VQS: Linking Segmentations to Questions and Answers for Supervised Attention in VQA and Question-Focused Semantic Segmentation

Chuang Gan[1] Yandong Li[2] Haoxiang Li[3] Chen Sun[4] Boqing Gong[2]
[1]IIIS, Tsinghua University, China [2]CRCV, University of Central Florida, USA
[3]Adobe Research, USA [4]Google Research, USA

Abstract

Rich and dense human labeled datasets are among the main enabling factors for the recent advance on vision-language understanding. Many seemingly distant annotations (e.g., semantic segmentation and visual question answering (VQA)) are inherently connected in that they reveal different levels and perspectives of human understandings about the same visual scenes — and even the same set of images (e.g., of COCO). The popularity of COCO correlates those annotations and tasks. Explicitly linking them up may significantly benefit both individual tasks and the unified vision and language modeling.

We present the preliminary work of linking the instance segmentations provided by COCO to the questions and answers (QAs) in the VQA dataset, and name the collected links visual questions and segmentation answers (VQS). *They transfer human supervision between the previously separate tasks, offer more effective leverage to existing problems, and also open the door for new research problems and models. We study two applications of the VQS data in this paper: supervised attention for VQA and a novel question-focused semantic segmentation task. For the former, we obtain state-of-the-art results on the VQA real multiple-choice task by simply augmenting the multilayer perceptrons with some attention features that are learned using the segmentation-QA links as explicit supervision. To put the latter in perspective, we study two plausible methods and compare them to an oracle method assuming that the instance segmentations are given at the test stage.*

1. Introduction

Connecting visual understanding with natural language has received extensive attentions in recent years. We have witnessed the resurgence of image captioning [42, 29, 18, 6, 40, 4, 47, 12, 33, 10] which is often addressed by jointly modeling visual and textual content with deep neural networks. However, image captions tend to be diverse and subjective — it is hard to evaluate the quality of captions generated by different algorithms [7, 41, 1], and tend to miss subtle details — in training, the models may be led to capturing the scene-level gist rather than fine-grained entities. In light of the premises and demerits of image captioning, visual question answering (VQA) [3, 51, 36, 13] and visual grounding [35, 15, 37, 28, 16, 43, 50] are proposed, in parallel, to accommodate automatic evaluation and multiple levels of focus on the visual entities (e.g., scene, object, activity, attribute, context, relationships, etc.).

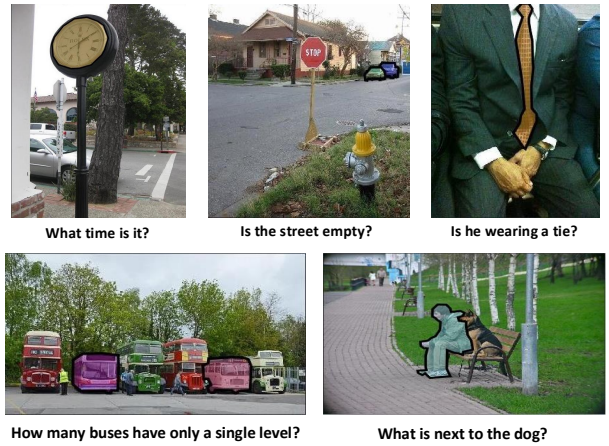


Figure 1. Taking as input an image and a question about the image, an algorithm for the question-focused semantic segmentation is desired to generate some segmentation mask(s) over the entities in the image that can visually answer the question.

Rich and dense human annotated datasets are arguably the main "enabler", among others, for this line of exciting works on vision-language understanding. COCO [25] is especially noticeable among them. It contains mainly classical labels (e.g., segmentations, object categories and instances, key points, etc.) and image captions. Many research groups have then collected additional labels of the COCO images for a variety of tasks. Agrawal et al. crowdsourced questions and answers (QAs) about a subset of the COCO images and abstract scenes [3]. Zhu et al. collected seven types of QAs in which the object mentions are as-

Code and data: https://github.com/Cold-Winter/vqs.

sociated with bounding boxes in the images [51]. Mao et al. [28] and Yu et al. [50] have users to give referring expressions that each pinpoints a unique object in an image. The Visual Genome dataset [22] also intersects with COCO in terms of images and provides dense human annotations, especially scene graphs.

These seemingly distant annotations are inherently connected in the sense that they reveal different perspectives of human understandings about the same COCO images. The popularity of COCO could strongly correlate those annotations — and even tasks. Explicitly linking them up, as we envision, can significantly benefit both individual tasks and unified vision-language understanding, as well as the corresponding approaches and models. One of our contributions in this paper is to initiate the preliminary work on this. In particular, we focus on linking the segmentations provided by COCO [25] to the QAs in the VQA dataset [3]. Displaying an image and a QA pair about the image, we ask the participant to choose the segmentation(s) of the image in order to visually answer the question.

Figure 1 illustrates some of the collected "visual answers". For the question "What is next to the dog?", the output is supposed to be the segmentation mask over the man. For the question "What time is it?", the clock should be segmented out. Another intriguing example is that the cars are the desired segmentations to answer "Is this street empty?", providing essential visual evidence for the simple text answer "no". Note that while many visual entities could be mentioned in a question, we only ask the participants to choose the target segmentation(s) that visually answer the question. This simplifies the annotation task and results in higher agreement between participants. Section 2 details the annotation collection process and statistics.

Two related datasets. Das et al. have collected some human attention maps for the VQA task [5]. They blur the images and then ask users to scratch them to seek visual cues that help answer the questions. The obtained attention maps are often small, revealing meaningful parts rather than complete objects. The object parts are also mixed with background areas and with each other. As a result, the human attention maps are likely less accurate supervision for the attention based approaches to VQA than the links we built between segmentations and QAs. Our experiments verify this hypothesis (cf. Section 3). While bounding boxes are provided in Visual7W [51] for object mentions in QAs, they do not serve for the purpose of directly answering the questions except for the "pointing" type of questions. In contrast, we provide direct visual answers in the form of segmentations to more question types.

