MMBERT: Multimodal BERT Pretraining for Improved Medical VQA

Yash Khare[⋆†] *Viraj Bagal*[⋆‡] *Minesh Mathew*[†]
Adithi Devi[††] *U Deva Priyakumar*[†] *CV Jawahar*[†]

[†] IIIT Hyderabad, India [‡]IISER Pune, India [††]Osmania Medical College, India

ABSTRACT

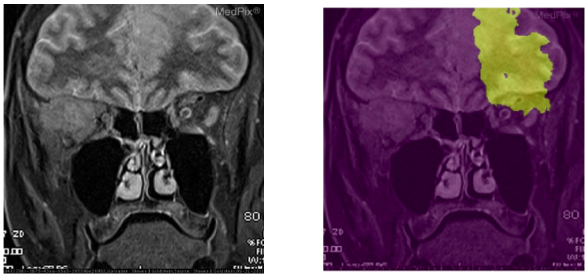
Images in the medical domain are fundamentally different from the general domain images. Consequently, it is infeasible to directly employ general domain Visual Question Answering (VQA) models for the medical domain. Additionally, medical image annotation is a costly and time-consuming process. To overcome these limitations, we propose a solution inspired by self-supervised pretraining of Transformer-style architectures for NLP, Vision, and Language tasks. Our method involves learning richer medical image and text semantic representations using Masked Vision-Language Modeling as the pretext task on a large medical image+caption dataset. The proposed solution achieves new state-of-the-art performance on two VQA datasets for radiology images – VQA-Med 2019 and VQA-RAD, outperforming even the ensemble models of previous best solutions. Moreover, our solution provides attention maps which help in model interpretability.

Index Terms— medical VQA, multimodal BERT, vision, and language

1. INTRODUCTION AND RELATED WORK

Visual Question Answering (VQA) on medical images aspires to build models that can answer diagnostically relevant natural language questions asked on medical images. It can provide valuable additional insights to medical professionals and can help the patients in the interpretation of their medical images. However, supervised learning algorithms require large labeled datasets for effective performance. A major drawback of VQA in the medical domain is the small size of existing datasets [1, 2]. Since the annotations on medical images require the help of an expert, it is difficult to crowdsource and annotation cost is high. This motivates the usage of self-supervised pretraining methods.

Self-supervised pretraining of BERT-like architectures has proven quite effective in Natural Language Processing (NLP) [3], Vision, and Language [4] space. The solution we propose - a Multimodal Medical BERT (MMBERT) is inspired by these approaches. We first pretrain our MMBERT model on a set of medical images and their corresponding captions with the Masked Vision-Language Modeling task. Later this model is finetuned for the VQA task.



Question: What imaging modality was used?
Answer: MR-T2 Weighted

Fig. 1: Example illustrating the attention map from our MMBERT model. For the given question, the model attends to grey matter, white matter, and cerebrospinal fluid (CSF) to predict the correct answer – 'MR-T2 Weighted'.

Yan et al. [5] who are the winners of the VQA-Med 2019 [2] challenge, use a Convolutional Neural Network (CNN) and BERT to extract image and question features respectively, followed by co-attention to fuse these features and a decoder to predict the answers. Ren et al. [6] propose a model called CG-MVQA that uses a multimodal transformer architecture, similar to the proposed MMBERT. Zhan et al.[7] use a conditional reasoning framework for medical VQA on VQA-RAD dataset. They train a model separately for both the open-ended and closed-ended questions in the dataset.

Although the prior methods obtain effective results, they do not use existing large multimodal medical datasets to learn better image and text representations. Our approach takes this into account and achieves better performance on two medical VQA datasets. Our MMBERT, with a single model for both the type of questions, yields better results than all the previous models on VQA-RAD[1]. It also achieves a 5% improvement in accuracy over the previous state-of-the-art model on VQA-Med 2019 dataset. Moreover, our model provides attention maps and as shown in Fig 1, it focuses on correct region (grey and white matter difference) to predict the modality of the image.

[⋆] Equal Contribution. Order decided by a coin toss. Work done during an internship at IIIT Hyderabad.

