Baghdad Science Journal

Volume 23 | Issue 4 Article 11

4-23-2026

Cross-Attention Mechanism for Medical Visual Question Answering

Nada Fadhil Mohammed
College of Information Technology, University of Babylon, Babylon, Iraq,
nada.mohammed@uobabylon.edu.iq

Israa H. Ali
College of Information Technology, University of Babylon, Babylon, Iraq, sraa_hadi@itnet.uobabylon.edu.iq

How to Cite this Article

Mohammed, Nada Fadhil and Ali, Israa H. (2026) "Cross-Attention Mechanism for Medical Visual Question Answering," *Baghdad Science Journal*: Vol. 23: Iss. 4, Article 11.
DOI: https://doi.org/10.21123/2411-7986.5267

BAGHDAD SCIENCE JOURNAL 2026;23(4):1270–1281

Scan the QR to view the full-text article on the journal website

RESEARCH ARTICLE

Cross-Attention Mechanism for Medical Visual Question Answering

Nada Fadhil Mohammed, Israa H. Ali

College of Information Technology, University of Babylon, Babylon, Iraq

ABSTRACT

Visual Question Answering (VQA) is a machine learning task that aims to create systems capable of answering natural language questions based on given images. Medical VQA systems, a domain-specific application of VQA, assist in understanding clinically relevant information from medical images. These systems leverage deep neural techniques to generate accurate answers to questions, which can be closed-ended or open-ended. This paper proposes a Cross-Attention Mechanism-based Medical VQA system. The proposed Medical VQA system achieves good performance through the integration of three-key components, each addressing critical challenges in medical visual question answering: The biomedical domain-specific pretraining of BioBERT enables extracting powerful contextual features from questions. It is highly effective in handling rare medical terminology, abbreviations and suitable for processing and grammatically complex medical questions compared to generic language models. Denoising Autoencoder model for visual features extraction enables extracting strong visual features and focusing on small object through slicing image into overlapping patches, thereby improving the localization of abnormalities within image. Finally propose a cross-attention mechanism which applied to model the hidden relationship between medical image and question and enhance the fused features vector. The attention mechanism comprises intramodal (within-modality) and intermodal (cross-modality) components, enabling the model to focus on relevant parts of image and question for answer generation. Experiments conducted on the VQA-RAD and Med-VQA 2019 datasets demonstrate that the proposed system achieves good results and accuracies was 76.5% and 78.3%, respectively, outperforming baseline models that use traditional attention mechanisms like the Bilinear attention networks (BAN) or Stacked Attention Networks (SAN).

Keywords: BAN, BioBERT, Computer vision, Cross-attention mechanism, DAE, Medical VQA, Natural language processing

Introduction

Visual Question Answering (VQA) is an advanced deep learning application that bridges computer vision and natural language processing (NLP). A VQA system analyzes visual content, images, or videos, and generates accurate responses to user questions posed in natural language. These queries can vary from open-ended and multiple-choice questions to common-sense reasoning tasks, making VQA highly versatile. By integrating visual and linguistic data, VQA enables a broad spectrum of real-world applications. [1,2]

VQA has a wide range of applications. One important application was in the medical domain. Medical VQA assists specialists by providing quick, accurate answers to questions about medical images, reducing workload, and supporting diagnosis decisions. [3] Also, patients sometimes use search engines to check and understand their condition, and this increases the risk of getting incorrect information. So, there is an increased need for a system that can help patient to understand their medical image. Medical VQA could fulfill this need since it takes a medical image and the user's natural language question about the image and returns an answer to that question.

