A Study on Effectiveness of BERT Models and Task-Conditioned Reasoning Strategy for Medical Visual Question Answering

Chau Nguyen[1(✉)], Tung Le[2], Nguyen-Khang Le[1], Trung-Tin Pham[1], and Le-Minh Nguyen[1]

[1] Japan Advanced Institute of Science and Technology, Nomi, Japan
{chau.nguyen,lnkhang,tinpham,nguyenml}@jaist.ac.jp
[2] University of Science - VNUHCM, Ho Chi Minh City, Vietnam
lttung@fit.hcmus.edu.vn

Abstract. Medical visual question answering task requires a framework to understand a medical question in natural language and examine the corresponding image to produce the answer to the question. The common framework consists of a language understanding module, a visual understanding module, a signal fusion module, and an answer prediction module. Most existing works employed recurrent neural network-based models for the language understanding module. However, these approaches may not produce robust text presentations and are hard to interpret. On the other hand, BERT models are more robust for text representation and can provide a clue for interpretability via the attention weights between the words. Besides, as the questions consist of closed-answer questions and open-answer questions, the task-conditioned reasoning strategy was proposed to handle each type of question separately while maintaining several modules in the framework to be shared. In this paper, we investigate the effectiveness of pre-trained BERT models and the task-conditioned reasoning strategy for the task of medical visual question answering on the VQA-RAD dataset. Experimental results demonstrate improvements when pre-trained BERT models are combined with the task-conditioned reasoning strategy.

Keywords: Medical visual question answering · Visual question answering · Task-conditioned reasoning · Conditional reasoning

1 Introduction

Medical visual question answering (medical VQA) is the task aiming to answer a natural language question on a medical image. Medical VQA requires a system to understand the natural language in the medical context (i.e., understand the question) as well as to pay attention to the areas in the image that contain the clues for the answer, then produce an answer based on the knowledge from both language signals and visual signals. The common framework consists of a

Y. Kambayashi et al. (Eds.): AICON 2022, LNICST 477, pp. 60–71, 2023.
https://doi.org/10.1007/978-3-031-29126-5_5

language understanding module, a visual understanding module, a signal fusion module, and an answer prediction module (see Fig. 1). In fact, medical VQA is a generation task. However, most existing systems [6,7,18,19,23,24] consider it as a classification task (i.e., the answer prediction module is an answer classifier where the set of answers is a collection of answers from the training set).

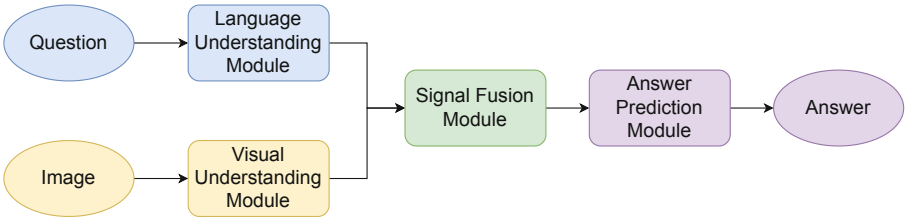

Fig. 1. The common framework for medical VQA

CMSA-MTPT [7] was proposed as a multi-task pre-training technique where the visual understanding module (which are ResNet [10] models) are pre-trained in a way that is compatible with the language understanding module (which is an LSTM (long short-term memory [12]) model). While LSTM is trained from scratch, BERT [5] models can be pre-trained so that the learned language patterns can be leveraged during the fine-tuning phase. In fact, BERT has become one of the dominant language embedders in the field of natural language processing (NLP). Besides, the attention weights between the words provided by BERT models also give a clue for the researcher for interpretability or for error analysis. Hence, employing BERT models as the language understanding module should strengthen the framework. [23] proposed task-conditioned reasoning (TCR) strategy which addresses different types of questions separately.

In our study, we replace LSTM in the CMSA-MTPT framework with pre-trained BERT models (BERT [5] and BioBERT [17] in particular) and apply the TCR strategy to this framework. We did experiments with many settings to investigate the effectiveness of pre-trained BERT models and TCR for the medical VQA task.