1.1. Applications of the segmentation-QA links

We call the collected links between the COCO segmentations [25] and QA pairs in the VQA dataset [3] *visual questions and segmentation answers (**VQS**)*. Such links transfer human supervision between the previously separate tasks, i.e., semantic segmentation and VQA. They enable us to tackle existing problems with more effective leverage than before and also open the door for new research problems and models for the vision-language understanding. We study two applications of our VQS dataset in this paper: supervised attention for VQA and a novel question-focused semantic segmentation (QFSS) task. For the former, we obtain state-of-the-art results on the VQA real multiple-choice task by simply augmenting the multilayer perceptrons (MLP) of [17] with attention features.

1.1.1 Supervised attention for VQA

VQA is designed to answer natural language questions about images in the form of short texts. The attention scheme is often found useful for VQA, by either attending particular image regions [48, 46, 45, 26, 24] or modeling object relationships [2, 27]. However, lacking explicit attention annotations, the existing methods opt for latent variables and use indirect cues (e.g., text answers) for inference. As a result, the machine-generated attention maps are poorly correlated with human attention maps [5]. This is not surprising since latent variables hardly match semantic interpretations due to the lack of explicit training signals; similar observations exist in other studies, e.g., object detection [8], video recognition [11] and text processing [49].

These phenomena highlight the need for explicit links between the visual and text answers, realized in this work as VQS. We show that, by supervised learning to attend different image regions using the collected segmentation-QA links, we can boost the simple MLP model [17] to very compelling performance on the VQA real multi-choice task.

1.1.2 Question-focused semantic segmentation (QFSS)

In addition to the supervised attention for better tackling VQA, VQS also enables us to explore a novel question-focused semantic segmentation (QFSS) task.

Since VQA desires only text answers, there exist potential shortcuts for the learning agent, e.g., to generate correct answers without accurately reasoning the locations and relations of different visual entities. While visual grounding (VG) avoids the caveat by placing bounding boxes [35, 37, 28, 16] or segmentations [15] over the target visual entities, the scope of the text expressions in existing VG works is often limited to the visual entities present in the image. In order to bring together the best of VQA and VG, we propose the QFSS task, whose objective is to produce pixel-wise segmentations in order to visually answer the questions about images. It effectively borrows the versatile questions from VQA and meanwhile resembles the design of VG in terms of the pixel-wise segmentations as



Figure 2. Some typical examples in our VQS dataset. From the left to right, the underlying tasks are respectively about object localization, semantic segmentation, understanding object relationships, fine-grained activity localization, and commonsense reasoning.

the desired output.

Given an image and a question about the image, we propose a mask aggregation approach to generating a segmentation mask as the visual answer. Since QFSS is a new task, to put it in perspective, we not only compare the proposed approach to competing baselines but also study an upper-bound method by assuming all instance segmentations are given as oracles at the test stage.

Hu et al.'s work [15] is the most related to QFSS. They learn to ground text expressions in the form of image segmentations. Unlike the questions used in this work that are flexible to incorporate commonsense and knowledge bases, the expressive scope of the text phrases in [15] is often limited to the visual entities in the associated images.

The rest of this paper is organized as follows. Section 2 details the collection process and analyses of our VQS data. In section 3, we show how to use the collected segmentation-QA links to learn supervised attention features and to augement the existing VQA methods. In section 3.2, we study a few potential frameworks to address the new question-focused semantic segmentation task. Section 4 concludes the paper.

2. Linking image segmentations to text QAs

In this section, we describe in detail how we collect the links between the semantic image segmentations and text questions and answers (QAs). We build our work upon the images and instance segmentation masks in COCO [25] and the QAs in the VQA dataset [3]. The COCO images are mainly about everyday scenes that contain common objects in their natural contexts, accommodating complex interactions and relationships between different visual entities. To avoid trivial links between the segmentations and QA pairs, we only keep the images that contain at least three instance segmentations in this work. The questions in VQA [3] are diverse and comprehensively cover various parts of an image, different levels of semantic interpretations, as well as commonsense and knowledge bases.

Next, we elaborate the annotation instructions and provide some analyses about the collected dataset.

2.1. Annotation instructions

We display to the annotators an image, its instance segmentations from the COCO dataset, and a QA pair about the image from the VQA dataset. The textual answer is given in addition to the question, to facilitate the participants to choose the right segmentations as the visual answer. Here are the instructions we give to the annotators (cf. the supplementary materials for the GUI):

- Please choose the right segmentation(s) in the image to answer the question. Note that the text answer is shown after the question.
- A question about the target entities may use other entities to help refer to the target. Choose the target entities only and nothing else (e.g., the purse for "What is on the bench next to woman?" in Figure 2(g)).
- A question may be about an activity. Choose all visual entities involved in the activity. Taking Figure 2(j) for instance, choose both the person and motorcycle for the question "what is the person doing?".
- Sometimes, in addition to the image regions covered by the segmentation masks, you may need other re-

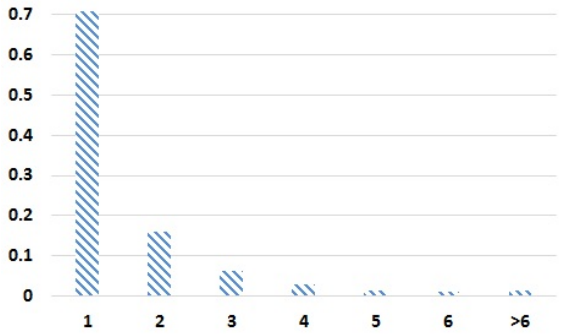


Figure 3. Distribution of the number of segmentations per question-image pair.



