Consistency-preserving Visual Question Answering in Medical Imaging

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 take an image and a natural-language question as input and infer the answer to the question. Recently, VQA systems in medical imaging have gained popularity thanks to potential advantages such as patient engagement and second opinions for clinicians. While most research efforts have been focused on improving architectures and overcoming data-related limitations, answer consistency has been overlooked even though it plays a critical role in establishing trustworthy models. In this work, we propose a novel loss function and corresponding training procedure that allows the inclusion of relations between questions into the training process. Specifically, we consider the case where implications between perception and reasoning questions are known a-priori. To show the benefits of our approach, we evaluate it on the clinically relevant task of Diabetic Macular Edema (DME) staging from fundus imaging. Our experiments show that our method outperforms state-of-the-art baselines, not only by improving model consistency, but also in terms of overall model accuracy. Our code and data are available at `https://github.com/sergiotasconmorales/consistency_vqa`.

Keywords: VQA · Consistency · Attention · Diabetic Macular Edema

1 Introduction

Visual Question Answering (VQA) models are neural networks that answer natural language questions about an image by interpreting the question and the image provided [1,7,11,22]. Specifying questions using natural language gives VQA models great appeal, as the set of possible questions one can ask is enormous and does not need to be identical to the set of questions used to train the models. Due to these advantages, VQA models for medical applications have also been proposed [6,8,12,13,24,28], whereby allowing clinicians to probe the model with subtle differentiating questions and contributing to build trust in predictions.

To date, much of the work in medical VQA has focused on building more effective model architectures [6,12,24] or overcoming limitations in medical VQA datasets [14,12,19,28]. Yet a critical component of VQA is the notion of *consistency* in the answers produced by a model. Here, consistency refers to a model's
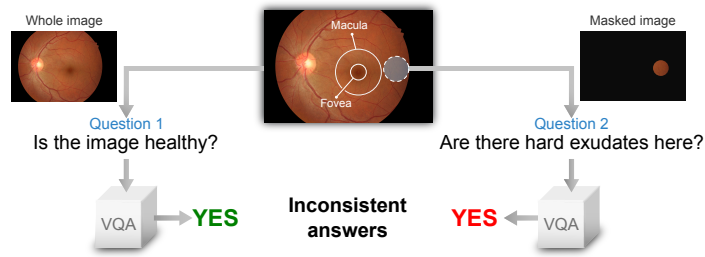
Fig. 1. VQA inconsistency in Diabetic Macular Edema staging from fundus photograph. While the VQA model correctly answers "Is the image healthy?" (left), it incorrectly answers yes to "Are there hard exudates here?" for a specified retinal region.

capacity to produce answers that are not self-contradictory. For instance, the task of staging diabetic macular edema (DME) from color fundus photograph illustrated in Fig. 1 involves identifying *perception* elements in the image (*e.g.*, "are there hard exudates visible near the macula?") to infer a disease stage, which can be expressed as a *reasoning* question (*e.g.*, "what is the stage of disease?"). Ultimately, for any VQA model to be trustworthy, it should be able to answer these without contradicting itself (*i.e.*, answer that the image is healthy, but also identify hard exudates in the periphery of the eye).

Consistency in VQA has been been studied in the broader computer vision context [4,5,16,18,21], where the relation between perception and reasoning questions is unconstrained. That is, the answers to perception questions do not necessarily imply any information with respect to the reasoning question and vice-versa. In these broad cases, some methods have modeled question implications [16,18] or rephrased questions [21] by generating tailored question-answer pairs (*e.g.*, consistent data-augmentation). Alternatively, [5,23,27] used relations between questions to impose constraints in the VQA's embedding space. To avoid needing to know the relation between questions, [20] proposed to enforce consistency by making attention maps of reasoning and perception questions similar to one another. However, even though these approaches tackle unconstrained question relations, the ensuring of VQA models' consistency remains limited and often reduces the overall performance [20].

