Open-Ended Medical Visual Question Answering Through Prefix Tuning of Language Models

Tom van Sonsbeek(✉), Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees G. M. Snoek, and Marcel Worring

University of Amsterdam, Amsterdam, The Netherlands
{t.j.vansonsbeek,m.m.derakhshani,i.najdenkoska,c.g.m.snoek, m.worring}@uva.nl

Abstract. Medical Visual Question Answering (VQA) is an important challenge, as it would lead to faster and more accurate diagnoses and treatment decisions. Most existing methods approach it as a multi-class classification problem, which restricts the outcome to a predefined closed-set of curated answers. We focus on open-ended VQA and motivated by the recent advances in language models consider it as a generative task. Leveraging pre-trained language models, we introduce a novel method particularly suited for small, domain-specific, medical datasets. To properly communicate the medical images to the language model, we develop a network that maps the extracted visual features to a set of learnable tokens. Then, alongside the question, these learnable tokens directly prompt the language model. We explore recent parameter-efficient fine-tuning strategies for language models, which allow for resource- and data-efficient fine-tuning. We evaluate our approach on the prime medical VQA benchmarks, namely, Slake, OVQA and PathVQA. The results demonstrate that our approach outperforms existing methods across various training settings while also being computationally efficient.

Keywords: Visual Question Answering · Language Models · Prompting · Prefix Tuning

1 Introduction

Images and text are inherently intertwined in clinical diagnosis and treatment. Having an automated approach that is able to answer questions based on images, giving insight to clinicians and patients, can be a valuable asset. In such a medical Visual Question Answering (VQA) setting the common approach is to treat VQA as a multi-class classification problem solved by neural networks. Given a joint encoded representation of the image and question, the model classifies it into a predefined set of answers. Although these approaches yield good performance [5, 18, 24, 33], they deal with closed-set predictions, which is not an ideal solution

T. van Sonsbeek, M. M. Derakhshani and I. Najdenkoska—Equal contribution.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14224, pp. 726–736, 2023.
https://doi.org/10.1007/978-3-031-43904-9_70

for VQA. For instance, medical VQA datasets commonly contain hundreds to thousands of free-form answers [11], which is suboptimal to be treated as a classification task. Moreover, the severe class imbalance and out-of-vocabulary answers further hinder the generalizability of these classification methods.

We believe that a possible solution can be found in the generative capability of language models, since they are able to produce free text, instead of being limited to closed-set predictions. However, leveraging language models for solving open-ended medical VQA is limited due to several challenges, such as finding ways to properly communicate the visual features and letting such large-scale models be employed on small-sized medical VQA datasets.

Inspired by recent image captioning models [22], we propose to use the medical images by converting them into a set of learnable tokens through a small-scale mapping network. These tokens can then be interpreted as a visual prefix for the language model [1,4,23]. Afterward, the visual prefix is used together with the question as input to the language model, which generates the answer token by token [25].

Furthermore, large-scale language models can generalize across domains while keeping their weights frozen [30]. This makes them very appealing for the medical domain, which inherently does not possess large quantities of labeled data required to train these models from scratch [29]. Models like BioGPT [21] and BioMedLM [31] are based on the generic GPT2 language model [26] and are trained on biomedical text corpora. They perform quite well compared to their general counterparts on specific biomedical language tasks, like question answering or relation extraction. We design our model in a flexible manner, which allows us to incorporate any of these pre-trained language models.

In summary, we contribute in three major aspects: (i) We propose the first large-scale language model-based method for open-ended medical VQA. (ii) We adopt parameter-efficient tuning strategies for the language backbone, which gives us the ability to fine-tune a large model with a small dataset without the danger of overfitting. (iii) We demonstrate through extensive experiments on relevant benchmarks that our model yields strong open-ended VQA performance without the need for extensive computational resources.

2 Related Works

To describe existing medical VQA methods, we make a distinction between classification methods and generative methods. The majority of methods are classification-based and make use of different types of encoders, such as CNNs or Transformers [3,9,11,17,32] followed by a classification layer.