Figure 4. The distribution of question types in the VQS dataset.

gions to answer the question. To include them, please draw tight bounding box(es) over the region(s).

- For the "How many" type of questions, the number of selected segments (plus bounding boxes) must match the answer. If the answer is greater than three, it is fine to put one bounding box around the entities being asked in the question.
- Please tick the black button under the question, if you think the question has to be answered by the full image.
- Please tick the gray button under the question, if you feel the question is ambiguous, or if you are not sure which segment/region to select to answer the question.

Occasionally, the visual answer is supposed to be only part of an instance segment given by COCO. For instance, the McDonald logo answers "What fast food restaurant can be seen?" in Figure 2(o) but there is no corresponding segmentation for the logo in COCO. Another example is the region of the ring that answers "Is the woman wearing ring?" (cf. Figure 2(c)). For these cases, we ask the participants to draw tight bounding boxes around them. If we segment them out instead, a learning agent for QFSS may never be able to produce the right segmentation for them unless we include more training images in the future, since these regions (e.g., McDonald logo, ring) are very fine-grained visual entities and show up only a few times in our data collection process.

Quality control. We tried AMTurk to collect the annotations at the beginning. While the inter-annotator agreement is high on the questions about objects and people, there are many inconsistent annotations for the questions referring to activities (e.g., "What sport is played?"). Besides, the AMTurk workers tend to frequently tick the black button, which says the full image is the visual answer, and the gray button, which tells the question is ambiguous. To obtain higher-quality annotations, we instead invited 10 undergraduate and graduate volunteers and trained them in person (we include some slides used for the training in the supplementary materials). To further control the annotation quality, each annotator was asked to finish an assignment of 100 images (around 300 question-answer pairs) before we met with them again to look over their annotations together — all the volunteers were asked to participate the discussion and jointly decide the expected annotations for every question. We also gradually increased the hourly payment rate from $12/hr to $14/hr as incentives for high-quality work.

2.2. Tasks addressed by the participants

Thanks to the rich set of questions collected by Agrawal *et al.* [3] and the complex visual scenes in COCO [25], the participants have to parse the question, understand the visual scene and context, infer the interactions between visual entities, and then pick up the segmentations that answer the questions. We find that many vision tasks may play roles in this process. Figure 2 shows some typical examples to facilitate the following discussion.

Object detection. Many questions directly ask about the properties of some objects in the images. In Figure 2(b), the participants are supposed to identify the *cup* in the cluttered scene for the question "What color is the coffee cup?".

Semantic segmentation. For some questions, the visual evidence to answers is best represented by semantic segmentations. Take Figures 2(j) and (k) for instance. Simply detecting the rider and/or the bike would be inadequate in expressing their spatial interactions.

Spatial relationship reasoning. A question like "What is on the bench next to the woman?" (Figure 2(g)) poses a challenge to the participants through the spatial relationship between objects including bench, woman, and the answer purse. Figure 2(i) is another example in this realm.

Fine-grained activity recognition. When the question is about an activity (*e.g.*, "What sport is being played?" in Figure 2(l)), we ask the participants to label all the visual entities (*e.g.*, person, tennis racket , and tennis ball) involved. In other words, they are expected to spot the fine-grained details of the activity.

Commonsense reasoning. Commonsense knowledge can help the participants significantly reduce the search space for the visual answers, e.g., the clock to answer "What time is it?" in Figure 1, and the McDonald logo to answer "What fast food restaurant can be seen?" in Figure 2(o).

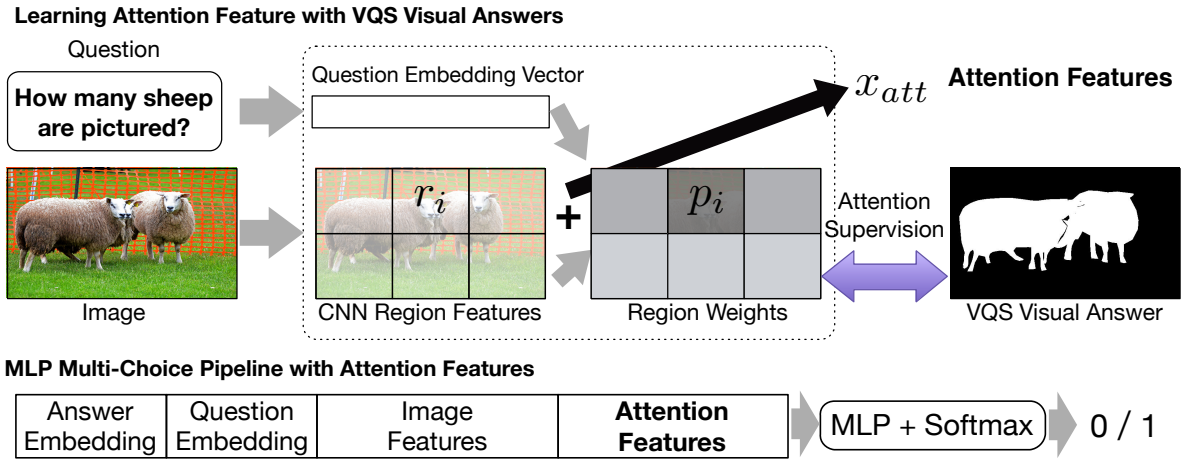


Figure 5. Supervised attention for VQA. To learn the attention features for each question-image pair, we use the corresponding segmentation mask as supervision to train the attention network. After that, we augment the MLP model [17] by the attention features.

