-mounted cameras, respectively. The indoor-scene dataset contains 91,479 human-annotated questions for 48,138 image sets, and the outdoor-scene dataset has 49,617 questions for 12,746 image sets. We analyze the properties of the two datasets, including question-and-answer distributions, types of questions, biases in dataset, and question-image dependencies. We also build new baseline models to investigate new research challenges in ISVQA.

1 Introduction

Answering natural-language questions about images requires understanding both linguistic and visual data. Since its introduction [4], Visual Question Answering (VQA) has attracted significant attention. Several related datasets [14,22,33] and methods [9,12,20] have been proposed.

In this paper, we introduce the new task of Image Set Visual Question Answering (ISVQA)[1]. It aims to answer a free-form natural-language question based on a set of images. The proposed ISVQA task requires reasoning over objects and concepts in different images to predict the correct answer. For example, for Fig. 1 (Left), a model has to find the relationship between the `bed` in the top-left image and the `mirror` in the top-right, via `pillows` which are common to both the images. This example shows the unique challenges associated with

[1] Project page: https://ankanbansal.com/isvqa.html.

This work was done when Ankan Bansal was an intern at AWS.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-58589-1_4) contains supplementary material, which is available to authorized users.

A. Vedaldi et al. (Eds.): ECCV 2020, LNCS 12366, pp. 51–67, 2020.
https://doi.org/10.1007/978-3-030-58589-1_4

image-set VQA. A model for solving this type of problems has to understand the question, find the connections between the images, and use those connections to relate objects across images. Similarly, in Fig. 1 (Right), the model has to avoid double-counting recurring objects in multiple images. These challenges associated with scene understanding have not been explored in existing single-image VQA settings but frequently happen in the real world.

Fig. 1. (Left) Given the set of images above, and the question "What is hanging above the bed?", it is necessary to connect the bed in the top-left image to the mirror in the top-right image. To answer this question a model needs to understand the concepts of "bed", "mirror", "above", "hanging", etc. and be able to relate the bed in the first image with the headrest and pillows in the third image. (Right) When asked the question "How many rectangles are on the interior doors?", the model should be able to provide the answer ("four") and avoid counting the same rectangles multiple times.

ISVQA reflects information retrieval from multiple images of relevance but with no obvious continuous correspondences. Such image sets can be any albums and images captured by multiple devices, *e.g.*, images under the same story on Facebook/Instagram, images of the same product on Craigslist and Amazon, pictures of the same house on real estate websites, and images from different car-mounted cameras. Other instances of the ISVQA task include answering questions about images taken at different times (*e.g.* like in camera trap photography), at different locations (*e.g.* multiple cameras from an indoor or outdoor location), or from different viewpoints (*e.g.* live sports coverage). Some of these settings contain images taken from the same scene, while others might involve images of a larger span. While ISVQA can be generally applied to any type of images, in this paper, we focus on images from multiple views of an environment, especially street and indoor scenes.

ISVQA may require finding the same objects in different images or determining the relationships between different objects within or across images. It can also entail determining which images are the most relevant for the question and then answering based only on them, ignoring the other images. ISVQA leads to new research challenges, including: a) How to use natural language to guide scene understanding across multiple views/images; and b) how to fuse information from relevant images to reason about relationships among entities.

To enable research into these problems, we built two datasets for ISVQA - one for indoor scenes and the other for outdoor scenes. The indoor scenes dataset comes from Gibson Environment [31] and contains 91,479 human-generated questions, each for a set of images - for a total of 48,138 image sets. Similarly, the outdoor scenes dataset comprises of 49,617 questions for 12,746 image sets. The images in the outdoor scenes dataset come from the nuScenes dataset [5]. We explain the data collection methodology and statistics in Sect. 3.

The indoor ISVQA dataset contains two parts: 1.) Gibson-Room - containing images from the same room; and 2.) Gibson-Building - containing images from different places in the same building. This is to facilitate spatial and semantic reasoning both in localized and extended regions in the same scene. The outdoor dataset contains image sets taken from mostly urban environments.

We propose two extensions of single-image VQA methods as baseline approaches to investigate the ISVQA task and the datasets. In addition, we also use an existing Video VQA approach as a simple baseline. Finally, we also propose to use use a transformer-based approach which can specifically target ISVQA. Such baselines meet significant difficulties in solving the ISVQA problem, and they reflect the particular challenges of the ISVQA task. We also present the statistics of the datasets, by analyzing the types of question, distributions of answers for different types of questions, and biases present in the dataset.

In summary, we make the following major contributions in this work.

- We propose ISVQA as a new setting for scene understanding via Q&A;
- We introduce two large-scale datasets for targeting the ISVQA problem. In total, these datasets contain 141,096 questions for 60,884 sets of images.
- We establish baseline methods on ISVQA tasks to recognize the challenges and encourage future research.

