Vision–Language Model for Visual Question Answering in Medical Imagery

Yakoub Bazi 1,* , Mohamad Mahmoud Al Rahhal 2 , Laila Bashmal 1 and Mansour Zuair 1
1 Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
2 Applied Computer Science Department, College of Applied Computer Science, King Saud University, Riyadh 11543, Saudi Arabia
Correspondence: ybazi@ksu.edu.sa; Tel.: +966-114696297

DOI: 10.3390/bioengineering10030380
Published: 20 March 2023

bioengineering

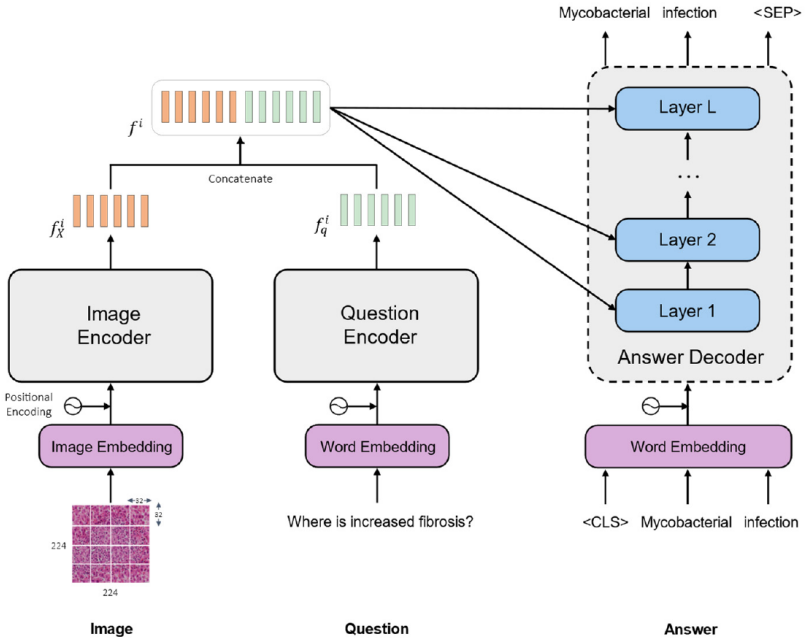
PMC PubMed Central®

Why?

- Medical images are crucial in diagnosis.
- Doctors often ask questions like: “Is there fluid in the lungs?”
- Automating this with AI can save time.

How does it work?

Takes a medical image and a text question as input
Extracts and fuses visual and language features
Generates a short, clinically relevant answer



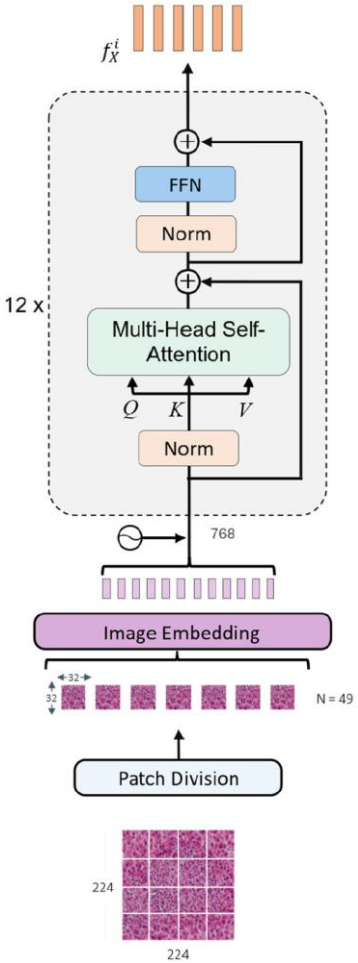
- Two types: Closed-ended and Open-ended