Received 2 May 2025; revised 12 July 2025; accepted 29 July 2025.
Available online 23 April 2026

* Corresponding author.
E-mail addresses: nada.mohammed@uobabylon.edu.iq (N. F. Mohammed), sraa_hadi@itnet.uobabylon.edu.iq (I. H. Ali).

https://doi.org/10.21123/2411-7986.5267

Recently, usage of deep learning in medical imaging has made great progress, especially in the broad research of algorithms for computer-aided diagnosis in diagnosis-related tasks, including image recognition, image classification, and so on. [4,5]

Medical VQA combines computer vision and natural language processing (NLP) to answer questions about radiology images, histopathology slides, and other medical imaging modalities. Medical VQA poses challenges due to the complexity of medical data, scarcity of annotated datasets, complex medical terminology requiring domain-specific NLP models, and Fine-grained localization of abnormalities (e.g., tumors, fractures). [3,5]

The medical VQA system consists of two phases: encoder and decoder. During the encoder phase, generating effective representations of medical images and clinical questions, and passing these inputs through a deep model to produce a co-dependent embedding vector. Decoder phase involves taking the fused features vector and then generating a response for the input question. [6] Visual features extraction is mostly done using one of the Convolution Neural network (CNN) pretrained models, such as VGG, [7] ResNet, [8] etc. For textual understanding, language models such as Recurrent Neural Network (RNN) architectures: Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and transformer-based language models (Bidirectional Encoder Representation (BERT) have demonstrated exceptional capability in capturing clinical semantics and context. After feature extraction is done, the fusion step is performed, which is regarded as the core component of the Medical VQA system. Different methods are used for feature fusion, such as concatenation, pointwise multiplication, Stack attention mechanism, [9] and Bilinear Attention Network. [10] The typical fusion method involves fusion with attention.

The primary objective of this paper is to design a reliable Medical VQA system capable of providing a 'second opinion' on medical case assessments. The proposed system incorporates advanced techniques for image and question feature extraction, enabling the generation of precise answers for clinical questions.

This paper proposes a medical VQA system for answering questions about medical images. BioBERT, a domain-specific pre-trained language model, is used for textual feature extraction, with a Medical DAE for image feature extraction. A cross-attention mechanism is proposed for feature fusion to model the hidden relationship between the question and the medical image,and pay different attention weights to different parts of the image to determine which part of the image is related to the input question words. With the help of the proposed attention mechanism, the model enables the system to provide reasonable answers while assisting clinicians in getting diagnostic information. The fused features are fed to the classifier to generate answers. Also, this paper delves into the critical aspects of medical VQA, exploring its importance, underlying mechanisms, challenges, limitations, and future directions.

The contributions of this study can be listed as follows:

- The proposed medical Visual question answering system gives better performance, comparable to several baselines, and across two datasets.
- The proposed medical VQA system adopts an effective method that extracts strong and important visual features from medical images and focuses on small objects through slicing the image into overlapping patches, thereby improving the localization of abnormalities within the image.
- Using BioBERT model for medical text encoding: unlike general language models, BioBERT is a biomedical language model that improves question understanding. It is highly effective in handling rare medical terms and abbreviations and is suitable for processing grammatically complex medical questions. The built-in attention layers inside BioBERT allow the model to automatically identify and focus on the most important parts of the questions.
- Propose a cross-attention mechanism to effectively fuse features. This model is designed to model the relationship. within the same modality and find the relationship between textual features and visual features. The fused vector captures the interaction between image and question features, thereby enhancing the accuracy of the answers in the medical domain.

Related works

Many studies have tried to solve the problem of medical visual question answering. Some of these studies will be explained here: In, [11] a proposed model that combines an image encoder based on EfficientNetV2 with a multimodal encoder based on the RealFormer architecture, Transformer is used for visual features and the question fusion. The source [12] proposed a "semantic-enhanced graph transformer-based textual representation learning approach, called SemTGT. They used EfficientNet and BiLSTM pre-training models as feature extractors to deal with medical images and questions, respectively. Concatenation is used for feature fusion. Next, they use a classifier to get answers to the questions, and the

results were 0. 42 of BLUE, and accuracy was 0.391 on the VQA-Med 2019 dataset. Paper[13] presented a VQA system that allows an image to be asked a written question. For image feature extraction, they use ResNet-152 and Skip-thought vectors to extract question features. They introduced a fusion mechanism called Question-Centric Cross Low-rank Bilinear (QCMLB), which fuses image and question features by enforcing high adherence to the query sentence. This model established results: 0.603 in accuracy in the Image CLEF 2019 VQA-Med dataset.

