Text-Enhanced Medical Visual Question Answering

Chih-Ying Liu [1] Fan Diao [1]

[1]Stanford University

Stanford | ENGINEERING
Computer Science

Background/Introduction

Medical VQA is a task that requires comprehending both text-based queries and medical images to produce accurate answers. This project investigated the impact of **augmenting additional medical textual knowledge** and **various designs of fusion modules** in the medical VQA system.

Problem Statement

- **Input and Output** We formulate medical VQA as a classification task. The input to our algorithm is an image and a question. We then use a model consisting of a CLIP base model, a multimodal fusion module and a classifier, to output a predicted answer from a pre-defined set of answer candidates.
- **Evaluation metrics** We conducted our experiments on the VQA-RAD dataset with both **binary** and **multi-class** settings. For experiments using only closed-ended questions, we evaluate the closed-ended testing accuracy. For experiments using all questions, we evaluate the overall, closed-ended, and open-ended testing accuracy.

Data

VQA-RAD

- Medical VQA dataset.
- Questions are closed-ended if the answer is yes/no, and otherwise open-ended.
- The training set includes 458 answer candidates.

Close-ended

Question
Are the lungs normal appearing?
Answer
no

Open-ended

Question
What is the condition of the patient?
Answer
blind loop syndrome

Figure 1. Examples of the VQA-RAD dataset.

MedMCQA

- Medical multiple-choice QA dataset.
- Augment with extra medical textual knowledge.
- Each sample includes a question text, 4 options, and a correct option.

Question Characterstic X Ray finding in ASD is:
Option A Enlarged left ventricle
Option B Enlarged left atria
Option C Pulmonary pletheora
Option D PAH
Answer Option C

Figure 2. Examples of the MedMCQA dataset.

Data	Size (train/val/test)
VQA-RAD (all)	2681/383/408
VQA-RAD (closed)	1498/215/266
MedMCQA	183k/4183/6150

Table 1. Data size.

Method

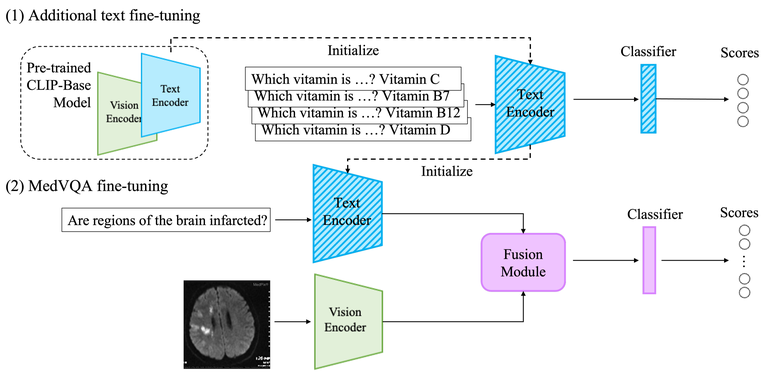


Figure 3. Overall framework.

Two-Stage Training

- **Stage 1: Additional Text Fine-tuning**: Fine-tune the text encoder on the MedMCQA dataset.
- **Stage 2: MedVQA Fine-tuning**: Fine-tune the whole system on the VQA-RAD dataset. Three components: (1) CLIP base model, (2) multi-modal fusion module, and (3) classifier.

Model Design

- **With/without text enhancement**
- **Base CLIP Model**: CLIP, PubMedCLIP
- **Multi-modal fusion module**

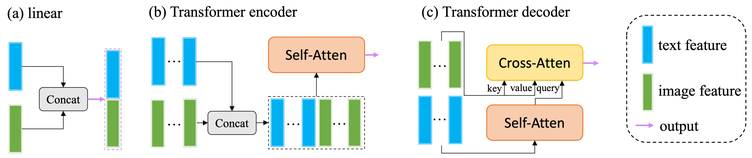


Figure 4. Multi-modal fusion module

PMC-CLIP: In addition to combining pre-trained CLIP and PubMedCLIP models with fusion modules, we also adapted PMC-CLIP [1], which features a fusion module jointly pre-trained with visual and text encoders on large-scale image-text pairs.