Instead, we propose a novel approach to enforce VQA consistency that is focused on cases where answers to the perception questions have explicit implications on reasoning question answers and vice-versa (*e.g.*, cancerous cells and severity of cancer found in H&E staining, or presence of hard exudates and DME staging). By focusing on this subset of question relations, our aim is to improve both the accuracy of our model and its consistency, without needing external data as in [18,14,4]. To do this, we allow questions to probe arbitrary image regions by masking irrelevant parts of the image and passing the masked image to the VQA model (see Fig. 1). To then enforce consistency, we propose a new loss function that penalizes incorrect perceptual predictions when reasoning ones

are correct for a given image. To validate the impact of our approach, we test it in the context of DME staging and show that it outperforms state-of-the-art methods for consistency, without compromising overall performance accuracy.

2   Method

We present here our approach which consists of using a simple VQA model with a training protocol that encourages consistency among pairs of perception and reasoning questions. Fig. 2 illustrates this VQA model and our training approach.

VQA model. Following [2], our VQA model, $f : \mathcal{I} \times \mathcal{Q} \to \mathcal{P}(\mathcal{A})$, takes a tuple containing an image, $\mathbf{x}$, and a question, $\mathbf{q}$, to produce a distribution, $\mathbf{p} = f(\mathbf{x}, \mathbf{q})$, over a finite set of possible answers $\mathcal{A}$ (see Fig. 2(Top)). After encoding the inputs, the VQA model combines visual (v) and textual (q) features through an attention module (k) [26] that selects the visual features relevant to the question (v'). The final classifier receives a combination of the relevant features and the text features through a fusion module to predict the final distribution.

In some cases, questions may consider asking about content related to specific regions of the image (*e.g.*, "are there hard exudates in this region?"). To process these cases, the input image is masked so that the visible area corresponds to the region mentioned in the question while the rest of the image is set to zero.

Training this model requires a dataset $\mathcal{T} = \{t^{(i)} = (\mathbf{x}^{(i)}, \mathbf{q}^{(i)}, a^{(i)})\}_{i=1}^{N} \subseteq \mathcal{I} \times \mathcal{Q} \times \mathcal{A}$ of images and questions annotated with their answers. The VQA loss is simply the cross-entropy between the predicted distribution and the real answer,

$$\ell_{\text{VQA}}(\mathbf{p}, a) = H(\mathbf{p}, a) = -\log \mathbf{p}_a. \tag{1}$$

While this loss alone is sufficient to reach a reasonable performance, it ignores the potentially useful interactions that may exist among training questions.

Consistency loss. We aim to improve the quality of our VQA model by exploiting the relationships between reasoning and perception questions at training time. To this end, we augment the training dataset with an additional binary relation $\prec$ over the set of questions $\mathcal{Q}$. Two questions are related, $\mathbf{q}^{(i)} \prec \mathbf{q}^{(j)}$, if $\mathbf{q}^{(i)}$ is a perception question associated to the reasoning question $\mathbf{q}^{(j)}$. From hence on, we refer to perception questions as *sub-questions* and reasoning questions as *main questions*.

Following the terminology in [20], an inconsistency occurs when the VQA model infers the main question correctly but the sub-question incorrectly. Using the entropy as a measurement of incorrectness, we propose to impose the consistency at training time by penalizing the cases with high $H^{(i)} = H(\mathbf{p}^{(i)}, a^{(i)})$ and low $H^{(j)} = H(\mathbf{p}^{(j)}, a^{(j)})$ when $\mathbf{q}^{(i)} \prec \mathbf{q}^{(j)}$. To do this, we use an adapted hinge loss that disables the penalty when $H^{(j)}$ is larger than a threshold $\gamma > 0$, but otherwise penalizes large values of $H^{(i)}$,

