DRAU: Dual Recurrent Attention Units for Visual Question Answering
SUPPLEMENTARY MATERIAL

Ahmed Osman[a,*], Wojciech Samek[a,*]

[a]Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, Berlin 10587, Germany

Abstract

In the supplementary material, we provide more experiments, qualitative results, and VQA discussions for the interested reader.

Keywords: Visual Question Answering, Attention Mechanisms, Multi-modal Learning, Machine Vision, Natural Language Processing

1. Introduction

In Section 2, we provide additional experiments of baselines and variants of the DRAU model. In Section 3, additional qualitative results with comparisons to MCB (Fukui et al., 2016), including a closer look at the counting questions, are presented. In Section 4, we discuss drawbacks related to our model and most VQA models, in general. Finally, we mention some criticisms about the current state of VQA datasets and discuss future improvements in Section 5.

2. Additional Experiments

We show experiments that we have done during our work on the VQA datasets. During early experiments, the VQA 2.0 dataset was not yet released. Thus, the baselines and early models were evaluated on the VQA 1.0 dataset. While building the final model, several parameters were changed, mainly, the learning rate, activation functions, dropout value, and other modifications which we discuss in this section. All baselines use ResNet image features unless explicitly stated. We will ocassionally refer to the modules in Figure 2 in the paper when describing the models' architectures.

2.1. Baselines

Early Baselines. We started by designing three baseline architectures. The first baseline produced predictions solely from the question while totally ignoring the image. The model used the same question representation in the main paper as described by Fukui et al. (2016) and passed the output to a softmax 3000-way classification layer. The goal of this architecture was to assess the extent of the language bias present in VQA.

The second baseline is a simple joint representation of the image features and the question representation. The representations were combined using the compact bilinear pooling from Gao et al. (2016). We chose this method specifically because it was shown to be effective by Fukui et al. (2016). The main objective of this model is to measure how a robust pooling method of multimodal features would perform on its own without a deep architecture or attention. We refer to this model as *Simple MCB*.

Table 1: Evaluation of the baseline models on the VQA 1.0 and 2.0 *validation* splits.

VQA 1.0 Validation Split				
Baselines	All	Y/N	Num.	Other
Language only	48.3	78.56	27.98	30.76
Simple MCB	54.82	78.64	32.98	39.79
Joint LSTM	59.34	**79.90**	**36.96**	49.58
VQA 2.0 Validation Split				
Baselines	All	Y/N	Num.	Other
Joint LSTM	56.00	72.04	37.95	48.58
Joint LSTM + PReLU	59.74	79.61	36.21	**50.77**
Joint LSTM + Pos.	57.71	79.68	36.52	46.59
Joint LSTM + Norm Pos.	**59.75**	79.69	36.36	50.69
Joint LSTM + High dropout	57.59	79.03	34.84	47.25
Joint LSTM + Extra FC	56.51	78.86	33.51	45.57

For the last baseline, we substituted the compact bilinear pooling from Simple MCB with an LSTM consisting of hidden states equal to the image size. A 1×1 convolutional layer followed by a *tanh* activation were used on the image features prior to the LSTM, while the question representation was replicated to have a common embedding size for both representations This model is referred to as *Joint LSTM*. Note that this model does not use attention.

We begin by testing our baseline models on the VQA 1.0 validation set. As shown in Table 1, the language-only baseline model managed to get 48.3% overall. More impressively, it scored 78.56% on Yes/No questions. The *Simple MCB* model further improves the overall performance, although little improvement is gained in the binary Yes/No tasks. Replacing MCB with our basic *Joint LSTM* embedding improves performance across the board.

Modifications to the Joint LSTM Model. We test several variations of the *Joint LSTM* baseline which are highlighted in Table 1. Using PReLU activations has helped in two ways. First, it reduced time for convergence from 240K iterations to 120K. Second, the overall accuracy has improved, especially in the *Other* category. Therefore, we include it in all the following models. The next modifications were inspired by the results from Kazemi and Elqursh (2017). We experimented with appending positional features (*Pos.*) which can be described as the coordinates of each

Table 2: Evaluation of the recurrent attention models on the VQA 2.0 *validation* split.

