Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?

Abhishek Das[1†] ABHSHKDZ@VT.EDU
Harsh Agrawal[1†] HARSH92@VT.EDU
C. Lawrence Zitnick[‡] ZITNICK@FB.COM
Devi Parikh[†] PARIKH@VT.EDU
Dhruv Batra[†] DBATRA@VT.EDU

[†]Virginia Tech, Blacksburg
[‡]Facebook AI Research, Menlo Park

Abstract

We conduct large-scale studies on 'human attention' in Visual Question Answering (VQA) to understand where humans choose to look to answer questions about images. We design and test multiple game-inspired novel attention-annotation interfaces that require the subject to sharpen regions of a blurred image to answer a question. Thus, we introduce the VQA-HAT (Human ATtention) dataset. We evaluate attention maps generated by state-of-the-art VQA models against human attention both qualitatively (via visualizations) and quantitatively (via rank-order correlation). We find that depending on the implementation used, machine-generated attention maps are either *negatively correlated* with human attention or have positive correlation worse than task-independent saliency. Overall, our experiments paint a bleak picture for the current generation of attention models in VQA.

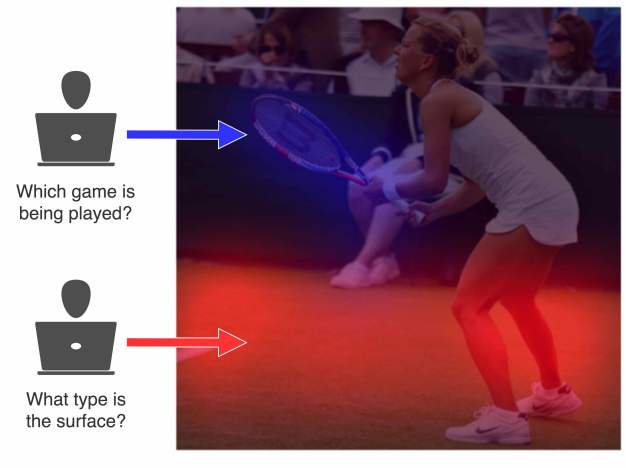


Figure 1. Different human attention regions based on question (best viewed in color).

1. Introduction

It helps to pay attention. Humans have the ability to quickly perceive a scene by selectively attending to parts of the image instead of processing the whole scene in its entirety (Rensink, 2000). Inspired by human attention, a recent trend in computer vision and deep learning is to build computational models of attention. Given an input signal, these models learn to attend to parts of it for further processing and have been successfully applied in machine translation (Bahdanau et al., 2014; Firat et al., 2016), object recognition (Ba et al., 2015; Mnih et al., 2014; Sermanet et al., 2014), image captioning (Xu et al., 2015; Cho et al., 2015) and visual question answering (Yang et al., 2015; Lu et al., 2016; Xu & Saenko, 2015; Xiong et al., 2016).

In this work, we study attention for the task of Visual Question Answering (VQA). Unlike image captioning, where a coarse understanding of an image is often sufficient for producing generic descriptions (Devlin et al., 2015), visual questions selectively target different areas of an image including background details and underlying context. This suggests that a VQA model may benefit from an explicit or implicit attention mechanism to answer a question correctly.

In this work, we are interested in the following questions: 1) Which image regions do humans choose to look at in order to answer questions about images? 2) Do deep VQA models with attention mechanisms attend to the same regions as humans?

[1]Denotes equal contribution.

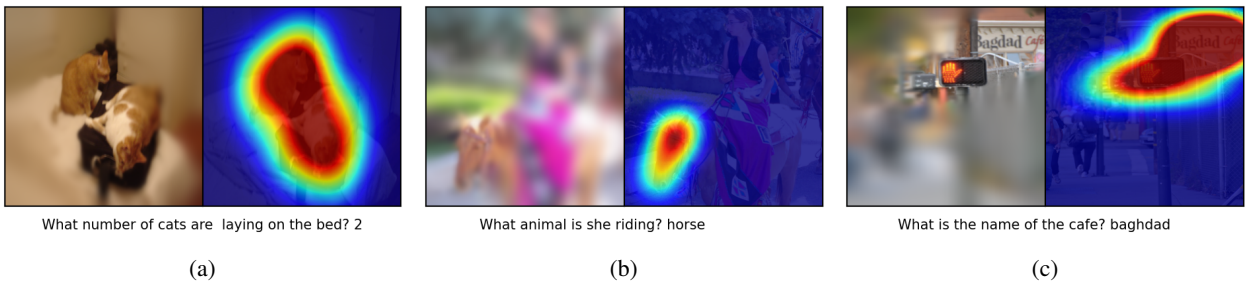


Figure 2. (a-c): Column 1 shows deblurred image, and column 2 shows human attention map.