Classification-Based VQA. We highlight a number of methods that showed good performance on current competitive medical VQA datasets. The Mixture Enhanced Visual Features (MEVF) [24] is initialized based on pre-trained weights from the Model-Agnostic Meta-Learning (MAML) model [7] in combination with image feature extraction from Conditional Denoising Auto-Encoders

(CDAE) to generate a joint question-answer representation using Bilinear (BAN) or Stacked (SAN) Attention Networks. Do *et al.* [5] create a similar embedding space by extracting annotations from multiple pre-trained meta-models, and learning meta-annotations by training each meta-model. Linear combinations [10] or question-conditioned selections [34] from this multi-modal embedding space can further enhance performance. The use of Transformer [14] and CLIP [6,25] encoders also results in strong VQA classification performance.

Open-Ended VQA. MedFuseNet [28] is one of the few methods performing and reporting open-ended visual question answering on recent public datasets. They do so by creating a BERT-based multi-modal representation of image and question and subsequently passing it through an LSTM decoder. Ren *et al.* [27] create open-ended answers by using the masked token prediction functionality of BERT. We aim to show that generative language models are more versatile and better suited for this task.

3 Methodology

3.1 Problem Statement

Given an input image $\mathbf{I}$ and an input question in natural language $\mathbf{Q}$, our method aims to sequentially generate an answer $\mathbf{A} = \{A_0, A_1, ..., A_N\}$ composed of N tokens, by conditioning on both inputs. From a model definition perspective, we aim to find the optimal parameters θ^* for a model by maximizing the conditional log-likelihood as follows:

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{N} \log p_\theta(\mathbf{A}_i | \mathbf{Q}, \mathbf{I}, \mathbf{A}_{i-1}). \tag{1}$$

3.2 Model Architecture

Our VQA model is designed as an encoder-decoder architecture, with a two-stream encoder and a language model (LM) as a decoder, as illustrated in Fig. 1. Specifically, the two streams encode the two input modalities, namely the image $\mathbf{I}$ and the question $\mathbf{Q}$. The language model is defined as a causal language Transformer [26], and it generates the answer $\mathbf{A}$ in an autoregressive manner. It closely follows the prefix tuning technique for prompting a language model to produce an output of a particular style [16], such as in our case an answer given a question and an image[1].

Vision Encoding Stream. For encoding the image, we employ a pre-trained vision encoder to extract visual features $\{x_1, x_2 ... x_{\ell_x}\}$. To use these features as input to the decoder, they should be mapped into the latent space of the

[1] Code available at: github.com/tjvsonsbeek/open-ended-medical-vqa.

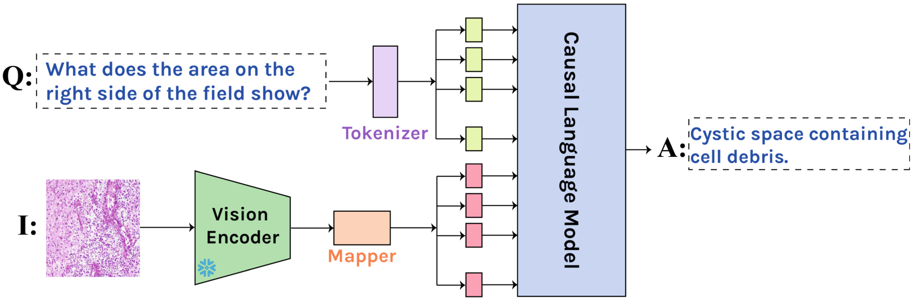

Q:
What does the area on the right side of the field show?
Tokenizer
I:
Vision Encoder
Mapper
Causal Language Model
A:
Cystic space containing cell debris.

Fig. 1. Model architecture of our proposed open-ended generative VQA method.

language decoder. Following [22], we define a mapping network $\mathbf{f_M}$, implemented as a three-layer MLP. This network maps the visual features into a visual prefix $\{v_1, v_2, \dots v_x\} \in \mathbb{R}^{\ell_x \times e}$ for the language model, where e is the embedding size.

Language Encoding Stream. Regarding the encoding of the textual part, firstly we utilize a standard tokenization process to obtain a sequence of tokens, both for the question $\mathbf{Q} = \{q_1, q_2 ... q_{\ell_q}\} \in \mathbb{R}^{\ell_q \times e}$ and answer $\mathbf{A} = \{a_1, a_2 ... a_{\ell_a}\} \in \mathbb{R}^{\ell_a \times e}$. This is followed by embedding the tokens using the embedding function of a pre-trained language model.