2 Related Works

VQA Settings. The basic VQA setting [4] involves answering natural language questions about images. The VisualGenome dataset [17] also contains annotations for visual question-answer pairs at both image and region levels. Visual7W [33] built upon the basic VQA setting and introduced visual grounding to VQA. Several other VQA settings target specific problems or applications. For example, VizWiz [13] was designed to help answer questions asked by blind people. RecipeQA [32] is targeted for answering questions about recipes from multi-modal cues. TallyQA [1], and HowMany-QA [30] specifically target counting questions for single images. Unlike these, the CLEVR [16] benchmark and dataset uses synthetically generated images of rendered 3D shapes and is aimed towards understanding the geometric relationships between objects. IQA [11] is also a synthetic setting where an agent is required to navigate a scene and reach the desired location in order to answer the question.

Unlike existing work, ISVQA targets scene understanding by answering questions which might require multiple images to answer. This important setting has not been studied before and necessitates a specialized dataset. Additionally,

answering most of the questions requires a model to ignore some of the images in the set. This capability is absent from many state-of-the-art VQA models.

We also distinguish our work from video VQA. Unlike many such datasets (*e.g.* TVQA+ [18], MovieQA [29]), our datasets do not contain any textual cues like scripts or subtitles. Also, videos are temporally continuous and are usually taken from a stationary view-point. This makes finding associations between objects across frames easy, even if datasets do not provide textual cues (*e.g.* tGIF-QA [15]). The image sets in ISVQA dataset are not akin to video frames. Also, unlike embodied QA [6], ISVQA does not have an agent interacting with the environment. ISVQA algorithms can use only the few given images, which resembles real-world applications. Embodied QA does not require sophisticated inference of the correspondence between images, as the frames that an agent sees are continuous. The agent can reach the desired location, and answer the question using only the final frame. In contrast, ISVQA often needs reasoning across images and an implicit understanding of a scene.

VQA Methods. Most of the recent VQA methods use attention mechanisms to focus on the most relevant image regions. For example, [3] proposed a bottom-up and top-down attention mechanism for answering visual questions. In addition, several methods which use co-attention (or bi-directional) attention over questions and images have been proposed. Such methods include [9,21], all of which use the information from one modality (text or image) to attend to the other. Somewhat different from these is the work from Gao *et al.* [10] which proposed the multi-modality latent interaction module which can model the relationships between visual and language summaries in the form of latent vectors.

Unlike these, [7] used reasoning modules over detected objects to answer questions about geometric relationships between objects. Similarly, Santoro *et al.* [24] proposed using Relation Networks to solve specific relational reasoning problems. Neither of these approaches used attention mechanisms. In this paper, we mostly focus on attention-based mechanisms to design the baseline models.

3 Dataset

The main goal of our data collection effort is to aid multi-image scene understanding via question answering. We use two publicly available datasets (Gibson [31] and nuScenes [5]) as the source of images to build our datasets. Gibson provides navigable 3D-indoor scenes. We use the Habitat API [25] to extract images from multiple locations and viewpoints from it. nuScences contains sets of images taken simultaneously from multiple cameras on a car.

3.1 Annotation Collection

Indoor Scenes. Gibson is a collection of 3D scans of indoor spaces, particularly houses and offices. It provides virtualized scans of real indoor spaces like houses and offices. Using the Habitat platform, we place an agent at different locations and orientations in a scene and store the views visible to the agent. We generate a

set of images by obtaining several views from the same scene. Therefore, together, each image set can be considered to represent the scene.

We collect two types of indoor scenes: 1.) Gibson-Building; and 2.) Gibson-Room. Gibson-Building contains multiple images taken from the same building by placing the agent at random locations and recording its viewpoint while Gibson-Room is collected by obtaining several views from the same room.

For Gibson-Building, we sample image sets by placing an agent at random locations in the scene. We show images from Gibson-Building sets to annotators and request them to ask questions about the scene.

We obtain question-answer annotations for a scene from several annotators using Amazon Mechanical Turk. We let each annotator ask a question about the scene and also provide the corresponding answer. We request that the annotators should ensure that their question can be answered using only the scene shown and no additional knowledge should be required.

From a pilot study, we observed that it is easier for humans to frame questions if they are shown the full 3D view of a scene, simulating the situation of them being present in the scene. Humans are able to frame better questions about locations of objects, and their relationships in such a setting. For Gibson-Room, we simulate such immersion by creating a 360° video from a room. We show these videos (see supplementary material for examples of how these videos are created) to the annotators and ask them to provide questions and answers about the scenes. This process helped annotators understand the entire scene more easily and enabled us to collect more questions requiring across-image reasoning. Videos are not used for annotating nuScene and Gibson-Building.