We design and conduct studies to collect "human attention maps". Fig. 1 shows human attention maps on the same image for two different questions. When asked 'What type is the surface?', humans choose to look at the floor, while attention for 'Which game is being played?' is concentrated around the player and racket.

These human attention maps can be used both for evaluating machine-generated attention maps and for explicitly training attention-based models.

Contributions. First, we design and test multiple game-inspired novel interfaces for collecting human attention maps of where humans choose to look to answer questions from the large-scale VQA dataset (Antol et al., 2015); this VQA-HAT (Human ATtention) dataset will be released publicly. Second, we perform qualitative and quantitative comparison of the maps generated by state-of-the-art attention-based VQA models (Yang et al., 2015; Lu et al., 2016) and a task-independent saliency baseline (Judd et al., 2009) against our human attention maps through visualizations and rank-order correlation. We find that machine-generated attention maps from the most accurate VQA model have a mean rank-correlation of 0.26 with human attention maps, which is worse than task-independent saliency maps that have a mean rank-correlation of 0.49. It is well understood that task-independent saliency maps have a 'center bias' (Tatler, 2007; Judd et al., 2009). After we control for this center bias in our human attention maps, we find that the correlation of task-independent saliency is poor (as expected), while trends for machine-generated VQA-attention maps remain the same (which is promising).

2. Related Work

Our work draws on recent work in attention-based VQA and human studies in saliency prediction. We work with the free-form and open-ended VQA dataset released by (Antol et al., 2015).

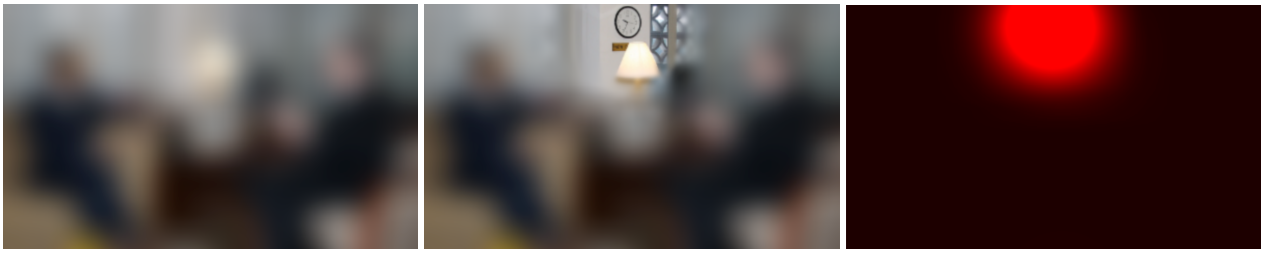
VQA Models. Attention-based models for VQA typically use convolutional neural networks to highlight relevant regions of image given a question. Stacked Attention Networks (SAN) proposed in (Yang et al., 2015) use LSTM encodings of question words to produce a spatial attention distribution over the convolutional layer features of the image. Hierarchical Co-Attention Network (Lu et al., 2016) generates multiple levels of image attention based on words, phrases and complete questions, and is the top entry on the VQA Challenge[2] as of the time of this submission. Another interesting approach uses question parsing to compose the neural network from modules, attention being one of the sub-tasks addressed by these modules (Andreas et al., 2016).

Note that all these works are *unsupervised* attention models, where "attention" is simply an intermediate variable (a spatial distribution) that is produced by the model to optimize downstream loss (VQA cross-entropy). The fact that some (it's unclear how many) of these spatial distributions end up being interpretable is simply fortuitous. In contrast, we study where humans choose to look to answer visual questions. These human attention maps can be used to evaluate unsupervised maps.

Human Studies. There's a rich history of work in collecting eye tracking data from human subjects to gain an understanding of image saliency and visual perception (Jiang et al., 2014; Judd et al., 2009; Fei-Fei et al., 2007; Yarbus, 1967). Eye tracking data to study natural visual exploration (Jiang et al., 2014; Judd et al., 2009) is useful but difficult and expensive to collect on a large scale. (Jiang et al., 2015) established mouse tracking as an accurate approach to collecting attention maps. They collected large-scale attention annotations for MS COCO (Lin et al., 2014) on Amazon Mechanical Turk (AMT). While (Jiang et al., 2015) studies natural exploration and collects task-independent human annotations by asking subjects to freely move the mouse cursor to anywhere they wanted to look on a blurred image, our approach is task-driven.

