Localized Questions in Medical Visual Question Answering

Sergio Tascon-Morales ✉, Pablo Márquez-Neila, Raphael Sznitman

University of Bern, Bern, Switzerland
{sergio.tasconmorales, pablo.marquez, raphael.sznitman}@unibe.ch

Abstract. Visual Question Answering (VQA) models aim to answer natural language questions about given images. Due to its ability to ask questions that differ from those used when training the model, medical VQA has received substantial attention in recent years. However, existing medical VQA models typically focus on answering questions that refer to an entire image rather than where the relevant content may be located in the image. Consequently, VQA models are limited in their interpretability power and the possibility to probe the model about specific image regions. This paper proposes a novel approach for medical VQA that addresses this limitation by developing a model that can answer questions about image regions while considering the context necessary to answer the questions. Our experimental results demonstrate the effectiveness of our proposed model, outperforming existing methods on three datasets. Our code and data are available at https://github.com/sergiotasconmorales/locvqa.

Keywords: VQA · Attention · Localized Questions

1 Introduction

Visual Question Answering (VQA) models are neural networks that answer natural language questions about an image [2,8,12,21]. The capability of VQA models to interpret natural language questions is of great appeal, as the range of possible questions that can be asked is vast and can differ from those used to train the models. This has led to many proposed VQA models for medical applications in recent years [9,16,14,23,25,7,24]. These models can enable clinicians to probe the model with nuanced questions, thus helping to build confidence in its predictions.

Recent work on medical VQA has primarily focused on building more effective model architectures [7,20,23] or developing strategies to overcome limitations in medical VQA datasets [18,15,19,4,23]. Another emerging trend is to enhance VQA performance by addressing the consistency of answers produced [22], particularly when considering entailment questions (*i.e.*, the answer to "Is the image that of a healthy subject?" should be consistent with the answer to "Is there a fracture in the tibia?"). Despite these recent advances, however, most VQA models restrict to questions that consider the entire image at a time. Specifically,

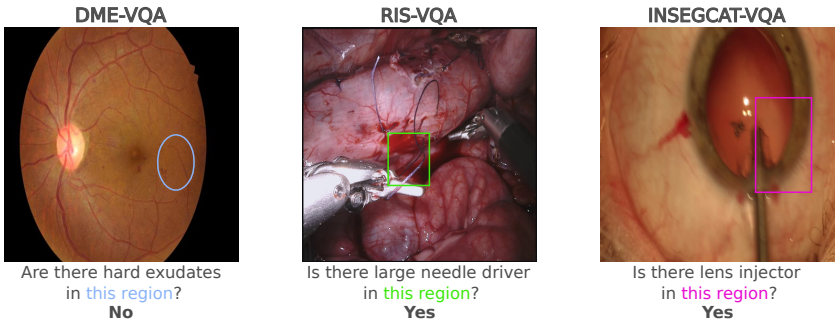

Fig. 1. Examples of localized questions. In some cases (RIS-VQA and INSEGCAT-VQA), the object mentioned in the question is only partially present in the region. We hypothesize that context can play an important role in answering such questions.

VQA typically uses questions that address content within an image without specifying where this content may or may not be in the image. Yet the ability to ask specific questions about regions or locations of the image would be highly beneficial to any user as it would allow fine-grained questions and model probing. For instance, Fig. 1 illustrates examples of such *localized questions* that combine content and spatial specifications. In the medical field, posing localized questions can significantly enhance the diagnostic process by providing second opinions to medical experts about suspicious regions. Additionally, this approach can improve trustworthiness by assessing the consistency between answers to both global and localized questions.

To this day, few works have addressed the ability to include location information in VQA models. In [17], localization information is posed in questions by constraining the spatial extent to a point within bounding boxes yielded by an object detector. The model then focuses its attention on objects close to this point. However, the method was developed for natural images and relies heavily on the object detector to limit the attention extent, making it difficult to scale in medical imaging applications. Alternatively, the approach from [23] answers questions about a pre-defined coarse grid of regions by directly including region information into the question (*e.g.*, "Is grasper in (0,0) to (32,32)?"). This method relies on the ability of the model to learn a spatial mapping of the image and limits the regions to be on a fixed grid. Localized questions were also considered in [22], but the region of interest was cropped before being presented to the model, assuming that the surrounding context is irrelevant for answering this type of question.

