Multimodal Integration of Human-Like Attention in Visual Question Answering

Ekta Sood[1], Fabian Kögel[1], Philipp Müller[2], Dominike Thomas,[1], Mihai Bâce[1],
Andreas Bulling[1]
[1]University of Stuttgart, Institute for Visualization and Interactive Systems (VIS), Germany
[2]German Research Center for Artificial Intelligence (DFKI)
{ekta.sood,fabian.koegel,dominike.thomas,mihai.bace,andreas.bulling}@vis.uni-stuttgart.de
{philipp.mueller}@dfki.de

Abstract

Human-like attention as a supervisory signal to guide neural attention has shown significant promise but is currently limited to unimodal integration – even for inherently multimodal tasks such as visual question answering (VQA). We present the Multimodal Human-like Attention Network (MULAN) – the first method for multimodal integration of human-like attention on image and text during training of VQA models. MULAN integrates attention predictions from two state-of-the-art text and image saliency models into neural self-attention layers of a recent transformer-based VQA model. Through evaluations on the challenging VQAv2 dataset, we show that MULAN is competitive to state of the art in its model class – achieving 73.98% accuracy on test-std and 73.72% on test-dev with approximately 80% fewer trainable parameters than prior work. Overall, our work underlines the potential of integrating multimodal human-like attention into neural attention mechanisms for VQA.

1. Introduction

Visual question answering (VQA) is an important task at the intersection of natural language processing (NLP) and computer vision (CV) that has attracted significant research interest in recent years [6, 27, 28, 29, 49]. One of its key challenges is, by its very nature, to jointly analyze and understand the language and visual input. State-of-the-art methods for VQA rely on neural attention mechanisms to encode relations between questions and images and, for example, focus processing on parts of the image that are particularly relevant for a given question [4, 18, 47, 49]. An increasing number of methods make use of multiple Transformer-based attention modules [45]: they enable attention-based text features to exert complex patterns of influence on the allocation of attention on images [18, 49].

Simultaneously, an increasing number of works have demonstrated the effectiveness of integrating human-like attention into neural attention mechanisms across a wide range of tasks, including image captioning [43], text comprehension [41], or sentiment analysis and grammatical error detection [7]. Integration into VQA, however, has focused only on images [10, 32, 34, 39, 46], despite the inherent multimodality of the VQA task. A method for predicting and integrating human-like attention on text, which has obtained state-of-the-art performance in downstream NLP tasks, has only recently been proposed [41] and multimodal integration remains unexplored.

We fill this gap by proposing the Multimodal Human-like Attention Network (MULAN) – the first method for multimodal integration of human-like attention into VQA. In contrast to previous unimodal integration methods on images alone, our method allows human attention information to act as a link between text and image input. We base MULAN on the MCAN VQA architecture [49] and integrate state-of-the-art human saliency models for text and images into the attention scoring functions of the self-attention layers. This way, human-like attention acts as an inductive bias directly modifying neural attention processes. To model human-like attention on text we make use of the recently proposed Text Saliency Model (TSM) [41] that we adapt to the VQA task while training the MCAN framework. On images we use the recent Multi-Duration Saliency (MDS) model [14] that also models the temporal dynamics of attention. We train MULAN on the VQAv2 dataset [16] and achieve results that are competitive to the state of the art in MCAN's model class with 73.98% accuracy on *test-std* and 73.72% on *test-dev*. Higher accuracy is commonly achieved by increased parameter count, large-scale pre-training and ensembling or more recently, generative models, which are beyond the scope of this work. Notably, given that our model is based on the MCAN small variant, we require significantly fewer trainable parameters than prior work.

Our contributions are three-fold: First, we propose a novel method to jointly integrate human-like attention on

text and image into the MCAN VQA framework [49]. Second, we evaluate our method on the challenging VQAv2 benchmark [16] and show that it obtains performance competitive with the state of the art on both *test-std* and *test-dev* while requiring about 80% fewer trainable parameters. Finally, through detailed analysis of success and failure cases we provide insights into how MULAN makes use of human attention information to correctly answer questions that are notoriously difficult, e.g. longer questions.

2. Related Work

Our work is related to previous works on 1) visual question answering, 2) using neural attention mechanisms, and 3) using human-like attention as a supervisory signal.

Visual Question Answering. Using natural language to answer a question based on a single image [6] has been a topic of increasing interest in recent years [27, 28]. Antol et al. [6] built the first, large-scale VQA dataset that provided open-ended, free-form questions created by humans. Given that models have been shown to exploit bias in datasets [1], Goyal et al. [16] expanded the VQA dataset by balancing it so that each question had two images, with two different answers to the same question. Tests on this new dataset (VQAv2) obtained significantly reduced performance for current models, showing high prevalence of answer bias. Another challenge in VQA remains the lack of inconsistency in answer predictions [35, 38, 50] and reduced performance for compositional questions [2, 5, 38] or linguistic variation [3, 36].