Related work

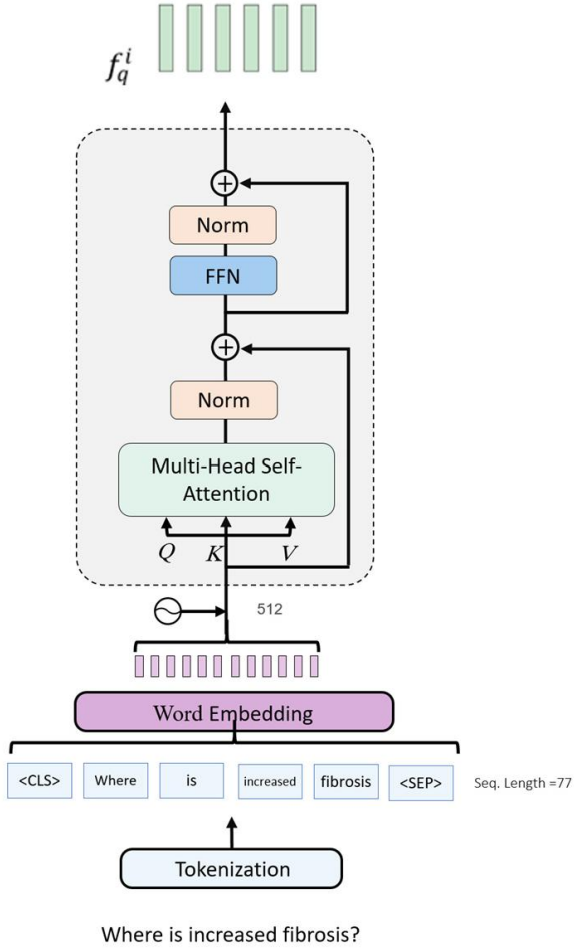
- Conditional reasoning (Zhan et al. 2020)
- CMSA (cross-modal self-attention) (Gong et al. 2021)
- Contrastive learning + teacher-student models (Do et al. 2022)

Limitations

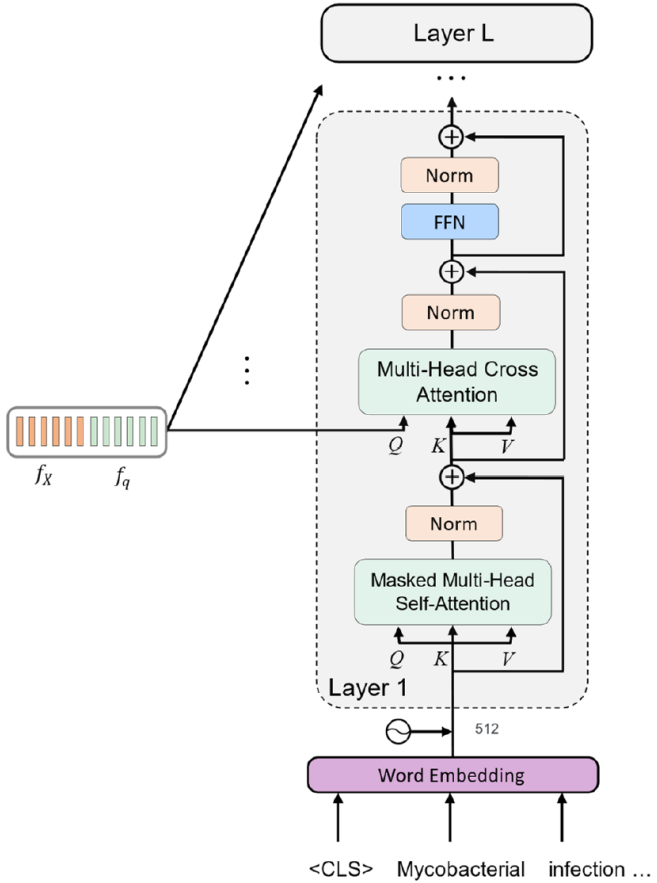
❌ Most used classifiers, not generative decoders → couldn't generate open-ended answers
❌ Shallow fusion of image and text (e.g., MLPs or attention pooling) → weak multimodal understanding
❌ Often relied on CNNs + RNNs, which are less effective than transformers
❌ Poor performance on open-ended questions (e.g., "Where is the abnormality?")



The architecture of the image encoder



The architecture of the text encoder

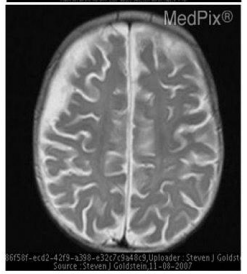


The architecture of the answer decoder

Datasets

Questions	Answers
What modality is used to take this image?	XR
Are the costophrenic angles blunted?	NO
Is there any blunting of the costophrenic angle(s)?	NO
Do you see cardiomegaly?	NO
Is there cardiomegaly present?	NO