Specifically, as described in 3, we collect ground truth attention annotations by instructing subjects to sharpen parts

[2] http://visualqa.org/challenge.html



(a) Initial blurred image (b) Regions sharpened by subject (c) Attention map

Figure 3. Deblurring procedure to collect attention maps.

of a blurred image that are important for answering the questions accurately. Section 4 covers evaluation of unsupervised attention maps generated by VQA models against our human attention maps.

3. VQA-HAT (Human ATtention) Dataset

We design and test multiple game-inspired novel interfaces for conducting large-scale human studies on AMT. Our basic interface design consists of a "deblurring" exercise for answering visual questions. Specifically, we present subjects with a blurred image and a question about the image, and ask subjects to sharpen regions of the image that will help them answer the question correctly, in a smooth, click-and-drag, 'coloring' motion with the mouse. The sharpening is gradual: successively scrubbing the same region progressively sharpens it. Fig. 3 shows intermediate steps in our attention annotation interface, from a completely blurry image to a deblurred attention map.

3.1. Attention Annotation Interface

Our interface starts by showing a low-resolution blurry version of the image. This is to convey a partial 'holistic' understanding of the scene to the subjects so they may intelligently choose which regions to sharpen. Gradual sharpening with strokes was aimed to capture initial exploration as they tried to get a better sense of the scene, and eventually focussed sharpening to answer the question. Next we describe the three variants of our attention annotation interface that we experimented with.

3.1.1. Blurred Image without Answer

In our first interface, subjects were shown a blurred image and a question without the answer, and were asked to deblur regions and enter the answer. We found that this interface sometimes resulted in 'exploratory attention', where the subject lightly sharpens large regions of an image to find salient regions that eventually lead them to the answer. However, subjects often ended up with 'incomplete' attention maps since they did not see the high-resolution image and the answer, so they did not know when to stop deblurring or exploring. For instance, for an image with 3 players playing a sport, if the question is "How many players are visible in the image?", the subject might sharpen a region that seems to have the players, count the 2 players in there and answer 2, and completely miss another region of the image that had 1 more. The resulting attention map in this case is incomplete since there are 3 players in the image. This effect of incomplete human attention maps was seen in counting ("How many ...") and binary ("Is there ...") types of questions, and as a result, the answers to these were often incorrect.

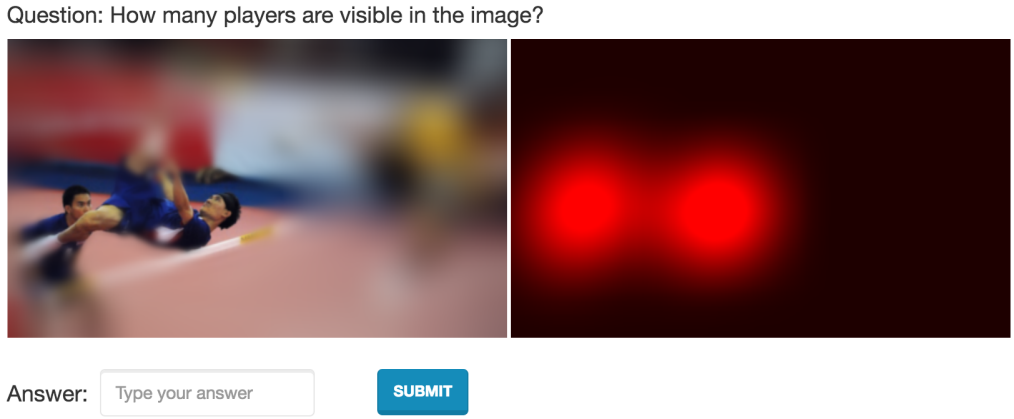
3.1.2. Blurred Image with Answer

In our second interface, subjects were shown the correct answer in addition to the question and blurred image. They were asked to sharpen as few regions as possible such that someone can answer the question just by looking at the blurred image with sharpened regions. This interface is shown in Fig. 4b. Providing the answer fixed the failure cases from the 1st interface, i.e. for counting and binary questions, since the subjects now knew the answer, they continued to explore till they found the answer region in the image.