2.3. Data statistics

After collecting the annotations, we remove the question-image pairs for which the users selected the black buttons (full image) or gray buttons (unknown) to avoid trivial and ambiguous segmentation-QA links, respectively. In total, we keep 37,868 images, 96,508 questions, 108,537 instance segmentations, and 43,725 bounding boxes. In the following, we do not differentiate the segmentations from the bounding boxes for the ease of presentation and also for the sake that the bounding boxes are tight, small, and much fewer than the segmentations.

Figure 3 counts the distribution of the possible number of instance segmentations selected per image in response to a question. Over 70% of questions are answered by one segmentation. On average, each question-image pair has 6.7 candidate segmentations, among which 1.6 are selected by the annotators as the visual answers.

In Figure 4, we visualize the distribution of question types. The most popular type is the "What" questions (46%). There are 31,135 "is/are" and "does/do" questions (32.1%). Note that although the textual answers to them are simply yes or no, in VQS, we ask the participants to explicitly demonstrate their understanding about the visual content by producing the semantic segmentation masks. In the third column of Table 3, we show the average number of segmentations chosen by the users out of the average number of candidates for each of the question types.

3. Applications of VQS

The user linked visual questions and segmentations, where the latter visually answers the former, are quite versatile. They offer better leverage than before for at least two problems, i.e., supervised attention for VQA and question-focused semantic segmentation (QFSS).

3.1. Supervised attention for VQA

VQA is designed to answer natural language questions about an image in the form of short texts. We conjecture that a learning agent can produce more accurate text answers given the privileged access to the segmentations that are user linked to the QAs in training. To verify this point, we design a simple experiment to augment the MLP model in [17]. The augmented MLP significantly improves upon the plain version and gives rise to state-of-the-art results on the VQA real multiple-choice task [3].

Experiment setup. We conduct experiments on the VQA Real Multiple Choices [3]. The dataset contains 248,349 questions for training, 121,512 for validation, and 244,302 for testing. Each question has 18 candidate answer choices and the learning agent is required to figure out the correct answer among them. We evaluate our results following the metric suggested in [3].

MLP for VQA Multiple Choice. Since the VQA multiple-choice task supplies candidate answers to each question, Jabri et al. propose to transform the problem to a stack of binary classification problems [17] and solve them by the multilayer perceptrons (MLP) model:

$$y = \sigma(\boldsymbol{W}_2 \max(0, \boldsymbol{W}_1 \boldsymbol{x}_{iqa}) + b) \quad (1)$$

where $\boldsymbol{x}_{iqa}$ is the concatenation of the feature representations of an image, a question about the image, and a candidate answer, and $\sigma(\cdot)$ is the sigmoid function. The hidden layer has 8,096 units and a ReLU activation. This model is very competitive, albeit simple.

3.1.1 Augmenting MLP by supervised attention

We propose to augment the MLP model by richer feature representations of the questions, answers, images, and es-

Table 1. Comparison results on both VQA TestDev and Standard for the Real Multiple Choice task.

Method	Dev	Standard
Two-layer LSTM [3]	62.7	63.1
Region selection [38]	62.4	62.4
DPPNet [32]	62.5	62.7
MCB [9]	65.4	–
Co-Attention [26]	65.9	66.1
MRN [19]	66.3	66.3
MLB [20]	–	68.9
MLP + ResNet [17]	67.4	–
MLP + ResNet +Atten.	68.9	–
MLP + Attri.	68.4	–
MLP + Attri. + Atten.	**69.5**	**69.8**
10 ensemble models	**70.5**	**70.5**

pecially by the supervised attention features detailed below.

Question and answer features $\boldsymbol{x}_q$&$\boldsymbol{x}_a$. For a question or answer, we represent it by averaging the 300D word2vec [30] vectors of the constituent words, followed by the l_2 normalization. This is the same as in [17].

Image features $\boldsymbol{x}_i$. We extract two types of features from an input image: ResNet [14] pool5 activation and attribute features [44], where the latter is the attribute detection scores. We implement an attribute detector by revising the output layer of ResNet. Particularly, given $C = 256$ attributes, we impose a sigmoid function for each attribute and then train the network using the binary cross-entropy loss. The training data is obtained from the COCO image captions [25]. We keep the most frequent 256 words as the attributes after removing the stop words.

Attention features $\boldsymbol{x}_{att}$. We further concatenate attention features $\boldsymbol{x}_{att}$ to the original input $\boldsymbol{x}_{iqa}$. The attention features are motivated by the weighted combination of image regional features and question features in [48, eq. (22)], where the the non-negative weight $p_i = f(Q, \{\boldsymbol{r}_i\})$ for each image region is a function of the question Q and regional features $\{\boldsymbol{r}_i\}$. We borrow the network architecture as well as code implementation from Yang et al. [48, Section 3.3] for this function, except that we train this network by a cross-entropy loss to match the weights $\{p_i\}$ to the "groundtruth" attentions derived from the segmentations in our VQS dataset. In particular, we down-sample the segmentation map associated with each question-image pair to the same size as the number of image regions, and then l_1 normalize it to a valid probability distribution. By training the network to match the weights $p_i = f(Q, \{\boldsymbol{r}_i\})$ toward such attentions, we enforce larger weights for the regions that correspond to the user selected segmentations.

The upper panel of Figure 5 illustrates the process of extracting the attention features, and the bottom panel shows the MLP model [17] augmented with our attention features for the VQA real multiple-choice task.

Table 2. Comparison results on VQA TestDev Real Multiple Choice task.

Method	Y/N	Num.	Others	All
Plain MLP [17]	80.11	38.88	64.17	67.49
HAT [5]	80.19	39.34	64.92	68.42
Bounding boxes	80.15	38.9	65.54	68.65
VQS	**80.60**	**39.41**	**65.73**	**68.94**

3.1.2 Experimental results

Table 1 reports the comparison results of the attention features augmented MLP with several state-of-the-art methods on the VQA real multiple-choice task. We mainly use the Test Dev for comparison. After determining our best single and ensemble models, we also submit them to the evaluation server to acquire the results on Test Standard.