Neural Attention Mechanisms. To imbue models with more reasoning capabilities, researchers started experimenting with human-inspired neural attention and showed that adding neural attention mechanisms improved performance for VQA. Shih et al. [37] added a region selection layer to pinpoint relevant areas of an image and improved over Antol et al. [6] by 5%. Similarly, Anderson et al. [4] demonstrated that using bottom-up attention was preferable to top-down attention, winning the first place in the 2017 VQA Challenge. Jiang et al. [19] further expanded on this work by optimizing the model architecture and won the 2018 VQA challenge. Follow-up works combined learned visual and language attention in order to narrow down which part of the image and question are relevant, first with alternating attention [25], dual attention [30], and finally multi-level attention [47]. The success of Transformers [45] in NLP tasks also inspired new work in VQA. Yu et al. [49] created the Transformer-inspired Modular Co-Attention Network (MCAN) that combines self-attention with guided-attention to leverage the interaction within and between modalities. Jiang et al. [18] further built on this architecture and won the 2020 VQA Challenge. Tan and Bansal [44] improved input encoding with a Transformer and transfer learning, while Li et al. [23] modified input encodings by adding an object tag to help align images and text semantically.

Supervision Using Human Attention. Despite its advantages, it was also demonstrated that neural attention may focus on the wrong area of an image [10, 11]. To rectify this, human attention was brought in as an additional supervisory signal. Researchers investigated differences between neural and human attention in VQA [10, 11] and created datasets containing human attention maps [10, 11, 14]. At the same time, integrating human attention supervision showed to be promising in closely related computer vision [21, 43] or NLP tasks [7, 40]. Sood et al. [41] proposed a novel text saliency model that, by combining a cognitive model of reading with human attention supervision, set a new state-of-the-art on paraphrase generation and sentence compression. For VQA tasks, Gan et al. [15] combined human attention on images and semantic segmentation of questions. Using ground truth human attention, Wu and Mooney [46] penalized networks for focusing on the wrong area of an image, while Selvaraju et al. [34] guided neural networks to look at areas of an image that humans judged as particularly relevant for question answering. Chen et al. [10] continued in this direction by using human attention to encourage reasoning behaviour from a model. Since obtaining ground truth human attention annotations is costly and time-consuming, Qiao et al. [32] trained a network on the VQA-HAT dataset to automatically generate human-like attention on unseen images, then used these saliency maps to create the enhanced Human-Like ATention (HLAT) dataset.

3. Method

The central contribution of our work is to propose MULAN, the first multimodal method to integrate human-like attention on both the image and text for VQA (see Figure 1 for an overview of the method). At its core, our method builds on the recent MCAN model [48, 49], which won the 2019 VQA challenge as well as being the basis of the 2020 winning method utilizing grid features [18]. We adapted the open source implementation[1] and trained the small variant of MCAN using grid features[2]. We first introduce the feature representations and the MCAN framework and subsequently explain our novel multimodal integration method[3].

Feature Representations. We represent the input images by spatial grid features, following the extraction methodology of Jiang et al. [18]. A Faster R-CNN with ResNet-50

[1] github.com/MILVLG/openvqa

[2] github.com/facebookresearch/grid-feats-vqa

[3] Code and other supporting material can be found at perceptual.ui.org/publications/sood23_gaze

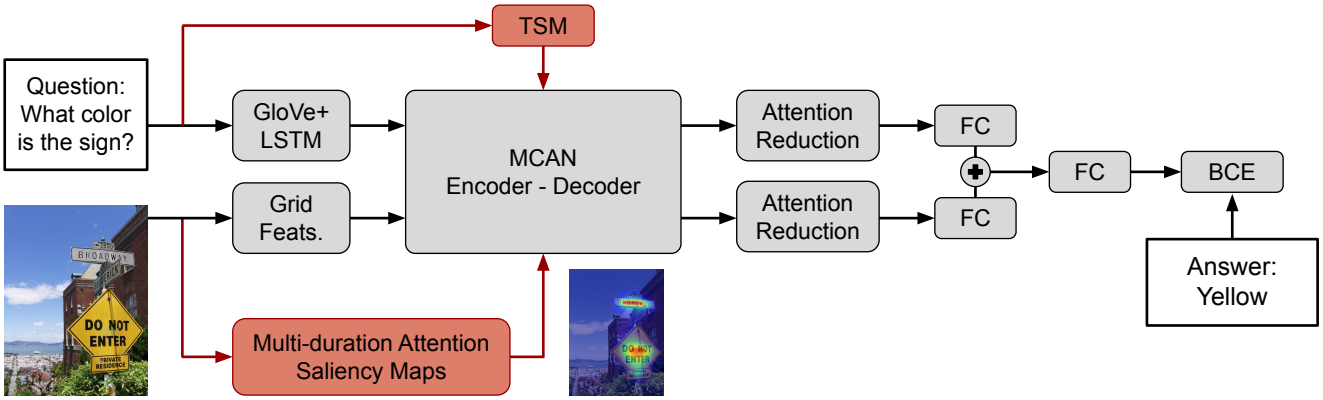

Figure 1. Overview of the Multimodal Human-like Attention Network (MULAN). Our method proposes multimodal integration of human-like attention on questions as well as images during training of VQA models. MULAN leverages attention predictions from two state-of-the-art text [41] and image saliency models [14].

backbone [17, 33] is pre-trained on ImageNet [12] and VG [22] and then the object proposal and RoI pooling (used for region features in Anderson et al. [4]) is removed. The remaining ResNet directly outputs the grid features. We obtain up to 19x32 features (depending on aspect ratio) per image. The final image representation is $X \in \mathbb{R}^{m \times d_x}$ with $m \in [192, 608]$, where m represents the number of features, and d_x the feature embedding dimension.