Table 1. Statistics on the dataset

Data split	**Train**	**Test**
# questions	3064	451
# questions with closed answer	1821	272
# questions with open answer	1243	179
# different answers	458	83
# different closed answers	56	13
# different open answers	431	77
# overlapped closed & open answers	29	7

2 Related Work

2.1 VQA-RAD Dataset

Among many medical VQA datasets [1–3,9,11,15,16], VQA-RAD [16] is popular as it is the only dataset which contains natural questions on a variety types of questions. Table 1 provides some statistics on the VQA-RAD dataset. The training data contains 3064 questions while the test data contains 451 questions on a total of 315 radiological images. The question set includes questions about plane, modality, organ system, abnormality, object/condition presence, position, color, size, attribute other, counting, and others. The answers to those questions are divided into closed questions (i.e., limited-choice questions such as yes/no or left/right) and open questions (i.e., non-limited-choice questions which may have multiple correct answers). The radiological images are images of the head, chest, and abdomen areas.

2.2 CMSA-MTPT Framework

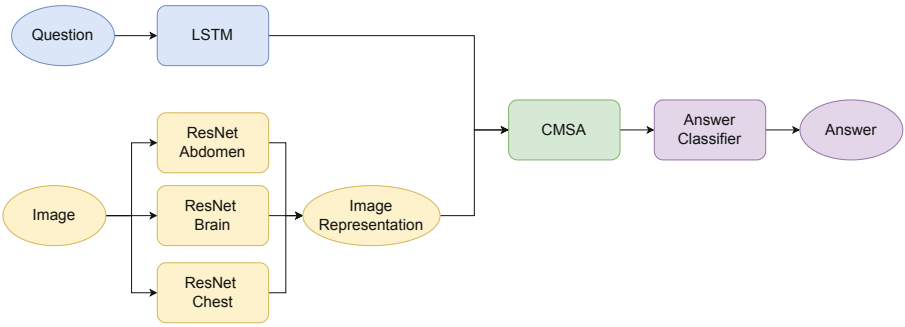

Fig. 2. The CMSA-MTPT framework

The CMSA-MTPT framework (Fig. 2) followed the common framework described above. Experiments were performed on the VQA-RAD dataset. This framework used LSTM as the language understanding module and employed pre-trained ResNet models as the visual understanding module. A cross-modal self-attention module (CMSA) is proposed as the signal fusion module, which outperforms SAN [22] and BAN [14]. The answer prediction module (answer classifier) is a simple 2-layer MLP (multi-layer perceptron) classifier. Because of the lack of training images in the VQA-RAD dataset (only 315 radiological images of 3 areas: head, chest, abdomen), it is impractical to apply state-of-the-art models in VQA [4,13] directly to medical VQA. CMSA-MTPT leveraged different available image datasets for each area for pre-training corresponding ResNet models, in a transfer learning manner [18,20]. Each ResNet model is

designed to embed the content of an image type (head, chest, or abdomen). Given a medical image, the 3 ResNet models will generate 3 vectors representing it. The final representation of the image is a soft combination of the 3 vectors based on an image type classifier.

Because of the labels provided for those external data, ResNet models for the head and chest are trained on a classification task while the ResNet model for the abdomen is trained on the image segmentation task. Besides, CMSA-MTPT also proposed a multi-task pre-training strategy so that the compatibility of the pre-trained ResNet models and the language understanding module is ensured. Specifically, the loss function for pre-training ResNet models not only contains classification/segmentation loss but also contains a loss specifying the compatibility score between the type of the image and the language representation of the input question. For the signal fusion module, the CMSA-MTPT framework proposed cross-modal self-attention (CMSA) to effectively fuse the representation of the question and the image (see Fig. 2).

2.3 Task-Condition Reasoning Strategy

The main idea of task-condition reasoning (TCR) [23] is to have two signal fusion modules and two answer prediction modules to handle closed-answer questions and open-answer questions separately while other modules (i.e., language understanding module and visual understanding module) are shared.