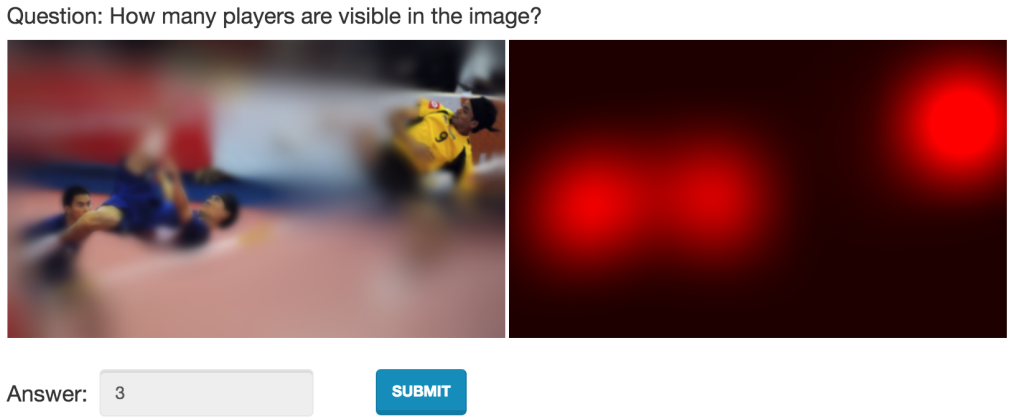
3.1.3. Blurred and Original Image with Answer

To encourage exploitation instead of exploration, in our third interface, subjects were shown the question-answer pair and full-resolution original image. In principle, seeing the original (full-resolution) image, the question, and answer provides most information to subjects, thus enabling them to provide the most 'accurate' attention maps. However, this task turns out to be fairly counter-intuitive – subjects are shown full-resolution images and the answer, and asked to imagine a scenario where someone else has to answer the question without looking at the original image.

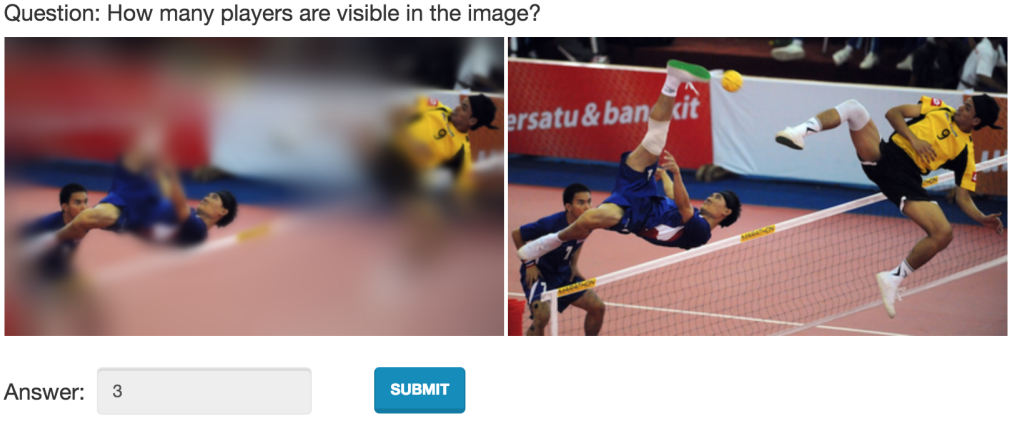
Fig. 4 shows screen-captures of the 3 attention annotation interfaces.



(a) Blurred Image without Answer



(b) Blurred Image with Answer



(c) Blurred & Original Image with Answer

Figure 4. Attention annotation interface variants. (a) In our first interface, subjects were shown a blurred image and a question without the answer, and were asked to deblur regions and enter the answer. (b) In our second interface, subjects were shown the correct answer in addition to the question and blurred image. They were asked to sharpen as few regions as possible such that someone can answer the question just by looking at the blurred image with sharpened regions. (c) To encourage exploitation instead of exploration, in our third interface, subjects were shown the question-answer pair and full-resolution original image. Out of the three interfaces, Blurred Image with Answer (b) struck the right balance between exploration and exploitation, and gives the highest accuracy on evaluation by humans as described in section 3.2.