The input questions are represented as in MCAN: tokenized at word-level, trimmed to $n \in [1, 14]$ tokens and represented using 300-D GloVe [31] word embeddings. The $n \times 300$ embeddings are further passed through a one-layer LSTM with hidden size d_y and all intermediate hidden states form the final question representation matrix $Y \in \mathbb{R}^{n \times d_y}$. Both representations are zero-padded to accommodate the varying number of grid features and question words.

Base model. In general, an attention function computes an alignment score between a query and key-value pairs and uses the score to re-weight the values. Attention methods differ in their choice of scoring function, whether they attend to the whole (global/soft) or only parts (local/hard) of the input, and whether queries and key-value pairs are projections from the same inputs (self-attention) or different inputs (guided attention). The Deep Modular Co-Attention Network (MCAN) for VQA [49] is a Transformer-based network [45] that runs multiple layers of multi-headed self-attention (SA) and guided-attention (GA) modules in an encoder-decoder architecture using the scaled dot-product score function.

A schematic of an SA module is shown in Figure 2 in gray. It consists of two sub-layers: the multi-headed attention and a feed-forward layer. Both are encompassed by a residual connection and layer normalization. The attention sub-layer projects the input feature embeddings into queries $Q \in \mathbb{R}^{n \times d}$, keys $K \in \mathbb{R}^{n \times d}$, and values $V \in \mathbb{R}^{n \times d}$ with a common hidden dimension d. For a query q the attended output is calculated with:

$$A(q, K, V) = \text{softmax}(\frac{qK^T}{\sqrt{d}})V \quad (1)$$

As in the Transformer [45], this is calculated for multiple queries at once with QK^T and the results of several heads with different projections for Q, K, V are combined.

The GA module is set up identically to the SA module except the queries and key-value pairs are provided by separate inputs. In this way, text features can guide attention on image features. Intuitively, the attention layer reconstructs the queries from a linear combination of the values, emphasizing interactions between them. The value space is projected from the input features, which in the GA case is a fusion space between the modalities.

The MCAN encoder stacks multiple layers of SA on text features $Y \in \mathbb{R}^{n_y \times d_y}$ before feeding the result of the last layer into the decoder. The decoder stacks modules with SA on image features $X \in \mathbb{R}^{n_x \times d_x}$ and GA between the encoder result and the SA output. After the last layer, the resulting feature matrices from both encoder and decoder are flattened to obtain the attended features $\tilde{y} \in \mathbb{R}^d$ and $\tilde{x} \in \mathbb{R}^d$ and fused by projecting them into the same space and adding them. The VQA task is formulated as classification, so a final projection into the answer dimension and a sigmoid function conclude the network. Jiang et al. [18] improved the performance of the original MCAN by replacing the region image features with spatial grid features. We use their model as a baseline for our experiments.

Human-Like Attention Integration. Although the importance of fusing both modalities has been underlined by many previous works and is the driving idea behind co-attention,

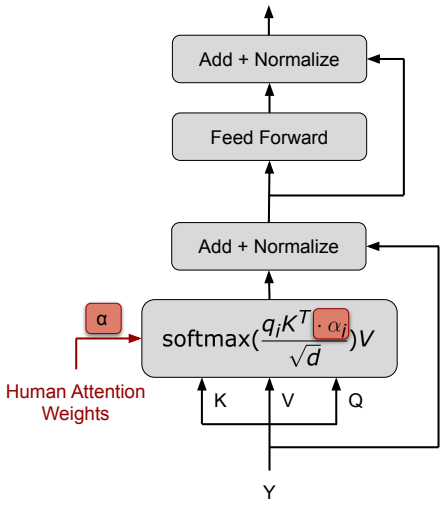

Figure 2. Schematic of a self-attention (SA) layer. The vanilla SA layer is shown in gray, while our human attention integration approach in red.

the integration of external guidance has only been explored in the image domain [10, 15, 32, 34, 39, 46].

We integrate human-like attention into both the text and image streams of the MCAN base model. For both streams, we use the same basic principle of integration into SA modules (see Figure 2). We propose a new attention function A_H for the SA layer, multiplying human-like attention weights $\alpha \in \mathbb{R}^n$ ($\mathbb{R}^m$ for image features) into the attention score. For the query q_i corresponding to the i-th input feature embedding, we calculate the i-th attended embedding:

$$A_H(q, K, V, \alpha) = \text{softmax}(\frac{q_i K^T \cdot \alpha_i}{\sqrt{d}})V \qquad (2)$$

We integrate human-like attention on the question text into the first SA module in the encoder part of MCAN (see Figure 1) and on the image after the first GA module that integrates text and image. This early integration is motivated by Brunner et al. [9] who investigated the token mixing that occurs in self-attention layers. They found that the contribution of the original input token to the embedding at the same position quickly decreases after the first layer, making the integration of re-weighting attention weights less targeted at later layers. We opted to integrate human-like image attention in the SA module *after* the first GA module (as opposed to before) because this allows text-attention dependent features to interact during integration of human-like attention on the image. To obtain human-like attention scores for questions and images, we make use of domain-specific attention networks that we discuss in the following.