In,[14] they proposed a model called New Classification and Generative Model for Medical VQA, by turning this problem into multiple sub-problems. To extract image features, they used the ResNet152 network and the BERT model to deal with questions. They used a concatenation for feature fusion. This model established results: 0.640 of classification accuracy and 0.659 of word matching for the VQA-Med 2019 dataset. Paper[15] proposed a conditional reasoning mechanism with a question-conditioned reasoning component and a type- conditioned reasoning strategy to learn effective reasoning skills for different Med-VQA tasks adaptively. Further, they used a pre-trained visual feature extractor for Med-VQA via contrastive learning on large amounts of unlabeled radiology images. They evaluate the performance of the proposed model on two benchmark datasets, VQA-RAD and SLAK. The use of ResNet-50 for visual features extraction, and the question embedding is generated by a GRU model. Question-Conditioned Reasoning (QCR) is used to enable the model to gain question-specific reasoning skills by leveraging question attention information to modulate cross-fusion features. They train a model on the VQA-RAD dataset. Overall accuracy is 72.5 (60 for open-ended and 79 for closed-ended).

In,[2] a model was proposed based on ResNet and BERT models with attention modules to focus on the relevant part of the medical images and questions. The model predicts the answer either by a classification or a generation head, depending on the type of question. They performed the qualitative analysis of MedFuseNet and compared its results to the ones from SAN and Hierarchical Co-Attention models BLEU score was 0.276. Researcher in,[16] Hierarchical Question Segregation based Visual Question Answering (HQS-VQA). They proposed a question segregation (QS) technique for VQAMed, integrated the QS model into the hierarchical deep multi-modal neural network to generate proper answers to the queries related to medical images, and they studied the impact of QS in Medical-VQA by comparing the performance of the proposed model with QS and a model without QS. InceptionResnet-v2 is used for image processing, and Bi-LSTM for question processing, and the Concatenation is used for image and question feature fusion, and their system records 0.411 BLEU.

Materials and methods

In this section, the proposed system shown in Fig. 1 is presented. The proposed system consists of several models for the following main stages: visual features extraction, textual features extraction, features fusion with an attention mechanism, and finally answer generation.

Data description

In order to train and evaluate the model, the publicly available medical VQA datasets are used: VQA-RAD and Med-VQA 2019. These datasets contain medical images paired with corresponding question-answer pairs. The following is a brief description of these:

The VQA-RAD[17] contains 3,515 questions about 315 images, where more than one question is asked about each. The questions about “Abnormality,” “Attribute,” “Color,” “Count,” “Modality,” “Organ,” “Other,” “Plane,” “Positional reasoning,” “Object/Condition Presence,” “Size”. The questions are of two types: closed-ended (“yes/no”) and open-ended questions. Med-VQA 2019[11] is a dataset that contains radiology images, and the categories of questions are: Modality, Plane, Organ system, and Abnormality. Fig. 2 shows an example from the VQA-RAD and Med-VQA 2019 dataset; each example from the dataset consists of a medical image, a question about this image, and the answer to the question.

Question preprocessing

Involves the conversion of question text into a suitable format that aligns with the model’s requirements. For the preprocessing of the questions, all questions and answers are converted into lowercase letters. Once the text is purified, the tokenization process is performed by splitting it into individual words and eliminating punctuation and symbols. In order to pad the question tokens to ensure that the input sequences are of the same length, the distribution of the question lengths is found. Questions have a different length, so they are padded with zero to make each encoded array have the same length before being passed as an input to the embedding layer. In the VQA-RAD dataset, each question length is set to 12 words and uses zero padding for lengths less than 12.