3.2. Dataset Evaluation

We ran pilot studies on AMT to experiment with the above described three interfaces. In order to quantitatively evaluate the interfaces, we conducted a second human study where (a second set of) subjects where shown the attention-sharpened images generated from each of the attention interfaces from the first experiment and asked to answer the question. The intuition behind this experiment is that if the attention map revealed too little information, this second set of subjects would answer the question incorrectly. Table 1 shows VQA accuracies of the answers given by human subjects under these 3 interfaces. We can see that the "Blurred Image with Answer" interface (section 3.1.2) gives the highest accuracy on evaluation by humans.

Since the payments structure on AMT encourage completing tasks as quickly as possible, this implicitly incentivizes subjects to deblur as few regions as possible, and our human study shows that humans can still answer questions. Thus, overall we achieve a balance between highlighting too little or too much.

Interface Type	Human Accuracy
Blurred Image without Answer	75.2
Blurred Image with Answer	78.7
Blurred & Original Image with Answer	71.2
Original Image	80.0

Table 1. Human accuracies to compare the quality of human attention maps collected by different interfaces. Subjects were shown deblurred images from each of these interfaces and asked to answer the visual question.

We collected human attention maps for 58475 train (out of 248349 total) and 1374 val (out of 121512 total) question-image pairs in the VQA dataset. Overall, we conducted approximately 20000 Human Intelligence Tasks (HITs) on AMT, among 800 unique workers. Fig. 2 shows examples of collected human attention maps. This VQA-HAT dataset will be released publicly.

To visualize the collected dataset, we cluster the human attention maps and visualize the average attention map and example questions falling in each of them for 6 selected clusters in Fig. 5.

4. Human Attention Maps vs Unsupervised Attention Models

Now that we have collected these human attention maps, we can ask the following question – do unsupervised attention models learn to predict attention maps that are similar to human attention maps? To rephrase, *do neural networks look at the same regions as humans to answer a visual question?*

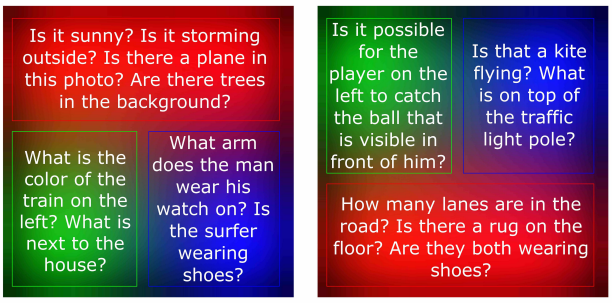


Figure 5. Visualization of 6 human attention map clusters – the average attention map for the cluster and example questions falling in each of them.

VQA Attention Models. We evaluate maps generated by the following unsupervised models:

- Stacked Attention Network (SAN) (Yang et al., 2015) with a single attention layer (SAN-1)[3].
- Hierarchical Co-Attention Network (HieCoAtt) (Lu et al., 2016) with word-level (HieCoAtt-W), phrase-level (HieCoAtt-P) and question-level (HieCoAtt-Q) attention maps; we evaluate all three maps[4].

Comparison Metric: Rank Correlation. We first scale both the machine-generated and human attention maps to 14x14, normalize them spatially and rank the pixels according to their spatial attention, and then compute correlation between these two ranked lists. We choose an order-based metric so as to make the evaluation invariant to absolute spatial probability values which can be made peaky or diffuse by tweaking a 'temperature' parameter.

Table 2 shows rank-order correlation averaged over all image-question pairs on the validation set. We compare with random attention maps and task-independent saliency maps generated by a model trained to predict human eye fixation locations where subjects are asked to freely view an image for 3 seconds (Judd et al., 2009). Interestingly, SAN-1 is *negatively* correlated to human attention maps. HieCoAtt attention maps are positively correlated with human attention maps, but not as strongly as task-independent Judd saliency maps. Our findings lead to two take-away messages with significant potential impact on future research in this active field. First, current VQA attention models do not seem to be 'looking' at the same regions as humans to produce an answer. Second, as attention-based VQA models become more accurate (58.9% SAN $\rightarrow 62.1\%$ HieCoAtt), they seem to be better correlated with humans in terms of where they look. Our dataset will allow

[3]Performance of our re-implementation is comparable to that reported in the original paper.