To overcome these limitations, we propose a novel VQA architecture that alleviates the mentioned issues. At its core, we hypothesize that by allowing the VQA model to access the entire images and properly encoding the region of interest, this model can be more effective at answering questions about regions. To achieve this, we propose using a multi-glimpse attention mechanism [3,23,22]

restricting its focus range to the region in question, but only after the model has considered the entire image. By doing so, we preserve contextual information about the question and its region. We evaluate the effectiveness of our approach by conducting extensive experiments on three datasets and comparing our method to state-of-the-art baselines. Our results demonstrate performance improvements across all datasets.

2 Method

Our method extends a VQA model to answer localized questions. We define a *localized question* for an image $\mathbf{x}$ as a tuple $(\mathbf{q}, \mathbf{m})$, where $\mathbf{q}$ is a question, and $\mathbf{m}$ is a binary mask of the same size as $\mathbf{x}$ that identifies the region to which the question pertains. Our VQA model p_θ, depicted in Fig. 2, accepts an image and a localized question as input and produces a probability distribution over a finite set $\mathcal{A}$ of possible answers. The final answer of the model $\hat{a}$ is the element with the highest probability,

$$\hat{a} = \arg\max_{a \in \mathcal{A}} p_\theta(a \mid \mathbf{q}, \mathbf{x}, \mathbf{m}). \tag{1}$$

The model proceeds in three stages to produce its prediction: input embedding, localized attention, and final classification.

Input embedding. The question $\mathbf{q}$ is first processed by an LSTM [11] to produce an embedding $\hat{\mathbf{q}} \in \mathbb{R}^Q$. Similarly, the image $\mathbf{x}$ is processed by a ResNet-152 [10] to produce the feature map $\hat{\mathbf{x}} \in \mathbb{R}^{C \times H \times W}$.

Localized attention. An attention mechanism uses the embedding to determine relevant parts of the image to answer the corresponding question. Unlike previous attention methods, we include the region information that the mask defines. Our *localized attention* module (Fig. 2 right) uses both descriptors and the mask to produce multiple weighted versions of the image feature map, $\hat{\mathbf{x}}' = \mathrm{att}(\hat{\mathbf{q}}, \hat{\mathbf{x}}, \mathbf{m})$. To do so, the module first computes an attention map $\mathbf{g} \in \mathbb{R}^{G \times H \times W}$ with G glimpses by applying unmasked attention [13,23] to the image feature map and the text descriptor. The value of the attention map at location (h, w) is computed as,

$$\mathbf{g}_{:hw} = \mathrm{softmax}\left(\mathbf{W}^{(g)} \cdot \mathrm{ReLU}\left(\mathbf{W}^{(x)}\hat{\mathbf{x}}_{:hw} \odot \mathbf{W}^{(q)}\hat{\mathbf{q}}\right)\right), \tag{2}$$

where the index $:hw$ indicates the feature vector at location (h, w), $\mathbf{W}^{(x)} \in \mathbb{R}^{C' \times C}$, $\mathbf{W}^{(q)} \in \mathbb{R}^{C' \times Q}$, and $\mathbf{W}^{(g)} \in \mathbb{R}^{G \times C'}$ are learnable parameters of linear transformations, and $\odot$ is the element-wise product. In practice, the transformations $\mathbf{W}^{(x)}$ and $\mathbf{W}^{(g)}$ are implemented with 1×1 convolutions and all linear transformations include a dropout layer applied to its input. The image feature maps $\hat{\mathbf{x}}$ are then weighted with the attention map and masked with $\mathbf{m}$ as,

$$\hat{\mathbf{x}}'_{cghw} = \mathbf{g}_{ghw} \cdot \hat{\mathbf{x}}_{chw} \cdot (\mathbf{m} \downarrow_{H \times W})_{hw}, \tag{3}$$