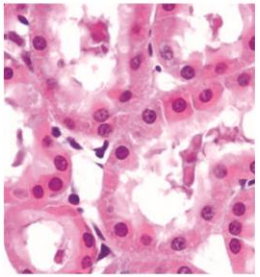
Questions	Answers
What organ system is evaluated primarily?	GI
What kind of scan is this?	CT
What does nodular liver suggest?	Cirrhosis
What causes hyper intensity in aorta?	Atherosclerotic calcification
Is the aorta size abnormal?	NO



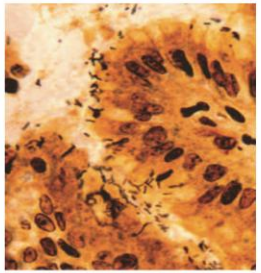
Questions	Answers
Is the cerebellum visible?	NO
Is this a MRI image?	YES
In which lobe is the enhancement?	Right frontal lobe
Are there fractures on the skull?	NO
What is the plane of this image?	Axial

VQA-RAD with radiology images and 3500+ questions

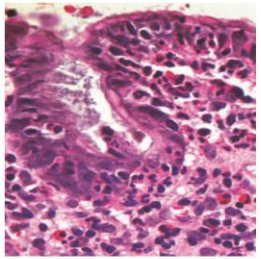
Datasets



Questions	Answers
What does ischemic injury show?	surface blebs
Does early ischemic injury show surface blebs, increase eosinophilia of cytoplasm, and swelling of occasional cells?	Yes
'What is showing increased eosinophilia of cytoplasm, and swelling of occasional cells?	early (reversible) ischemic injury
Did early (reversible) ischemic injury increase eosinophilia of cytoplasm, and swelling of occasional cells?	NO



Questions	Answers
What are abundant within surface mucus?	organisms
What are organisms abundant within?	surface mucus
Are organisms abundant within surface mucus?	YES
Are iron deposits shown by a special staining process abundant within surface mucus	NO



Questions	Answers
What are prominent?	intraepithelial and lamina propria neutrophils
Are intraepithelial and lamina propria neutrophils prominent?	YES
Are histologic appearance in active takayasu aortitis illustrating destruction and fibrosis of the arterial media prominent?	NO

PathVQA with over 32,000 Q&A pairs related to pathology images

Experimental setup

- CLIP encoders frozen (ViT + BERT are not updated during training)
- Only the decoder is trained (autoregressive transformer)

- Decoder layers: tested with 1–4 layers
- Best: 4 decoder layers
- Embedding dimension: 512
- Max sequence length: 77 tokens
- Dropout: 0.1 in all fully connected layers
- Tokenizer: BERT-base-uncased (vocab size = 49,408)

- Optimizer: Adam
- Learning rate: 0.001
- Batch size: 50
- Epochs: 50
- Data augmentation: Random horizontal flip with probability = 0.2
- Image shuffling at each epoch

Results

Dataset	Evaluation Metric						
	B1	B2	B3	B4	Closed	Open	All
VQA-RAD	71.03 ± 0.90	70.81 ± 0.95	67.01 ± 0.99	64.43 ± 0.99	82.47 ± 1.43	71.49 ± 0.73	75.41 ± 0.98
PathVQA	61.78 ± 0.03	61.16 ± 0.04	59.28 ± 0.02	58.19 ± 0.02	84.63 ± 0.83	58.29 ± 0.73	67.05 ± 0.58

BLEU-1, 2, 3, 4, closed-ended, open-ended, and overall accuracies

#Decoder Layers	VQA-RAD Dataset			PathVQA Dataset		
	Closed	Open	All	Closed	Open	All
1	82.06 ± 1.21	70.98 ± 0.39	74.99 ± 0.85	83.13 ± 0.49	57.90 ± 0.49	66.17 ± 0.75
2	82.47 ± 1.43	71.49 ± 0.73	75.41 ± 0.98	84.63 ± 0.83	58.29 ± 0.73	67.05 ± 0.58
3	84.46 ± 1.43	72.51 ± 0.99	76.78 ± 1.09	86.90 ± 0.21	62.58 ± 0.04	70.65 ± 0.07
4	84.99 ± 1.15	72.97 ± 1.49	77.27 ± 1.37	86.83 ± 0.31	62.37 ± 0.22	70.54 ± 0.25

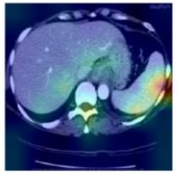
Sensitivity analysis with respect to the number of decoder layers.