Text Attention Model. For text, we make use of the recently introduced Text Saliency Model (TSM) [41] that yields an attention weight for every token in the question. TSM is pre-trained on synthetic data obtained from a cognitive reading model as well as on real human gaze data. Sood et al. [41] proposed a joint training approach in which TSM predictions are integrated into the Luong attention layer [26] of a downstream NLP task and fine-tuned while training for this downstream task. We follow a similar methodology and fine-tune the TSM while training our VQA network.

Table 1. Results showing *test-std* and *test-dev* accuracy scores of our model (trained on *train+val+vg*) and ablated versions over different datasets. MULAN achieves competitive with state-of-the-art for its model class on both benchmarks.

Model	*test-std*	*test-dev*
MULAN (multimodal)	**73.98%**	**73.72%**
Text only (TSM)	73.77%	73.52%
Image only (MDS)	73.67%	73.39%
No Integration	73.65%	73.39%
Li et al. (2020)	73.82%	73.61%
Jiang et al. (2020)	72.71%	72.59%

Image Attention Model. For images, we obtain human-like attention using the state-of-the-art Multi-Duration Saliency (MDS) method [14]. MDS predicts human attention allocation for viewing durations of 0.5, 3, and 5 seconds. Because our integration approach requires a single attention map per image, we use the output of MDS for the 3 second viewing duration as suggested by the original authors. The input images are of different aspect ratios which leads to black borders in the obtained fixed-size MDS maps after re-scaling. However, the grid features are extracted from images with their original aspect-ratios. Therefore, to obtain a single attention weight per feature, we remove the borders from the MDS maps, overlay the grid features and sum the pixel values in every grid cell. The values are then normalized over the total sum to produce a distribution.

Implementation Details. We trained the network using the basic configuration and hyperparameters of MCAN_small. The input features were set to $d_y = 512$ and after the change to grid features, $d_x = 2048$. The hidden dimension d inside the Transformer heads was kept at 512. The increased dimensions in the MCAN_large configuration did not bring performance advantages in our preliminary experiments, so we opted for fewer parameters. In the added TSM model we set the hidden dimension for both BiLSTM and Transformer heads to 128. We used 4 heads and one layer. Even including the trainable parameters added by the TSM model, our full MULAN model has significantly fewer parameters than MCAN_large (MULAN: 58M, MCAN_large: 203M).

We kept the MCAN Adam Solver and the corresponding learning rate schedule and trained over 12 epochs with batch size 64. The pre-trained TSM model is trained jointly with the MCAN. Results on *test-std* were obtained after training on *train* and *val* splits and a subset of Visual Genome *vg*, for results on *val* we trained on *train*. We trained on single Nvidia Tesla V100-SXM2 GPUs with 32GB RAM. Average runtime was 2.6h for models trained on *train* (36h for convergence) and 5.3h for models trained on *train+val+vg* (68h for convergence).

4. Experiments

We used VQAv2[4], the balanced version [16] of the VQA dataset [6], for all our experiments. VQAv2 is among the most popular benchmark datasets in the field and contains 1.1M human-annotated questions on 200K images from MS COCO [24], split into *train*, *val*, and *test-dev*/*test-std* sets. The annotations for the test splits have been held back for online evaluation of models submitted to the annual challenge. The standard evaluation metric[5] is simple overall accuracy, for which agreement with three out of the 10 available annotator answers is considered as achieving an accuracy of 100%. In practice the machine accuracy is calculated as mean over all 9 out of 10 subsets.

The standard evaluation is misleading because the same overall accuracy can be achieved by answering very different sets of questions. To address this issue, we evaluated overall accuracy for comparison with other works, but also utilized a "per question-type" binning approach [20, 42] to compensate for class imbalance and answer bias skewing the evaluation. For this, we used their question types that categorize questions by the task they solve [20]. In addition, we used the *reading* category proposed by Sood et al. [42] for questions that are answered by text on the image. Furthermore, because Kafle and Kanan [20] only labelled about 8% of VQAv2 *val*, Sood et al. [42] extended the annotation on the full *val* set using a BiLSTM network pre-trained on TDIUC (1.6M samples) and hand-crafted regular expressions (245K samples). These additional annotations enabled us to assess the performance changes of our model in more detail.

We performed a series of experiments to evaluate the performance of our proposed method. To shed more light on the importance of multimodal integration, we first compared different ablated versions of our method on *test-dev* and *test-std*. Specifically, we compared multimodal integration with text-only, image-only, and integration of human-like attention. Afterwards, we validated our hypothesis in Section 3 that integration in early layers is more beneficial. To do this, we trained multiple models integrating human-like attention at different layers of the Transformer network. Finally, we evaluated more fine-grained performance values of the multimodal, unimodal, and no attention integration method for the different question types. For all of these experiments, we used fixed hyperparameters and report accuracy on *test-std* with training on the union of *train*, *val* and *vg* sets.

[4] visualqa.org/download.html
[5] visualqa.org/evaluation.html

Table 2. Layer-wise integration ablation study results, on *test-std*. We integrate human-like attention at different layer combinations. TSM question attention weights are integrated into encoder SA modules, MDS image attention weights into decoder SA modules.