Prompt Structure. To create a structured prompt, following existing QA methods using language models [2,26], we prepend the question, image, and answer tokens with tokenized descriptive strings, namely `question:`, `context:` and `answer:`. By placing the embeddings of the question before the visual tokens we mitigate the problem of fixation of the language model on the question [21,22]. As an example this would yield the following prompt template: $p =$`[question: What does the right side of the field show? context:` $v_1, v_2, \dots v_x$ `answer: ]` which is fed as input to the language model.

Language Model. Following standard language modeling systems, we treat VQA as a conditional generation of text, and we optimize the standard maximum likelihood objective during training. The language model receives the prompt sequence p as input and outputs the answer $\mathbf{A}$, token by token. Specifically, at each time step i, the output of the model are the logits parametrizing a categorical distribution $p_\theta(\mathbf{A})$ over the vocabulary tokens. This distribution is represented as follows:

$$\log p_\theta(\mathbf{A}) = \sum_{l_a} \log\ p_\theta(a_i | q_1, ... q_{\ell_q}, v_1, ... v_x, a_1, ... a_{i-1}). \tag{2}$$

The parameters of the language model are initialized from a pre-trained model, which has been previously pre-trained on huge web-collected datasets.

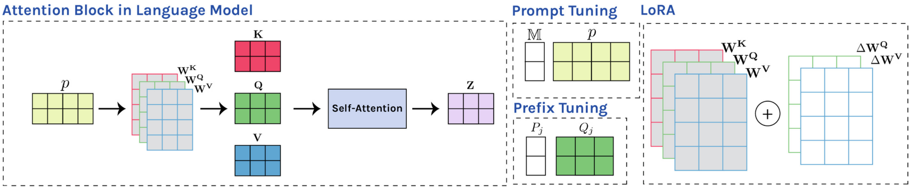


Fig. 2. Parameter-efficient language model fine-tuning strategies used in our method.

3.3 Parameter-Efficient Strategies for Fine-Tuning the Language Model

Standard fine-tuning of language models can hurt the generalization capabilities of the model, especially if small, domain-specific datasets are used as in our case. Therefore, we consider four different parameter-efficient strategies that adapt the attention blocks of language models, as illustrated in Fig. 2 and outlined below:

Frozen Method: the parameters of the language model are kept entirely frozen during training, following [30]. In this setting, only the mapping network is updated through backpropagation. **Prompt Tuning**: we prepend a set of m learnable tokens $\mathbb{M} \in \mathbb{R}^{m \times e}$ to the input prompt sequence, which yields $[\mathbb{M}, p]$ [15] as input to the frozen language model. Besides updating the mapping network, this approach also involves updating these learnable tokens through backpropagation. **Prefix Tuning**: we prepend a learnable prefix P_j to the query Q_j of each attention block j in the Transformer, such that $Q_j^{ft} = [P_j, Q_j]$ [16]. Similar as in prompt tuning, we update both the mapping function and the learnable prefixes of the queries. **Low-Rank Adaptation (LoRA)**: We add learnable weight matrices to the query Q and value V of the attention blocks in each layer of the frozen language model as $\mathbf{W} + \Delta\mathbf{W}$ following [12]. Again, the mapping function is trained together with the learnable weight matrices.

4 Experimental Setup

Datasets. The three datasets used for the evaluation of our method are Slake [20], PathVQA [11], and OVQA [13]. These three datasets are the current most suitable VQA datasets given their large variety in answers and the manual curation of answers by domain experts. Each dataset is split 50/50 between 'yes/no' and open-set answers. See the datatset details in Table 1. We use the official train/validation/test splits across all three datasets.

Evaluation Protocol. We evaluate our approach using the conventional metrics BLEU-1 and F1 Score. Additionally, we measure the contextual capturing of information with BERTScore [35] this method can handle synonyms. Lastly to allow for comparison against existing classification-based methods we also report accuracy and F1 score.

Table 1. Statistics of the medical VQA datasets used in this paper.