Method	Evaluation Metric			
	BLEU-1	BLEU-2	BLEU-3	F1 (%)
GRU + Faster R-CNN [40]	32.4	22.8	17.4	24.0
CNN + LSTM [40]	13.3	9.5	6.8	12.5
SAN+ CNN + LSTM [40]	19.2	17.9	15.8	19.7
SAN+ CNN + LSTM+ Faster R-CNN [40]	24.7	19.1	16.5	21.2
SAN+ CNN + LSTM+ ResNet [40]	19.9	18.0	16.0	19.8
Proposed	71.03 ± 0.90	70.81 ± 0.95	67.01 ± 0.99	72.85 ± 0.95

BLEU-1, 2, 3 and F1 on open-ended questions (with the main work that introduces the PathVQA dataset).

Method	VQA-RAD Dataset			PathVQA Dataset		
	Closed	Open	All	Closed	Open	All
Zhan, L.M. et al. [6]	79.3	60.0	68.5	-	-	-
Nguyen, B.D. et al. [7]	75.1	43.9	62.6	81.4	8.1	44.8
Gong, H. et al. [45]	77.8	52.8	67.9	-	-	-
Do, T. et al. [43]	72.4	52.0	64.3	82.1 ± 0.5	11.8 ± 0.6	47.1 ± 0.4
Gong, H. et al. [64]	69.7	38.2	57.1	75.3	5.4	40.5
Gong, H. et al. [64]	72.4	49.6	63.3	81.3 ± 0.3	9.1 ± 0.5	45.3 ± 0.4
Gong, H. et al. [64]	79.6	56.6	70.4	83.5 ± 0.2	13.4 ± 0.6	48.6 ± 0.3
Moon, J.H. [65]	77.7 ± 0.71	59.5 ± 0.32	-	-	-	-
Proposed	82.47 ± 1.43	71.49 ± 0.73	75.41 ± 0.98	85.61 ± 0.83	71.49 ± 0.73	66.68 ± 0.58

Comparison with state-of-the-art methods

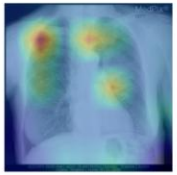


Question: was the arterial contrast phase selected

Answer
True: yes
Predicted: yes

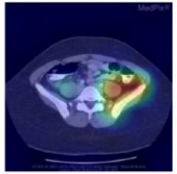
Question: Are the branches of the superior mesenteric arteries filled with contrast

Answer
True: yes
Predicted: yes



Question: Are the lungs normal appearing

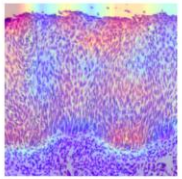
Answer
True: yes
Predicted: yes



Is/Are there air in the patient's peritoneal cavity

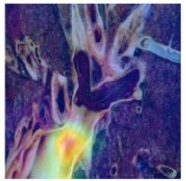
Answer
True: yes
Predicted: No

Attention maps obtained for sample images from RAD-VQA dataset



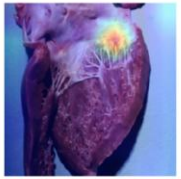
Question: Do interstitial fibrosis and tubular atrophy show that the entire thickness of the epithelium is replaced by atypical dysplastic cells

Answer
True: yes
Predicted: yes



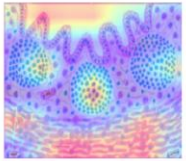
Question: Is embolus derived from a lower-extremity deep venous thrombus lodged in a pulmonary artery branch

Answer
True: yes
Predicted: no



Question: Does acute lymphocytic leukemia show normal mitral valve

Answer
True: yes
Predicted: yes



Question: Are the lumen present necrosis of mucosa and periappendicitis

Answer
True: parathyroid adenoma
Predicted: excellent example of perineural invasion typical for lesion

Attention maps obtained for sample images from PathVQA dataset

Main Contributions

- Introduced the first full transformer encoder–decoder model for medical VQA
- Designed a simple but effective fusion strategy using concatenated ViT + BERT outputs, enabling autoregressive generation
- The model handles both closed and open-ended questions with strong BLEU and accuracy scores
- Outperforms existing methods on VQA-RAD and PathVQA across all metrics, especially in open-ended answer generation

Limitations

- Struggles with very complex, long open-ended questions requiring deep reasoning
- Performance still limited by the small size of medical VQA datasets

Thanks

Any questions?