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

Anonymous ICCV submission

Paper ID 667

1. Annotation Interface

Figure 1 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.

2. VQS *vs.* VQA-HAT

Figure 2 contrasts the human attention maps in VAQ-HAT [1] 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

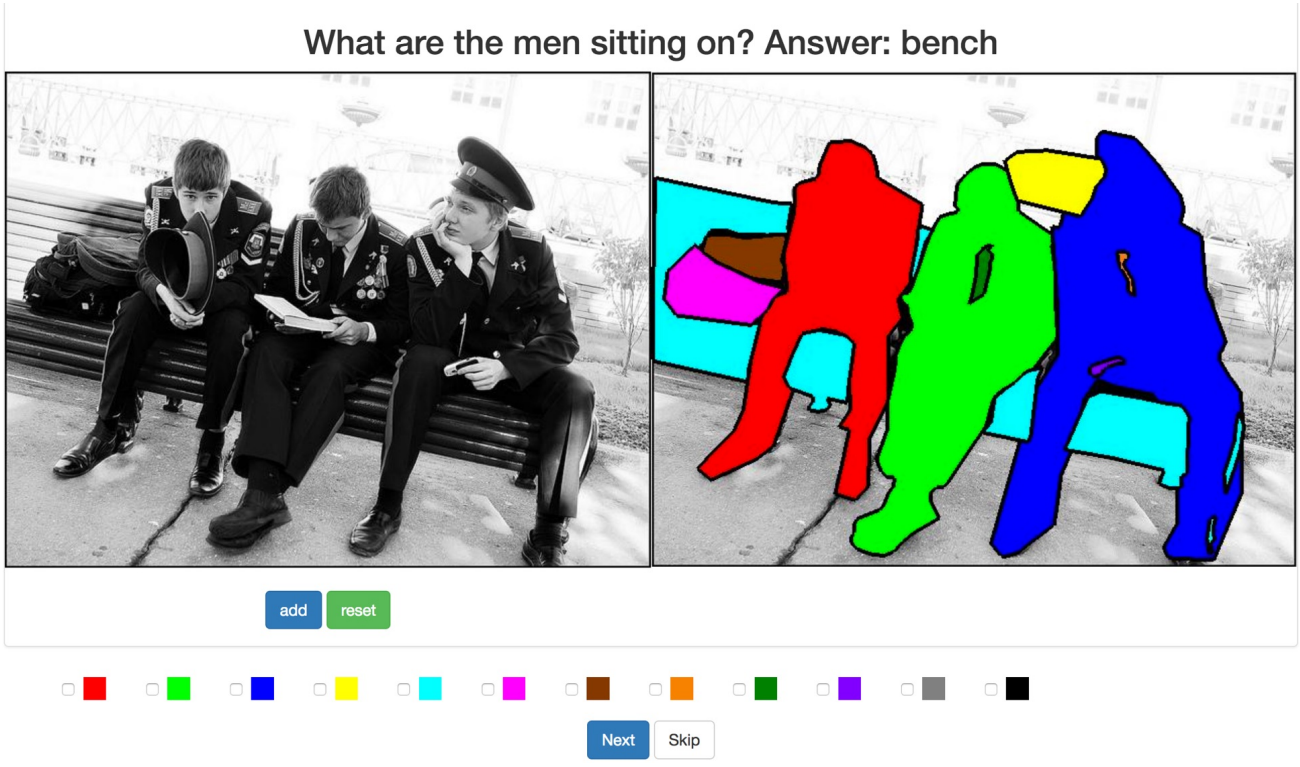


Figure 1. GUI we used to collect the links between image segmentations to questions and answers (VQS).

Table 1. 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	68.36

Table 2. Comparison results of different language embeddings for VQS.

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

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.

3. 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 1, 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.

4. 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 [2] 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.

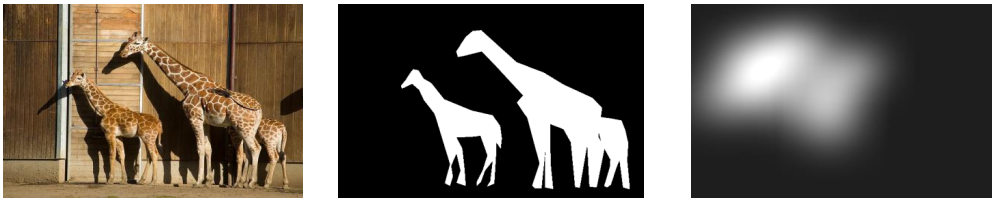
References

[1] 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. 1, 3

[2] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

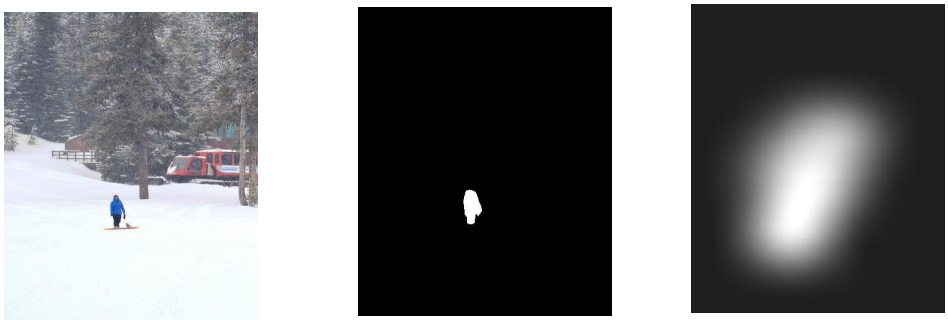
What is the guy on the right doing? Answer: catching

What color horse is closer to the camera? Answer: black

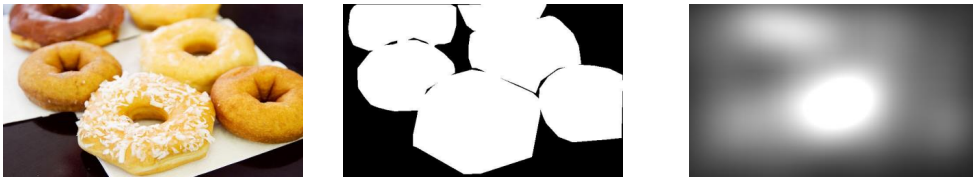


How many giraffes are there? Answer: 3

What color is the ball? Answer: yellow



What color coat in the person wearing? Answer: blue



How many donuts are here? Answer: 6

Figure 2. Comparing the segmentation annotations we collected for VQS with the human attention maps in VQA-HAT [1].