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

Abhishek Das[1*] **Harsh Agrawal**[1*] **C. Lawrence Zitnick**[2] **Devi Parikh**[1,3] **Dhruv Batra**[1,3]
[1]Virginia Tech [2]Facebook AI Research [3]Georgia Institute of Technology
{abhshkdz, harsh92, parikh, dbatra}@vt.edu, zitnick@fb.com

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). Overall, our experiments show that current VQA attention models do not seem to be looking at the same regions as humans.

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., 2015; 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., 2016; Lu et al., 2016; Xu and 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?

We design and conduct studies to collect "human attention maps". Figure 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.

[*]Denotes equal contribution.

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 932–937,
Austin, Texas, November 1-5, 2016. ©2016 Association for Computational Linguistics

(a) (b) (c)

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

Contributions. First, we design 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 is publicly available at our project webpage[1] Second, we perform qualitative and quantitative comparison of the maps generated by state-of-the-art attention-based VQA models (Yang et al., 2016; 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, 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 confirms our key finding that current VQA attention models do not seem to be looking at the same regions as humans.

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., 2016) 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 alternative to eye tracking for collecting attention maps. They collected large-scale attention annotations for MS COCO (Lin et al., 2014) on Ama-

[1] http://computing.ece.vt.edu/~abhshkdz/vqa-hat

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

zon 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. (Jia Deng and Jonathan Krause and Li Fei-Fei, 2013; Deng et al., 2015) leverage crowdsourcing to help computers select discriminative features for fine-grained recognition. They introduce a novel gamified setting where the humans can reveal regions with certain penalty which ensures discriminative regions with assured quality. Related to this is the work of (von Ahn and Dabbish, 2004) who explore gamification to locate objects in an image. To the best of our knowledge, this is the first work to collect human attention maps for VQA.

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

We experiment with multiple variants of the data collection interface. Analysis of the interfaces as well as details of the human evaluation studies conducted to converge on the final interface used for results in this main document have been included in the supplement. The human evaluation studies consisted of showing these attention-sharpened images to humans and asking them to answer the question. Based on these human studies, we pick the "Blurred Image with Answer" interface, where subjects were shown the correct answer in addition to the question and blurred image, and asked to deblur as few regions as possible such that someone can answer the question just by looking at the sharpened regions. Since the payment structure on AMT encourage completing tasks as quickly as possible, this implicitly incentivizes subjects to deblur as few regions as possible. Our followup human studies on these collected maps show that other subjects are able to answer questions based on these collected maps (details in supplement). Thus, overall we achieve a balance between highlighting too little or too much.

Note that the "Blurred Image with Answer" interface used to collect attention maps is a verification task as opposed to actual question answering. We show subjects an answer and ask them to sharpen regions that will help them answer the question correctly, as opposed to showing them just the question and asking them for the answer as well as relevant sharpened regions in the image ("Blurred Image without Answer" interface). Attention maps collected via this verification task "Blurred Image with Answer" are more informative (in terms of human VQA accuracy) than those collected for "Blurred Image without Answer" – 78.7% vs. 75.2%.

We collected human attention maps for 58475 train (out of 248349 total) and 1374 val (out of 121512 total) question-image pairs from the VQA dataset. This dataset is publicly available[1]. Overall, we conducted approximately 20000 Human Intelligence Tasks (HITs) on AMT, among 800 unique workers. Figure 2 shows examples of collected human attention maps.

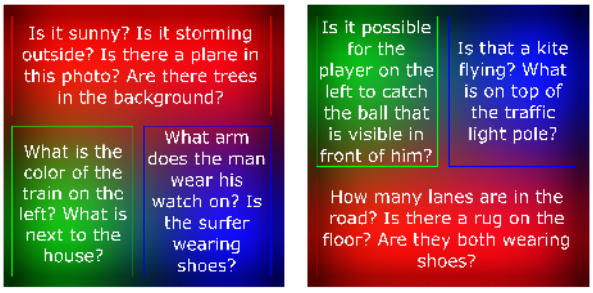

Figure 3

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 Figure 3.

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?*

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

- Stacked Attention Network (SAN) (Yang et al., 2016) with two attention layers (SAN-2)[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, 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 1 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). Both SAN-2 and 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 (slightly) better correlated with humans in terms of where they look. Our dataset will allow for a more thorough validation of this observation as future attention-based VQA models are proposed. Figure 4 shows examples of human and machine-generated attention maps with their rank-correlation coefficients.

Model	Rank-correlation
SAN-2 (Yang et al., 2016)	0.249 ± 0.004
HieCoAtt-W (Lu et al., 2016)	0.246 ± 0.004
HieCoAtt-P (Lu et al., 2016)	0.256 ± 0.004
HieCoAtt-Q (Lu et al., 2016)	0.264 ± 0.004
Random	0.000 ± 0.001
Judd et al. (Judd et al., 2009)	0.497 ± 0.004
Human	0.623 ± 0.003

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

Model	Rank-correlation
SAN-2 (Yang et al., 2016)	0.038 ± 0.011
HieCoAtt-W (Lu et al., 2016)	0.062 ± 0.012
HieCoAtt-P (Lu et al., 2016)	0.048 ± 0.010
HieCoAtt-Q (Lu et al., 2016)	0.114 ± 0.012
Judd et al. (Judd et al., 2009)	-0.063 ± 0.009

Table 2: Correlation on the reduced set without center bias goes down significantly for Judd saliency since they have a strong center bias. Relative trends among SAN-2 & HieCoAtt are similar to those over the whole validation set (reported in Table 1).