	Slake	OVQA	PathVQA
Number of images	642	2,001	4,998
Number of questions	14,028	19,020	32,799
Mean length of questions	4.52	8.98	6.36
Mean length of answers	1.21	3.31	1.80
Number of unique answers	461	641	3,182

Implementation Details. We extract the visual features using a pre-trained CLIP model with ViT backbone [25], having a dimensionality of 512. The MLP layers of the mapping network $\mathbf{f_M}$ have sizes $\{512, (\ell_x \cdot e)/2, \ell_x \cdot e\}$. The length of ℓ_x is set at 8. The lengths ℓ_q and ℓ_a are dataset dependent and defined by the mean number of tokens in the train set plus three times its standard deviation. Zero padding is added to the right side of the sequence for batch-wise learning.

We use the following language models: GPT2-XL [26], a causal language model with 1.5B parameters trained on WebText [26]. BioMedLM [31] and BioGPT [21] are both GPT2-based models, pre-trained on PubMed and biomedical data from The Pile [8], with a size of 1.5B and 2.7B parameters, respectively. All models are able to train on a single NVIDIA RTX 2080ti GPU (average training time ≈ 3 h). We use the AdamW optimizer with 600 warmup steps and a learning rate of 5e-3 and apply early stopping with a tolerance of 3.

5 Results

Benefits of Parameter-Efficient Fine-Tuning. The evaluation of our method across various language models and fine-tuning settings in Table 2 shows

Table 2. Performance across different language models and fine-tuning strategies, measured in BLEU1 (BL1), BERTScore (BS), F1 and accuracy. Params% is the amount of trainable parameters in the language model. Our method using GPT2 in combination with LoRA yields the best performance across all datasets.

	LM fine-tuning	LM size	Params%	Slake				OVQA				PathVQA			
				BL1	BS	F1	Acc.	BL1	BS	F1	Acc.	BL1	BS	F1	Acc.
MedFuseNet [28]				-	-	-	-	-	-	-	-	60.5	-	38.1	-
Ours w/ BioGPT	Frozen	1.5B	0%	64.5	69.9	57.7	66.5	32.4	71.9	52.5	53.5	36.9	57.6	31.0	45.3
	Prefix [16]		0.487%	58.1	74.1	54.1	67.4	37.9	65.0	46.1	53.2	53.6	61.8	34.8	46.7
	Prompt [15]		0.001%	44.2	**75.6**	47.6	53.7	47.5	62.9	34.6	50.3	28.0	58.7	**43.8**	33.2
	LoRA [12]		0.311%	**59.2**	72.2	**63.1**	**71.9**	**41.0**	**68.5**	**57.7**	**57.3**	**57.8**	**62.9**	40.4	**47.9**
Ours w/ BioMedLM	Frozen	2.7B	0%	70.2	77.8	47.8	66.0	55.2	72.9	54.2	61.1	61.2	66.1	52.4	53.0
	Prefix [16]		0.753%	64.3	79.4	60.9	63.3	49.1	76.9	51.5	60.1	59.7	60.7	48.9	52.3
	Prompt [15]		0.009%	44.6	73.5	38.8	41.6	48.9	72.8	44.3	59.5	51.9	59.8	38.9	49.3
	LoRA [12]		0.101%	**72.3**	**80.6**	**62.4**	**71.7**	**59.0**	76.2	**62.6**	**67.8**	**67.9**	**76.0**	**54.4**	**57.2**
Ours w/ GPT2	Frozen	1.5B	0%	65.1	83.3	57.7	71.2	60.2	79.8	59.4	66.1	64.2	74.6	47.5	58.1
	Prefix [16]		0.492%	70.0	86.5	66.3	74.1	61.2	83.9	65.5	68.9	67.5	76.2	52.5	60.5
	Prompt [15]		0.003%	57.8	80.3	49.9	60.0	57.8	78.3	55.2	63.1	54.4	72.0	38.1	46.6
	LoRA [12]		0.157%	**78.6**	**91.2**	**78.1**	**83.3**	**61.8**	**85.4**	**69.1**	**71.0**	**70.3**	**78.5**	**58.4**	**63.6**

Table 3. Comparison of the accuracy between open-ended VQA against classification-based VQA methods, split between yes/no and open-set answers. Our method performs particularly well on both types of answers compared to the state-of-the-art methods.