TSM	MDS	*test-std*
1	2	**(ours) 73.98%**
2	2	73.64%
1, 3, 5	2	73.73%
1	1–6	71.55%
1–3	2	73.49%
1–6	2	73.73%
1–6	2–6	73.50%

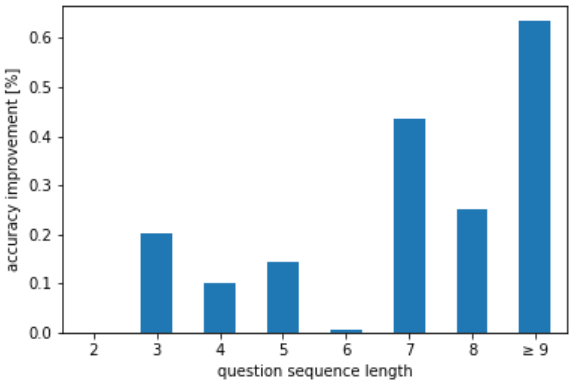

Figure 3. Performance improvements of our multimodal integration method (MULAN) relative to the baseline (*No Integration*), depending on the question length. Results show a significant increase in accuracy for longer questions, in particular when questions have seven or more tokens.

5. Results and Discussion

Overall Performance. Table 1 shows results of our MULAN model along with current state-of-the-art approaches and ablations of our method. Our method obtains competitive with current state of the art models [23], reaching accuracy scores of 73.98% on *test-std* and 73.72% on *test-dev*, as compared to 73.82% and 73.61%. Our model also uses approximately 80% less trainable parameters than Li et al. [23]. Notably, we observe a systematic increase in performance as a result of human-like attention integration. Our base model without any integration reaches 73.65% on *test-*

Table 3. Performance on VQAv2 *val* split in terms of per-question-type accuracy of the proposed multimodal integration method (MULAN) and the unimodal ablations *text-only* or *image-only* and no integration of human attention (*No Integration*). Because the online evaluation of VQAv2 only returns overall accuracy, we cannot obtain fine-grained accuracy for *test-std* or *test-dev*. A star indicates statistically significant p at $p < 0.05$.

Question type	Bin Size	No Integration	text-only	image-only	**MULAN**
reading	31 K	**42.46**	42.28	42.40	42.30
activity recognition	15 K	74.55	74.72	74.59	**75.01**
positional reasoning	26 K	61.74	61.97	61.85	**62.01**
object recognition	28 K	82.59	82.50	82.49	**82.68**
counting	24 K	59.77	59.70	59.44	**59.82**
object presence	17 K	86.45	86.47	**86.59**	86.57
scene recognition	15 K	79.19	79.10	**79.20**	79.19
sentiment understanding	14 K	83.59	83.77	83.53	**83.92**
color	25 K	**80.56**	80.52	80.31	**80.56**
attribute	4 K	69.36	69.09	69.37	**69.65**
utility affordance	11 K	66.33	66.40	**66.64**	66.42
sport recognition	6 K	85.39	85.38	**85.93**	85.60
Overall VQAv2 *val* Accuracy:		70.06	70.09	70.03	**70.28***

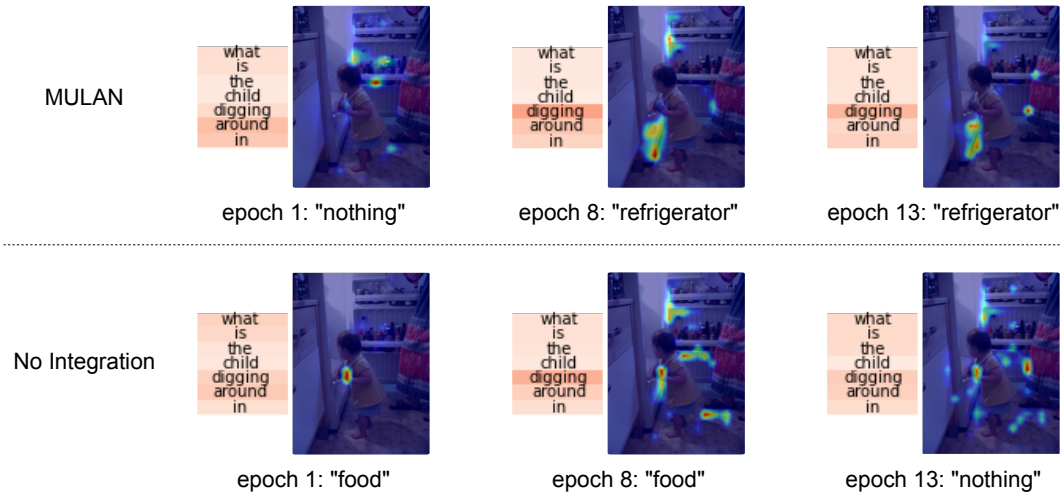

Figure 4. Visualization of the weights in the attention reduction modules for text and image features. We compared MULAN and the baseline model (No Integration) at epochs 1, 8, and 13. The input question was "What is the child digging around in?", the correct answer is "fridge". Classification outputs are given for each considered epoch.

std. While the integration of human-like attention on only images (73.67%) or text (73.77% on *test-std*) can lead to an increase in performance, our full MULAN model employing multimodal integration is the best performing approach. This further underlines the importance of jointly integrating human-like attention on both text and images for the VQA task, which is inherently multimodal.