Note that the ISVQA problem and datasets do not have videos. The images for Gibson-Room still came from random views as previously described. It is possible that the image set has less coverage of the scene than the video. Just using the image set, it might not be possible to answer the questions collected on the video. We prune out those cases by by asking other human-annotators to verify if the question can be answered using the provided image set.

Outdoor Scenes. We collect annotations for the nuScenes dataset similar to the Gibson-Building setting. We show the annotators images from an image set. These represent a 360° view of a scene. We, again, ask them to write questions and answers about the scene as before.

Refining Annotations. We showed all the image sets in our datasets and the associated questions obtained from the previous step to up to three other annotators. We asked them to provide an answer to the question based only on the image set shown. We also asked them to say "Not possible to answer" if the question cannot be answered. This step increases confidence about an answer if there is a consensus among the annotators. This step has the added benefit of ensuring that the question can be answered using the image set.

In addition, we also asked the annotators at this stage to mark the images which are required to answer the given question. This provides us information about which images are the most salient for answering a question.

Train and Test Splits. After refinement, we divided the datasets into train and test splits. The statistics of these splits are given in Table 1. For test splits, we have select samples for which at least two annotators agreed on the answer. We also ensured that the train and test sets have the same set of answers.

Table 1. Statistics of train and test splits of the datasets.

Dataset	#Train sets	#Test sets	#Unique answers
Indoor - Gibson (Room + Building)	69,207	22,272	961
Outdoor - nuScenes	33,973	15,644	650

3.2 Dataset Analysis

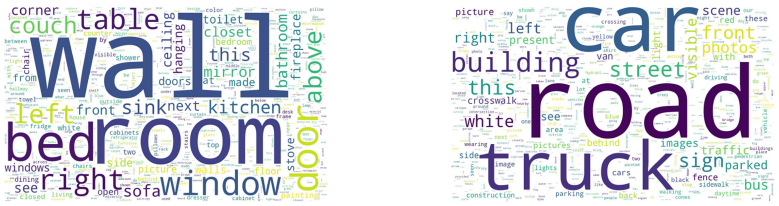


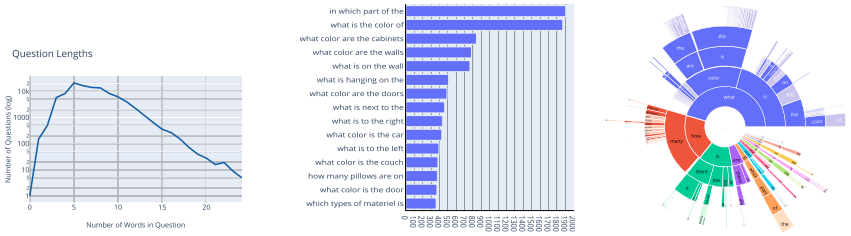
Fig. 2. Question wordclouds for Gibson (left) and nuScenes (right) datasets.

Question Word Distributions. The question word clouds for datasets are shown in Fig. 2. We have removed the first few words from each question before plotting these. This gives us a better picture of which objects people are interested in. Clearly, for outdoor scenes, people are most interested in objects commonly found on the streets and their properties (types, colors, numbers). On the other hand, for indoor scenes, the most frequent questions are about objects hanging on walls and kept on beds, and the room layouts in general.

Types of Questions. Figure 3a shows the distribution of question lengths for the dataset. We observe that a large chunk of the questions contain between 5 and 10 words. Further, Fig. 3b shows the numbers of the most frequent types of questions for the dataset. We observe that the most frequent questions are about properties of objects, and spatial relationships between different entities.

To understand the types of questions in the dataset, we plot the distribution of the most frequent first five words of the questions in the whole dataset in Fig. 3c. Note that a large portion of the questions are about the numbers of different kinds of objects. Another major subset of the questions are about geometric relationships between objects in a scene. A third big part of the dataset contains questions about colors of objects in scenes. Answering questions about

the colors of things in a scene requires localization of the object of interest. Depending on the question, this might require reasoning about the relationships between objects in different images. Similarly, counting the number of a particular type of object requires keeping track of previously counted objects to avoid double counting if the same object appears in different images.



(a) Question lengths (b) Frequent questions (c) Types of questions

Fig. 3. (Left) Distribution of questions over no. of words. (Middle) Most frequent types of questions in the dataset. (Right) First five words of the questions.