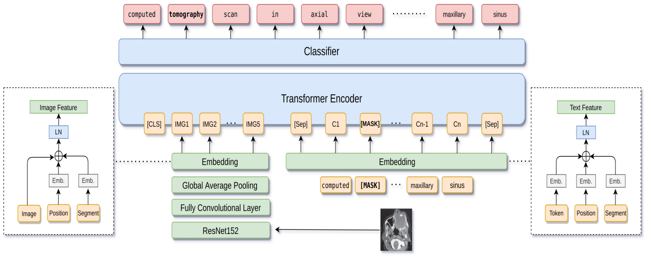


Fig. 2: Model architecture for Masked Vision-Language Modeling on image caption data. The image features are extracted from ResNet152. The caption is tokenized and the keywords are masked with [Mask] tokens. To distinguish, we use embeddings of 0 and 1 as segment embeddings for image and text features respectively. We use embeddings of enumeration of image and text features as position embeddings for image and text respectively. Features obtained by adding the input, position, and segment embeddings are passed as input to the model.

2. METHOD

Transfer learning is quite popular in machine learning. However, a shift in image data distribution might result in suboptimal performance when using pretrained weights from the general domain. Moreover, there are changes in cooccurences of words in the medical text compared to the general domain text. These factors motivate the need for learning semantic representations of medical images and texts from scratch. Owing to the attention operation, we use the Transformer encoder for learning effective representations.

2.1. Self-Attention

Self-Attention allows attention to intra-modality and intermodality features, thus enhancing the semantics of the intermediate representations. It involves mapping a query vector to the weighted addition of the value vectors where the weights are obtained by scaling the dot product of the query and the key vectors [8]. The query, key, and value vectors are represented together in the matrices Q, K, and V respectively. The dot product of Q and K is scaled inversely by $\sqrt{d_k}$, where d_k is the dimension of query and key vectors.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

Instead of performing a single self-attention, the Transformer Encoder performs multiple self-attentions (multi-head attention) in parallel and concatenates the output. Multi-head attention provides better representations by attending to different representation subspaces at different positions.

2.2. Pretraining and Finetuning

A schematic of the MMBERT pretraining is shown in Figure 2. For image features, similar to the CGMVQA [6] we use ResNet152 [10] and extract features from different convolution layers. This helps in retaining information from different resolutions. We use the BERT WordPiece tokenizer [3] for text tokenization. The sequence of 5 image features and the caption token embeddings are together provided as input to the BERT-like model. Unlike BERT$_{\text{BASE}}$ [3] our model has only 4 BERT layers and a total of 12 attention heads.

We use masked vision-language modeling as the pretraining task. In masked vision-language modeling, the task is to predict the original token in place of a [MASK] token with the usage of text and image features. To ensure that the model learns to predict medical words from the context, we mask only medical keywords (provided with the dataset) from the captions and leave the common words untouched.

We load the model with weights from pretraining and finetune it further on the train split of the respective medical VQA dataset. Instead of using [CLS] (Classification) token representation from the last layer of the Transformer, we average the representation of each token obtained from the last layer, and further pass it through dense layers for classification.

3. EXPERIMENTS AND RESULTS

3.1. Data and Experiments

Radiology Objects in COntext (ROCO) [11] dataset contains over 81,000 radiology images with several medical imaging modalities. For pretraining, we use all the images, their corresponding captions and use the keywords for masking. VQA-Med 2019 [2] is a challenge dataset introduced as part of the ImageCLEF-VQA Med 2019 challenge. It contains radiology images and has four main categories of questions: modality, plane, organ system and abnormality. All the samples having yes/no as the ground truth are considered as yes/no category. The dataset includes a training set of 3200 medical images with 12,792 Question-Answer (QA) pairs, a validation set of 500 medical images with 2000 QA pairs, and a test set of 500 medical images with 500 QA pairs. VQA-RAD has 315 images and 3515 questions of 11 types. 58% of questions are close-ended while the rest are open-ended.

In our study, we primarily experiment with three different settings for the MMBERT: (i) MMBERT General: a model pretrained on ROCO and finetuned on all samples in the train split of the respective VQA dataset. (ii) MMBERT Exclusive: an initial model pretrained on ROCO, which is further finetuned separately for different question categories. For example, in the case of VQA-Med 2019, we learn 5 different models, one for each question category. (iii) MMBERT Non-Pretrained (NP): Dedicated models for each question category but without pretraining on the ROCO dataset. At the time of the inference, for settings where there are dedicated models for each ques-