Layer-Wise Integration Experiments. We evaluated the integration of human-like attention on questions and images for different layers in the MCAN encoder-decoder architecture (see Table 2). We investigated the integration of TSM outputs into different SA layers of the encoder, and the integration of MDS outputs into different SA modules of the decoder. Among all investigated combinations, the initial integration into the first layer of the encoder and the second layer of the decoder performed best (73.98% accuracy). Integrating TSM predictions into the second encoder layer decreased the overall accuracy to 73.64%, which is in line with the reasoning discussed in Brunner et al. [9], where with layer depth feature embeddings are increasingly mixed and therefore less attributable to the input word token at the same position. The TSM predicts attention weights for specific word tokens. We further investigated the integration of

TSM and MDS predictions at multiple layers in the MCAN architecture. However all options resulted in decreased performance in comparison to MULAN. Our results indicate that early integration of human-like attention at a single point for both text and image is optimal for the VQA task.

Category-Specific Performance. To obtain a deeper understanding of our improvements over baseline approaches, we categorized question types into 12 fine-grained bins, similarly to Kafle and Kanan [20]. Table 3 shows a detailed breakdown of accuracy results by category type. We used the validation set, rather than the test set, since we needed access to the ground truth annotations to calculate the per-category accuracy. For the same reason, we can only perform the paired t-test on the full validation set. As can be seen, all ablated models obtain inferior performance to our full model on the validation set (statistically significant at the 0.05 level). For most categories, MULAN achieves the highest accuracy. Moreover, in comparison to the baseline, our method is the best performing one in 10 out of 12 categories with especially clear improvements in activity recognition and sentiment analysis categories. MULAN expectedly reduces accuracy on reading questions that other models can most likely only answer by bias exploitation, and improves on small bins like attribute. The distances between the models are small in absolute terms, but given the vastly different bin sizes the relative improvements are large. This underlines the robustness of improvements with human-like attention integration and, in particular, multimodal integration.

Sequence-Length Analysis. Previous works have shown that VQA models often converge to the answer after processing only the first words of a question, a behavior that has been characterized as "jumping to conclusions" [1]. Human-like attention integration might be useful to combat this effect as the TSM was previously shown to successfully predict human-like attention distributions across all salient words in the input sequence [41]. As this effect might be especially pronounced for longer questions, we investigated whether human-like attention integration in MULAN can especially improve on those questions [1]. Figure 3 shows the results of an evaluation where we analyzed the improvements of our system relative to the baseline model, depending on the question length. We find that while MULAN improves for all questions independent of their length, its advantage is especially significant for questions that contain seven tokens or more (relative improvements of 0.3% or more), indicating that MULAN can improve upon the above-described challenge.

Attention Visualizations. To further investigate MULAN's ability to answer longer questions than the baseline model, we visualized the attention weights in the attention reduction modules after the encoder/decoder, which merge the text and image features to one attended feature vector each. These weights represent the final impact of the transformed features. Figure 4 shows examples from the validation set with the corresponding predictions comparing our method to the baseline at epochs 1, 8 and 13. The input question was "What is the child digging around in?" and the correct answer is "fridge". Our method it able to correctly predict that the child is digging in the fridge as opposed to the baseline that outputs "nothing". MULAN focuses on both the token "digging" as well as the location, which is in front of the child. In contrast, the attention of the baseline model is more spread out, failing to focus on the relevant cues. Interestingly, the focus of attention in the baseline evolved over several epochs of training, unlike MULAN which quickly converged to a stable attention distribution. This indicates that initial human-like attention maps on the image are indeed adapted using attention-based information extracted from the question text. Figure 5 shows three additional examples of our method compared to the baseline from the final epoch. The top and middle examples show how our method is able to correctly answer the question, while the baseline fails. The bottom example shows an error case of our method.

Limitations Our work is limited in that we do not fully explore the possible methods for integration in transformer networks. Recently works have investigate that specific layers are more conducive to human gaze [8, 13]. Our layer-wise experiments explore this direction, however in the future we see the need for further investigation of more appropriate layer specific integration strategies. We also identified a some risks and ethical concerns. By aiming to integrate human data into and neural attention layers of deep learning models, we allow for the potential of user biases exploitation. Perhaps, if there is work which uses our method, one could develop a tool to discriminate against particular users based on their attentive behaviors.

6. Conclusion

In this paper, we propose the first method for multimodal integration of human-like attention on both image and text for visual question answering. Our Multimodal Human-like Attention Network (MULAN) method integrates state-of-the-art text and image saliency models into neural self-attention layers by modifying attention scoring functions of transformer-based self-attention modules. Evaluations on the challenging VQAv2 dataset show that our method not only achieves competitive with state-of-the-art performance in its model class (73.98% on *test-std* and 73.72% on *test-dev*), but also does so with significantly fewer trainable parameters than current models. As such, our work provides further

Figure 5. Visualization of attention distributions for MULAN and the baseline (No Integration). The upper examples show improvements of MULAN over the baseline, while the bottom shows a failure case.

evidence for the potential of integrating human-like attention as a supervisory signal in neural attention mechanisms.