Answer Distributions. Figure 4 shows the distribution of answers in the dataset for frequently occurring questions types. On one hand, due to human bias in asking questions, dominant answers exist for a few types of questions, such as "can you see the" (usually for an object in the image) and "what is this" (usually a large object). In ISVQA and other VQA datasets humans' tendency to only ask questions about objects that they can see leads to a higher frequency of "yes" answers. On the other hand, most question types do not have a dominant answer. Of particular note are the questions about relative locations and orientations of objects, *e.g.* "What is next to", and questions about the numbers of objects *e.g.* "How many chairs are" etc. This means that it is difficult for a model to perform well by lazily exploiting the statistics of question types.

Number of Images Required. While refining the annotations, we also collect annotations for which images are required to answer the given question. In Fig. 5, we plot the number of images required to answer each question for all datasets. For the plot in Fig. 5, we only consider those image sets for which at least 2 annotators agree about the images which are needed. We observe that one-third of the samples (about 7,000/21,000) in Gibson-Room require at least two images to answer the question. As expected, this ratio is lower for Gibson-Building dataset. However, for all three datasets, we have a large number of questions which require more than one image to answer. The large number of samples in both cases enable the study of both across-image reasoning and image-level focusing. In particular, the latter case also involves rejecting most of the images in the image set and focusing only on one image. In theory, such questions can potentially be answered by using existing single-image VQA models. However, this would require the single-image VQA model to say "Not possible to answer"

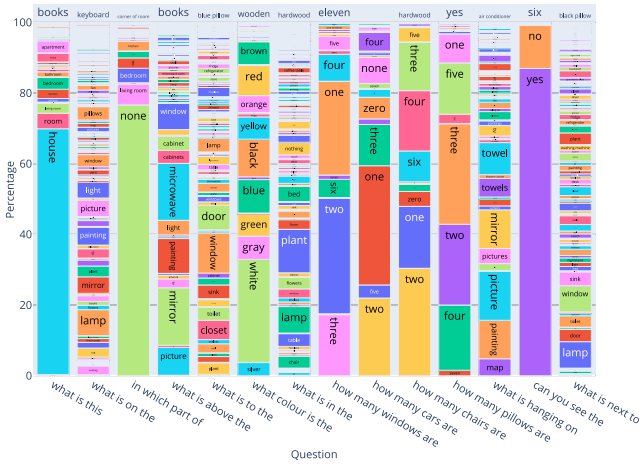
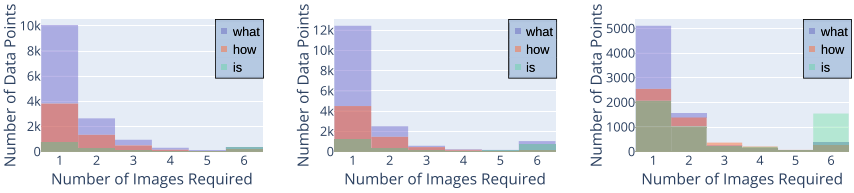


Fig. 4. Answer distributions for several types of questions in the whole dataset.

for all the irrelevant images and finding only the most relevant one. Current VQA models do not have the ability to do this in many cases. (see supplementary material for examples).



(a) Gibson-Room (b) Gibson-Building (c) nuScenes

Fig. 5. Number of images required to answer different types of questions.

4 ISVQA Problem Formulation and Baselines

4.1 Problem Definition

Refer to Fig. 6 for some examples of the ISVQA setting. Given a set of images, $S = \{I_1, I_2, \ldots, I_n\}$, and a natural language question, $Q = \{v_1, v_2, \ldots, v_T\}$, where v_i is the i^{th} word in the question, the task is to provide an answer, $a = f(S, Q)$, which is true for the given question and image set. The function f can either output a probability distribution over a pre-defined set of possible answers, $\mathcal{A}$, or select the best answer from several choices which are input along with the

question, i.e., $a = f(S, Q, C_Q)$, where C_Q is the list of choices associated with Q. The former is usually called open-ended QA and the latter is called multiple-choice QA. In this work, we mainly deal with the open-ended setting. Another possible setting is to actually generate the answer using a text generation method similar to image-captioning. But, most existing VQA works focus on either of the first two settings and therefore, we also consider the open-ended setting in this work. We leave the harder problem of generating answers to future work.

Fig. 6. Some examples from our dataset which demonstrate the ISVQA problem setting. In each case, input is a set of images and a question.

4.2 Model Definitions

Now, we describe some baselines for the ISVQA problem. These baselines directly adapt single image VQA models. The first of these processes each image separately and concatenates the features obtained from each image to predict the answer. The second baseline directly adapts VQA methods by simply stitching the images and using single image VQA methods to predict the answer.

We also propose an approach to address the special challenges in ISVQA. A fundamental direction to solve ISVQA problem is to enable finer-grained and across-image interactions in a VQA model, where self-attention-based transformers can fit well. In particular, we adapt LXMERT [28], which is designed for cross-modality learning, to both cross-modality and cross-image scenarios.