First of all, we note that there is an 1.5% absolute improvement over the plain MLP model (MLP + ResNet) by simply augmenting it using the learned attention features (MLP + ResNet + Atten.). Second, the attribute features for the images are actually quite effective. We gain 1.0% improvement over the plain MLP by replacing the ResNet image features with the attribute features (cf. the row of MLP + Attri. vs. MLP + ResNet). Nonetheless, by appending attention features to MLP + Attri., we can still observe 1.1% absolute gain. Finally, with an ensemble of five MLP + ResNet + Atten. models and five MLP + Attri. + Atten. models, our submission to the evaluation server was ranked to the second on Test Standard for the VQA real multiple-choice task, as of the paper submission date.

3.1.3 What is good supervision for attention in VQA?

In this section, we contrast the VQS data to the human attention maps (HAT) [5] and bounding boxes that are placed tightly around the segmentations in VQS. The comparison results, reported in Table 4, are evaluated on the TestDev dataset of VQA Real Multiple Choice. We can see that the segmentaitons linked to QAs give rise to a little better results than bounding boxes, which further outperform HAT. These confirm our conjecture that HAT might be suboptimal for the supervised learning of attentions in VQA, since they reveal usually small parts of objects and contain large proportions of background. However, we believe it remains interesting to examine VQS for more generic attention-based VQA models [48, 46, 45, 26, 24, 2, 27].

In the supplementary materials, we describe the detailed implementation for the ensemble model. We also present additional results studying how different resolutions of the segmentation masks influence the VQA results.

3.2. Question-focused semantic segmentation

This section explores a new task, question-focused semantic segmentation (QFSS), which is feasible thanks to the collected VQS that connects two previously separate tasks (i.e., segmentations and VQA). Given a question about

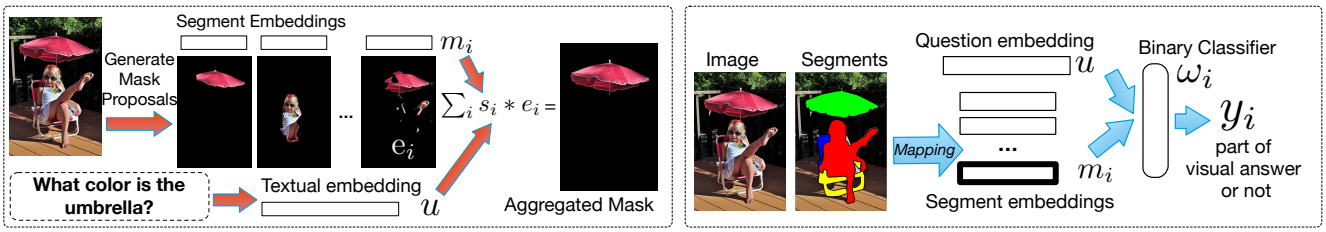


Figure 6. Mask aggregation method for QFSS (left) and the method to estimate its upper bound performance (right).

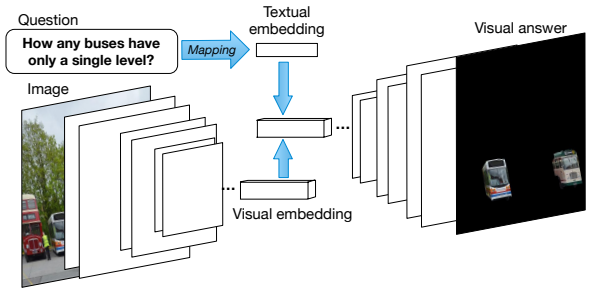


Figure 7. Our DeconvNet baseline for QFSS.

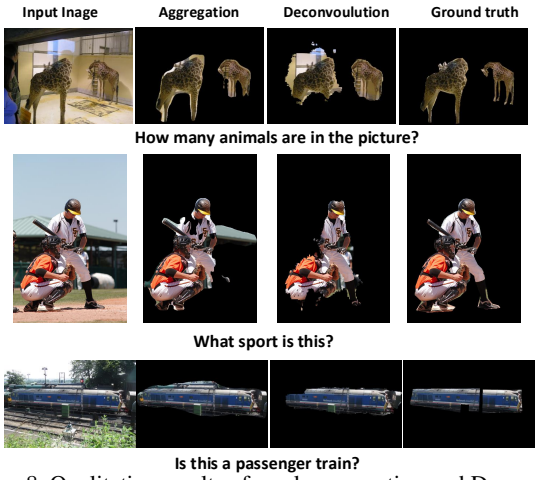


Figure 8. Qualitative results of mask aggregation and DeconvNet.

an image, QFSS expects the learning agent to output a visual answer by semantically segment the right visual entities out of the image. It is designed in a way similarly to the segmentation from natural language expressions [15], with possible applications to robot vision, photo editing, etc.

In order to put the new task in perspective, we propose a mask aggregation approach to QFSS, study a baseline, and also investigate an upper bound method by assuming all instance segmentations are given as oracles at the test stage.

3.2.1 Mask aggregation for QFSS

We propose a mask aggregation approach to tackling QFSS. The modeling hypothesis is that the desired output segmentation mask can be composed from high-quality segmentation proposals. In particular, we use $N = 25$ segmentation proposals $e_1, e_2, \cdots, e_N$ generated by SharpMask [34] given an image. Each proposal is a binary segmentation mask of the same size as the image.