3 Our Framework

3.1 Overview

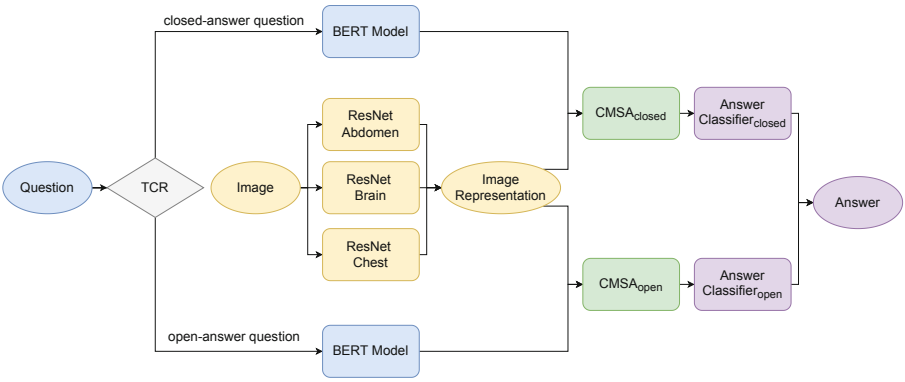

Fig. 3. An overview of our framework

Figure 3 provides an overview of our framework. Our framework is based on the CMSA-MTPT framework [7]. The two main adjustments are (i) to enhance the

language understanding module (replacing LSTM with BERT models: BERT [5] and BioBERT [17]) and (ii) to employ the task-conditioned reasoning strategy [23].

Intuitively, BERT models (especially BioBERT - which is pre-trained on the corpora of the medical domain) are pre-trained on huge text corpora, which is supposed to adapt promptly to understand natural language in medical fields. Besides, as BERT deploys the self-attention architecture, the attention weights between the words can be leveraged as a clue for error analysis and interpretability of the model.

Table 2. Length of the questions by different tokenizers

Data split	Value	Word tokenizer (# words)	BERT tokenizer (# tokens)	BioBERT tokenizer (# tokens)
Train	max	21	28	28
	min	3	5	6
	average	6.43	11.76	12.06
	median	6	11	11
Test	max	22	32	33
	min	3	5	6
	average	6.89	11.51	11.88
	median	7	11	11

Table 2 contains the information about the length of the questions by different tokenizers: word tokenizer, BERT tokenizer, and BioBERT tokenizer. On average, the length (by words) of a question is less than 7 although some of them can reach the length of 22 words. In the original CMSA-MTPT framework, they truncate the question to a maximum of 12 words before feeding it to LSTM. In our case, as we use the BERT tokenizer and BioBERT tokenizer, the average number of tokens is around 12 tokens while there are some questions that reach 33 tokens. We performed experiments with many values of the maximum number of tokens to investigate which value should be the most appropriate. In particular, we experimented with the set 12, 20, 25, 30, and 40.

3.2 BERT Models as the Language Understanding Module

For the BERT models, we experimented with BERT base uncased[1] and BioBERT base cased version 1.2[2]. As BERT models contain a big number of trainable parameters, fine-tuning all those parameters makes the training process longer. Hence, for each BERT model, we tried with 2 settings: (i) frozen

[1] https://huggingface.co/bert-base-uncased.
[2] https://huggingface.co/dmis-lab/biobert-base-cased-v1.2.

BERT model: freeze all BERT layers and only use the BERT model as a frozen text embedder, and (ii) unfrozen BERT model: freeze all BERT layers except for the last one (i.e., only the parameters of the last layer of BERT model are trained along with other modules).

The attention weights provided BERT models in different heads demonstrate behaviors which seems to be related to the sentence's structure [21]. Although those behaviors are not fully investigated, they still provide a clue to determine how the model may work. Figure 4 shows an example of a visualization of a head produced by a BERT model (see Fig. 4).

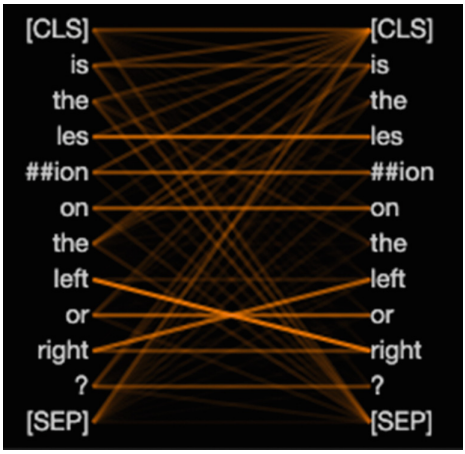

Fig. 4. Visualization of the attention weights of layer 2's head 12

3.3 Setting of the Task-Conditioned Strategy

Previous frameworks [6,7,18] trained a fusion module to learn the fused representation of all samples and classified to 458 answers (458 is the number of different answers in the training data, see Table 1). However, TCR first determines the type of the question (question with closed answer or question with open answer), then employs a model to learn the fusion of each type before feeding it into the answer classifiers. Here, the answer classifier for closed-answer questions has only 56 classes and the answer classifier for open-answer questions has 431 classes. There are 29 overlapped answers between the two sets.