Concatenate-Feature. Starting from a given set of n images $S = \{I_1, I_2, \ldots, I_n\}$, we use a region proposal network (RPN) to extract region proposals $R_i, i = 1, 2, \ldots, n$ and the corresponding RoI-pooled features (fc6). With some abuse of notation, we denote the region features obtained from each image as $R_i \in \mathbb{R}^{p \times d}, i = 1, 2, \ldots, n$, where p is the number of region features obtained from each image and d is the dimension of the features. We are also given a natural language question $Q = \{v_1, v_2, \ldots, v_T\}$, where v_i is the i^{th} word, encoded as a one-hot vector over a fixed vocabulary V of size d_V. For all the models, we first obtain question token embeddings $E = \{W_w^T v_i\}_{i=1}^T$, where $W_w \in \mathbb{R}^{d_V \times d_q}$ is a continuous word-vector embedding matrix. We obtain the question embedding feature using an LSTM-attention module, i.e., $q = \text{AttentionPool}(\text{LSTM}(E)) \in \mathbb{R}^{d_q}$.

Figure 7 shows an outline of the model. For each image, I_i, we obtain the image embedding, x_i by attending over the corresponding region features R_i using the question embedding q.

$$x_i = \text{AttentionPool}(\text{Combine}(R_i, q)) \tag{1}$$

Fig. 7. Concatenate-Feature Baseline. This method adapts a single-image VQA model to an image set $S = \{I_1 \ldots I_n\}$. We first extract region proposals, R_i from each image I_i. The model attends over the regions in each image separately using the question embedding q. Pooling the region features gives a representation of an image as x_i. These are concatenated and combined (element-wise multiplied) by the question embedding to give the joint scene representation x. We use fully-connected layers to predict the final answer a.

where, we use element-wise multiplication (after projecting to suitable dimensions) as the Combine layer and AttentionPool is a combination of an Attention module over the region features which is calculated through a softmax operation and a Pool operation. The region features are multiplied by the attention and added to obtain the pooled image representation. For a single image, this model is an adaptation of the recent Pythia model [27] without its OCR functionality. We concatenate the image features x_i and element-wise multiply by the question embedding to obtain the joint embedding

$$x = \text{Combine}(\text{Concat}(x_1, x_2, \ldots, x_n), q) \tag{2}$$

where the Combine layer is again an element-wise multiplication. This is passed through a small MLP to obtain the distribution over answers, $P_A = \text{MLP}(x)$.

Stitched Image. Our next baseline is also an adaptation of existing single-image VQA methods. We start by stitching all the images in an image set into a mosaic, similar to the ones shown in Fig. 6. Note that the ISVQA setting does not require the images in an image set to follow an order. Therefore, the stitched image obtained need not be panoramic. We train the recent Pythia [27] model on the stitched images and report performance in Table 2.

Video VQA. To highlight the differences between Video VQA and ISVQA, we adapt the recent state-of-the-art method HME-VideoQA [8]. This model consists of heterogeneous memory module which can potentially learn global context information. We consider images in the image set as frames of a video. Note that, the images in an image set in ISVQA do not necessarily constitute the frames of a video. Therefore, it is reasonable to expect such Video VQA methods to not provide any advantages over our baselines.

Using these baselines, we show that ISVQA is not a trivial extension of VQA. Solving ISVQA requires development of specialized methods.

Transformer-Based Method. We utilize the power of transformers and adapt the LXMERT model [28] to both cross-modality and cross-image scenarios. The transformer can summarize the relevant information within an image set and also model the across-image finer-grained dependencies. Here, we briefly described the original LXMERT model and then describe our modifications.

LXMERT learns cross-modality representations between regions in an image and sentences. It first uses separate visual and language encoders to obtain visual and semantic embeddings. The visual encoder consists of several self-attention sub-layers which help in encoding the relationships between objects. Similarly, the language encoder consists of multiple self-attention sub-layers and feed-forward sub-layers which provide a semantic embedding for the sentence or question. The visual and semantic embeddings are then used to attend to each other via cross-attention sub-layers. This helps the LXMERT model learn final visual and language embeddings which can tightly couple the information from visual and semantic domains. These coupled embeddings can be seen as the joint representations of the image and sentence and are used for inference.

Instead of using features from only a single image as input to the object-relationship encoder, we propose to use the region features from each image in our image-set. As described above, we start by extracting p region proposals and the corresponding features from each of the n images in the image set. We pass the $p \times n$ region features as inputs to the object-relationship encoder in LXMERT. We note that this enables the our model to encode relationships between objects across different images.