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 0.623. Lastly, all reported correlations are averaged over 3 trials by adding random noise (order of 10^{-14}) to 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 cen-

[3] https://github.com/zcyang/imageqa-san

[4] https://github.com/jiasenlu/HieCoAttenVQA

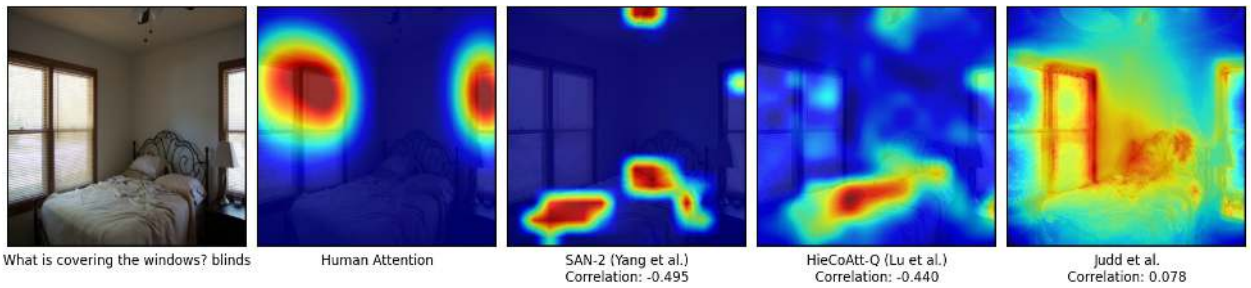

Figure 4: Random example of human attention (column 2) v/s machine-generated attention (columns 3-5)

ter bias still exists albeit not as severely as in eye-tracking. A potential source of center bias is the fact that the VQA dataset was human-generated by subjects looking at images. Thus, salient objects in the center of the image are likely to be potential subjects of 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 0.877 and human attention maps have a mean rank-correlation of 0.458 on the validation set.

To eliminate the effect of center bias in this evaluation, we removed human attention maps that have 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 2. Mean correlation goes down significantly for Judd saliency maps since they have a strong center bias. Relative trends among SAN-2 & HieCoAtt are similar to those over the whole validation set (reported in Table 1). HieCoAtt-Q now has higher correlation with human attention maps than Judd saliency. Thus 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[1]. 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?

Acknowledgements

We thank Jiasen Lu and Rama Vedantam for helpful suggestions. This work was supported in part by the National Science Foundation CAREER awards to DB & DP, Army Research Office YIP awards to DB & DP, ICTAS Junior Faculty awards at VT to DB & DP, Army Research Lab grant W911NF-15-2-0080 to DP & DB, Office of Naval Research (ONR) YIP award to DP, ONR grant N00014-14-1-0679 to DB, Alfred P. Sloan Fellowship to DP, Paul G. Allen Family Foundation Allen Distinguished Investigator award to DP, Google Faculty Research award to DP & DB, AWS in Education Research grant to DB, and NVIDIA GPU donation to DB. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the US Government or any sponsor.

References

[Andreas et al.2016] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In *NAACL HLT*. 2

[Antol et al.2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*. 2

[Ba et al.2015] Jimmy Lei Ba, Volodymyr Mnih, and Koray Kavukcuoglu. 2015. Multiple Object Recognition With Visual Attention. In *ICLR*. 1

[Bahdanau et al.2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*. 1

[Cho et al.2015] KyungHyun Cho, Aaron C. Courville, and Yoshua Bengio. 2015. Describing Multimedia Content using Attention-based Encoder-Decoder Networks. volume abs/1507.01053. 1

[Deng et al.2015] Jia Deng, Jonathan Krause, Michael Stark, and Li Fei-Fei. 2015. Leveraging the Wisdom of the Crowd for Fine-Grained Recognition. *PAMI*. 3

[Devlin et al.2015] Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. 2015. Exploring nearest neighbor approaches for image captioning. volume abs/1505.04467. 1

[Fei-Fei et al.2007] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. 2007. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7(1):10. 2

[Firat et al.2016] Orhan Firat, KyungHyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. volume abs/1601.01073. 1

[Jia Deng and Jonathan Krause and Li Fei-Fei2013] Jia Deng and Jonathan Krause and Li Fei-Fei. 2013. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *CVPR*. 3

[Jiang et al.2014] Ming Jiang, Juan Xu, and Qi Zhao. 2014. Saliency in Crowd. In *ECCV*. 2

[Jiang et al.2015] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. 2015. Salicon: Saliency in context. In *CVPR*. 2, 3

[Judd et al.2009] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look. In *ICCV*. 2, 4

[Lin et al.2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*. 2

[Lu et al.2016] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *NIPS*. 1, 2, 4

[Mnih et al.2014] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. In *NIPS*. 1

[Rensink2000] Ronald A. Rensink. 2000. The dynamic representation of scenes. *Visual Cognition*, 7(1-3):17–42. 1

[Sermanet et al.2014] Pierre Sermanet, Andrea Frome, and Esteban Real. 2014. Attention for Fine-Grained Categorization. volume abs/1412.7054. 1

[Tatler2007] Benjamin W. Tatler. 2007. 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. 2, 4

[von Ahn and Dabbish2004] Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *CHI*. 3

[Xiong et al.2016] Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *ICML*. 1

[Xu and Saenko2015] Huijuan Xu and Kate Saenko. 2015. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. volume abs/1511.05234. 1

[Xu et al.2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *ICML*. 1

[Yang et al.2016] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked Attention Networks for Image Question Answering. In *CVPR*. 1, 2, 4

[Yarbus1967] A. L. Yarbus. 1967. *Eye Movements and Vision*. Plenum. New York. 2