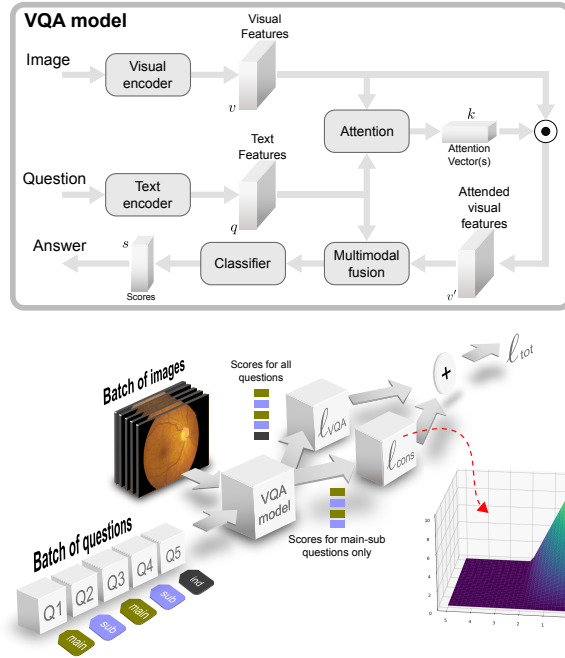$$\ell_{\text{cons}}(H^{(i)}, H^{(j)}) = H^{(i)} \max\{0, \gamma - H^{(j)}\}. \tag{2}$$

Fig. 2. *Top:* VQA model architecture. *Bottom:* Visualization of the training process with the proposed loss. The total loss, ℓ_{tot}, is based on two terms: the conventional VQA loss, ℓ_{VQA} and our proposed consistency loss term, ℓ_{cons}. The latter is computed only for pairs of main (reasoning) and sub (perception) questions. Training mini-batches consist of main and sub questions at the same time, whereby sub-questions can consider specific regions of the image. Unrelated questions (denoted with "ind") can also be included in training batches, but do not contribute to ℓ_{cons}.

The final cost function then minimizes the expected value of the VQA loss (1) for the elements of the training dataset and the consistency loss (2) for the pairs of training samples with $\prec$-related questions,

$$\mathbb{E}_{t\sim\mathcal{T}}[\ell_{\text{VQA}}(\mathbf{p}, a)] + \lambda\mathbb{E}_{(t^{(i)},t^{(j)})\sim\mathcal{T}^2}[\ell_{\text{cons}}(H^{(i)}, H^{(j)}) \mid \mathbf{x}^{(i)} = \mathbf{x}^{(j)}, \mathbf{q}^{(i)} \prec \mathbf{q}^{(j)}],\tag{3}$$

where $\lambda > 0$ controls the relative strength of both losses and $\mathcal{T}^2$ is the Cartesian product of $\mathcal{T}$ with itself, that is, all pairs of training samples.

To train, this cost is iteratively minimized approximating the expectations with mini-batches. The two expectations of Eq. (3) suggest that two mini-batches are necessary at each iteration: one mini-batch sampled from $\mathcal{T}$ and a second mini-batch of $\prec$-related pairs sampled from $\mathcal{T}^2$. However, in practice a single mini-batch is sufficient as long as we ensure that it contains pairs of $\prec$-related questions. While this biased sampling could in turn bias the estimation of the

first expectation, we did not observe a noticeable impact in our experiments. Fig. 2(Bottom) illustrates this training procedure.

3 Experiments and results

DME staging. Diabetic Macular Edema (DME) staging from color fundus images involves grading images on a scale from 0 to 2, with 0 being healthy and 2 being severe (see Fig. 3). Differentiation between the grades relies on the presence of hard exudates present in different locations of the retina. Specifically, a grade of 0 implies that no hard exudates are present at all, a grade of 1 implies that hard exudates are located in the retina periphery (*i.e.*, outside a circular region centered at the fovea center with radius of one optic disc diameter), and a grade of 2 when hard exudates are in the macular region [17].

