Is Unimodal Bias Always Bad for Visual Question Answering? A Medical Domain Study with Dynamic Attention

1st Zhongtian Sun
Department of Computer Science
Durham University
Durham, UK
zhongtian.sun@durham.ac.uk

2nd Anoushka Harit
Department of Computer Science
Durham University
Durham, UK
anoushka.harit@durham.ac.uk

3rd Alexandra I. Cristea
Department of Computer Science
Durham University
Durham, UK
alexandra.i.cristea@durham.ac.uk

4th Jialin Yu
Department of Computer Science
Durham University
Durham, UK
jialin.yu@durham.ac.uk

5th Noura Al Moubayed
Department of Computer Science
Durham University
Durham, UK
noura.al-moubayed@durham.ac.uk

6th Lei Shi
Department of Computer Science
Durham University
Durham, UK
lei.shi@durham.ac.uk

***Abstract*—Medical visual question answering (Med-VQA) is to answer medical questions based on clinical images provided. This field is still in its infancy due to the complexity of the trio formed of questions, multimodal features and expert knowledge. In this paper, we tackle, a 'myth' in the Natural Language Processing area - that unimodal bias is always considered undesirable in learning models. Additionally, we study the effect of integrating a novel dynamic attention mechanism into such models, inspired by a recent graph deep learning study.**

Unlike traditional attention, dynamic attention scores are conditioned on different query words in a question and thus enhance the representation learning ability of texts. We propose that some questions are answered more accurately with a reinforcement of question embedding after fusing multimodal features. Extensive experiments have been implemented on the VQA-RAD datasets and demonstrate that our proposed model, reinforCe unimOdal dynamiC Attention (COCA), outperforms the state-of-the-art methods overall and performs competitively at open-ended question answering.

***Index Terms*—Healthcare data, medical application, multimedia, visual question answering, feature representation learning**

I. Introduction

Medical Visual Question Answering (Med-VQA) is a domain-specific multimodal challenging task widely studied by research communities in computer vision and natural language processing. Med-VQA aims to answer clinical questions in text form, based on medical image and language information, as a sub-domain of the question answering task in natural language processing (NLP) [1]. Therefore, it is a multimodal learning task related to both computer vision (CV) and NLP [2], including a variety of sub-tasks, as shown in Fig 1 [3].

In practice, doctors are required to have a profound understanding of the problems indicated by medical images and perform explicit reasoning to confirm a diagnosis [4], the process of which may be lengthy and costly. Instead, the medical visual question answering task can better assist the doctor in the diagnosis and alleviate the imbalanced medical resource status [5], by significantly reducing misdiagnosis and improving accuracy [4].

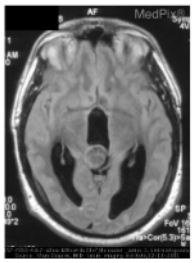
CV tasks	NLP tasks
Object recognition	Tokenization
Object detection	Sentence boundary detection
Scene classification	Semantics
Attribute classification	Synatax parsing
Activity recognition	Word sense disambiguation
Common reasoning	Sentence generation
Knowledge-based reasoning	Opinion mining
Counting	Summarizing
Spatial relations among objects	Aggregation

Fig. 1. Learning tasks involved in VQA, cited from [3]

This paper aims to enhance the learning ability of models in Med-VQA for the social good. The input for the Med-VQA models is a pair of questions and images, and the output is the answer to the question based on the image, shown in Figure 2. The dataset we use is VQA-RAD, proposed by [6], which contains 11 topics of questions. There may be more than one question-answer pair for one image. Questions can be divided into open-ended - (free-form text) and close-ended answers (limited answers, mostly "yes" or "no").

Compared with general VQA tasks, Med-VQA requires higher accurate prediction, due to the safety concern. I.e., it is necessary to enhance the systems' recognition and reasoning

A image



B image

	A	B
Question	what is the location of the mass?	Can the optic nerve be visualized in this MRI image?
Question Topic	Position	Pres
Answer	Head of the pancreas	Yes
Answer Type	Open	Closed

Fig. 2. Example of Med-VQA Data

skills, to support correct clinical diagnosis [1]. The prime focus is to improve the representation learning ability of models for Med-VQA tasks. Our contributions are summarised as follows:

- We propose a novel pipeline to reinforce the text bias after fusing multimodal features, combined with dynamic attention, named the reinforCe unimOdal dynamiC Attention model (COCA), which can be applied universally for visual question answering tasks.
- To the best of our knowledge, we are the *first* to question an overall accepted 'myth' that unimodal biases in medical VQA should be avoided [7] and, moreover, prove that adding unimodal bias after feature fusion under certain conditions can improve the prediction accuracy under specific circumstances.
- Experimental results on a real-world dataset show the superior performance of the proposed COCA model compared with the state-of-the-art.