	Slake			OVQA			PathVQA		
	Open-set	Yes/no	All	Open-set	Yes/no	All	Open-set	Yes/no	All
MEVF-SAN [24]	75.3	78.4	76.5	36.9	72.8	58.5	6.0	81.0	43.6
MEVF-BAN [24]	77.8	79.8	78.6	36.3	76.3	60.4	8.1	81.4	44.8
MEVF-SAN+VQAMix [10]	-	-	-	-	-	-	12.1	84.4	48.4
MEVF-BAN+VQAMix [10]	-	-	-	-	-	-	13.4	83.5	48.6
MMQ-SAN [5]	-	-	-	56.9	76.2	68.5	9.6	83.7	46.8
MMQ-BAN [5]	-	-	-	48.2	76.2	65.0	11.8	82.1	47.1
QCR-BAN [34]	78.8	82.0	80.0	52.6	77.7	67.7	-	-	-
CRPD-BAN [19]	81.2	84.4	82.1	-	-	-	-	-	-
MMBERT [14]	-	-	-	37.9	80.2	63.3	-	-	-
QCR-CLIP [6]	78.4	82.5	80.1	-	-	-	-	-	-
Ours w/ BioGPT (LoRA)	71.1	72.7	71.9	48.3	66.5	57.3	30.2	65.5	47.9
Ours w/ BioMedLM (LoRA)	72.1	71.4	71.7	55.3	80.3	67.8	34.1	80.4	57.2
Ours w/ GPT2 (LoRA)	**84.3**	82.1	**83.3**	**62.6**	**84.7**	**71.0**	**40.0**	**87.0**	**63.6**

that language models can perform open-ended medical VQA. Specifically, we outperform the only existing method MedFuseNet [28] that does open-ended VQA, due to the capability of pre-trained language models to capture long-term dependencies when generating free-form answers. Additionally, prefix [16] and prompt tuning [15] do not improve the performance of the model as much as using LoRA [12] which directly adapts the Q and V weight matrices of the attention blocks. Moreover, larger datasets show the most consistent performance gain of parameter-efficient fine-tuning across all metrics.

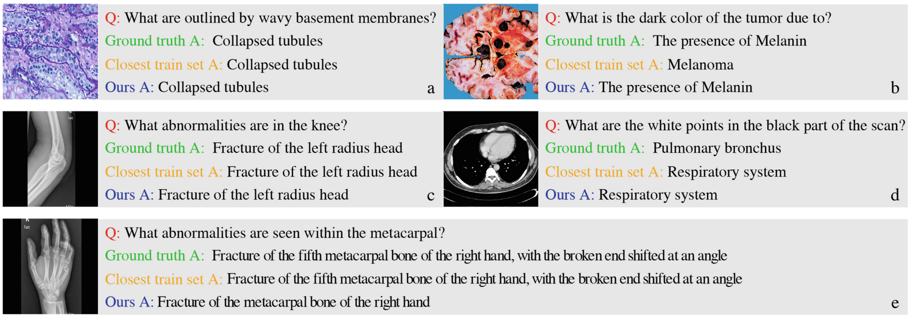
Comparison Between Standard and Medical LMs. Using a language model pre-trained on a general text corpus, such as GPT2 [26], improves the overall performance compared to its medically-trained models (e.g. BioGPT or BioMedLM), as can be observed in Table 2. BioGPT and BioMedLM could be overoptimized to their medical text corpora, which leads to lack of generalization to different downstream domains.

As mentioned in [21,31], these models require full fine-tuning on the respective downstream tasks, to achieve the desired performance. On the other hand, GPT2 benefits from observing diverse data during pre-training which also encompasses medically oriented text. This enables GPT2 models to generalize easily to other domains, which is relevant for our different VQA datasets.

Benefit of Open-Ended Answer Generation. Our method is performing significantly better on the open-set answering, in comparison to classification-based methods, as shown in Table 3. We also confirm that CLIP based image embeddings perform well in the medical domain [6] compared to the conventional use of CNNs. Since our approach is generative, it is not bounded by the class imbalance issue, which is considered a bottleneck of classification-based VQA

Table 4. Effect of using different prompt structures. Note that **Q** and **I** denote the question and image respectively. The regular setting with the question embeddings followed by the visual prefix (Fig. 1) leads to the best overall performance.