[4]Code available at https://github.com/jiasenlu/HieCoAttenVQA

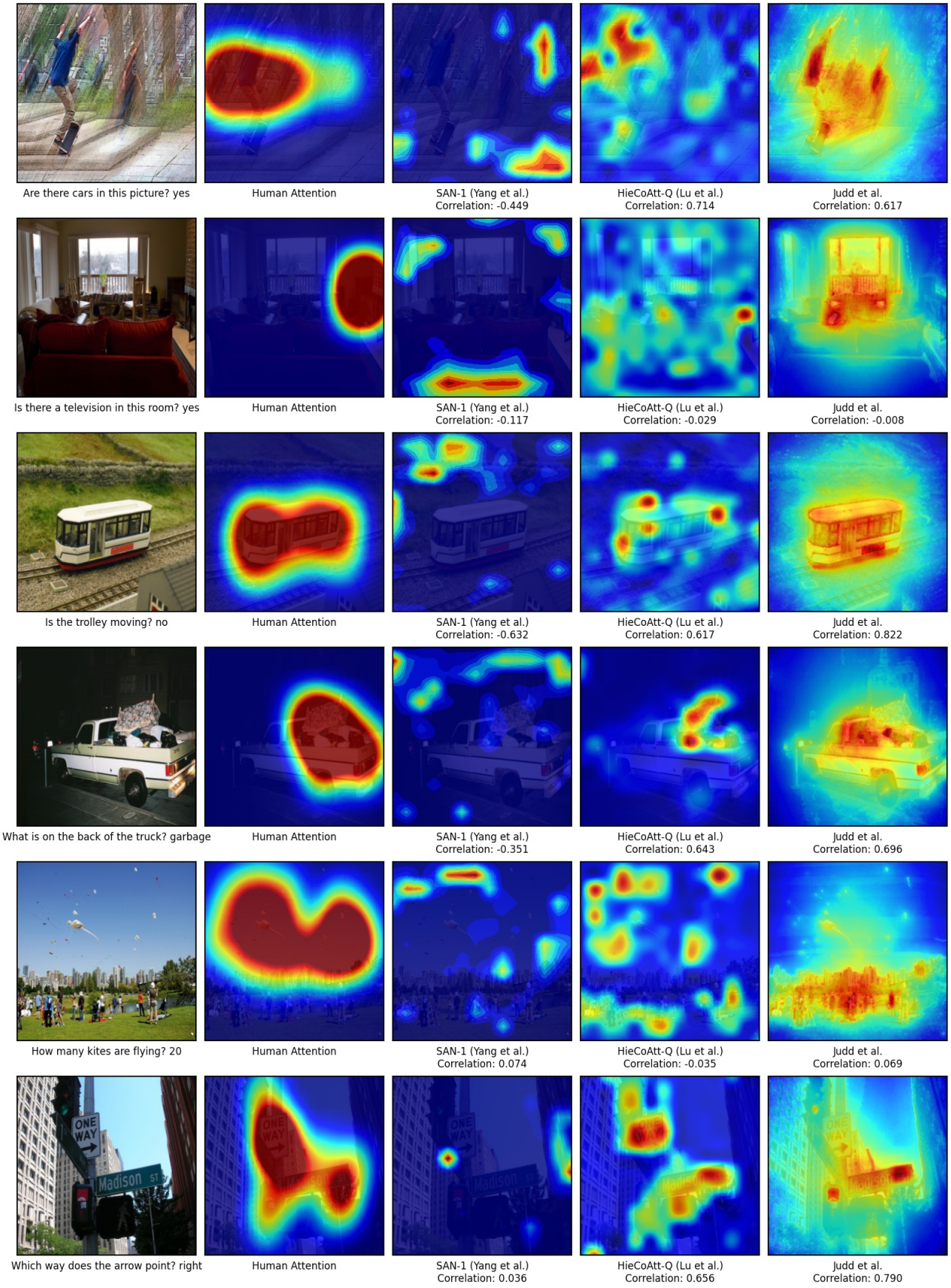


Figure 6. Random samples of human attention (column 2) v/s machine-generated attention (columns 3-5).

Model	Rank-correlation
SAN-1 (Yang et al., 2015)	-0.088 $\pm$ 0.003
HieCoAtt-W (Lu et al., 2016)	0.246 $\pm$ 0.004
HieCoAtt-P (Lu et al., 2016)	0.256 $\pm$ 0.004
HieCoAtt-Q (Lu et al., 2016)	0.264 $\pm$ 0.004
Random	0.000 $\pm$ 0.001
Judd et al. (Judd et al., 2009)	0.497 $\pm$ 0.004
Human	0.623 $\pm$ 0.003

Table 2. Mean rank-correlation coefficients (higher is better); error bars show standard error of means. We can see that SAN-1 attention maps are negatively correlated to human attention maps, while HieCoAtt attention maps are positively correlated with human attention maps, but not as strongly as task-independent Judd saliency maps.