II. Related Work

Medical Visual Question Answering Medical VQA systems are developed based on deep learning, to automatically extract the information from medical images and assist in clinical diagnosis [8]. Current medical VQA mainly consists of a visual representation learning module, language modelling for question module, and a multimodal feature fusion module. Attention mechanisms, including Bilinear Attention Networks (BAN) [9] and Stacked Attention Networks (SAN) [10] are also deployed, to enhance the relation caption of visual and textual information. Recent advances include reasoning [1], to consider closed - and open-ended questions, multi-view attention [4], to correlate questions with images and texts with more attention, and Multiple Meta-model Quantifying (MMQ) [11], to enhance meta-data, by auto-annotation with noisy labels.

Unimodal Biases of VQA. According to [12], the data collection process may lead to inherent biases in real-world datasets, and models will learn the biases during training [13]. In general-domain VQA, unimodal bias refers to a model answering questions without considering another modality, like images or videos. Unimodal bias will undermine VQA model performance when there is a dataset shift, and thus reducing unimodal biases becomes a goal of state-of-the-art learning strategies [14]. However, there are cases when reducing unimodal bias will not benefit model performance for other textually biased VQA datasets [15]. Due to the small scale of VQA datasets in the medical domain [7], this is less covered and some questions indeed do not need to view the related image before answering. In such cases, reinforcing unimodal biases can improve prediction performance, particularly after fusing multimodal features, which we study in this paper.

Attention Mechanism. The attention mechanism plays an indispensable role in the existing VQA tasks for feature extraction and multimodal fusion [1], [16], [17]. It assists models in understanding the relations between questions and images deeply and is necessary when the questions are complex or difficult to answer. The use of attention can significantly improve the performance of models on VQA tasks [4]. [1] proposed a question-conditioned reasoning model, which selectively fused the multimodal features, according to their importance, and learned more semantic information from the question representations. [4] then designed the multi-view attention model, including image-to-question and word-to-text attention, to correlate questions with images. In this paper, we apply a different expressive attention method, known as dynamic graph attention, introduced by [18], which we show to outperform the state-of-the-art.

III. Methodology

In this section, we introduce our model, inspired by [1], with dynamic attention mechanism and unimodal biases enhancement, as well as the basic blocks of models for VQA tasks, shown in Fig 3.

A. Basic Blocks

A traditional VQA model consists of three sections: a) image features caption module, b) text features caption module, and c) multimodal feature fusion and classification module. The structure is as follows:

Image Feature Extraction. Due to the small size of the medical VQA dataset, in our model, the image features are learnt the same way as in [19], which initialised the pre-trained weights using Model-Agnostic Meta-Learning (MAML) and a Collaborative Denoising Auto-Encoder (CDAE). MAML is trained to solve new learning tasks with a small number of samples [20]. It is suitable for the medical VQA task, which has limited available training data [21]. CDAE can be applied

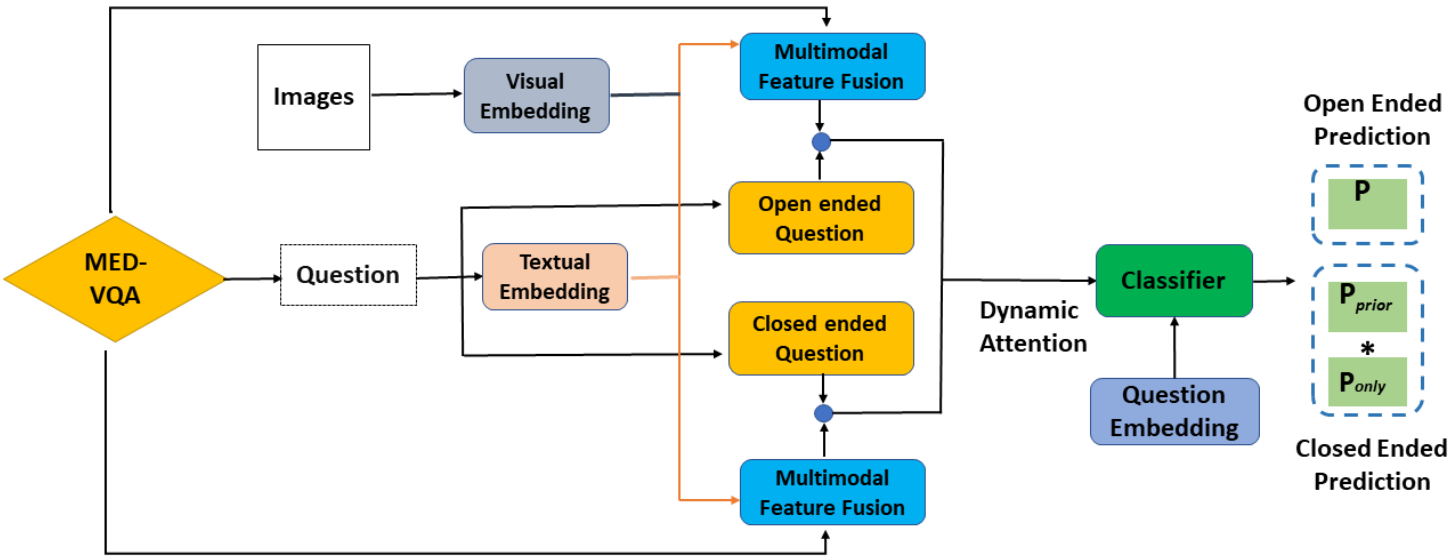


Fig. 3. Framework of the proposed COCA Method including feature extraction, fusion and unimodal bias reinforcing module