Setting	Slake				OVQA				PathVQA			
	B1	BS	F1	Acc.	B1	BS	F1	Acc.	B1	BS	F1	Acc.
w/o **Q**	29.4	48.4	14.3	22.1	33.2	41.9	18.1	27.6	42.9	43.8	18.3	24.6
w/o **I**	54.8	79.3	49.5	50.9	45.5	77.1	49.5	54.4	65.9	72.8	47.2	46.3
Swap **Q** and **I**	73.3	88.7	73.2	74.9	60.0	84.2	67.3	66.9	70.2	78.0	57.2	58.7
Regular	**78.6**	**91.2**	**78.1**	**83.3**	**61.8**	**85.4**	**69.1**	**71.0**	**70.3**	**78.5**	**58.4**	**63.6**



Q: What are outlined by wavy basement membranes?
Ground truth A: Collapsed tubules
Closest train set A: Collapsed tubules
Ours A: Collapsed tubules

Q: What is the dark color of the tumor due to?
Ground truth A: The presence of Melanin
Closest train set A: Melanoma
Ours A: The presence of Melanin

Q: What abnormalities are in the knee?
Ground truth A: Fracture of the left radius head
Closest train set A: Fracture of the left radius head
Ours A: Fracture of the left radius head

Q: What are the white points in the black part of the scan?
Ground truth A: Pulmonary bronchus
Closest train set A: Respiratory system
Ours A: Respiratory system

Q: What abnormalities are seen within the metacarpal?
Ground truth A: Fracture of the fifth metacarpal bone of the right hand, with the broken end shifted at an angle
Closest train set A: Fracture of the fifth metacarpal bone of the right hand, with the broken end shifted at an angle
Ours A: Fracture of the metacarpal bone of the right hand

Fig. 3. Outputs of our open-ended VQA method, with GPT2 and LoRA fine-tuning, using data samples from PathVQA (a, b, d) and OVQA (c, e).

models. Our method performs especially well compared to other method on PathVQA, which relatively has the largest class imbalance, accentuating this effect. Even on the simple 'yes/no' questions, the performance is better, showing that this simple yet effective method provides a more natural way of doing VQA.

It worth noting that the comparison of accuracy as a metric for exact matches, between classification and generation methods is not in favor of generative methods. Despite that, we outperform existing methods on all datasets and metrics, which is a testament to the benefit of phrasing VQA as an open-ended generation problem.

In Fig. 3(a–c), we show qualitative examples of capability of the language model to successfully predict the correct answer. However, in Fig. 3 (d, e) we show cases where our method predicts a factually correct answer which is not specific enough.

Effect of Using Different Prompt Structures. We also investigate the influence of the prompt structure on the overall performance, demonstrated in Table 4. It can be observed that the performance largely decreases when the question is removed, compared to when the visual information is removed. This suggests that the question plays a more important role in answer generation.

Interestingly, the model is sensitive the order of the elements in the prompt, as the swapping of the question embeddings and the visual prefix yields decreases the performance. The reason for this is that the language model conveys lower to no importance the visual information if it is located in front of the question. In this situation the language model basically generates blind answers. This highlights the importance of prompt structure.

6 Conclusion

In this paper, we propose a new perspective on medical VQA. We are using generative language models to generate answers in an open-ended manner, instead of performing a closed-set classification. Additionally, by using various parameter-efficient fine-tuning strategies we are able to use language models with billions of parameters, even though dataset sizes in this domain are small. This leads to excellent performance compared to classification-based methods. In conclusion, our approach offers a more accurate and efficient solution for medical VQA.

Acknowledgements. This work is financially supported by the Inception Institute of Artificial Intelligence, the University of Amsterdam and the allowance Top consortia for Knowledge and Innovation (TKIs) from the Netherlands Ministry of Economic Affairs and Climate Policy.