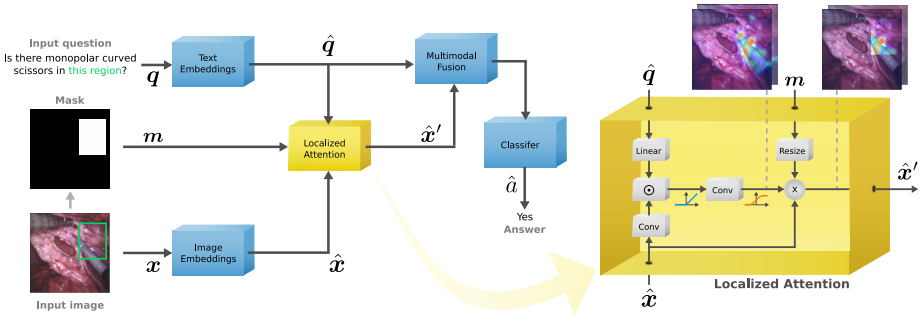

Fig. 2. Left: Proposed VQA architecture for localized questions. The Localized Attention module allows the region information to be integrated into the VQA while considering the context necessary to answer the question. **Right:** Localized Attention module.

where c and g are the indexes over feature channels and glimpses, respectively, (h, w) is the index over the spatial dimensions, and $\mathbf{m}\downarrow_{H\times W}$ denotes a binary downsampled version of $\mathbf{m}$ with the spatial size of $\hat{\mathbf{x}}$. This design allows the localized attention module to compute the attention maps using the full information available in the image, thereby incorporating context into them before being masked to constrain the answer to the specified region.

Classification. The question descriptor $\hat{\mathbf{q}}$ and the weighted feature maps $\hat{\mathbf{x}}'$ from the localized attention are vectorized and concatenated into a single vector of size $C \cdot G + Q$ and then processed by a multi-layer perceptron classifier to produce the final probabilities.

Training. The training procedure minimizes the standard cross-entropy loss over the training set updating the parameters of the LSTM encoder, localized attention module, and the final classifier. The training set consists of triplets of images, localized questions, and the corresponding ground-truth answers. As in [2], the ResNet weights are fixed with pre-trained values, and the LSTM weights are updated during training.

3 Experiments and results

We compare our model to several baselines across three datasets and report quantitative and qualitative results. Additional results are available in the supplementary material.

3.1 Datasets

We evaluate our method on three datasets containing questions about regions which we detail here. The first dataset consists of an existing retinal fundus VQA dataset with questions about the image's regions and the entire image.

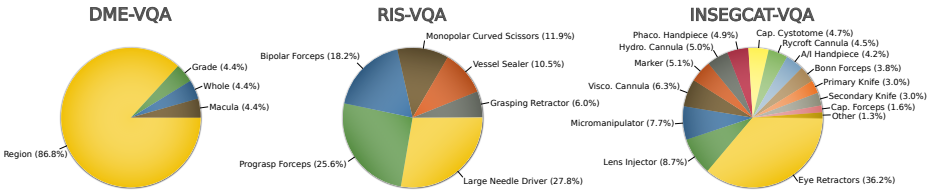

Fig. 3. Distribution by question type (DME-VQA) and by question object (RIS-VQA and INSEGCAT-VQA).

The second and third datasets are generated from public segmentation datasets but use the method described in [23] to generate a VQA version with region questions.

DME-VQA [22]. 679 fundus images containing questions about entire images (*e.g.*, "what is the DME risk grade?") and about randomly generated circular regions (*e.g.*, "are there hard exudates in this region?"). The dataset comprises 9'779 question-answer (QA) pairs for training, 2'380 for validation, and 1'311 for testing.

RIS-VQA. Images from the 2017 Robotic Instrument Segmentation dataset [1]. We automatically generated binary questions with the structure "is there [instrument] in this region?" and corresponding masks as rectangular regions with random locations and sizes. Based on the ground-truth label maps, the binary answers were labeled "yes" if the region contained at least one pixel of the corresponding instrument and "no" otherwise. The questions were balanced to maintain the same amount of "yes" and "no" answers. 15'580 QA pairs from 1'423 images were used for training, 3'930 from 355 images for validation, and 13'052 from 1'200 images for testing.

INSEGCAT-VQA. Frames of cataract surgery videos from the InSegCat 2 dataset [5]. We followed the same procedure as in RIS-VQA to generate balanced binary questions with masks and answers. The dataset consists of 29'380 QA pairs from 3'519 images for training, 5'306 from 536 images for validation, and 4'322 from 592 images for testing.

Fig. 3 shows the distribution of questions in the three datasets.

3.2 Baselines and metrics

We compare our method to four different baselines, as shown in Fig. 4:

No mask: no information is provided about the region in the question.
Region in text [23]: region information is included as text in the question.
Crop region [22]: image is masked to show only the queried region, with the area outside the region set to zero.
Draw region: region is indicated by drawing its boundary on the input image with a distinctive color.