to large unlabeled datasets, to improve the robustness of the model [4]. It can be formulated as follows:

$$s_i = f_\theta(v) = \sigma(Wv + b) \tag{1}$$

$$y_i = f_{\theta'}(s_i) = \sigma(W's_i + b') \tag{2}$$

where the image features are represented as vector v with 128 dimensions, and the auto-encoder is deployed, by finding the latent representation s of image i first to denoise the inputs in equation 1. From the latent representation s_i, it reconstructs the original input, as in equation 2 [22]. Images will be passed to the MAML and CDAE models separately, and the learned features will be concatenated for further processing, as per Figure 4. The rest of the model parameters are similar to [1].

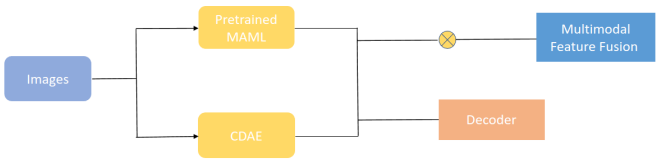


Fig. 4. Image processing

Text Feature Extraction. Each question q consists of n words, and the maximum length of a question is 12. Every word in a question is represented by 300 dimensions embedding W_{emb}, initialised from Glove [23]. Next, the word embedding W_{emb} is passed to the Long Short Term Memory (LSTM) model [24] in equation 4 [25], to obtain semantic information Q_{emb}.

$$W_{emb} = Wordembedding(q) \tag{3}$$

$$\begin{aligned}
i_t &= \sigma(W_{ii} \cdot x^t + b_{ii} + W_{hi} h_{t-1} + b_{hi}), \\
f_t &= \sigma(W_{if} \cdot x^t + b_{if} + W_{hf} ht - 1 + b_{hf}), \\
g_t &= tanh(W_{ig} \cdot x^t + b_{ig} + W_{hg} ht - 1 + b_{hg}), \\
o_t &= \sigma(W_{io} \cdot x^t + b_{io} + W_{ho} ht - 1 + b_{ho}), \\
c_t &= f_t \cdot c_{t-1} + i_t \cdot g_t, \\
h_t &= o_t \cdot tanh(c_t).
\end{aligned} \tag{4}$$

where x is the word embedding obtained from equation 3 and h_t is the hidden state of word at time t . The dimension of each hidden state is 1024, similar to [1]. i_t, f_t, g_t and o_t are gates for input, forget, cell and output. c_t is the updated cell state.

$$Q_{emb} = LSTM(W_{emb}) \tag{5}$$

This Q_emb is then separated into closed and open-ended question embeddings Q_emb^{cl} and Q_emb^{op}, respectively. Highlighting the critical words in questions, [1] proposed an attention mechanism: concatenate W_{emb} and Q_{emb} (for simplification, we omit the close cl and open op notations here) to obtain a comprehensive representation of the text T_{emb} and then use its dot product for further attention calculation:

$$T_{new} = tanh(W_1 T_{emb}) \odot \sigma(W_2 T_{emb}) \tag{6}$$

$$a = softmax(W_a T_{new}), Q_{new} = aQ_{emb} \tag{7}$$

where W_1,W_2, W_a are weights to be trained, $\odot$ is a Hadamard product, σ denotes the sigmoid activation, and Q_{new} is the new question embedding, which considers the importance of different words, and will be fed into the question type attention layer.

Multimodal Feature Fusion and Classification. The open - and close-ended question features will be updated considering each type of question attention. The final features of open - and close-ended questions will be fed into a linear prediction layer for classification:

$$C = \begin{cases} MLP(F(v^{op}, Q^{op}_{new}) \cdot a^{op}_{type}) \text{ for open question} \\ MLP(F(v^{cl}, Q^{cl}_{new}) \cdot a^{cl}_{type}) \text{ for closed question} \end{cases}$$

where F is a multimodal feature fusion function. The new fused embedding will then multiply a question type attention a_{type}. Then, an MLP is applied for the final classification C.