Dataset and questions. To validate our method, we make use of two publicly available datasets: the Indian Diabetic Retinopathy Image Dataset (IDRiD) [15] and the e-Ophta dataset [3]. From the IDRiD dataset, we use images from the segmentation and grading tasks, which consist of 81 and 516 images, respectively. Images from the segmentation task include segmentation masks for hard exudates and images from the grading task only have the DME grade. On the other hand, the e-Ophta dataset comprises 47 images with segmentation of hard exudates and 35 images without lesions. Combining both datasets yields a dataset of 128 images with segmentation masks for hard exudates and 128 images without any lesions, plus 423 images for which only the DME risk grade is available.
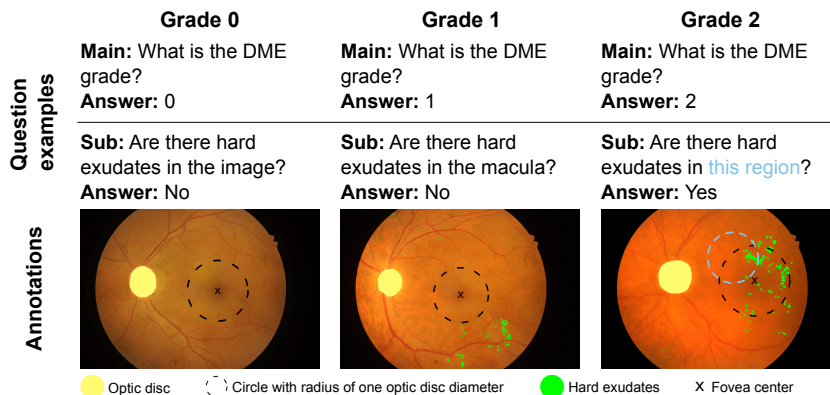

Fig. 3. DME risk grading Grade 0 is assigned if there are no hard exudates present in the whole image. Grade 1 is assigned if there are hard exudates, but only located outside a circle centered at the fovea with radius of one optic disc diameter. Grade 2 is assigned if there are hard exudates located within the circle. Examples of main and sub-questions are provided for each grade.

In this context, we consider main questions to be those asking "What is the DME risk grade?" when considering the entire image. Sub-questions were then defined as questions asking about the presence of the hard exudates. For instance, as shown in Fig. 3(Right), "Are there hard exudates in this region?" where the region designated contains the macula. In practice, we set three types of sub-questions: "are there hard exudates in this image?", "are there hard exudates in the macula?" and "are there hard exudates in this region?". We refer to these three questions as *whole*, *macula* and *region* questions, respectively. For the region sub-questions, we consider circular regions that can be centered anywhere, or centered on the fovea, depending on availability of fovea center location annotations. As mentioned in Sec. 1, to answer questions about regions, images are masked so that only the region is visible.

The total number of question-answer pairs in our dataset consist of 9779 for training (4.4% main, 21.4% sub, 74.2% ind), 2380 for validation (4.5% main, 19.2% sub, 76.3% ind) and 1311 for testing (10% main, 46.1% sub, 43.9% ind), with images in the train, validation and test sets being mutually exclusive.

Baselines, implementation details and evaluation metrics: We compare our approach to a baseline model that does not use the proposed ℓ_{cons} loss, equivalent to setting $\lambda = 0$. In addition, we compare our method against the attention-matching method, SQuINT [20], as it is a state-of-the-art alternative to our approach that can be used with the same VQA model architecture.

Our VQA model uses an ImageNet-pretrained ResNet101 [9] with input image of 448×448 pixels and an embedding of 2048 dimensions for the image encoding. For text encoding, a single-layer LSTM [10] network processes the input question with word encoding of length 300 and produces a single question embedding of 1024 dimensions. The multi-glimpse attention mechanism [26] uses 2 glimpses and dropout rate 0.25, and the multimodal fusion stage uses standard concatenation. The final classifier is a multi-layer perceptron with hidden layer of 1024 dimensions and dropout rate of 0.25. Hyperparameters λ and γ were empirically adjusted to 0.5 and 1.0, respectively.