Method	Dedicated	Modality		Plane		Organ		Abnormality		Yes/No		Overall	
	Models	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU	Acc.	BLEU
VGG16+BERT [5]	-	-	-	-	-	-	-	-	-	-	-	62.4	64.4
CGMVQA [6]	✓	80.5	85.6	80.8	81.3	72.8	76.9	1.7	1.7	75.0	75.0	60.0	61.9
CGMVQA Ens. [6]	✓	81.9	**88.0**	**86.4**	**86.4**	**78.4**	79.7	4.40	7.60	78.1	78.1	64.0	65.9
MMBERT General	✗	77.7	81.8	82.4	82.9	73.6	76.6	5.20	6.70	85.9	85.9	62.4	64.2
MMBERT NP	✓	80.6	85.6	81.6	82.1	71.2	74.4	4.30	5.70	78.1	78.1	60.2	62.7
MMBERT Exclusive	✓	**83.3**	86.2	**86.4**	**86.4**	76.8	**80.7**	**14.0**	**16.0**	**87.5**	**87.5**	**67.2**	**69.0**

Table 1: Results of VQA-Med 2019 dataset. Our method outperforms all previous methods that include methods with ensemble models in overall accuracy and BLEU score. Ens. refers to ensemble models.

Method	Dedicated	Accuracy		
	Models	Open	Closed	Overall
MEVF+SAN [9]	-	40.7	74.1	60.8
MEVF+BAN [9]	-	43.9	75.1	62.7
CR [7]	✓	60.0	**79.3**	71.6
MMBERT General	✗	**63.1**	77.9	**72.0**

Table 2: Results of VQA-RAD dataset. Our method with single model for both open and closed-ended question types outperforms all previous methods including methods with dedicated models for each question type in overall accuracy.

tion category we first predict the question category using a $\text{BERT}_{\text{BASE}}$ classifier.

3.2. Results and Analysis

We use accuracy and BiLingual Evaluation Understudy (BLEU) score to evaluate the VQA performance. BLEU score is the percentage of uniformly weighted 4-grams in the predicted answer that are shared with the ground truth. Table 1 reports results on the VQA-Med 2019 dataset. Our MMBERT Exclusive achieves state-of-the-art results on the overall accuracy and BLEU score, even surpassing CGMVQA Ens. which is an ensemble of 3 dedicated models for each category. Even our MMBERT General performs better than the CGMVQA Ens. on the abnormality and yes/no categories. Additionally, our MMBERT General outperforms single dedicated CGMVQA models in all the categories but modality.

In the organ category, MMBERT Exclusive outperforms CGMVQA Ens. in BLEU but not in accuracy. BLEU score is calculated by counting matching 1-gram in the predicted answer to the 1-gram in the ground truth. The comparison is made regardless of the order. This suggests that even though our model could not perfectly predict right answers, it could predict answers closer to the ground truth than the CGMVQA Ens. We find the opposite behaviour in the modality category. When compared to MMBERT NP, we find that the pretraining increases the accuracy and BLEU score by 7.2 and 9 points respectively.

Table 2 reports results of the VQA-RAD dataset. MMBERT General, which is a single model for both the question types in the dataset, outperforms the existing approaches including the ones which have a dedicated model for each question type.