VQA 2.0 Validation Split				
Model	All	Y/N	Num.	Other
Joint LSTM	56.00	72.04	37.95	48.58
RVAU	59.02	74.59	37.75	**52.81**
RVAU$_{multilabel}$	53.67	**77.53**	36.05	40.18
DRAU$_{Hadamard\ fusion}$	59.58	76.62	**38.92**	52.09
DRAU$_{answer\ vocab\ =\ 5k}$	59.27	76.33	38.21	51.85
DRAU$_{ReLU}$	54.11	72.69	34.92	45.05
DRAU$_{no\ final\ dropout}$	58.69	77.02	38.26	50.17
DRAU$_{high\ final\ dropout}$	**59.71**	76.47	38.71	52.52

Table 3: Evaluation of DRAU-based models on the VQA 2.0 *test-dev* split.

VQA 2.0 Test-Dev Split				
Model	All	Y/N	Num.	Other
DRAU$_{Hadamard\ fusion}$	62.24	78.27	40.31	53.57
DRAU$_{small}$	60.03	77.53	38.78	49.93
DRAU$_{no\ genome}$	61.88	79.63	39.55	51.81
DRAU$_{MCB\ fusion}$	63.41	78.97	40.06	55.47

pixel to the depth/feature dimension of the image representation. When unnormalized with respect to the other features, it worsened results significantly, dropping the overall accuracy by over 2 points. Normalizing positional features (*Norm Pos.*) did not have enough of a noticeable improvement (0.01 points overall) to warrant its effectiveness. Next, increasing dropout values from 0.3 to 0.5 deteriorated the network's accuracy, particularly in the Number and Other categories. The final modification was inserting a fully connected layer with 1024 hidden units before the classifier, which surprisingly dropped the accuracy massively.

2.2. RAU Networks

RVAU Evaluation. Our first network with explicit visual attention, *RVAU*, uses the standard image and question representation in the paper (IR and QR) while only having one attention branch (RVAU). The output of RVAU is sent to the reasoning module (RM). RVAU shows an accuracy jump by almost 3 points compared to the Joint LSTM model in Table 2. This result highlights the importance of attention for good performance in VQA. Training the *RVAU* network as a multi-label task (*RVAU$_{multilabel}$*), i.e. using all available annotations at each training iteration, drops the accuracy horribly. This is the biggest drop in performance so far. This might be caused by the variety of annotations in VQA for each question which makes the task for optimizing all answers at once much harder.

DRAU Evaluation. The addition of RTAU marks the creation of our *DRAU* network. As mentioned in the paper, the output of both RAUs is combined using fusion operation. All subsequent models use Hadamard fusion unless explicitly stated. In Table 2, the *DRAU* model shows favorable improvements over the *RVAU* model. Adding textual attention improves overall accuracy by 0.56 points. Substituting the PReLU activations with ReLU (*DRAU$_{ReLU}$*) massively drops performance. While further training might have helped the model improve, PReLU offers much faster

convergence. Increasing the value of the dropout layer after the fusion operation (*DRAU$_{high\ final\ dropout}$*) improves performance by 0.13 points, in contrast to the results of the *Joint LSTM model* on VQA 1.0. Note that on the VQA 1.0 tests, we changed the values of all layers that we apply dropout on, but here we only change the last one after the fusion operation. Totally removing this dropout layer worsens accuracy (*DRAU$_{no\ final\ dropout}$*). This suggests that the optimal dropout value should be tuned per-layer.

Next, we test a few variations of DRAU on the Test-Dev set. For this test set, we use the *Train*, *Validation*, and Visual Genome data for training. We can observe that VQA benefits from more training data; the same DRAU network performs better (62.24% vs. 59.58%) thanks to the additional data. Reducing the image feature size from $2048 \times 14 \times 14$ to $2048 \times 7 \times 7$ adversely affects accuracy as shown in Table 3 (*DRAU$_{small}$*). Removing the extra data from Visual Genome hurts the model's accuracy (*DRAU$_{no\ genome}$*). That supports the fact that VQA is very diverse and that extra data helps the model perform better. Finally, substituting Hadamard product of MCB in the final fusion operation boosts the network's accuracy significantly by 1.17 points (*DRAU$_{MCB\ fusion}$*).