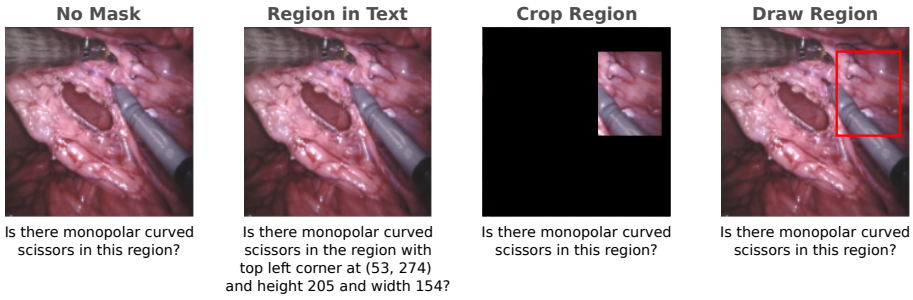

Fig. 4. Illustration of evaluated baselines for an example image.

We evaluated the performance of our method using accuracy for the DME-VQA dataset and the area under the receiver operating characteristic (ROC) curve and Average Precision (AP) for the RIS-VQA and INSEGCAT-VQA datasets.

Implementation details: Our VQA architecture uses an LSTM [11] with an output dimension 1024 to encode the question and a word embedding size of 300. We use the ResNet-152 [10] with ImageNet weights to encode images of size 448×448, generating feature maps with 2048 channels. In the localized attention block, the visual and textual features are projected into a 512-dimensional space before being combined by element-wise multiplication. Following [6,13], the number of glimpses is set to $G = 2$ for all experiments. The classification block is a multi-layer perceptron with a hidden layer of 1024 dimensions. A dropout rate of 0.25 and ReLU activation are used in the localized attention and classifier blocks.

We train our models for 100 epochs using an early stopping condition with patience of 20 epochs. Data augmentation consists of horizontal flips. We use a batch size of 64 samples and the Adam optimizer with a learning rate of 10^{-4}, which is reduced by a factor of 0.1 when learning stagnates. Models implemented in PyTorch 1.13.1 and trained on an Nvidia RTX 3090 graphics card.

3.3 Results

Our method outperformed all considered baselines on the DME-VQA (Table 1), the RIS-VQA, and the INSEGCAT-VQA datasets (Table 2), highlighting the importance of contextual information in answering localized questions. Context proved to be particularly critical in distinguishing between objects of similar appearance, such as the bipolar and prograsp forceps in RIS-VQA, where our method led to an 8 percent point performance improvement (Table 3). In contrast, the importance of context was reduced when dealing with visually distinct objects, resulting in smaller performance gains as observed in the INSEGCAT-VQA dataset. For example, despite not incorporating contextual information, the baseline *crop region* still benefited from correlations between the location of

Method	Accuracy (%)				
	Overall	Grade	Whole	Macula	Region
No Mask	61.1 ± 0.4	80.0 ± 3.7	85.7 ± 1.2	**84.3 ± 0.5**	57.6 ± 0.4
Region in Text [23]	60.0 ± 1.5	57.9 ± 12.5	85.1 ± 1.9	83.2 ± 2.4	57.7 ± 1.0
Crop Region [22]	81.4 ± 0.3	78.7 ± 1.3	81.3 ± 1.7	82.3 ± 1.4	81.5 ± 0.3
Draw Region	83.0 ± 1.0	79.6 ± 2.5	77.0 ± 4.8	84.0 ± 1.9	83.5 ± 1.0
Ours	**84.2 ± 0.6**	**82.8 ± 0.4**	**87.0 ± 1.2**	83.0 ± 1.5	**84.2 ± 0.7**

Table 1. Average accuracy for different methods on the DME-VQA dataset. The results shown are the average of 5 models trained with different seeds.