B. Dynamic Attention

Based on the basic blocks for multimodal learning, we have the following improvement. The improved attention mechanism is mainly inspired by [18], who proposed that dynamic attention rather than static attention in graph attention network (GAT) [26] should be deployed in graph representation learning. According to [27], a graph neural network with multi-head attention can be considered a transformer. To be more specific, we select one question out of the dataset: *Is there a rib fracture?*; and analyse it from both graph and transformer aspects, depicted in (a) and (b) in Figure 6. In Transformer, the attention mechanism will measure the relative importance of all other words to each word in a sentence [28]. From the graph perspective, it indicates that each word/node is connected with all other words/nodes, and the attention between any pair of nodes will be calculated for every edge. In other words, graph attention can be seen as a shared linear transformation for all words [29]. Therefore, sentences can be considered fully-connected word graphs and there is no need to build an extra graph structure or adjacency matrix.

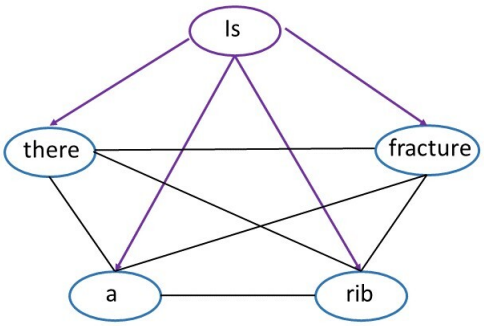


Fig. 5. Graph view of a sentence

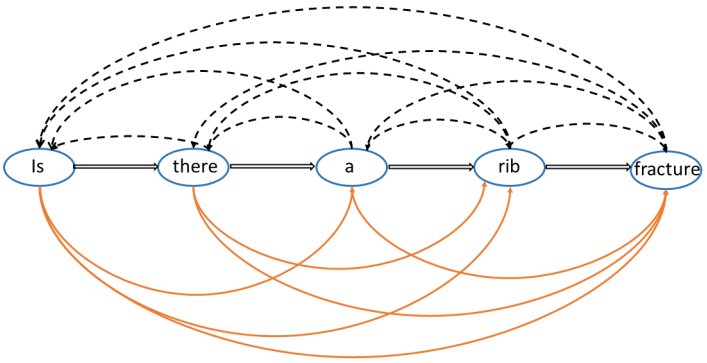


Fig. 6. Transformer view of a sentence

In such a case, applying graph dynamic attention learning to texts is reasonable. Each question sentence is viewed as a graph containing different nodes/words. The attention $\alpha_{i,j}$ of node j to node i in graph representation learning is normally obtained by calculating a score for every edge $e_{i,j}$ followed by a softmax function:

$$\begin{aligned} e(h_i, h_j) &= LeakyReLU(a^T \cdot [Wh^i || Wh^j]) \\ \alpha_{ij} &= \frac{exp(e_{ij})}{\sum_{j \in N_i} exp(e_{ij'})} \end{aligned} \tag{8}$$

where $||$ denotes concatenation, a^T is a learnable vector. LeakyReLU is an activation applied later and $j \in N_i$ refers to all neighbours of node i. According to [18], the above method is typical *Static Attention*, as there is a highest scoring key k_{j_f}, for every query q_i, $f(q_i, k_{j_f}) \geq f(q_i, k_{j_{else}})$. In which case, a specific key will always be considered and the attention value is always the same, regardless of the query, due to the monotonicity of softmax and LeakyReLU.

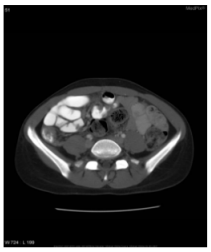
The ideal attention should be *Dynamic Attention* that for any query q_i, $f(q_i, k_{j_f}) > f((q_i, k_{j_{else}}))$ and the formula is:

$$e(h_i, h_j) = a^T LeakyReLU(W \cdot [h^i || h^j]) \tag{9}$$

followed by the same softmax function in eq 8. The new attention score will be different, conditioned on the query node. Inspired by [18], we apply this dynamic attention to the original question representation q to calculate the attention a_t of closed and open-ended questions by adding a LekyReLu activation, which is then applied to an MLP based on the existing attention, obtained in eq. 7. Next, the attention multiplies with the updated embeddings after the multimodal feature fusion function. Note that we only change the computing order within the attention mechanism and keep the same Glove and LSTM model for text feature extraction, as addressed in subsection III-A.

C. Reinforcing Unimodal Biases

Inspired by [14], we input question features directly into the classifier layer, to obtain a unimodal prediction, which



Same image

Question	Is there evidence of small bowel obstruction on this image?	Is there small bowel thickening present?
Question Topic	Pres	Pres, Size
Answer	Yes	Yes
Answer Type	Closed	Closed

Fig. 7. Multiple questions for one image

will be multiplied by the previous prediction to reinforce the unimodal biases. The idea is that some questions can be theoretically answered or assumed as logic-based, e.g., *How would you measure the length of the kidneys? What is located immediately inferior to the right hemidiaphragm?* The answers to the above questions are *coronal plane* and *the liver*, respectively, which do not require screening first. Additionally, as there are multiple questions for one image, as shown in 7, if we reinforce the unimodal bias (text), the model may better understand the question for the same image.