7. Acknowledgments

E. Sood was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. A. Bulling was funded by the European Research Council (ERC; grant agreement 801708). P. Müller was funded by the German Ministry for Education and Research (BMBF; grant number 01IS20075). M. Bâce was funded by a Swiss National Science Foundation (SNSF) Early Postdoc.Mobility Fellowship (grant number 199991).
We would like to especially thank Simon Tannert and Prajit Dhar for their valuable insights and support as well as to Pavel Denisov and Manuel Mager for their helpful suggestions. Lastly, we would like to thank the anonymous reviewers for their useful feedback.

References

[1] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the Behavior of Visual Question Answering Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1955–1960. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1203. URL https://www.aclweb.org/anthology/D16-1203. 2, 7

[2] Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-VQA: A Compositional Split of the Visual Question Answering (VQA) v1.0 Dataset. In *arXiv:1704.08243*, 2017. URL https://arxiv.org/abs/1704.08243. 2

[3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Agrawal_Dont_Just_Assume_CVPR_2018_paper.html. 2

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 6077–6086, 2018. doi: 10.1109/CVPR.2018.00636. URL https://ieeexplore.ieee.org/document/8578734. 1, 2, 3

[5] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural Module Networks. In *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 39–48, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Andreas_Neural_Module_Networks_CVPR_2016_paper.html. 2

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. doi: 10.1109/ICCV.2015.279. URL https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Antol_VQA_Visual_Question_ICCV_2015_paper.pdf. 1, 2, 5

[7] Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. Sequence Classification with Human Attention. In *Proceedings of the ACL SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 302–312, 2018. doi: http://dx.doi.org/10.18653/v1/K18-1030. URL https://www.aclweb.org/anthology/K18-1030. 1, 2

[8] Joshua Bensemann, Alex Peng, Diana Prado, Yang Chen, Neset Tan, Paul Michael Corballis, Patricia Riddle, and

Michael J Witbrock. Eye gaze and self-attention: How humans and transformers attend words in sentences. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 75–87, 2022. 7

[9] Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On Identifiability in Transformers. In *International Conference on Learning Representations (ICLR)*, pages 1–35, 2020. URL https://openreview.net/forum?id=BJg1f6EFDB. 4, 6

[10] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. AiR: Attention with Reasoning Capability. In *European Conference on Computer Vision (ECCV)*, 2020. doi: https://doi.org/10.1007/978-3-030-58452-8_6. URL https://link.springer.com/chapter/10.1007/978-3-030-58452-8_6. 1, 2, 4

[11] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 932–937, June 2016. doi: 10.18653/v1/D16-1092. URL https://www.aclweb.org/anthology/D16-1092. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A Large-Scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. 3

[13] Oliver Eberle, Stephanie Brandl, Jonas Pilot, and Anders Søgaard. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, 2022. 7

[14] Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. How Much Time Do You Have? Modeling Multi-Duration Saliency. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 4473–4482, 2020. doi: 10.1109/CVPR42600.2020.00453. URL https://www.computer.org/csdl/proceedings-article/cvpr/2020/716800e472/1m3njRPpHdC. 1, 2, 3, 4

[15] Chuang Gan, Yandong Li, Haoxiang Li, Chen Sun, and Boqing Gong. VQS: Linking Segmentations to Questions and Answers for Supervised Attention in VQA and Question-Focused Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1811–1820, 2017. doi: 10.1109/ICCV.2017.201. URL https://ieeexplore.ieee.org/document/8237463. 2, 4

[16] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334, 2017. doi: 10.1109/CVPR.2017.670. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Goyal_Making_the_v_CVPR_2017_paper.html. 1, 2, 5

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. doi: 10.1109/CVPR.2016.90. 3

[18] Huaizu Jiang, Ishan Misra, Marcus Rohrbach, Erik Learned-Miller, and Xinlei Chen. In Defense of Grid Features for Visual Question Answering. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/papers/Jiang_In_Defense_of_Grid_Features_for_Visual_Question_Answering_CVPR_2020_paper.pdf. 1, 2, 3

[19] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the Winning Entry to the VQA Challenge 2018. 2018. URL http://arxiv.org/abs/1807.09956. 2

[20] Kushal Kafle and Christopher Kanan. An Analysis of Visual Question Answering Algorithms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1983–1991, 2017. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.217. URL http://ieeexplore.ieee.org/document/8237479/. 5, 7

[21] Nour Karessli, Zeynep Akata, Bernt Schiele, and Andreas Bulling. Gaze Embeddings for Zero-Shot Image Classification. In *Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4525–4534, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/papers/Karessli_Gaze_Embeddings_for_CVPR_2017_paper.pdf. 2

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision*, 2016. doi: 10.1007/s11263-016-0981-7. 3

[23] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-Semantics

Aligned Pre-training for Vision-Language Tasks. In *European Conference on Computer Vision (ECCV)*, volume 12375, pages 121–137, 2020. doi: 10.1007/978-3-030-58577-8_8. URL http://link.springer.com/10.1007/978-3-030-58577-8_8. 2, 5

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48. 5

[25] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pages 1–9, 2016. URL https://papers.nips.cc/paper/2016/hash/9dcb88e0137649590b755372b040afad-Abstract.html. 2