Let us denote the image features as $R = [R_1; R_2; \ldots; R_n] \in \mathbb{R}^{pn \times d}$, where R_i are the region features obtained from I_i. We also have the corresponding position encodings of each region in the images $P = [P_1; P_2; \ldots; P_n] \in \mathbb{R}^{pn \times 4}$, where P_i contains the bounding box co-ordinates of the regions in I_i. We combine the region features and position encodings to obtain position-aware embeddings, $S \in \mathbb{R}^{pn \times d'}$, where $S = \text{LayerNorm}(\text{FC}(R)) + \text{LayerNorm}(\text{FC}(P))$. Within- and across-image object relationships are encoded by applying N_R layers of the object relationship encoder. The l-th layer can be represented as

$$x_l = \text{FC}(\text{FC}(\text{SelfAttention}(x_{l-1}))) \tag{3}$$

where, $x_0 = S$, and $X\,(=x_{N_R})$ is the final visual embedding of the object-relationship encoder.

Similarly, given the word embeddings of the question, E, and the index embeddings of each word in the question, $E' = \{\text{IdxEmbed}(1), \ldots, \text{IdxEmbed}(T)\}$, the index-aware word embedding of the i-th word is obtained as $H_i = \text{LayerNorm}(E_i + E_i^{'})$. Note that the index embedding, IdxEmbed, is just a projection of the position of the word to a vector using fully-connected layers. We apply a similar operation as Eq. 3 N_L times to the word embeddings $H = [H_1; H_2; \ldots; H_T] \in \mathbb{R}^{T \times d_q}$ to give the question embedding, L.

Finally, LXMERT consists of N_X cross-modality encoders stacked one-after-another. Each encoder consists of two operations: 1.) language to vision cross attention, $X = \text{FC}(\text{SelfAttention}(\text{CrossAttention}_{LV}(X, L)))$; and 2.) vision to language cross attention, $L = \text{FC}(\text{SelfAttention}(\text{CrossAttention}_{VL}(L, X)))$. The final output of the N_X encoders are used to predict the answer.

Evaluating Biases in the Datasets. We also evaluate the following prior-based baselines to reveal and understand the biases present in the datasets.
Naïve Baseline. The model always predicts the most frequent answer from the training set. For nuScenes, it always predicts "yes", while for Gibson it predicts "white". Ideally, this should set a minimum performance bar.
Hasty-Student Baseline. In this baseline, a model simply finds the most frequent answer for each type of question. In this case, we define a "question type" as the first two words of a question. For example, a hasty-student might always answer "one" for all "How many" questions. This is similar to the hasty-student baseline used in [19] (MovieQA).
Question-Only Baseline. In this model, we ignore the visual information and only use question text to train a model. Our implementation takes as input only the question embedding, q which is passed through several fully-connected layers to predict the answer distribution. This baseline is meant to reveal the language-bias present in the dataset.

5 Experiments

5.1 Human Performance

An ideal image-set question answering system should be able to reach at least the accuracy achieved by humans. We evaluate the human performance using the annotations with the standard VQA-accuracy metric described below. For the outdoor scenes dataset, humans obtain a VQA-accuracy of 91.88% and for the indoor scenes they obtain 88.80%. Comparing this with Table 2 shows that ISVQA is extremely challenging and requires specialized methods. The reason for the human performance being lower than 100% is that, in many cases an annotator has given an answer which is not exactly similar to the other two but is still semantically similar. For example, the majority answer might have been "black and white" but the third annotator answered "white and black".

5.2 Implementation Details

We start by using Faster R-CNN in Detectron to extract the region proposals and features R_i for each image. Each region feature is 2048-D and we use the top 100 region proposals from each image. To obtain the word-vector embeddings we use 300-D GloVe [23] vectors. The joint visual-question embedding, x is taken to be 5000-D. For evaluation, we use the VQA-Accuracy metric [4]. A predicted answer is given a score of one if it matches at least two out of the three annotations. If it matches only one annotations, it is given a score of 0.5. All of our VQA models are implemented in the Pythia framework [26] and are trained on two NVIDIA V100 GPUs for 22,000 iterations with a batch size of 32. The initial learning rate is warmed up to 0.01 in the first 1,000 iterations. The learning rate is dropped by a factor of 10 at iterations 12,000 and 18,000. For the HME-VideoQA model, we use the implementation provided by the authors. We train the model for 22,000

iterations with a batch size of 32 and a starting learning rate of 0.001. For the transformer-based model, we use $N_L = 9, N_R = 5, N_X = 5$. We use a batch-size of 32, learning rate of 0.00005, and we train the model for 20 epochs. All feature dimensions are kept the same as LXMERT.