We then threshold a convex combination of these masks $E = \sum_i s_i e_i$ as the final output in response to a question-image pair, where the i-th combination coefficient s_i is determined by the question features $\boldsymbol{x}_q$ and the representations $\boldsymbol{z}_i$ of the i-th segmentation proposal through a softmax function, i.e., $s_i = \text{softmax}(\boldsymbol{x}_q^T A \boldsymbol{z}_i)$. We learn the model parameters A by minimizing an l_2 loss between the the user selected segmentations $E^\star$ and the model generated segmentation mask E. Our current model is "shallow" but it is straightforward to make it deep, e.g., by stacking its output with the original input following the prior practice (e.g., memory network [45] and stacked attention network [48]).

An oracle upper bound. We devise an upper bound to the proposed method by 1) replacing the segmentation proposals with all the instance segmentations released by MS COCO, assuming they are available as oracles at testing, and 2) using a binary classifier to determine whether or not an instance segmentation should be included into the visual answer. The results can be considered an upper bound for our approach because the segmentations are certainly more accurate than the machine generated proposals, and the binary classification is arguably easier to solve than aggregating multiple masks. We re-train the MLP (eq. 1) for the binary classifier here; it now takes as input the concatenated features of a segmentation and a question.

Figure 6 depicts the proposed approach and the upper-bound method with a concrete question-image example.

A baseline using deconvolutional network. Finally, we study a competitive baseline which is motivated by the text-conditioned FCN [15]. As Figure 7 shows, it contains three components, a convolutional neural network (CNN) [23], a deconvolutional neural network (DeconvNet) [31], and a question embedding to attend the feature maps in CNN. All the images are resized to 224×224. The convolutional and deconvolutional nets follow the specifications in [31]. Namely, a VGG-16 [39] is trimmed till the last convolutional layer, followed by two fully connected layers, and then mirrored by DeconvNet. For the input question, we use an embedding matrix to map it to the same size as the feature map of the last convolutional layer. The question embedding is then element-wsie multiplied with the feature map. We train the network with an l_2 loss between the output mask and the groundtruth segmentation mask.

Table 3. Comparison results on QFSS (evaluated by IOU, the higher the better). For the question representations, we consider the bag-of-words features (B) and the word embedding based features (W).

Type	Num.	#seg ans/candts	Aggre. (B)	Aggre. (W)	DeconvNet (B)	DeconvNet (W)	Upper
All	14875	1.6/6.1	**0.3256**	0.3174	0.2687	0.2979	0.5709
does/do	561	1.6/6.0	0.3294	**0.3321**	0.2751	0.3297	0.5346
how many	1814	2.2/6.3	**0.3697**	0.3645	0.3147	0.3370	0.6394
is/are	4238	1.7/5.9	**0.3672**	0.3573	0.3061	0.3548	0.6169
what color	1631	1.1/6.0	**0.2596**	0.2568	0.1940	0.1677	0.5557
what is	2464	1.3/5.9	**0.2472**	0.2328	0.2030	0.2003	0.4987
what (other)	2722	1.6/6.1	**0.3332**	0.3235	0.2556	0.2809	0.5482
where	433	1.4/6.2	0.1996	**0.2040**	0.1716	0.1896	0.5707
which	202	1.4/5.9	**0.2419**	0.2339	0.1695	0.2012	0.4504
who	144	1.3/5.9	**0.2573**	0.2527	0.2004	0.2164	0.2912
why	124	1.9/6.3	0.3453	**0.3594**	0.2430	0.2917	0.4781
others	542	1.6/6.1	**0.3578**	0.3354	0.3097	0.3534	0.5267

3.2.2 Experiments on QFSS

Features. In addition to representing the questions using the word embedding features $\boldsymbol{x}_q$ as in Section 3.1.1, we also test the bag-of-words features. For each instance segmentation or proposal, we mask out all the other pixels in the image with 0's and then extract its features from the last pooling layer of a ResNet-152 [14].

Dataset Split. The SharpMask we use is learned from the training set of MS COCO. Hence, we split our VQS data in such a way that our test set does not intersect with the training set for SharpMask. Particularly, we use 26,995 images and correspondingly 68,509 questions as our training set. We split the remaining images and questions to two parts: 5,000 images and associated questions for validation, and 5,873 images with 14,875 questions as the test set.

Results. Table 3 reports the comparison results on QFSS, evaluated by intersection-over-union (IOU). In addition, the first three columns are about the number of different types of questions and the average numbers of user selected segmentations per question type. On average, more than one segmentations are selected for any of the question types.

First of all, we note that the proposed mask aggregation outperforms the baseline DeconvNet and yet is significantly worse than its upper bound method. The mask aggregation is superior over DeconvNet partially because it has actually used extra supervised information beyond our VQS data; namely, the SharpMask is trained using all the instance segmentations in the training set of MS COCO. The upper bound results indicate there is still large room for the mask aggregation framework to improve; one possibility is make it deep in the future work.

Besides, we find that the two question representations, bag-of-wrods (B) and word embedding (W), give rise to distinguishable results for either mask aggregation or DeconvNet. This observation is intriguing since it implies that the QFSS task is responsive to the question representation schemes. It is thus reasonable to expect that QFSS will both benefit from and advance the progress on joint vision and language modeling methods.

Finally, Figure 8 shows some qualitative segmentation results. Note the two separate instance segmentations in the first row that visually answer the "How many" question.

4. Conclusion

In this paper, we propose to link the instance segmentations provided by COCO [25] to the questions and answers in VQA [3]. The collected links, named visual questions and segmentation answers (**VQS**), transfer human supervision between the individual tasks of semantic segmentation and VQA, thus enabling us to study at least two problems with better leverage than before: supervised attention for VQA and a novel question-focused semantic segmentation task. For the former, we obtain state-of-the-art results on the VQA real multiple-choice task by simply augmenting multilayer perceptrons with some attention features. For the latter, we propose a new approach based on mask aggregation. To put it in perspective, we study a baseline method and an upper-bound method by assuming the instance segmentations are given as oracles.