2.3. DRAU variants vs. state-of-the-art

After examining the results from the previous experiments, we evaluate some variants of the DRAU model on the test/test-dev splits of VQA 1.0/2.0 in Tables 4 and 5. Contrary to Kim et al. (2017) results, we found that MCB performs significantly better in our model. This seems to be consistent since we are able to repeat these results in both VQA 1.0 and 2.0 test splits. Furthermore, we can see the performance uplift from switching ResNet global image features to the object-level visual features from Ren et al. (2015).

3. Additional qualitative results

In this section, we discuss the strengths and weaknesses of our DRAU network. Then, we mention some drawbacks of the VQA dataset.

To do so, we compare DRAU with the MCB model in a subset of VQA questions and offer some qualitative comparisons of the attention maps created by each model. Next, we discuss some drawbacks of our architecture and show some situations where DRAU fails. Finally, we provide instances where the inter-human consensus evaluation of VQA causes false evaluations.

3.1. DRAU versus MCB

Closer look at counting questions. The strength of RAUs is notable in tasks that require sequentially processing the image or relational/multi-step reasoning. In the same setting, DRAU outperforms MCB in counting questions. Since annotations for the test sets are not publicly available, we train both networks using the *Train* and Visual Genome sets and test on the validation set. This DRAU model uses Hadamard product for the fusion operation and ResNet for the image features. Around 10% of the questions start with "*how many*" which can be considered to require solving a counting task. Figure 1 shows a comparison of both models on the above-mentioned type of questions as well as the two largest subsets "*how many people are*" and "*how many people are in*".

It is clear that DRAU outperforms MCB in all three subsets. Since we use the same input representation to MCB (Fukui

Table 4: DRAU models compared to the state-of-the-art on the VQA 1.0 dataset. N corresponds to the number of models used for prediction. WE indicates whether the method uses a pre-trained word embedding. VG indicates whether the method uses external data from the Visual Genome dataset.

				VQA 1.0 Open Ended Task							
					Test-dev				Test-standard		
Model	N	WE	VG	All	Y/N	Num.	Other	All	Y/N	Num.	Other
SAN (Yang et al., 2016)	-	-	-	58.7	79.3	36.6	46.1	-	-	-	58.9
DMN+ (Xiong et al., 2016)	1	-	-	60.3	80.5	36.8	48.3	60.4	-	-	-
MRN (Kim et al., 2016)	-	✗	-	61.68	82.28	38.82	49.25	61.84	82.39	38.23	49.41
HieCoAtt (Lu et al., 2016)	1	-	-	61.8	79.7	38.7	51.7	62.1	-	-	-
RAU (Noh and Han, 2016)	1	-	-	63.3	81.9	39.0	53.0	63.2	81.7	38.2	52.8
DAN (Nam et al., 2017)	1	-	-	64.3	83.0	39.1	53.9	64.2	82.8	38.1	54.0
MCB (Fukui et al., 2016)	7	✓	✓	66.7	83.4	39.8	58.5	66.47	83.24	39.47	58.00
MLB (Kim et al., 2017)	1	✓	✗	-	-	-	-	65.07	84.02	37.90	54.77
MLB (Kim et al., 2017)	7	✓	✓	66.77	84.57	39.21	57.81	66.89	84.61	39.07	57.79
MUTAN (Ben-younes et al., 2017))	5	✓	✓	67.42	**85.14**	39.81	58.52	67.36	84.91	**39.79**	58.35
MFH (Yu et al., 2017)	1	✓	✓	**67.7**	84.9	**40.2**	**59.2**	67.5	84.91	39.3	**58.7**
DRAU$_{ResNet + Hadamard fusion}$	1	✗	✗	64.3	82.73	38.18	54.43	-	-	-	-
DRAU$_{ResNet + MCB fusion}$	1	✗	✗	65.1	82.44	38.22	56.30	65.03	82.41	38.33	55.97
DRAU$_{FRCNN + MCB fusion}$	1	✓	✗	**66.86**	**84.92**	**39.16**	**57.70**	**67.16**	**84.87**	**40.02**	**57.91**

Table 5: DRAU models compared to the current submissions on the VQA 2.0 dataset. N corresponds to the number of models used for prediction. WE indicates whether the method uses a pre-trained word embedding. VG indicates whether the method uses external data from the Visual Genome dataset.