Model	Rank-correlation
SAN-1 (Yang et al., 2015)	-0.020 $\pm$ 0.009
HieCoAtt-W (Lu et al., 2016)	0.062 $\pm$ 0.012
HieCoAtt-P (Lu et al., 2016)	0.048 $\pm$ 0.010
HieCoAtt-Q (Lu et al., 2016)	0.114 $\pm$ 0.012
Judd et al. (Judd et al., 2009)	-0.063 $\pm$ 0.009

Table 3. Mean rank-correlation coefficients (higher is better) on the reduced set without center bias; error bars show standard error of means. We can see that correlation goes down significantly for Judd saliency maps since they have a strong center bias. Relative trends among SAN-1 & HieCoAtt are similar to those over the whole validation set (reported in Table 2).

for a more thorough validation of this observation as future attention-based VQA models are proposed. Fig. 6 shows examples of human attention and machine-generated attention maps with corresponding rank-correlation coefficients.

To put these numbers in perspective, we computed inter-human agreement on the validation set by collecting 3 human attention maps per image-question pair and computing mean rank-correlation, which is 0.62289. Lastly, all reported correlation values are averaged over 3 trials by adding random noise (order of 10^{-14}) to the human attention maps to account for ranking variations in case of uniformly weighted regions.

Center Bias. Judd saliency maps aim to predict human eye fixations during natural visual exploration. These tend to have a strong center bias (Tatler, 2007; Judd et al., 2009). Although our human attention maps dataset is not an eye tracking study, the center bias still exists albeit not as severe. One potential source of this center bias is the fact that the VQA dataset was human-generated by subjects looking at the images. Thus, salient objects in the center of the image are likely be potential subjects of the questions. We compute rank-correlation of a synthetically generated central attention map with Judd saliency and human attention maps. Judd saliency maps have a mean rank-correlation of 0.87659 and human attention maps have a mean rank-correlation of 0.45781 on the validation set.

To eliminate the effect of center bias in this evaluation, we removed human attention maps that have a positive rank-correlation with the center attention map. We compute rank-correlation of machine-generated attention with human attention on this reduced set. See Table 3. Mean correlation goes down significantly for Judd saliency maps since they have a strong center bias. Relative trends among SAN-1 & HieCoAtt are similar to those over the whole validation set (reported in Table 2). HieCoAtt-Q now has a higher correlation with human attention maps than Judd saliency. This demonstrates that discounting the center bias, VQA-specific machine attention maps correlate better with VQA-specific human attention maps than task independent machine saliency maps.

5. Conclusion & Discussion

We introduce and release the VQA-HAT dataset. This dataset can be used to evaluate attention maps generated in an unsupervised manner by attention-based VQA models, or to explicitly train models with attention supervision for VQA. We quantify whether current attention-based VQA models are 'looking' at the same regions of the image as humans do to produce an answer.

Necessary vs Sufficient Maps. Are human attention maps 'necessary' and/or 'sufficient'? If regions highlighted by the human attention maps are sufficient to answer the question accurately, then so is any region that is a superset. For example, if attention mass is concentrated on a 'cat' for 'What animal is present in the picture?', then an attention map that assigns weights to any arbitrary-sized region that includes the 'cat' is sufficient as well. On the contrary, a *necessary* and sufficient attention map would be the smallest visual region sufficient for answering the question accurately. It is an ill-posed problem to define a necessary attention map in the space of pixels; random pixels can be blacked out and chances are that humans would still be able to answer the question given the resulting subset attention map. Our work thus poses an interesting question for future work – what is the right *semantic* space in which it is meaningful to talk about necessary and sufficient attention maps for humans?

6. Acknowledgements

We thank Jiasen Lu and Ramakrishna Vedantam for helpful suggestions and discussions. This work was supported in part by the following: National Science Foundation CAREER awards to DB and DP, Army Research Office YIP awards to DB and DP, ICTAS Junior Faculty awards to DB and DP, Army Research Lab grant W911NF-15-2-0080 to DP and DB, Office of Naval Research grant N00014-14-1-0679 to DB, Paul G. Allen Family Foundation Allen Distinguished Investigator award to DP, Google Faculty Research award to DP and DB, AWS in Education Research grant to DB, and NVIDIA GPU donation to DB.