Our work is inspired upon observing the popularity of COCO [25]. We suspect that the existing and seemingly distinct annotations about MSCOCO images are inherently connected. They reveal different levels and perspectives of human understandings about the same visual scenes. Explicitly linking them up can significantly benefit not only individual tasks but also the overarching goal of unified vision-language understanding. This paper just scratches the surface. We will explore more types of annotations and richer models in the future work.

Acknowledgement This work is supported in part by the NSF award IIS #1566511, a gift from Adobe Systems Inc., and a GPU from NVIDIA. C. Gan is partially supported by the National Basic Research Program of China 2011CBA00300 & 2011CBA00301, and the National Natural Science Foundation of China 61033001 & 61361136003.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398, 2016. 1

[2] J. Andreas, M. Rohrbach, T. Darrell, and K. Dan. Neural module networks. *CVPR*, 27:55–56, 2016. 2, 6

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. VQA: Visual question answering. In *ICCV*, pages 2425–2433, 2015. 1, 2, 3, 4, 5, 6, 8

[4] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, pages 2422–2431, 2015. 1

[5] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *arXiv preprint arXiv:1606.03556*, 2016. 2, 6, 10, 11

[6] 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*, pages 2625–2634, 2015. 1

[7] D. Elliott and F. Keller. Comparing automatic evaluation measures for image description. *ACL*, 452(457):457, 2014. 1

[8] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *T-PAMI*, 32(9):1627–1645, 2010. 2

[9] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *EMNLP*, 2016. 6

[10] C. Gan, Z. Gan, X. He, J. Gao, and L. Deng. Stylenet: Generating attractive visual captions with styles. CVPR, 2017. 1

[11] C. Gan, N. Wang, Y. Yang, D.-Y. Yeung, and A. G. Hauptmann. DevNet: A deep event network for multimedia event detection and evidence recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2577, 2015. 2

[12] Z. Gan, C. Gan, X. He, Y. Pu, K. Tran, J. Gao, L. Carin, and L. Deng. Semantic compositional networks for visual captioning. *CVPR*, 2017. 1

[13] 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. In *NIPS*, pages 2296–2304, 2015. 1

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016. 6, 8

[15] R. Hu, M. Rohrbach, and T. Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 1, 2, 3, 7

[16] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *CVPR*, 2016. 1, 2

[17] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *ECCV*, pages 727–739. Springer, 2016. 2, 5, 6

[18] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 1

[19] J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In *NIPS*, pages 361–369, 2016. 6

[20] J.-H. Kim, K.-W. On, J. Kim, J.-W. Ha, and B.-T. Zhang. Hadamard product for low-rank bilinear pooling. *ICLR*, 2017. 6

[21] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 11

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

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012. 7

[24] R. Li and J. Jia. Visual question answering with question representation update (qru). In *Advances in Neural Information Processing Systems*, pages 4655–4663, 2016. 2, 6

[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*, pages 740–755, 2014. 1, 2, 3, 4, 6, 8

[26] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. *NIPS*, 2016. 2, 6

[27] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *NIPS*, 2015. 2, 6

[28] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. *arXiv preprint arXiv:1511.02283*, 2015. 1, 2

[29] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-RNN). *ICLR*, 2015. 1

[30] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *NIPS*, 26:3111–3119, 2013. 6

[31] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. *ICCV*, pages 1520–1528, 2015. 7

[32] H. Noh, P. Hongsuck Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In *CVPR*, pages 30–38, 2016. 6

[33] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, pages 4594–4602, 2016. 1

[34] P. O. Pinheiro, T. Y. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. *ECCV*, 2016. 7

[35] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 1, 2

[36] M. Ren, R. Kiros, and R. Zemel. Exploring models and data for image question answering. In *NIPS*, pages 2953–2961, 2015. 1

[37] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *ECCV*, 2016. 1, 2

[38] K. J. Shih, S. Singh, and D. Hoiem. Where to look: Focus regions for visual question answering. In *CVPR*, pages 4613–4621, 2016. 6

[39] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 7

[40] C. Sun, C. Gan, and R. Nevatia. Automatic concept discovery from parallel text and visual corpora. In *ICCV*, 2015. 1

[41] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 1

[42] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 1

[43] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5005–5013, 2016. 1

[44] Q. Wu, C. Shen, L. Liu, A. Dick, and A. van den Hengel. What value do explicit high level concepts have in vision to language problems? In *CVPR*, pages 203–212, 2016. 6

[45] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. *ICML*, 2016. 2, 6, 7

[46] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *ECCV*, 2016. 2, 6

[47] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015. 1

[48] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. *CVPR*, 2016. 2, 6, 7

[49] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *ICML*, pages 1169–1176. ACM, 2009. 2

[50] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 1, 2

[51] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. *CVPR*, 2016. 1, 2

Appendices

A. Annotation Interface

Figure 9 shows the annotation user interface we used to collect the VQS dataset. Given a question about an image, the participants are asked to tick the colors of the corresponding segmentations to visually answer the question. The participants can also click the "Add" button to draw bounding box(es) over the image in order to answer the question, in addition to choosing the segments. For more information please see the attached slides which we used to train the annotators.

B. VQS *vs.* VQA-HAT

Figure 10 contrasts the human attention maps in VAQ-HAT [5] with our collected image segmentations that are linked by the participants to the questions and answers. We observe that the HAT maps are rough comparing to the segmentation masks. For example, to answer the question "what color is the ball?", our VQS dataset will provide a very accurate segmentation mask of the ball without including any background. We expect that such accurate annotations are more suitable for visual grounding tasks. Moreover, while segmentation is the desired final output in VQS, the HAT maps mainly serve to analyze and potentially improve VQA models that output/choose text answers.