4 Experiments

4.1 Training ResNet Models

As mentioned above, the ResNet models are trained so that they are compatible with the corresponding language understanding module. Because in our case,

we replace LSTM with BERT models, so we also need to re-train the corresponding ResNet models for the BERT models. Table 3 shows the experimental results. As mentioned above, all models are trained not only with a loss function indicating the compatibility of the image and the corresponding question, but also be trained with another task (a segmentation task for abdomen data and a classification task for brain data and chest data). We train each model for 200 epochs.

As shown in Table 3, we can roughly reproduce the ResNet models for the CMSA-MTPT with LSTM. The other experiments on CMSA-MTPT with BERT models demonstrate the same results in most cases, except for the compatibility accuracy of CMSA-MTPT with unfrozen BERT models (59.26% compared to 78.70% as reported). It may be because, with unfrozen BERT models, the models need to fine-tune on more parameters.

Table 3. Results of ResNet models for the CMSA-MTPT with LSTM. *Comp. acc.* means compatibility accuracy. *Cls. acc.* means classification accuracy. Values in **bold** indicate the highest values in the columns.

Model	Max words/max tokens	Abdomen data		Brain data		Chest data	
		mIOU	Comp. acc.	Cls. acc.	Comp. acc.	Cls. acc.	Comp. acc.
CMSA-MTPT with LSTM (reported)	12	71.00	78.70	98.40	89.10	98.70	83.60
CMSA-MTPT with LSTM (reproduced)	12	81.98	73.15	**100.00**	90.62	97.84	86.21
CMSA-MTPT with frozen BERT base	12	77.02	78.70	98.44	90.62	96.55	86.21
	20	76.73	79.63	96.88	90.62	96.55	87.93
	25	74.22	75.00	96.88	90.62	96.98	**90.09**
	30	75.88	77.78	98.44	90.62	97.84	88.79
	40	73.06	75.00	98.44	87.50	96.12	87.07
CMSA-MTPT with unfrozen BERT base	12	74.23	81.48	98.44	78.12	98.71	62.93
	20	82.58	59.26	**100.00**	78.12	98.71	62.93
	25	**83.32**	59.26	98.44	78.12	98.71	62.93
	30	82.44	59.26	98.44	78.12	**99.14**	62.93
	40	81.66	59.26	98.44	78.12	**99.14**	62.93
CMSA-MTPT with frozen BioBERT base	12	71.15	77.78	96.88	**96.88**	96.55	85.78
	20	74.31	75.93	96.88	93.75	97.41	86.21
	25	69.58	78.70	98.44	95.31	98.28	86.21
	30	72.71	82.41	96.88	89.06	98.28	87.07
	40	73.21	**83.33**	95.31	95.31	97.84	86.64
CMSA-MTPT with unfrozen BioBERT base	12	82.90	59.26	98.44	78.12	98.71	86.21
	20	82.38	59.26	**100.00**	78.12	96.98	88.79
	25	76.03	78.70	96.88	78.12	98.28	62.93
	30	83.03	59.26	**100.00**	78.12	97.84	89.66
	40	82.89	59.26	98.44	78.12	98.28	89.66

4.2 CMSA-MTPT with BERT Models

Experimental Settings. We did experiments on the VQA-RAD dataset. Each model is trained on an NVIDIA A40 GPU. Table 4 shows the experimental settings. Following [7], we also apply warmup steps [8]. We train for 250 epochs.

Results. Table 5 shows the experimental results. Here, "Open" means model accuracy on only open-answer questions, "Closed" means model accuracy on

Table 4. Parameter settings

Parameter	Value
# epochs	250
initial learning rate (LR)	0.005
LR decay step	48
LR decay rate	0.75
batch size	32

Table 5. Results of CMSA-MTPT with BERT models. Values in **bold** indicate the highest values in the columns.