5.3 Results

We report the VQA-accuracy for all methods in Table 2. The accuracy achieved by both of the VQA-based baselines is only around 50−54% and the Video VQA model achieves only 39.88% for the indoor dataset and 52.14% for the outdoor dataset. This highlights the need for advanced models for ISVQA.

Comparison Between Baselines. Table 2 shows that the naïve baseline reaches a VQA-Accuracy of only 8.6% for the indoor scenes dataset compared to 47.57% given by the Concatenate-Feature baseline and 50.53% given by the Stitched-Image baseline model. This shows that single-image VQA methods are not enough to overcome the challenges presented by ISVQA. On the other hand, our proposed transformer-based model performs the best for both indoor and outdoor scenes out-performing other models by over 10%.

Language Biases. Recent works (*e.g.* [2]) show that high performance in VQA could be achieved using only the language components. Deep networks can easily exploit biases in the datasets to find short-cuts for answering questions. We observe that most VQA-based baselines perform much better than the question-only baseline. This shows that the ISVQA datasets are less biased compared to many existing VQA datasets [2] and validates the utility of developing ISVQA models that can utilize both the visual and language components simultaneously.

Table 2. Results for both indoor and outdoor datasets.

	Method	VQA-accuracy (%)	
		Gibson	nuScenes
Prior-based baselines	Naïve	8.61	22.46
	Hasty-student	27.22	41.65
	Question-Only	40.26	46.06
Approaches	Video-VQA	39.88	52.14
	Concatenate-feature	47.57	53.66
	Stitched-image	50.53	54.32
	Transformer-based	61.58	64.91
Human performance		88.80	91.88

Performance by Question Type. Figure 8 shows the accuracy bar-chart of our single-image VQA-based baselines for various types of questions. Using this chart, we have the following observations and hypotheses:

Single-Image VQA Baselines can Predict Single-Object Attributes. Both baseline models can answer questions about colors of single objects well (black and gray bars). This is expected because no cross-image dependency is needed.

General Cases May Need Cross-image Inference. A large portion of questions involve multiple objects, which may appear in different images. The two VQA baselines using simple attention do not perform well on such questions. The across-image transformer-based approach performs much better.

Stitched-Image Captures Cross-image Dependency Better. The Stitched-Image baseline allows direct pooling from regions in all images, which may capture across-image dependency better. It also outperforms the Concatenate-Feature baseline for most question types, except for the counting questions. The Stitched-Image cannot avoid double counting. The transformer-based approach has all the advantages of the Stitched-Image and can do more sophisticated inference.

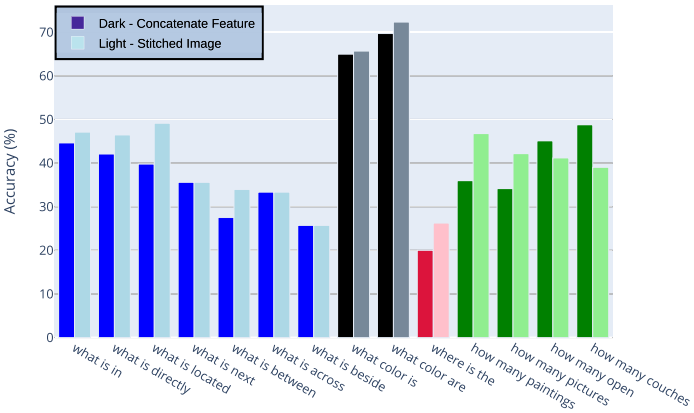


Fig. 8. Performance of the two VQA-based baselines for different types of questions for the combined Gibson test set. Dark colors represent the performance for Concatenate-Feature baseline and light colors for Stitched Image baseline. **Blue** is used for geometric relationship questions, **green** for counting questions, **red** for location, and **black** for color questions. We notice that the VQA-based baselines are able to answer simple questions like those about colors of single objects very well. However, questions involving spatial reasoning between objects in one image or across images are extremely challenging for such methods. (Color figure online)

6 Conclusion and Discussion

We proposed the new task of image-set visual question answering (ISVQA). This task can lead to new research challenges, such as language-guided cross-image attentions and reasoning. To establish the ISVQA problem and enable its

research, we introduced two ISVQA datasets for indoor and outdoor scenes. Large-scale annotations were collected for questions and answers with novel ways to present the scene to the annotators. We performed bias analysis of the datasets to set up performance lower bounds. We also extended a single-image VQA method to two simple attention-based baseline models and showed the performance of state-of-the-art Video VQA model. Their limited performance reflects the unique challenges of ISVQA, which cannot be solved trivially by the capabilities of existing models. Approaches for solving the ISVQA problem may need to pass information across images in a sophisticated way, understand the scene behind the image set, and attend the relevant images. Another potential direction could be to create explicit maps of the scenes. However, depending on the complexity of the scene, different techniques might be required to explicitly construct a coherent map. Where such maps can be obtained accurately, reconstruction-based ISVQA solutions can be more accurate than the baselines. Meanwhile, humans do not have to do exact scene reconstruction to answer questions. So, in this paper, we have focused on methods that can model across-image dependencies implicitly.