All experiments were implemented using PyTorch 1.10.1 and run on a Linux machine with an NVIDIA RTX 3090 graphic card using 16 GB of memory and 4 CPU cores. All methods use the weighted cross-entropy as the base VQA loss function. Batch size was set to 64, and we used Adam for optimization with a learning rate of 10^{-4}. Maximum epoch number was 100 and we use early stopping policy to prevent overfitting, with a patience of 20 epochs.

We report accuracy and consistency [20] performances, using two different definitions of consistency for comparison. Consistency, C1, is the percentage of sub-questions that are answered correctly when the main question was answered correctly. Consistency, C2, is the percentage of main questions that are answered correctly when all corresponding sub-questions were answered correctly.

Results: Table 1 shows the results. We compare these results to the case in which the value of λ is 0, which corresponds to the baseline in which no additional

Case	Accuracy					Consistency	
	overall	grade	whole	macula	region	C1	C2
Baseline (no att.)	77.54	73.59	81.37	83.37	76.66	81.70	91.86
Baseline (att.)	81.46	80.23	83.13	**87.18**	80.58	89.21	96.92
Baseline (att.) + SQuINT [20]	80.58	77.48	82.82	85.34	80.02	88.17	94.62
Baseline (att.) + Ours ($\lambda = 0.5, \gamma = 1$)	**83.49**	**80.69**	**84.96**	**87.18**	**83.16**	**94.20**	**98.12**

Table 1. Average test accuracy and consistency values for the different models. Results shown are averaged over 10 models trained with different seeds. Accuracy values are presented for all questions (overall), for main questions (grade) and for sub-questions (whole, macula and region). Both measures of consistency are shown as well.

Fig. 4. Qualitative examples from the test set. Inconsistent sub-answers are highlighted in red. Additional examples are shown in the supplementary material.

loss term is used. For each case, we present the overall accuracy and the accuracy for each type of question, as well as the consistency values. Fig. 4 illustrates the performance of each method with representative qualitative examples.

λ	γ	Accuracy					Consistency	
		overall	grade	whole	macula	region	C1	C2
0	-	81.46	80.23	83.13	87.18	80.58	89.21	96.92
0.2	0.5	82.01	80.38	83.59	86.56	81.36	90.93	97.38
0.2	1	82.65	79.77	83.97	86.64	82.30	93.22	97.51
0.2	1.5	83.05	81.22	84.27	87.33	82.53	93.23	97.56
0.3	0.5	82.34	79.92	83.59	87.71	81.74	92.32	97.31
0.3	1	83.27	80.53	84.58	87.25	82.91	94.01	98.10
0.3	1.5	83.28	80.84	84.43	87.48	82.86	93.28	98.29
0.4	0.5	82.87	80.69	84.89	87.02	82.30	92.66	96.66
0.4	1	82.97	80.15	83.97	86.72	82.69	93.91	98.23
0.4	1.5	83.33	80.08	84.20	86.87	83.17	93.96	97.77
0.5	0.5	82.54	81.07	83.66	88.02	81.81	91.87	97.73
0.5	1	83.49	80.69	84.96	87.18	83.16	94.20	98.12
0.5	1.5	83.25	79.92	84.58	86.95	83.01	94.20	98.12

Table 2. Average test accuracy and consistency values for different values of the parameters λ and γ. The first row ($\lambda = 0$) corresponds to the baseline.

In general, we observe that our proposed approach yields increases in accuracy and consistency when compared to both the baseline and SQuINT. Importantly, this increase in consistency is not at the expense of overall accuracy. Specifically, this indicates that our loss term causes the model to be correct about sub-questions when it is correct about main questions. The observed increase in accuracy also indicates that our approach is not synthetically increasing consistency by reducing the number of correct answers on main questions [20]. We note that SQuINT results in a reduction in accuracy and consistency, which can be partially explained by the presence of region questions that are not associated to any main question. These questions, which exceed the number of main questions, may affect the constraint in the learned attention maps.