References

1. Barraco, M., Cornia, M., Cascianelli, S., Baraldi, L., Cucchiara, R.: The unreasonable effectiveness of CLIP features for image captioning: an experimental analysis. In: CVPR, pp. 4662–4670 (2022)
2. Brown, T., et al.: Language models are few-shot learners. NeurIPS **33**, 1877–1901 (2020)
3. Cong, F., Xu, S., Guo, L., Tian, Y.: Caption-aware medical VQA via semantic focusing and progressive cross-modality comprehension. In: ACM Multimedia, pp. 3569–3577 (2022)
4. Derakhshani, M.M., et al.: Variational prompt tuning improves generalization of vision-language models. arXiv:2210.02390 (2022)
5. 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
6. Eslami, S., de Melo, G., Meinel, C.: Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? arXiv:2112.13906 (2021)
7. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: ICLR, pp. 1126–1135 (2017)
8. Gao, L., et al.: The pile: an 800 GB dataset of diverse text for language modeling. arXiv:2101.00027 (2020)
9. 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: ICMR, pp. 456–460 (2021)

10. Gong, H., Chen, G., Mao, M., Li, Z., Li, G.: Vqamix: conditional triplet mixup for medical visual question answering. IEEE Trans. Med. Imaging (2022)
11. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv:2003.10286 (2020)
12. Hu, E.J., et al.: Lora: low-rank adaptation of large language models. arXiv:2106.09685 (2021)
13. Huang, Y., Wang, X., Liu, F., Huang, G.: OVQA: A clinically generated visual question answering dataset. In: ACM SIGIR, pp. 2924–2938 (2022)
14. Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U.D., Jawahar, C.: MMBERT: multimodal BERT pretraining for improved medical VQA. In: ISBI, pp. 1033–1036. IEEE (2021)
15. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: EMNLP, pp. 3045–3059 (2021)
16. Li, X.L., Liang, P.: Prefix-tuning: optimizing continuous prompts for generation. In: ACL, pp. 4582–4597 (2021)
17. Li, Y., et al.: A bi-level representation learning model for medical visual question answering. J. Biomed. Inf. **134**, 104183 (2022)
18. Lin, Z., et al.: Medical visual question answering: a survey. arXiv:2111.10056 (2021)
19. Liu, B., Zhan, L.-M., Wu, X.-M.: Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12902, pp. 210–220. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87196-3_20
20. 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: ISBI, pp. 1650–1654. IEEE (2021)
21. Luo, R., et al.: BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings Bioinformat. **23**(6) (2022)
22. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: clip prefix for image captioning. arXiv:2111.09734 (2021)
23. Najdenkoska, I., Zhen, X., Worring, M.: Meta learning to bridge vision and language models for multimodal few-shot learning. In: ICLR (2023)
24. Nguyen, B.D., Do, T.-T., Nguyen, B.X., Do, T., Tjiputra, E., Tran, Q.D.: Overcoming data limitation in medical visual question aswering. 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
25. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763. PMLR (2021)
26. Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
27. Ren, F., Zhou, Y.: Cgmvqa: a new classification and generative model for medical visual question answering. IEEE Access **8**, 50626–50636 (2020)
28. Sharma, D., Purushotham, S., Reddy, C.K.: MedFuseNet: an attention-based multimodal deep learning model for visual question answering in the medical domain. Sci. Rep. **11**(1), 19826 (2021)
29. Taylor, N., Zhang, Y., Joyce, D., Nevado-Holgado, A., Kormilitzin, A.: Clinical prompt learning with frozen language models. arXiv:2205.05535 (2022)
30. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. NeurIPS **34**, 200–212 (2021)
31. Venigalla, A., Frankle, J., Carbin, M.: BioMedLM: a domain-specific large language model for biomedicine. www.mosaicml.com/blog/introducing-pubmed-gpt (2022). Accessed 06 Mar 2022

32. Wang, J., Huang, S., Du, H., Qin, Y., Wang, H., Zhang, W.: MHKD-MVQA: multimodal hierarchical knowledge distillation for medical visual question answering. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 567–574. IEEE (2022)
33. Wu, Q., Wang, P., Wang, X., He, X., Zhu, W.: Medical VQA. In: Visual Question Answering: From Theory to Application, pp. 165–176. Springer, Singapore (2022). https://doi.org/10.1007/978-981-19-0964-1_11
34. Zhan, L.M., Liu, B., Fan, L., Chen, J., Wu, X.M.: Medical visual question answering via conditional reasoning. In: ACM Multimedia, pp. 2345–2354 (2020)
35. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: evaluating text generation with bert. In: ICLR (2020)