|				VQA 2.0 Open Ended Task							
					Test-dev				Test-standard		
Model	N	WE	VG	All	Y/N	Num.	Other	All	Y/N	Num.	Other
neural-vqa-attention (Yang et al., 2016)	-	-	-	55.35	70.1	35.39	47.32	55.28	69.77	35.65	47.18
CRCV_REU	-	-	-	60.65	73.91	36.82	54.85	60.81	74.08	36.43	54.84
VQATeam_MCB (Goyal et al., 2017)	1	✓	✓	61.96	78.41	38.81	53.23	62.27	78.82	38.28	53.36
DCD_ZJU(Lin et al., 2017)	-	✗	-	62.47	79.84	38.72	53.08	62.54	79.85	38.64	52.95
VQAMachine (Wang et al., 2016)	-	-	-	62.62	79.4	40.95	53.24	62.97	79.82	40.91	53.35
UPMC-LIP6 (Ben-younes et al., 2017)	5	✓	✓	65.57	81.96	41.62	57.07	65.71	82.07	41.06	57.12
HDU-USYD-UNCC (Yu et al., 2017)	9	✓	✓	68.02	84.39	45.76	59.14	68.09	84.5	45.39	59.01
Adelaide-Teney (Teney et al., 2017)	1	✓	✓	**65.32**	**81.82**	**44.21**	**56.05**	**65.67**	**82.20**	**43.90**	**56.26**
Adelaide-Teney (Anderson et al., 2017)	30	✓	✓	-	-	-	-	70.34	86.60	48.64	61.15
FAIR A-STAR (Jiang et al., 2018)	1	✓	✓	70.01	-	-	-	70.24	-	-	-
FAIR A-STAR (Jiang et al., 2018)	30	✓	✓	72.12	87.82	51.54	63.41	72.25	87.82	51.59	63.43
DRAU$_{ResNet + Hadamard fusion}$	1	✗	✗	62.24	78.27	40.31	53.58	62.66	78.86	39.91	53.76
DRAU$_{ResNet + MCB fusion}$	1	✗	✗	63.41	78.97	40.06	55.48	63.71	79.27	40.15	55.55
DRAU$_{FRCNN + MCB fusion}$	1	✓	✗	**66.45**	**82.85**	**44.78**	**57.4**	**66.85**	**83.35**	**44.37**	**57.63**
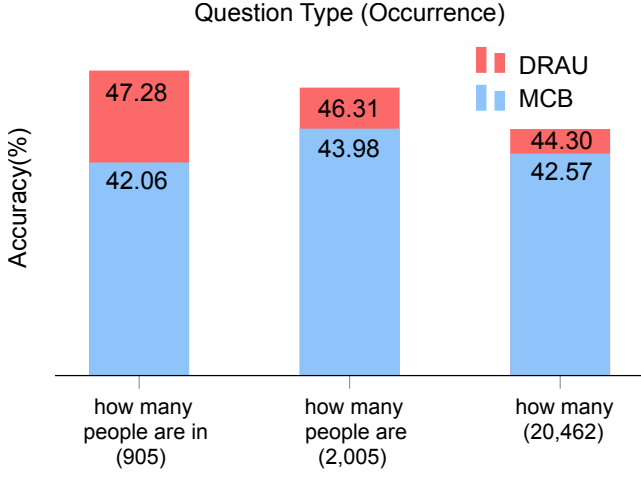
Figure 1: Results on questions that require counting in the VQA 2.0 validation set.

et al., 2016), the performance gain can be attributed to the capability of recurrent attention units to capture and accumulate local information while scanning the input.

Comparison of Attention maps. Figure 2 provides some qualitative results between DRAU's and MCB's attention mechanisms.