Model	Max words/max tokens	Accuracy		
		Open	Closed	All
CMSA-MTPT with LSTM (reproduced)	12	61.45	77.21	70.95
CMSA-MTPT with frozen BERT base	12	50.84	78.31	67.41
	20	54.75	76.47	67.85
	25	59.22	80.15	71.84
	30	56.42	75.74	68.07
	40	57.54	80.51	71.40
CMSA-MTPT with unfrozen BERT base	12	60.34	79.41	71.84
	20	60.34	79.04	71.62
	25	60.34	80.51	**72.51**
	30	58.10	77.57	69.84
	40	57.54	77.57	69.62
CMSA-MTPT with frozen BioBERT base	12	53.63	77.21	67.85
	20	56.98	80.15	70.95
	25	61.45	79.78	**72.51**
	30	59.22	**80.88**	72.28
	40	59.22	79.04	71.18
CMSA-MTPT with unfrozen BioBERT base	12	60.89	79.41	72.06
	20	**62.01**	76.84	70.95
	25	60.34	78.31	71.18
	30	59.78	78.68	71.18
	40	59.22	76.47	69.62

only closed-answer questions, and "All" means model accuracy on all questions. While the reported accuracy is 73.17%, our reproduced accuracy is only 70.95%. We compare the reproduced accuracy with our methods.

The experimental results show that the CMSA-MTPT model with unfrozen BERT base and the CMSA-MTPT model with frozen BioBERT base produce the highest accuracy on the whole test set with 72.51%. There is an improvement of 1.56% (from 70.95% to 72.51%) when replacing the LSTM language understanding module with a BERT model.

4.3 Task-Conditioned Reasoning CMSA-MTPT with BERT Models

We use the same settings as in training CMSA-MTPT with BERT models. Table 6 shows the experimental results when applying the task-condition reasoning strategy on CMSA-MTPT with BERT models. TCR helps improve the accuracy to 73.17% on the setting of TCR CMSA-MTPT with frozen BioBERT base.

Table 6. Results of TCR CMSA-MTPT with BERT models. Values in **bold** indicate the highest values in the columns.

Model	Max words/ max tokens	Accuracy		
		Open	Closed	All
TCR CMSA-MTPT with frozen BERT base	12	51.96	79.41	68.51
	20	53.07	79.41	68.96
	25	53.07	80.51	69.62
	30	60.34	78.68	71.40
	40	53.07	76.10	66.96
TCR CMSA-MTPT with unfrozen BERT base	12	61.45	79.04	72.06
	20	61.45	77.21	70.95
	25	58.66	77.21	69.84
	30	59.78	78.68	71.18
	40	58.10	77.94	70.07
TCR CMSA-MTPT with frozen BioBERT base	12	55.31	81.99	71.40
	20	59.22	79.78	71.62
	25	62.01	79.04	72.28
	30	58.66	**82.72**	**73.17**
	40	52.51	81.62	70.07
TCR CMSA-MTPT with unfrozen BioBERT base	12	59.78	79.04	71.40
	20	**62.57**	77.57	71.62
	25	60.89	76.10	70.07
	30	60.34	81.25	72.95
	40	55.31	77.94	68.96

TCR is helpful because it can separately train different modules for different types of questions where the answer classifiers need to deal with fewer classifiers: instead of dealing with 458 classes, the closed-answer classifier only deals with 56 classes, and the open-answer classifier deals with 431 classes (Table 1). Table 1 also shows that the number of answers in the test set is much fewer than those in the training set: 13 different closed answers and 77 different open answers. Hence, we tried an experiment where the models are only trained with the answer set in the test set instead of the answer set in the training set. It is an "oracle" setting so that we can see how the models perform if they are only trained on the questions that have the answers in the answer set of test data. Table 7 shows the results of this experiment.

The highest oracle performance is achieved with 76.72% by the oracle TCR CMSA-MTPT with frozen BioBERT base model with the max tokens set to 40. This model has significantly improved the accuracy of answering open-answer questions (increase 6.15%, from 62.57% to 68.72%). However, the accuracy on

Table 7. Results of oracle TCR CMSA-MTPT with BERT models. Values in **bold** indicate the highest values in the columns.