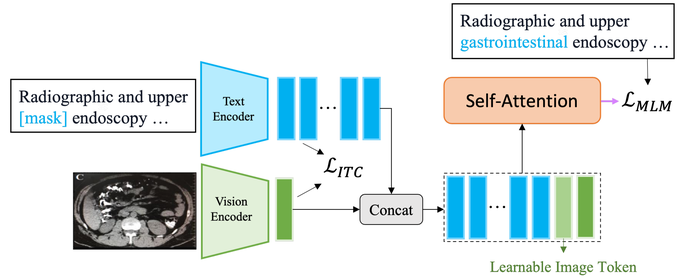


Figure 5. The architecture of PMC-CLIP

Binary Classification Loss

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(s_i) + (1-y_i)\log(1-s_i)]$$

Multi-class Classification Loss

$$\mathcal{L} = -\sum_{i=1}^{N}\log\left(\frac{e^{\mathbf{s}_i}}{\sum_j e^{\mathbf{s}_j}}\right)$$

Result/Analysis

Performance

Model	Method: Fusion	Method: Text++	VQA-RAD (%): Binary Closed	VQA-RAD (%): Multi-class Overall	VQA-RAD (%): Multi-class Closed	VQA-RAD (%): Multi-class Open
CLIP	Linear	✗	76.5	45.8	75.5	0.0
		✓	77.7	44.9	72.9	0.0
	TransEnc	✗	77.3	**66.7**	78.5	47.8
		✓	80.9	64.5	78.9	41.4
	TransDec	✗	80.5	61.3	75.7	37.6
		✓	**81.7**	62.8	77.7	40.1
PubMed CLIP	Linear	✗	75.3	44.6	72.5	0.0
		✓	76.9	44.4	72.1	0.0
	TransEnc	✗	78.8	65.4	72.3	**48.4**
		✓	78.8	65.2	77.7	45.9
	TransDec	✗	80.5	65.7	**80.1**	42.7
		✓	78.1	63.7	78.9	40.1
PMC -CLIP	Pre-Trans	✗	80.5	58.1	78.1	26.1
		✓	76.5	55.6	76.9	21.7

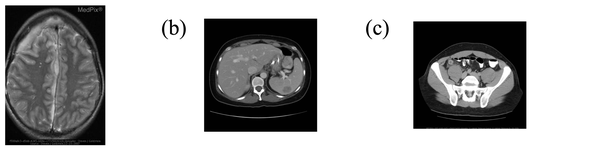
Table 2. Accuracy on the VQA-RAD dataset.

Base Model	MedMCQA (%)
CLIP	36.6
PubMedCLIP	37.0
PMC-CLIP	37.5

Table 3. Accuracy on the MedMCQA dataset.

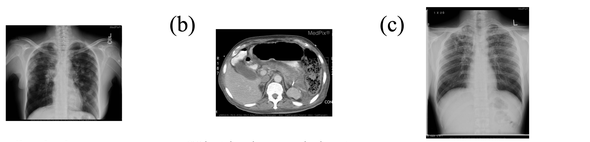
- Achieves the highest accuracy: 81.7% in the binary setting and 66.7% in the multi-class setting.
- Text enhancement improves accuracy in the binary setting.
- Transformer-based fusion methods outperform linear fusion, particularly benefiting open-ended questions.

Qualitative Result



	(a)	(b)	(c)
Question	Can fluids be highlighted with this modality?	How many lesions are in the spleen?	Where is the ascending colon?
Answer	yes	one	posterior to the appendix
CLIP + Linear	no ✗	no ✗	no ✗
CLIP + TransEnc	yes ✓	one ✓	posterior to the appendix ✓
CLIP + TransDec	yes ✓	one ✓	posterior to the appendix ✓
PMC-CLIP	yes ✓	one ✓	axial ✗

Figure 6. Examples from the VQA-RAD dataset.



	(a)	(b)	(c)
Question	Describe the lung abnormalities?	What is abnormal about the pancreas?	What kind of image is this?
Answer	pulmonary nodules	enlarged	x-ray
CLIP + Linear	hemidiaphragm ✗	no ✗	yes ✗
CLIP + TransEnc	bilateral ✗	fatty infiltration ✗	chest x-ray ✗
CLIP + TransDec	yes ✗	fatty infiltration ✗	chest x-ray ✗
PMC-CLIP	choroid plexus ✗	pancreas ✗	no ✗

Figure 7. Examples from the VQA-RAD dataset that all models fail.

Problem of Skewed Training Labels

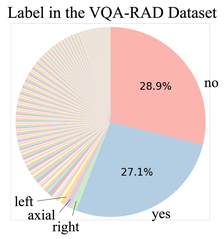


Figure 8. Label distribution of the VQA-RAD dataset.

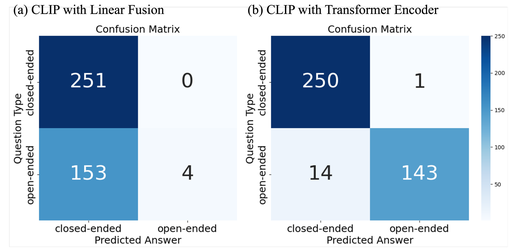


Figure 9. Answer distribution against close/open-ended questions.

Conclusion/Future Work

- Transformer-based fusion modules outperform linear fusion.
- Text enhancement is beneficial in binary classification tasks and when the training dataset size is small.
- Solutions to address the imbalanced label problem: weighted loss function, sampling strategy based on label frequency. Another direction is shifting to generative methods.

Reference

[1] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.