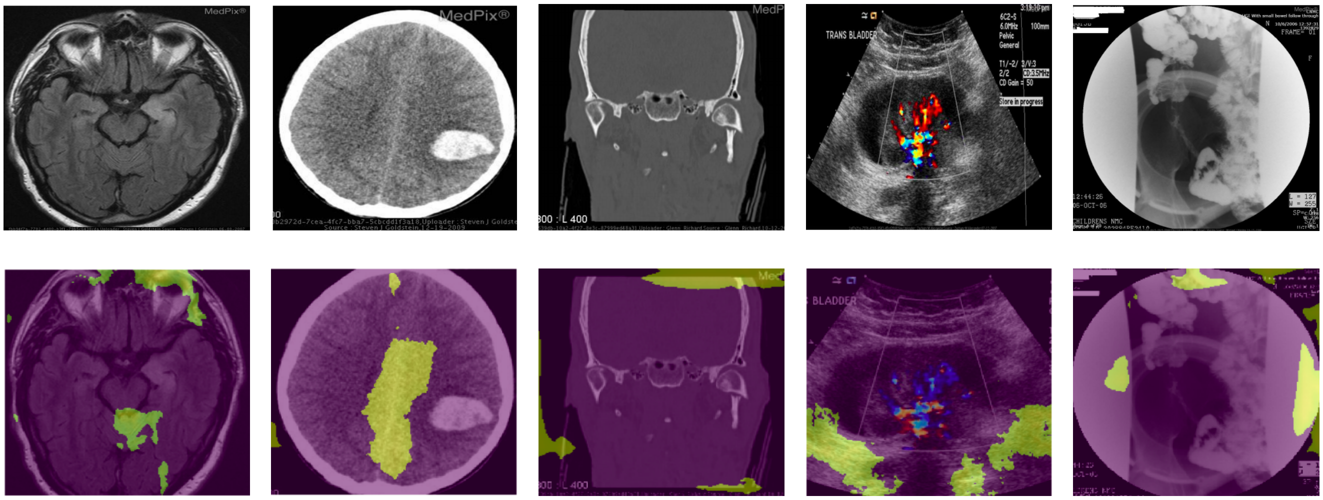
3.3. Qualitative Analysis

Fig. 3 shows the category-wise qualitative results from MMBERT Exclusive. The top row comprises the original images while the bottom row comprises the attention maps obtained from our model. The attention maps highlight the regions in the image which contribute the most to the prediction. In the organ and yes/no category, the model rightly attends to the skull (the bony part) and its contents (brain tissues) to predict the right answer. In the plane category, the model attends to the longitudinal fissure which is the key visual cue in identifying the axial plane as it separates the brain into right and left hemispheres. In the modality category, the model attends to the soft tissue and fluid part of the image and is able to correctly predict that it is an ultrasound image. However, it fails to attend to the visual cue of the Doppler (the colour region) and hence fails to correctly answer. In the abnormality category, our model predicts a better answer than the ground truth, for the attention map result of the ”Fluoroscopic evaluation of small bowel in Crohn’s ileitis”, it simultaneously predicts the modality (fluoroscopy), organ (bowel), and the abnormality (Crohn’s ileitis).

Medical experts find it difficult to make a correct diagnosis of abnormalities from a single image. They often resort to multiple sections (slices), planes, and other evidences. On closely analyzing our results we see that our model predicts abnormalities which could have also been a differential diagnosis for a human expert. However, our quantitative evaluation protocol does not take this into consideration.

4. CONCLUSION

In this work, we prospose to pretrain Multimodal Medical BERT (MMBERT) on ROCO dataset with masked language modeling using image features for medical VQA. We finetune it on VQA-RAD and VQA-Med 2019 datasets and achieve new progressive results on these datasets. Moreover, qualitative results show that our models can rightly attend to the image regions for prediction.



C: Organ
Q: What organ is the image of?
GT: skull and contents
ME: skull and contents

C: Plane
Q: What is the plane of the image?
GT: axial
ME: axial

C: Yes/No
Q: Is this an MRI image?
GT: no
ME: no

C: Modality
Q: What imaging method was used?
GT: us-d - doppler ultrasound
ME: us - ultrasound

C: Abnormality
Q: What is abnormal in the image?
GT: crohn's disease
ME: fluoroscopic evaluation of small bowel in crohn's ileitis

Fig. 3: ME and GT refers to MMBERT Exclusive & ground truth. The bottom row comprises attention maps for the corresponding top row images. In the Organ and Yes/No category, the model rightly attends to the bony part and soft tissue content to predict the right answer. In the Plane category, the model attends to the longitudinal fissure that is the key visual cue of the axial plane. The model fails to attend to the visual cue of the Doppler effect (colorful regions) in the Modality category. The Abnormality model surprisingly predicts a better answer than the ground truth by simultaneously predicting the modality, organ, and abnormality.

5. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by O. Pelka et al.[11], Asma Ben Abacha et al.[2] and Lau et al.[1]. Ethical approval was not required as confirmed by the license attached with the open access data.

6. ACKNOWLEDGMENTS

No funding was received for conducting this study. The authors have no relevant financial or non-financial interests to disclose.

7. REFERENCES

[1] Lau et al., "A dataset of clinically generated visual questions and answers about radiology images," *Scientific data*, 2018.

[2] Abacha et al., "VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019," in *CLEF 2019 Working Notes*, 2019, CEUR Workshop Proceedings.

[3] Devlin et al., "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[4] Yen-Chun et al., "UNITER: Universal image-text representation learning," in *ECCV*, 2020.

[5] Yan et al., "Zhejiang university at ImageCLEF 2019 visual question answering in the medical domain.," in *CLEF (Working Notes)*, 2019.

[6] Fuji Ren et al., "CGMVQA: A New Classification and Generative Model for Medical Visual Question Answering," in *IEEE Access*, 2020.

[7] Zhan et al., "Medical visual question answering via conditional reasoning," in *ACM*, 2020.

[8] Vaswani et al., "Attention is all you need," in *Neural Information Processing Systems (NIPS)*, 2017.

[9] Nguyen et al., "Overcoming data limitation in medical visual question answering," in *MICCAI*, 2019.

[10] He et al., "Deep residual learning for image recognition," in *CVPR*, 2016.

[11] Pelka et al., "Radiology objects in context (ROCO): A multimodal image dataset," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, 2018.