Therefore, we add the function for close-ended questions:

$$P^{cl}_{only} = MLP(Q^{cl}_{new}) \tag{10}$$

$$P^{cl}_{final} = P^{cl}_{prior} * \sigma(P^{cl}_{only}) \tag{11}$$

this P^{cl}_{only} denotes the question-only prediction based on unimodal question embeddings. P^{cl}_{prior} is the previous prediction obtained by the classifier layer using fused multimodal features. It will be updated by multiplying with the question-only prediction P^{cl}_{only} after a sigmoid function to obtain the final prediction P^{cl}_{final}. Then the final prediction P_{final} is passed to the cross-entropy loss function:

$$Loss = BCE(P^{cl}_{final}, Y^{cl}) + BCE(P^{op}, Y^{op}) \tag{12}$$

where BCE is the binary cross-entropy loss function and Y is the ground truth.

It should be noted that reinforcing bias does not mean we only consider single modular data. In our model, we reinforce the unimodal bias after the multimodal feature fusion step, shown in Figure 3, which means we have considered both images and questions earlier. Such reinforcement enables COCA to better understand the difference between the current question and other questions. Therefore, reinforcing unimodal bias would not hurt performance and can be applied to other datasets - where most examples require fusing between multiple modal features. The effectiveness of reinforcing unimodal bias is illustrated in the Ablation study.

TABLE I
MEDICAL-VQA RAD DATA SUMMARY [6]

	Training Set		Test Set	
	Close-ended	Open-ended	Close-ended	Open-ended
Modality	77	77	17	16
Plane	47	47	12	14
Organ	15	34	2	8
Abnormal	191	126	38	18
Presence	965	267	122	45
Position	67	439	8	52
Attribute	79	70	14	4
Color	67	17	3	0
Count	17	22	4	2
Size	244	25	41	5
Other	52	119	11	15
Total	1821	1243	272	179

IV. EXPERIMENTS

A. Dataset

We use the same benchmark dataset, VQA-RAD, as in [1]. It has been divided into training (3,064) and test (451) sets, with 3,515 question-answer pairs and 315 radiology images. There are eleven question topics, such as modality, plane, position, and colour, as shown in Table I. Questions can be divided into two types: open-ended and close-ended. The answers of the former are free-form [6], and the answers of the latter are mainly binary - in the form of ”yes” or ”no” and other limited choices.

B. Baselines and Experiment Settings

Baselines. We consider seven widely adopted -, and the-state-of-the-art methods, as baselines.

- Multimodal Compact Bilinear (MCB) pooling aims to reduce cost in feature fusion by projecting the outer product to lower the dimensional space [1], [30].
- Bilinear Attention Networks (BAN) deploy a low-rank bilinear pooling mechanism to reduce the computational cost in multimodal feature fusion [9].
- Stacked Attention Networks (SAN) focus on relevant areas in images in a multi-step reasoning manner, based on a stacked attention model [10].

- The mixture of Enhanced Visual Features (MEVF + BAN) combines the MEVF framework with a separate attention model to fuse multimodal features [19].
- Conditional Reasoning (CR) considers the influence of type via an attention mechanism based on BAN and MAML [1].
- Multi-view attention model (MuVAM) applies a dual attention mechanism for image-to-question and word-to-text and combines a composite loss to improve the similarity between visual and textual cross-modal features [4].
- Multiple Meta-model Quantifying (MMQ) proposes to use meta-annotation, leverages meaningful features for the Med-VQA task, and is the state-of-the-art [11].

Experimental Settings. We implement the proposed framework with Pytorch 1.11.0 and CUDA 11.3 on a Linux system (Ubuntu 20.04) with a GPU NVIDIA RTX 2080Ti. The hyper-parameters of the base model are mainly borrowed from the experiments of Conditional Reasoning [1]. The number of hidden units is set as 1024, batch size 64, question length 12, learning rate 0.005 and the Adam optimiser is deployed. A MAML with CDAE module [22] is implemented and pre-trained as in [19], to capture the image features, and the size is 128. For question features, a 1024 dimension hidden states LSTM model with Glove [23] is applied to initialise word embeddings.

C. Experimental Results

Accuracy is calculated as the proportion of the total amount of questions that the model classifies correctly:

$$Acc = \frac{P_{correct}}{P_{total}} \tag{13}$$

TABLE III
THE AVERAGE PREDICTION ACCURACY OF MODELS IN VQA-RAD DATASET

Methods	Open-ended	Close-ended	Overall
MCB	25.4	60.6	46.2
SAN	24.2	57.2	44.2
BAN	28.4	67.9	52.3
MEVF+BAN	49.2	77.2	66.1
MMQ	53.7	75.8	67.0
CR	60	79.3	71.6
MuVAM	**63.3**	81.1	72.2
Ours	61.7	**81.9**	**73.3**

Table III shows the classification performance of different models on the VQA-RAD datasets. The proposed COCA model outperforms most baselines, except for the open-ended questions, compared with MuVAM. The possible reason can be that we do not deploy extra attention for vision-to-questions, yet, we still obtain a competitive result - improved by 1.7% compared to the base model CR. This demonstrates the effectiveness of the proposed COCA model on Med-VQA tasks.