Table 2 shows the effect of λ and γ on the performance metrics. As expected, we notice that when λ increases, the consistency of our approach increases as well and will occasionally deteriorate overall accuracy. The impact of γ however is less evident, as no clear trend is visible. This would imply that the exact parameter value used is moderately critical to performances.

4   Conclusions

In this work, we presented a novel method for improving consistency in VQA models in cases where answers to sub-questions imply those of main questions and vice-versa. By using a tailored training procedure and loss function that measures the level of inconsistency, we show on the application of DME staging, that our approach provides important improvements in both VQA accuracy and consistency. In addition, we show that our method's hyperparameters are

relatively insensitive to model performance. In the future, we plan to investigate how this approach can be extended to the broader case of unconstrained question relations.

References

1. 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)
2. Cadene, R., Dancette, C., Cord, M., Parikh, D., et al.: Rubi: Reducing unimodal biases for visual question answering. Advances in neural information processing systems **32**, 841–852 (2019)
3. Decenciere, E., Cazuguel, G., Zhang, X., Thibault, G., Klein, J.C., Meyer, F., Marcotegui, B., Quellec, G., Lamard, M., Danno, R., et al.: Teleophta: Machine learning and image processing methods for teleophthalmology. Irbm **34**(2), 196–203 (2013)
4. Goel, V., Chandak, M., Anand, A., Guha, P.: IQ-VQA: Intelligent visual question answering. In: International Conference on Pattern Recognition. pp. 357–370. Springer (2021)
5. Gokhale, T., Banerjee, P., Baral, C., Yang, Y.: VQA-LOL: Visual question answering under the lens of logic. In: European conference on computer vision. pp. 379–396. Springer (2020)
6. 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)
7. 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)
8. 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)
9. 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)
10. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
11. 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)
12. 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)

13. 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)

14. 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)

15. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (idrid) (2018). https://doi.org/10.21227/H25W98, `https://dx.doi.org/10.21227/H25W98`

16. Ray, A., Sikka, K., Divakaran, A., Lee, S., Burachas, G.: Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. arXiv preprint arXiv:1909.04696 (2019)

17. Ren, F., Cao, P., Zhao, D., Wan, C.: Diabetic macular edema grading in retinal images using vector quantization and semi-supervised learning. Technology and Health Care **26**(S1), 389–397 (2018)

18. Ribeiro, M.T., Guestrin, C., Singh, S.: Are red roses red? evaluating consistency of question-answering models. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 6174–6184 (2019)

19. Sarrouti, M.: Nlm at vqa-med 2020: Visual question answering and generation in the medical domain. In: CLEF (Working Notes) (2020)

20. Selvaraju, R.R., Tendulkar, P., Parikh, D., Horvitz, E., Ribeiro, M.T., Nushi, B., Kamar, E.: Squinting at vqa models: Introspecting vqa models with sub-questions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10003–10011 (2020)

21. Shah, M., Chen, X., Rohrbach, M., Parikh, D.: Cycle-consistency for robust visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6649–6658 (2019)

22. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)

23. Teney, D., Abbasnejad, E., Hengel, A.v.d.: On incorporating semantic prior knowledge in deep learning through embedding-space constraints. arXiv preprint arXiv:1909.13471 (2019)

24. 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)

25. Wang, P., Liao, R., Moyer, D., Berkowitz, S., Horng, S., Golland, P.: Image classification with consistent supporting evidence. In: Machine Learning for Health. pp. 168–180. PMLR (2021)

26. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. pp. 2048–2057. PMLR (2015)

27. Yuan, Y., Wang, S., Jiang, M., Chen, T.Y.: Perception matters: Detecting perception failures of vqa models using metamorphic testing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16908–16917 (2021)

28. 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)