In Figure 2, it is clear that the recurrence helps the model attend to multiple targets as apparent in the difference of the attention maps between the two models. The top 4 examples give an insight to why DRAU noticeably performs better in the number category in VQA, because it creates more precise attention that helps the model count better. MCB seems to struggle focusing on full-body photos of humans that leads to an incorrect prediction. DRAU seems to have a higher quality visual attention that aids it in counting the correct number of people. For the question "*How many animals are there?*", it is likely that MCB has mistaken the boulders for a large animal. DRAU does not seem to have any problem in correctly attending in a multitude of sceneries and attending to multiple objects. This property also translates to questions that require relational reasoning. The question "*Are the larger elephant's eye open?*" illustrates that DRAU can attend to all possible elephants and successfully process the question. Note that MCB's attention is not focused on any discernible parts. In the next question "*Why is the cat sitting on the toilet?*", DRAU attends to both the cat and toilet in a cluttered environment and is not easily fooled by environmental bias. MCB attends to the sink which might explain its prediction of "drinking". The question "*What animal does the vase depict?*" demonstrates how DRAU can attend the location required to answer the question based on the textual and visual attention maps. Note the high weight given to the word "vase". It is important to re-state that these qualitative results were predicted by a DRAU model which was trained in the same fashion as MCB and thus, it does not represent our best network.

4. DRAU drawbacks

There is still a lot to improve in the DRAU network. First, the image representation might not be adequate to capture all the objects in an image. This is due to the fact that the image features were trained for image classification. Thus, inter-object relations are not explicitly captured by the image features.

Another drawback of DRAU — and almost all current VQA models — is the use of limited vocabulary. In Figure 3, DRAU fails to predict the correct answer due to its limited vocabulary of the most 3000 frequent answers seen in training. It is worth noting though that DRAU predicts a fairly close answer. In the first example, DRAU manages to predict the time to be "*3:20*" which is close the correct answers "*3:17*" and "*3:18*". The two other examples in Figure 3 can not be attributed to the lack of vocabulary particularly, since DRAU's vocabulary contains some of these words like "*burger, cellphone, food, menu, pig, hot, dog*". But rather to the naive way of predicting the answer by just a 3000-way classifier. To handle such questions, a model will be required to "create" answers by itself. A good candidate could be a compositional model that first decides the type of task as in (Hu et al., 2017), then the model would choose a different reasoning method depending on the task. For example, if the task requires counting, the model would keep a counter, process the input, and increase the counter whenever appropriate. If the task requires description such as in the second example of Figure 3, then it would use a reasoning method that is similar to image captioning methods to describe the contents of the plate.

VQA often demands specific answers, thus, general predictions will often fail to be answer the question. For example, if a question asks about the identity of a celebrity in a photo: "*Is this Obama?*", the model is incapable of solving this task without seeing this person during training. Since it's infeasible to account for all possible objects, zero-shot learning techniques are crucial. Zero-shot learning refers to the capability of a model to solve a task despite not seeing any examples of such task in training. Teney and van den Hengel (2016) propose test-time exemplar retrieval by using search engine services to retrieve images about every word in the question and compute global image features which are then embedded and passed down the architecture pipeline. However, this approach partially solves the problem. Because if there are no unique words that identify that task (e.g. "*Who is this?*" is a vague question), then the exemplar retrieval would not offer any valuable information.

5. VQA dataset drawbacks

According to the results in Section 2.1 and recent surveys (Kafle and Kanan, 2017; Wu et al., 2016), there exists a strong bias in VQA. This is can be attributed to the fact that when presented with an image, the humans creating questions for VQA tend to ask simple questions that often require basic scene understanding or querying about the presence of a certain object. Harder questions that require complex relational reasoning are less common. This can be confusing when evaluating different models since all questions are weighed equally. Kafle and Kanan (2017) suggest evaluating every type of question separately and calculate the accuracy as the mean across the question types rather than calculating the mean across all questions.