D. Ablation Study

We conduct an ablation study to demonstrate the effectiveness of our COCA model. As shown in Table IV, reinforcing the unimodal bias - the close question embedding (CQM) improves the overall accuracy from 71.6% to 72.7%. For dynamic attention (with DA), the closed-ended questions, overall and open-ended questions' accuracy improve by 1.9%, 1.6% and 1.1%, respectively. We also test our model with open question embedding (OQM) reinforcement for open question prediction P^{op}_{final}, similar to equation 11:

$$P^{op}_{final} = P^{op}_{prior} * \sigma(P^{op}_{only}) \tag{14}$$

which significantly downgrades the performance. It demonstrates that there are certain contexts to reinforce unimodal biases in VQA tasks.

As in this VQA-RAD dataset, one image can link to multiple questions, in which case some questions ask about the abnormal object in an image, while others ask for the normal area. The result validates the point that dynamic attention enables the model to better understand the differences between those questions, assisting in the classification.

TABLE IV
ABLATION STUDY OF OUR COCA MODEL

Methods	Open-ended	Close-ended	Overall
CR	60.0	79.3	71.6
COCA without CQM	61.1	81.2	73.2
COCA without DA	60.6	80.8	72.7
COCA with OQM	51.1	79.3	68.1
COCA	**61.7**	**81.9**	**73.3**

E. Visualisation Study

The visualisation evaluation of our model, COCA, on the VQA-RAD dataset in three organs: head, chest and abdomen, is shown in Fig 8. P represents the predicted answer, A refers to the ground truth, and ABD is short for the abdomen. The green colour of prediction means the inferred answers are correct, and red denotes the wrong prediction. Generally, COCA can identify the keywords of the question and correctly