C. The influence of VQS segmentation mask resolution on the supervised attention in VQA

The attention features we studied in Section 3.1.1 of the main text weigh the feature representations of different regions according to the question about the image. The number of regions per image indicate the attention resolutions. The more regions (the higher resolution) we consider, the more accurate the attention model could be. Of course, too small regions would also result in trivial solutions since the visual cues in each region would be too subtle then.

In the table 4, we report the VQA Real Multiple-Choice results on the Test-Dev by using different resolutions of the segmentation masks. We can observe that higher resolution leads to better VQA results. In some spirit, this implies the necessity of the accurate segmentation annotations for the supervised attention in VQA.

Table 4. Comparison results of segmentation mask resolutions for supervised attention in VQA.

Method	Y/N	Num.	Others	All
VQS (14×14)	**80.60**	**39.41**	**65.73**	**68.94**
VQS (11×11)	80.18	38.93	64.9	68.36
VQS (7×7)	79.49	38.08	63.71	67.41

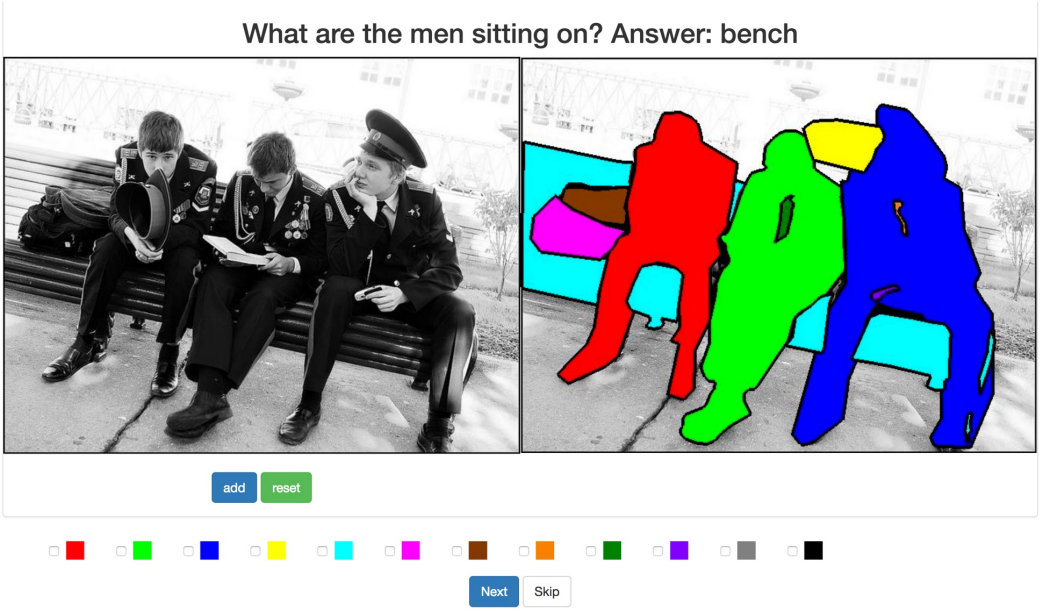


Figure 9. GUI we used to collect the links between image segmentations to questions and answers (VQS).

Table 5. Comparison results of different language embeddings for VQS.

DeconvNet (B)	DeconvNet (W)	DeconvNet (L)
0.2687	0.2979	**0.3144**

D. Some implementation details in the VQA and VQS experiments

We use an ensemble of 10 models in our experiments for the VQA Real Multiple-Choice task (cf. Table 1 of the main text). Among them, five are trained using the attribute feature representations of the images and the other five are based on the ResNet features. We use the validation set to select the best 10 models as well as how to combine them by a convex combination of their decision values. After that, we test the ensemble on Test-Dev and Test-Standard, respectively.

For the VQS experiments, we use the ADAM [21] gradient descent to train the whole network with the learning rate 0.001 and batch size 16. It takes about one week on one Titan X GPU machine to converge after 15 epochs. We also report some additional results in Table 3 for our exploration of the LSTM language embedding in the DeconvNet approach. We observe that the LSTM language embedding model (L) gives rise to about 0.02 improvement over the bag-of-words (B) and word2vec embedding (W) on the challenging VQS task.

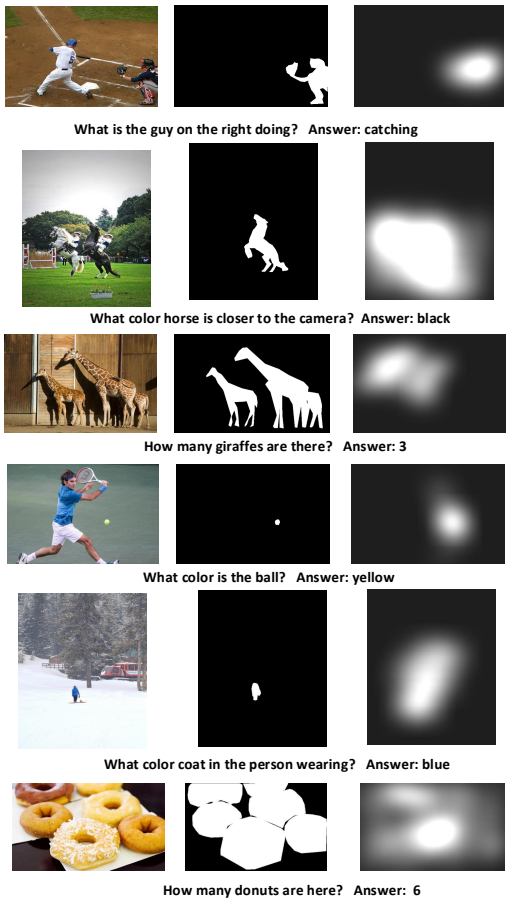


Figure 10. Comparing the segmentation annotations we collected for VQS with the human attention maps in VQA-HAT [5].