Dataset	Method	AUC	AP
RIS-VQA	No Mask	0.500 ± 0.000	0.500 ± 0.000
	Region in Text [23]	0.677 ± 0.002	0.655 ± 0.003
	Crop Region [22]	0.842 ± 0.002	0.831 ± 0.002
	Draw Region	0.835 ± 0.003	0.829 ± 0.003
	Ours	**0.885 ± 0.003**	**0.885 ± 0.003**
INSEGCAT-VQA	No Mask	0.500 ± 0.000	0.500 ± 0.000
	Region in Text [23]	0.801 ± 0.012	0.793 ± 0.014
	Crop Region [22]	0.901 ± 0.002	0.891 ± 0.003
	Draw Region	0.910 ± 0.003	0.907 ± 0.005
	Ours	**0.914 ± 0.002**	**0.915 ± 0.002**

Table 2. Average test AUC and AP for different methods on the RIS-VQA and INSEGCAT-VQA datasets. The results shown are the average over 5 seeds.

the region and the instrument mentioned in the question (*e.g.*, the eye retractor typically appears at the top or the bottom of the image), enabling it to achieve competitive performance levels that are less than 2 percent points lower than our method (Table 2, bottom).

Similar to our method, the baseline *draw region* incorporates contextual information when answering localized questions. However, we observed that drawing regions on the image can interfere with the computation of guided attention maps, leading to incorrect predictions (Fig. 5, column 4). In addition, the lack of masking of the attention maps often led the model to wrongly consider areas beyond the region of interest while answering questions (Fig. 5, column 1).

When analyzing mistakes made by our model, we observe that they tend to occur when objects or background structures in the image look similar to the object mentioned in the question (Fig. 5, column 3). Similarly, false predictions were observed when only a few pixels of the object mentioned in the question were present in the region.

Method	Instrument Type					
	Large Needle Driver	Monopolar Curved Scissors	Vessel Sealer	Grasping Retractor	Prograsp Forceps	Bipolar Forceps
No Mask	0.500 ±0	0.500 ±0	0.500 ±0	0.500 ±0	0.500 ±0	0.500 ±0
Region in Text [23]	0.717 ±0.003	0.674 ±0.001	0.620 ±0.011	0.616 ±0.020	0.647 ±0.008	0.645 ±0.003
Crop Region [22]	0.913 ±0.002	0.812 ±0.003	0.752 ±0.009	0.715 ±0.015	0.773 ±0.003	0.798 ±0.004
Draw Region	0.915 ±0.003	0.777 ±0.003	0.783 ±0.004	0.709 ±0.012	0.755 ±0.004	0.805 ±0.005
Ours	**0.944 ±0.001**	**0.837 ±0.005**	**0.872 ±0.008**	**0.720 ±0.031**	**0.834 ±0.006**	**0.880 ±0.003**

Table 3. Average test AUC for different methods on the RIS-VQA dataset as a function of instrument type. Results are averaged over 5 models trained with different seeds. The corresponding table for INSEGCAT-VQA is available in the Supplementary Materials.

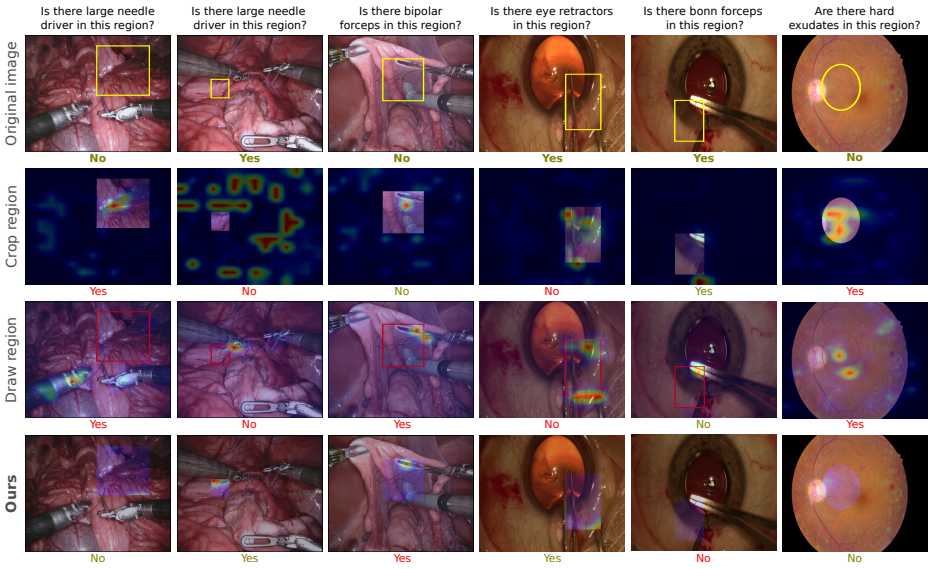

Fig. 5. Qualitative examples on the RIS-VQA dataset (columns 1-3), INSEGCAT-VQA (columns 4-5), and DME-VQA (last column). Only the strongest baselines were considered in this comparison. The first row shows the image, the region, and the ground truth answer. Other rows show the overlaid attention maps and the answers produced by each model. Wrong answers are shown in red. Additional examples are available in the Supplementary Materials.