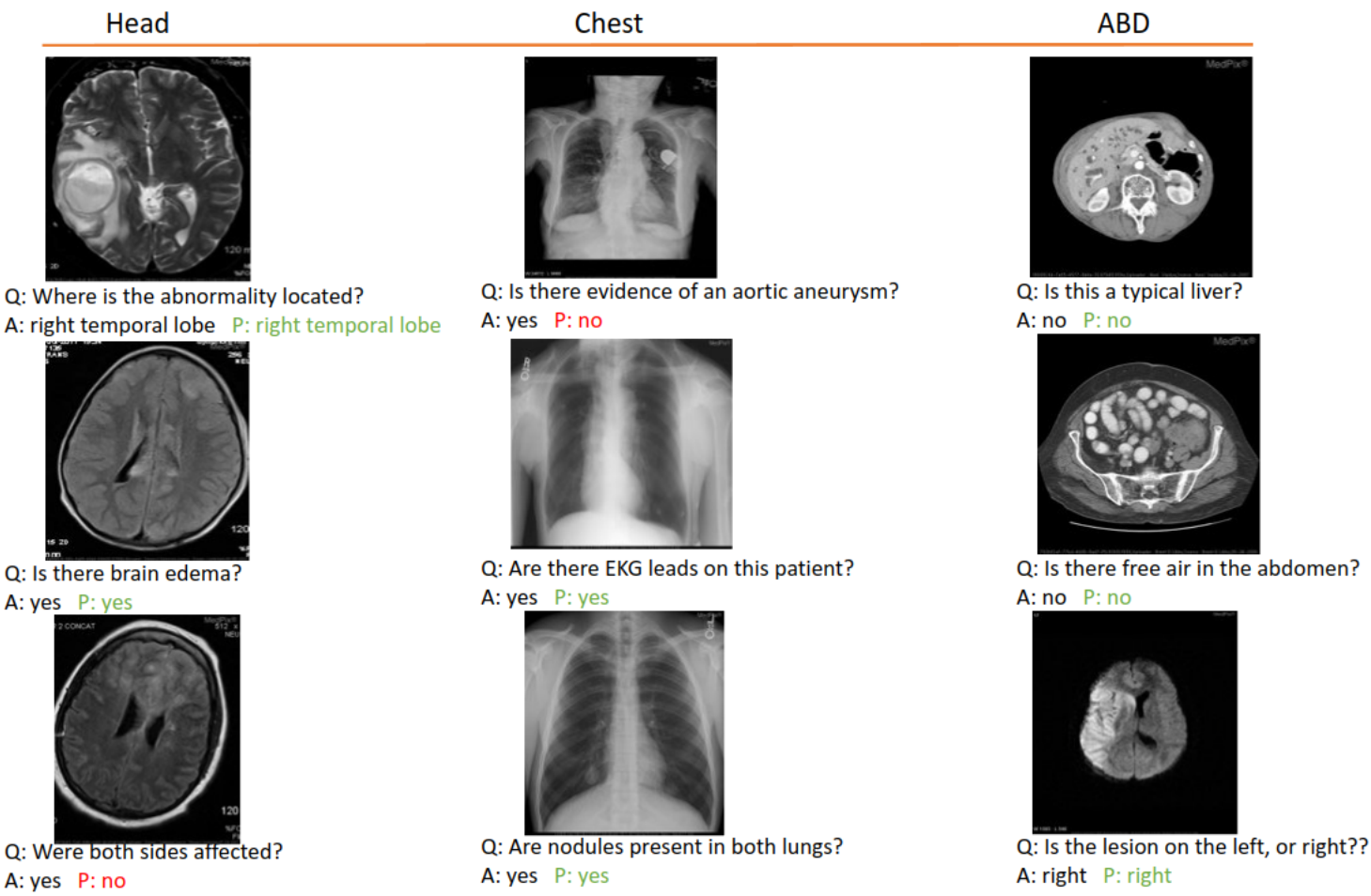


Fig. 8. Visualisation of the proposed model, COCA

identify the corresponding visual information. For instance, COCA accurately answers the position-type question in the first open-ended question related to the head, and understands the keyword "abnormality". Note that in radio-graph images, the actual location is opposite of what we observe [4]. Our method can understand if the image is a typical description of an organ, e.g. liver, in the first image of ABD. It gives the correct answer, "no", which indicates it remembers what a typical liver looks like.

However, our model cannot always give correct answers, as the first example for the chest shows. The question requires to be answered by more reasoning skills and the professional experience of doctors, such as what phenomena can be counted as evidence of an aortic aneurysm and how it can be shown on the image. It is hard to answer the question without external knowledge of aortic aneurysms, such as knowledge graphs. Additionally, the proposed model requires more specific descriptions in the questions to give correct answers. In the third sample for the head, it could be clearer to ask if both sides were affected, as affection includes various illnesses, and the method cannot know all of them, making it impossible to predict correctly. Therefore, we point out possible further improvements in section VI.

V. Conclusion

This paper takes on, for the first time, to the best of our knowledge, the accepted 'myth' that unimodal biases in medical VQA should be avoided - and proposes a novel and effective pipeline for question-answering tasks, based on condition reasoning methods. We demonstrate that dynamic attention can enhance the text representation of the language model, and adding unimodal question biases after fusing multimodal features improves the prediction accuracy for the closed questions. Extensive experiments illustrate that the proposed method, COCA, outperforms the state-of-the-art.

VI. Future Work

There are several promising future research directions to explore:

- One is to design an indicator to measure the unimodal portion of multimodal datasets and automatically apply our methods to those with high unimodal portions.
- The representation learning ability for images can be enhanced. For instance, a state-of-the-art model, Vision-Transformer [31], can be deployed for image feature

extraction, to improve the current results. Additionally, as there is a limited amount of training data available in medical VQA, we can apply graph generative methods [32], to enhance the generalisation ability of models.

- Graph representation learning methods can be introduced to the question embeddings, such as heterogeneous graph neural networks for different words [33], [34], [35].
- External knowledge, such as knowledge graphs, can be considered [36], [37], so that the model can understand questions and implement the inference, by connecting questions to knowledge graphs.
- Our proposed COCA framework can be applied to other visual question-answering tasks, such as image retrieval [38], cultural heritage [39] advertising [40], surveillance [41] and personal assistant [42], or other multimodal learning, including human-computer interaction [43] and communication [44], [45], to enhance the prediction performance of models.

REFERENCES

[1] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu, "Medical visual question answering via conditional reasoning," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2345–2354.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[3] S. Barra, C. Bisogni, M. De Marsico, and S. Ricciardi, "Visual question answering: Which investigated applications?" *Pattern Recognition Letters*, vol. 151, pp. 325–331, 2021.

[4] H. Pan, S. He, K. Zhang, B. Qu, C. Chen, and K. Shi, "Muvam: A multi-view attention-based model for medical visual question answering," *arXiv preprint arXiv:2107.03216*, 2021.

[5] F. Ren and Y. Zhou, "Cgmvqa: A new classification and generative model for medical visual question answering," *IEEE Access*, vol. 8, pp. 50 626–50 636, 2020.

[6] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, "A dataset of clinically generated visual questions and answers about radiology images," *Scientific data*, vol. 5, no. 1, pp. 1–10, 2018.

[7] Z. Lin, D. Zhang, Q. Tac, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge, "Medical visual question answering: A survey," *arXiv preprint arXiv:2111.10056*, 2021.