References

Andreas, Jacob, Rohrbach, Marcus, Darrell, Trevor, and Klein, Dan. Learning to compose neural networks for question answering. *CoRR*, abs/1601.01705, 2016. URL http://arxiv.org/abs/1601.01705. 2

Antol, Stanislaw, Agrawal, Aishwarya, Lu, Jiasen, Mitchell, Margaret, Batra, Dhruv, Zitnick, C. Lawrence, and Parikh, Devi. Vqa: Visual question answering. In *ICCV*, 2015. 2

Ba, Jimmy Lei, Mnih, Volodymyr, and Kavukcuoglu, Koray. Multiple Object Recognition With Visual Attention. *Iclr-2015*, 2015. 1

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. URL http://arxiv.org/abs/1409.0473. 1

Cho, KyungHyun, Courville, Aaron C., and Bengio, Yoshua. Describing multimedia content using attention-based encoder-decoder networks. *CoRR*, abs/1507.01053, 2015. URL http://arxiv.org/abs/1507.01053. 1

Devlin, Jacob, Gupta, Saurabh, Girshick, Ross, Mitchell, Margaret, and Zitnick, C. Lawrence. Exploring Nearest Neighbor Approaches for Image Captioning. *arXiv preprint*, 2015. 1

Fei-Fei, Li, Iyer, Asha, Koch, Christof, and Perona, Pietro. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):10, 2007. doi: 10.1167/7.1.10. URL +http://dx.doi.org/10.1167/7.1.10. 2

Firat, Orhan, Cho, KyungHyun, and Bengio, Yoshua. Multi-way, multilingual neural machine translation with a shared attention mechanism. *CoRR*, abs/1601.01073, 2016. URL http://arxiv.org/abs/1601.01073. 1

Jiang, Ming, Xu, Juan, and Zhao, Qi. Saliency in Crowd. *ECCV*, 2014. 2

Jiang, Ming, Huang, Shengsheng, Duan, Juanyong, and Zhao, Qi. Salicon: Saliency in context. In *CVPR*, June 2015. 2

Judd, Tilke, Ehinger, Krista, Durand, Frédo, and Torralba, Antonio. Learning to predict where humans look. In *ICCV*, 2009. 2, 5, 7

Lin, Tsung-Yi, Maire, Michael, Belongie, Serge, Hays, James, Perona, Pietro, Ramanan, Deva, Dollr, Piotr, and Zitnick, C. Lawrence. Microsoft COCO: Common objects in context, 2014. 2

Lu, J., Yang, J., Batra, D., and Parikh, D. Hierarchical Co-Attention for Visual Question Answering. *ArXiv e-prints*, May 2016. 1, 2, 5, 7

Mnih, Volodymyr, Heess, Nicolas, Graves, Alex, and Kavukcuoglu, Koray. Recurrent Models of Visual Attention. *arXiv preprint*, 2014. 1

Rensink, Ronald A. The dynamic representation of scenes. *Visual Cognition*, 7(1-3):17–42, 2000. doi: 10.1080/135062800394667. URL http://dx.doi.org/10.1080/135062800394667. 1

Sermanet, Pierre, Frome, Andrea, and Real, Esteban. Attention for fine-grained categorization. *CoRR*, abs/1412.7054, 2014. URL http://arxiv.org/abs/1412.7054. 1

Tatler, Benjamin W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7 (14):4, 2007. doi: 10.1167/7.14.4. URL +http://dx.doi.org/10.1167/7.14.4. 2, 7

Xiong, Caiming, Merity, Stephen, and Socher, Richard. Dynamic memory networks for visual and textual question answering. *CoRR*, abs/1603.01417, 2016. URL http://arxiv.org/abs/1603.01417. 1

Xu, Huijuan and Saenko, Kate. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. *CoRR*, abs/1511.05234, 2015. URL http://arxiv.org/abs/1511.05234. 1

Xu, Kelvin, Ba, Jimmy, Kiros, Ryan, Cho, Kyunghyun, Courville, Aaron C., Salakhutdinov, Ruslan, Zemel, Richard S., and Bengio, Yoshua. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044, 2015. URL http://arxiv.org/abs/1502.03044. 1

Yang, Zichao, He, Xiaodong, Gao, Jianfeng, Deng, Li, and Smola, Alexander J. Stacked attention networks for image question answering. *CoRR*, abs/1511.02274, 2015. URL http://arxiv.org/abs/1511.02274. 1, 2, 5, 7

Yarbus, A. L. *Eye Movements and Vision*. Plenum. New York., 1967. 2