Model	Max words/max tokens	Accuracy		
		Open	Closed	All
Oracle TCR CMSA-MTPT with frozen BERT base	12	58.10	79.78	71.18
	20	60.34	79.41	71.84
	25	62.01	77.21	71.18
	30	60.89	81.25	73.17
	40	64.25	79.04	73.17
Oracle TCR CMSA-MTPT with unfrozen BERT base	12	64.25	79.41	73.39
	20	64.25	81.62	74.72
	25	63.69	80.88	74.06
	30	62.01	77.94	71.62
	40	62.01	79.04	72.28
Oracle TCR CMSA-MTPT with frozen BioBERT base	12	60.34	77.57	70.73
	20	63.13	79.78	73.17
	25	62.01	81.62	73.84
	30	63.69	81.62	74.50
	40	**68.72**	**81.99**	**76.72**
Oracle TCR CMSA-MTPT with unfrozen BioBERT base	12	60.89	80.88	72.95
	20	64.25	77.57	72.28
	25	64.25	79.04	73.17
	30	60.89	79.78	72.28
	40	64.80	76.84	72.06

closed-answer questions decreases minorly. These observations indicate that the future model should focus more on open-answer questions as its accuracy is still low and can be increased significantly.

4.4 A Clue for Interpretability in BERT Models

Figure 4 shows the visualization of the attention weights of layer 2's head 12 for the question *"is the lesion on the left or right?"*. The model chosen for visualization is the BioBERT model in the TCR CMSA-MTPT model with frozen BioBERT base (max tokens = 30). In this closed-answer question, the answer should be *left* or *right*. As shown in Fig. 4, while many tokens (i.e., *les*, *##ion*, *on*) focus the most on itself, the token *left* focus the most on token *right*, and *vice versa*. It may indicate that this head learned the pattern of extracting the answer from the similar question in the training dataset.

5 Conclusion

In this paper, we investigate the effectiveness of pre-trained BERT models and the task-conditioned reasoning strategy for the task of medical visual question answering on the VQA-RAD dataset. Via comprehensive experiments, it is demonstrated that pre-trained BERT models (i.e., BERT base and BioBERT base) are suitable for this task and can replace LSTM as the language understanding module. Besides, the task-conditioned reasoning strategy also demonstrates improvements when employed in the framework. It is suggested that future research should focus more on the open-answer questions.

References

1. Abacha, A.B., Datla, V.V., Hasan, S.A., Demner-Fushman, D., Müller, H.: Overview of the VQA-med task at ImageCLEF 2020: visual question answering and generation in the medical domain. In: CLEF (Working Notes) (2020)
2. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-med: overview of the medical visual question answering task at ImageCLEF 2019. In: CLEF (Working Notes), vol. 2 (2019)
3. Allaouzi, I., Ahmed, M.B., Benamrou, B.: An encoder-decoder model for visual question answering in the medical domain. In: CLEF (Working Notes) (2019)
4. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Do, T., Nguyen, B.X., Tjiputra, E., Tran, M., Tran, Q.D., Nguyen, A.: Multiple meta-model quantifying for medical visual question answering. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12905, pp. 64–74. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87240-3_7

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, P., et al.: Accurate, large minibatch SGD: training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
9. Hasan, S.A., Ling, Y., Farri, O., Liu, J., Müller, H., Lungren, M.: Overview of ImageCLEF 2018 medical domain visual question answering task. Technical report, 10–14 September 2018 (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. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: PathVQA: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
13. Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.: Pythia v0. 1: the winning entry to the VQA challenge 2018. arXiv preprint arXiv:1807.09956 (2018)
14. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: Advances in Neural Information Processing Systems, vol. 31 (2018)
15. Kovaleva, O., et al.: Towards visual dialog for radiology. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, pp. 60–69 (2020)
16. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Sci. Data **5**(1), 1–10 (2018)
17. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (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., et al. (eds.) MICCAI 2019. LNCS, vol. 11767, pp. 522–530. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32251-9_57
19. Pan, H., He, S., Zhang, K., Qu, B., Chen, C., Shi, K.: MuVAM: a multi-view attention-based model for medical visual question answering. arXiv preprint arXiv:2107.03216 (2021)
20. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: understanding transfer learning for medical imaging. In: Advances in Neural Information Processing Systems, vol. 32 (2019)
21. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
22. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29 (2016)
23. 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)
24. Zhou, Y., Kang, X., Ren, F.: Employing inception-ResNet-v2 and bi-LSTM for medical domain visual question answering. In: CLEF (Working Notes) (2018)