Human consensus offers a good balance between allowing multiple correct answers and ease of evaluation. However, we experienced encounters when this does not hold. In Figure 4, we point out some of the examples. The first and second example highlight that synonyms will not be taken into account if they don't appear in the annotation. This unnecessarily punishes model for choosing a semantically identical answer and might compound the bias problem in VQA.
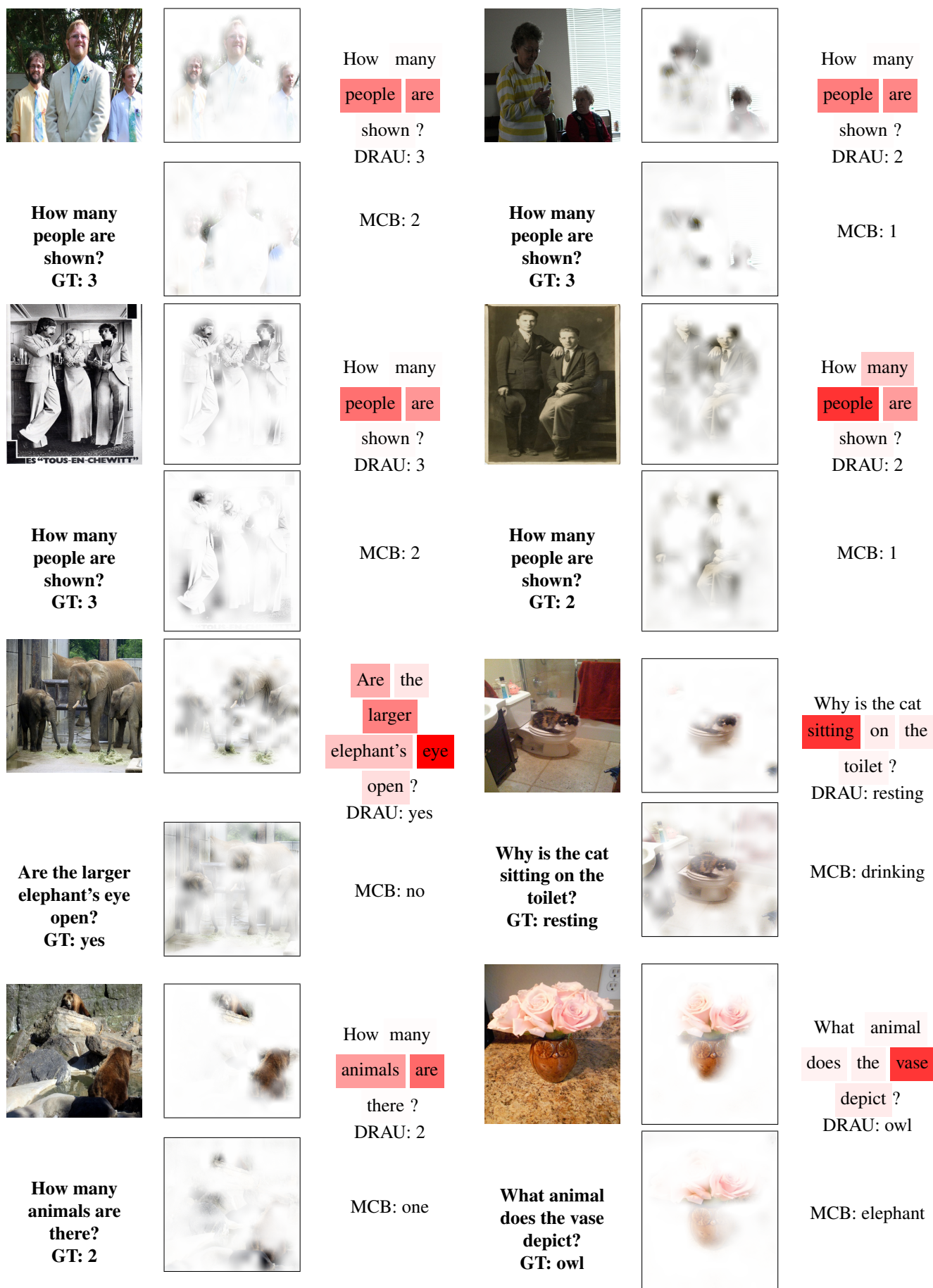
Figure 2: DRAU vs. MCB Qualitative examples. Attention maps for both models shown, only DRAU has textual attention.

Image	Question	Prediction	Annotations
	What time does the clock say?	"3:20"	• 3: 8 • 3:18 • 3:17 • 3:19 • 3:17 • 3:17 • 3:18 • 3:18 • 3:18 • 3:17
	What is on the table?	"plate"	• burger with fries • tablecloth cups cell phone food and menu • onion rings • bbq brisket sandwich onion rings bbq sauce menu cell phone • onion rings and burger • onion rings and burger • onion rings and burger • sandwich with onion rings • onion rings and hamburger • onion rings and hamburger
	What kind of food is this?	"pastry"	• pig in blanket • pastry's • corn dog • crescent dogs • pastry • pigs in blankets • pastry • pig in blanket • pigs in blanket • croissants with hot dogs inside

Figure 3: Examples where DRAU fails. DRAU generalizes because of the lack of expressive vocabulary.