References

1. Acharya, M., Kafle, K., Kanan, C.: TallyQA: answering complex counting questions. In: AAAI Conference on Artificial Intelligence (2019)
2. Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. In: Empirical Methods in Natural Language Processing (2016)
3. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Conference on Computer Vision and Pattern Recognition (2018)
4. Antol, S., et al.: VQA: Visual question answering. In: International Conference on Computer Vision (2015)
5. Caesar, H., et al.: nuScenes: a multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
6. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: Conference on Computer Vision and Pattern Recognition (2018)
7. Desta, M.T., Chen, L., Kornuta, T.: Object-based reasoning in VQA. In: Winter Conference on Applications of Computer Vision (2018)
8. Fan, C., Zhang, X., Zhang, S., Wang, W., Zhang, C., Huang, H.: Heterogeneous memory enhanced multimodal attention model for video question answering. In: Conference on Computer Vision and Pattern Recognition (2019)
9. Gao, P., et al.: Dynamic fusion with intra- and inter-modality attention flow for visual question answering. In: Conference on Computer Vision and Pattern Recognition (2019)
10. Gao, P., You, H., Zhang, Z., Wang, X., Li, H.: Multi-modality latent interaction network for visual question answering. In: International Conference on Computer Vision (2019)
11. Gordon, D., Kembhavi, A., Rastegari, M., Redmon, J., Fox, D., Farhadi, A.: IQA: visual question answering in interactive environments. In: Conference on Computer Vision and Pattern Recognition (2018)

12. Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: International Journal of Computer Vision (2019)
13. Gurari, D., et al.: Vizwiz grand challenge: answering visual questions from blind people. In: Conference on Computer Vision and Pattern Recognition (2018)
14. Hudson, D.A., Manning, C.D.: GQA: a new dataset for real-world visual reasoning and compositional question answering. In: Conference on Computer Vision and Pattern Recognition (2019)
15. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In: Conference on Computer Vision and Pattern Recognition. Institute of Electrical and Electronics Engineers Inc. (2017)
16. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: a diagnostic dataset for compositional language and elementary visual reasoning. In: Conference on Computer Vision and Pattern Recognition (2017)
17. Krishna, R., et al.: Visual genome: connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
18. Lei, J., Yu, L., Berg, T.L., Bansal, M.: TVQA+: spatio-temporal grounding for video question answering. arXiv preprint arXiv:1904.11574 (2019)
19. Liang, J., Jiang, L., Cao, L., Li, L.J., Hauptmann, A.: Focal visual-text attention for visual question answering. In: Conference on Computer Vision and Pattern Recognition (2018)
20. Lin, X., Parikh, D.: Don't just listen, use your imagination: leveraging visual common sense for non-visual tasks. In: Conference on Computer Vision and Pattern Recognition (2015)
21. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances In Neural Information Processing Systems (2016)
22. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: OK-VQA : a visual question answering benchmark requiring external knowledge. In: Conference on Computer Vision and Pattern Recognition (2019)
23. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Conference on Empirical Methods in Natural Language Processing (2014)
24. Santoro, A., et al.: A simple neural network module for relational reasoning. Advances in Neural Information Processing Systems (2017)
25. Savva, M., et al.: Habitat: a platform for embodied AI research. arXiv preprint arXiv:1904.01201 (2019)
26. Singh, A., et al.: Pythia-a platform for vision & language research. In: SysML Workshop, NeurIPS (2018)
27. Singh, A., et al.: Towards VQA models that can read. In: Conference on Computer Vision and Pattern Recognition (2019)
28. Tan, H., Bansal, M.: Lxmert: learning cross-modality encoder representations from transformers. In: Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (2019)
29. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: MovieQA: understanding stories in movies through question-answering. In: Conference on Computer Vision and Pattern Recognition (2016)

30. Trott, A., Xiong, C., Socher, R.: Interpretable counting for visual question answering. In: International Conference on Learning Representations (2018)
31. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson ENV: real-world perception for embodied agents. In: Conference on Computer Vision and Pattern Recognition (2018)
32. Yagcioglu, S., Erdem, A., Erdem, E., Ikizler-Cinbis, N.: Recipeqa: a challenge dataset for multimodal comprehension of cooking recipes. arXiv preprint arXiv:1809.00812 (2018)
33. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7W: grounded question answering in images. In: Conference on Computer Vision and Pattern Recognition (2016)