[26] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective Approaches to Attention-Based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015. doi: 10.18653/v1/D15-1166. URL https://www.aclweb.org/anthology/D15-1166. 4

[27] Mateusz Malinowski and Mario Fritz. A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 27, pages 1682–1690, 2014. URL https://proceedings.neurips.cc/paper/2014/file/d516b13671a4179d9b7b458a6ebdeb92-Paper.pdf. 1, 2

[28] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask Your Neurons: A Neural-Based Approach to Answering Questions about Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–9, 2015. doi: 10.1109/ICCV.2015.9. 1, 2

[29] Sruthy Manmadhan and Binsu C. Kovoor. Visual question answering: A state-of-the-art review. *Artif. Intell. Rev.*, 53(8):5705–5745, dec 2020. ISSN 0269-2821. doi: 10.1007/s10462-020-09832-7. URL https://doi.org/10.1007/s10462-020-09832-7. 1

[30] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual Attention Networks for Multimodal Reasoning and Matching. In *Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 2156–2164, 2017. doi: 10.1109/CVPR.2017.232. URL https://www.computer.org/csdl/proceedings-article/cvpr/2017/0457c156/12OmNBDQbn4. 2

[31] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162. 3

[32] Tingting Qiao, Jianfeng Dong, and Duanqing Xu. Exploring Human-Like Attention Supervision in Visual Question Answering. In *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 7300–7307, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16485. 1, 2, 4

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems*, volume 28, pages 91–99, 2015. URL https://proceedings.neurips.cc/paper/2015/file/14bfa6bb14875e45bba028a21ed38046-Paper.pdf. 3

[34] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: Leveraging Explanations to Make Vision and Language Models More Grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. doi: 10.1109/ICCV.2019.00268. URL https://ieeexplore.ieee.org/document/9009041. 1, 2, 4

[35] Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar. SQuINTing at VQA Models: Introspecting VQA Models with Sub-Questions. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10003–10011, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Selvaraju_SQuINTing_at_VQA_Models_Introspecting_VQA_Models_With_Sub-Questions_CVPR_2020_paper.html. 2

[36] Meet Shah, Xinlei Chen, Marcus Rohrbach, and Devi Parikh. Cycle-Consistency for Robust Visual Question Answering. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6642–6651. IEEE, 2019. doi: 10.1109/CVPR.2019.00681. URL https://ieeexplore.ieee.org/document/8954214/. 2

[37] Kevin J Shih, Saurabh Singh, and Derek Hoiem. Where to Look: Focus Regions for Visual Question Answering. In *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4613–4621, 2016. doi: 10.1109/CVPR.2016.499. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Shih_Where_to_Look_CVPR_2016_paper.pdf. 2

[38] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Answer Them All! Toward Universal Visual Question Answering Models. In *arXiv:1903.00366*, April 2019. URL http://arxiv.org/abs/1903.00366. 2

[39] Robik Shrestha, Kushal Kafle, and Christopher Kanan. A Negative Case Analysis of Visual Grounding Methods for VQA. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8172–8181. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.727. URL https://www.aclweb.org/anthology/2020.acl-main.727. 1, 4

[40] Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. Interpreting Attention Models with Human Visual Attention in Machine Reading Comprehension. In *Proceedings of the ACL SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 12–25, 2020. doi: 10.18653/v1/P17. 2

[41] Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. Improving Natural Language Processing Tasks with Human Gaze-Guided Neural Attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–15, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/460191c72f67e90150a093b4585e7eb4-Abstract.html. 1, 2, 3, 4, 7

[42] Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. Vqa-mhug: A gaze dataset to study multimodal neural attention in vqa. In *Proc. ACL SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 27–43. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.conll-1.3. 5

[43] Yusuke Sugano and Andreas Bulling. Seeing with Humans: Gaze-Assisted Neural Image Captioning. In *arXiv:1608.05203*, 2016. URL http://arxiv.org/abs/1608.05203. 1, 2

[44] Hao Tan and Mohit Bansal. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5100–5111, 2019. doi: 10.18653/v1/D19-1514. URL https://www.aclweb.org/anthology/D19-1514/. 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf. 1, 2, 3

[46] Jialin Wu and Raymond J. Mooney. Self-Critical Reasoning for Robust Visual Question Answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–11, 2019. URL https://papers.nips.cc/paper/2019/file/33b879e7ab79f56af1e88359f9314a10-Paper.pdf. 1, 2, 4

[47] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-Level Attention Networks for Visual Question Answering. In *Proceedings of the 2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4187–4195, 2017. doi: 10.1109/CVPR.2017.446. 1, 2

[48] Zhou Yu, Yuhao Cui, Zhenwei Shao, Pengbing Gao, and Jun Yu. OpenVQA. https://github.com/MILVLG/openvqa, 2019. 2

[49] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep Modular Co-Attention Networks for Visual Question Answering. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6281–6290, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Yu_Deep_Modular_Co-Attention_Networks_for_Visual_Question_Answering_CVPR_2019_paper.html. 1, 2, 3

[50] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and Answering Binary Visual Questions. In *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5014–5022, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Zhang_Yin_and_Yang_CVPR_2016_paper.html. 2