In other cases, the human consensus fails to provide a reliable annotation. In the last example in Figure 4, the annotators have not agreed on any common annotation. Thus, any method can never get 100% accuracy in such questions. This appears to be a common theme for questions that require more effort than usual from the annotators. It is worth noting that this problem is not an inherent attribute of human consensus evaluation, but rather specific to the VQA dataset. We believe it's possible to alleviate this problem by using more than ten annotations and providing a well-structured guideline to help annotators.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2017. Bottom-Up and Top-Down Attention for Image Captioning and VQA. arXiv:1707.07998 .

Ben-younes, H., Cadene, R., Cord, M., Thome, N., 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. arXiv:1705.06676 .

Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M., 2016. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding, in: EMNLP, pp. 457–468.

Gao, Y., Beijbom, O., Zhang, N., Darrell, T., 2016. Compact bilinear pooling, in: CVPR, pp. 317–326.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D., 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, in: CVPR, pp. 6904–6913.

Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K., 2017. Learning to Reason: End-to-End Module Networks for Visual Question Answering. arXiv:1704.05526 [cs] arXiv:1704.05526.

Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D., 2018. Pythia v0.1: The Winning Entry to the VQA Challenge 2018. ArXiv180709956 Cs arXiv:1807.09956.

Kafle, K., Kanan, C., 2017. Visual Question Answering: Datasets, Algorithms, and Future Challenges. Comput. Vis. Image Underst. doi:10.1016/j.cviu.2017.06.005, arXiv:1610.01465.

Kazemi, V., Elqursh, A., 2017. Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering. arXiv:1704.03162 .

Kim, J.H., Lee, S.W., Kwak, D., Heo, M.O., Kim, J., Ha, J.W., Zhang, B.T., 2016. Multimodal Residual Learning for Visual QA, in: NIPS, pp. 361–369.

Kim, J.H., On, K.W., Kim, J., Ha, J.W., Zhang, B.T., 2017. Hadamard Product for Low-Rank Bilinear Pooling, in: ICLR.

Lin, Y., Pang, Z., Wang, D., Zhuang, Y., 2017. Task-Driven Visual Saliency and Attention-Based Visual Question Answering. arXiv:1702.06700 .

Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering, in: NIPS, pp. 289–297.

Nam, H., Ha, J.W., Kim, J., 2017. Dual Attention Networks for Multimodal Reasoning and Matching, in: CVPR, pp. 299–307.

Noh, H., Han, B., 2016. Training Recurrent Answering Units with Joint Loss Minimization for VQA. arXiv:1606.03647 .

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. arXiv:1506.01497 .

Teney, D., Anderson, P., He, X., van den Hengel, A., 2017. Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge. arXiv:1708.02711 .

Teney, D., van den Hengel, A., 2016. Zero-Shot Visual Question Answering. arXiv:1611.05546 [cs] arXiv:1611.05546.

Wang, P., Wu, Q., Shen, C., van den Hengel, A., 2016. The VQA-Machine: Learning How to Use Existing Vision Algorithms to Answer New Questions. arXiv:1612.05386 .

Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A., van den Hengel, A., 2016. Visual Question Answering: A Survey of Methods and Datasets. arXiv:1607.05910 [cs] arXiv:1607.05910.

Xiong, C., Merity, S., Socher, R., 2016. Dynamic Memory Networks for Visual and Textual Question Answering, in: ICML, pp. 2397–2406.

Yang, Z., He, X., Gao, J., Deng, L., Smola, A., 2016. Stacked Attention Networks for Image Question Answering, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21–29.

Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D., 2017. Beyond Bilinear: Generalized Multi-Modal Factorized High-Order Pooling for Visual Question Answering. arXiv:1708.03619 .

Image	Question	Prediction	Annotations
	How many people are there?	"none"	0 0 0 0 0 0 0 0 0 0
	How many people are there?	"one"	1 2 3 1 3 1 1 1 2 1
	How many people are there?	"many"	150 - yes 134 - many 100s 200 200 - crowd - lots 67

Figure 4: Examples where VQA annotations can punish good predictions