[8] H. Gong, G. Chen, S. Liu, Y. Yu, and G. Li, "Cross-modal self-attention with multi-task pre-training for medical visual question answering," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021, pp. 456–460.

[9] J.-H. Kim, J. Jun, and B.-T. Zhang, "Bilinear attention networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[10] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[11] T. Do, B. X. Nguyen, E. Tjiputra, M. Tran, Q. D. Tran, and A. Nguyen, "Multiple meta-model quantifying for medical visual question answering," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 64–74.

[12] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*. IEEE, 2011, pp. 1521–1528.

[13] V. Manjunatha, N. Saini, and L. S. Davis, "Explicit bias discovery in visual question answering models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9562–9571.

[14] R. Cadene, C. Dancette, M. Cord, D. Parikh *et al.*, "Rubi: Reducing unimodal biases for visual question answering," *Advances in neural information processing systems*, vol. 32, 2019.

[15] T. Winterbottom, S. Xiao, A. McLean, and N. A. Moubayed, "On modality bias in the tvqa dataset," *arXiv preprint arXiv:2012.10210*, 2020.

[16] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 10 313–10 322.

[17] M. H. Vu, T. Löfstedt, T. Nyholm, and R. Sznitman, "A question-centric model for visual question answering in medical imaging," *IEEE transactions on medical imaging*, vol. 39, no. 9, pp. 2856–2868, 2020.

[18] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" *arXiv preprint arXiv:2105.14491*, 2021.

[19] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran, "Overcoming data limitation in medical visual question answering," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 522–530.

[20] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 1126–1135.

[21] D. Sharma, S. Purushotham, and C. K. Reddy, "Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Scientific Reports*, vol. 11, no. 1, pp. 1–18, 2021.

[22] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International conference on artificial neural networks*. Springer, 2011, pp. 52–59.

[23] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[25] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[26] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[27] C. Joshi, "Transformers are graph neural networks," *The Gradient*, p. 5, 2020.

[28] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu, "Do transformers really perform bad for graph representation?" *arXiv preprint arXiv:2106.05234*, 2021.

[29] J. Kim, T. D. Nguyen, S. Min, S. Cho, M. Lee, H. Lee, and S. Hong, "Pure transformers are powerful graph learners," *arXiv preprint arXiv:2207.02505*, 2022.

[30] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[32] Z. Sun, A. Harit, J. Yu, A. I. Cristea, and N. Al Moubayed, "A generative bayesian graph attention network for semi-supervised classification on scarce data," in *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2021, pp. 1–7.

[33] H. Linmei, T. Yang, C. Shi, H. Ji, and X. Li, "Heterogeneous graph attention networks for semi-supervised short text classification," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4821–4830.

[34] H. Wang, S. Li, T. Wang, and J. Zheng, "Hierarchical adaptive temporal-relational modeling for stock trend prediction," 2021.

[35] Z. Sun, A. Harit, A. I. Cristea, J. Yu, L. Shi, and N. Al Moubayed, "Contrastive learning with heterogeneous graph attention networks on short text classification," in *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022, pp. 1–6.

[36] J. Park, Y. Cho, H. Lee, J. Choo, and E. Choi, "Knowledge graph-based question answering with electronic health records," in *Machine Learning for Healthcare Conference*. PMLR, 2021, pp. 36–53.

[37] I. Tiddi and S. Schlobach, "Knowledge graphs as tools for explainable machine learning: A survey," *Artificial Intelligence*, vol. 302, p. 103627, 2022.

[38] A. Bansal, Y. Zhang, and R. Chellappa, "Visual question answering on image sets," in *European Conference on Computer Vision*. Springer, 2020, pp. 51–67.

[39] P. Bongini, F. Becattini, A. D. Bagdanov, and A. Del Bimbo, "Visual question answering for cultural heritage," in *IOP Conference Series: Materials Science and Engineering*, vol. 949, no. 1. IOP Publishing, 2020, p. 012074.

[40] Y. Zhou, S. Mishra, M. Verma, N. Bhamidipati, and W. Wang, "Recommending themes for ad creative design via visual-linguistic representations," in *Proceedings of The Web Conference 2020*, 2020, pp. 2521–2527.

[41] A. S. Toor, H. Wechsler, and M. Nappi, "Biometric surveillance using visual question answering," *Pattern Recognition Letters*, vol. 126, pp. 111–118, 2019.

[42] B. Sreedha and P. R. Nair, "Multimodal visual question answering using vizwiz data; a visual assistant for the blind," in *International Conference on Electrical and Electronics Engineering*. Springer, 2022, pp. 365–372.

[43] P. Xu, T. M. Hospedales, Q. Yin, Y.-Z. Song, T. Xiang, and L. Wang, "Deep learning for free-hand sketch: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[44] S. Ren, Y. Du, J. Lv, G. Han, and S. He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13 325–13 333.

[45] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Sign language transformers: Joint end-to-end sign language recognition and translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 023–10 033.