4 Conclusions

In this paper, we proposed a novel VQA architecture to answer questions about regions. We compare the performance of our approach against several baselines and across three different datasets. By focusing the model's attention on the region after considering the evidence in the full image, we show how our method brings improvements, especially when the complete image context is required to answer the questions. Future works include studying the agreement between answers to questions about concentric regions, as well as the agreement between questions about images and regions.

Acknowledgments. This work was partially funded by the Swiss National Science Foundation through grant 191983.

References

1. Allan, M., Shvets, A., Kurmann, T., Zhang, Z., Duggal, R., Su, Y.H., Rieke, N., Laina, I., Kalavakonda, N., Bodenstedt, S., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
3. Ben-Younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2612–2620 (2017)
4. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple meta-model quantifying for medical visual question answering. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. pp. 64–74. Springer (2021)
5. Fox, M., Taschwer, M., Schoeffmann, K.: Pixel-based tool segmentation in cataract surgery videos with mask R-CNN. In: de Herrera, A.G.S., González, A.R., Santosh, K.C., Temesgen, Z., Kane, B., Soda, P. (eds.) 33rd IEEE International Symposium on Computer-Based Medical Systems, CBMS 2020, Rochester, MN, USA, July 28-30, 2020. pp. 565–568. IEEE (2020). https://doi.org/10.1109/CBMS49503.2020.00112, https://doi.org/10.1109/CBMS49503.2020.00112
6. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
7. Gong, H., Chen, G., Liu, S., Yu, Y., Li, G.: Cross-modal self-attention with multi-task pre-training for medical visual question answering. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. pp. 456–460 (2021)
8. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6904–6913 (2017)

9. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Lungren, M., Müller, H.: Overview of the ImageCLEF 2018 medical domain visual question answering task. In: CLEF2018 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org <http://ceur-ws.org>, Avignon, France (September 10-14 2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
12. Hudson, D.A., Manning, C.D.: Gqa: a new dataset for compositional question answering over real-world images. arXiv preprint arXiv:1902.09506 **3**(8) (2019)
13. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv preprint arXiv:1610.04325 (2016)
14. Liao, Z., Wu, Q., Shen, C., Van Den Hengel, A., Verjans, J.: Aiml at vqa-med 2020: Knowledge inference via a skeleton-based sentence mapping approach for medical domain visual question answering (2020)
15. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1650–1654. IEEE (2021)
16. Liu, F., Peng, Y., Rosen, M.P.: An effective deep transfer learning and information fusion framework for medical visual question answering. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 238–247. Springer (2019)
17. Mani, A., Yoo, N., Hinthorn, W., Russakovsky, O.: Point and ask: Incorporating pointing into visual question answering. arXiv preprint arXiv:2011.13681 (2020)
18. Nguyen, B.D., Do, T.T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question answering. In: Shen, D., Liu, T., Peters, T.M., Staib, L.H., Essert, C., Zhou, S., Yap, P.T., Khan, A. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. pp. 522–530. Springer International Publishing, Cham (2019)
19. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology objects in context (roco): a multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3. pp. 180–189. Springer (2018)
20. Ren, F., Zhou, Y.: Cgmvqa: A new classification and generative model for medical visual question answering. IEEE Access **8**, 50626–50636 (2020)
21. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
22. Tascon-Morales, S., Márquez-Neila, P., Sznitman, R.: Consistency-preserving visual question answering in medical imaging. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII. pp. 386–395. Springer (2022)
23. Vu, M.H., Löfstedt, T., Nyholm, T., Sznitman, R.: A question-centric model for visual question answering in medical imaging. IEEE transactions on medical imaging **39**(9), 2856–2868 (2020)

24. Yu, Y., Li, H., Shi, H., Li, L., Xiao, J.: Question-guided feature pyramid network for medical visual question answering. Expert Systems with Applications **214**, 119148 (2023)
25. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2345–2354 (2020)