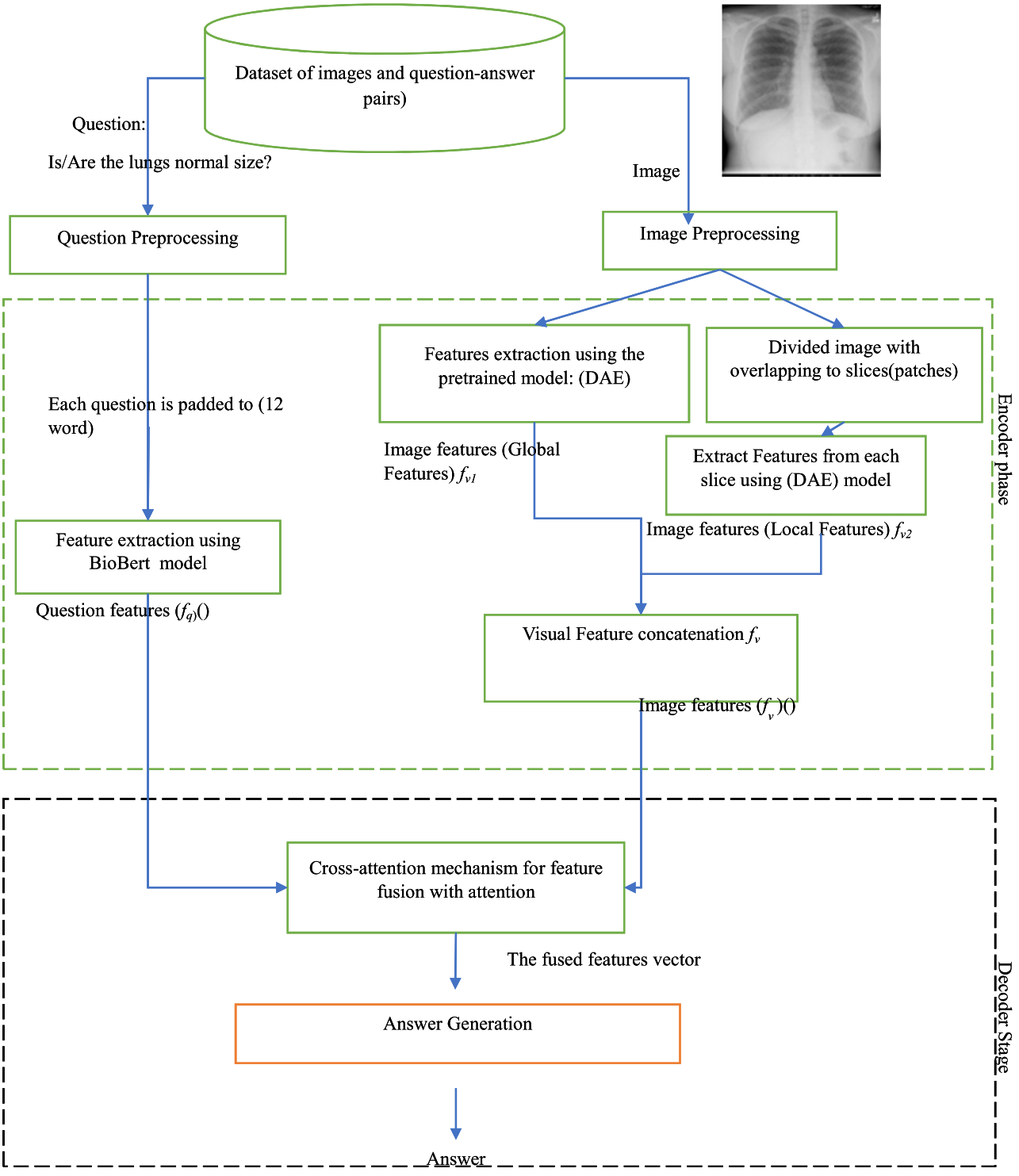


Fig. 1. The proposed medical question answering system.

Image preprocessing

The dataset used consists of different types of images and different formats of questions. The data is distributed over the different categories. The image sizes vary a lot; all images are resized to 224*224 (resize was performed while preserving aspect ratio).

Image feature extraction using denoising auto encoder model (DAE)

DAE is an unsupervised CNN model that is used for denoising medical images since medical images may contain various degrees of noise.[18] CNN-based approaches have achieved state-of-the-art results

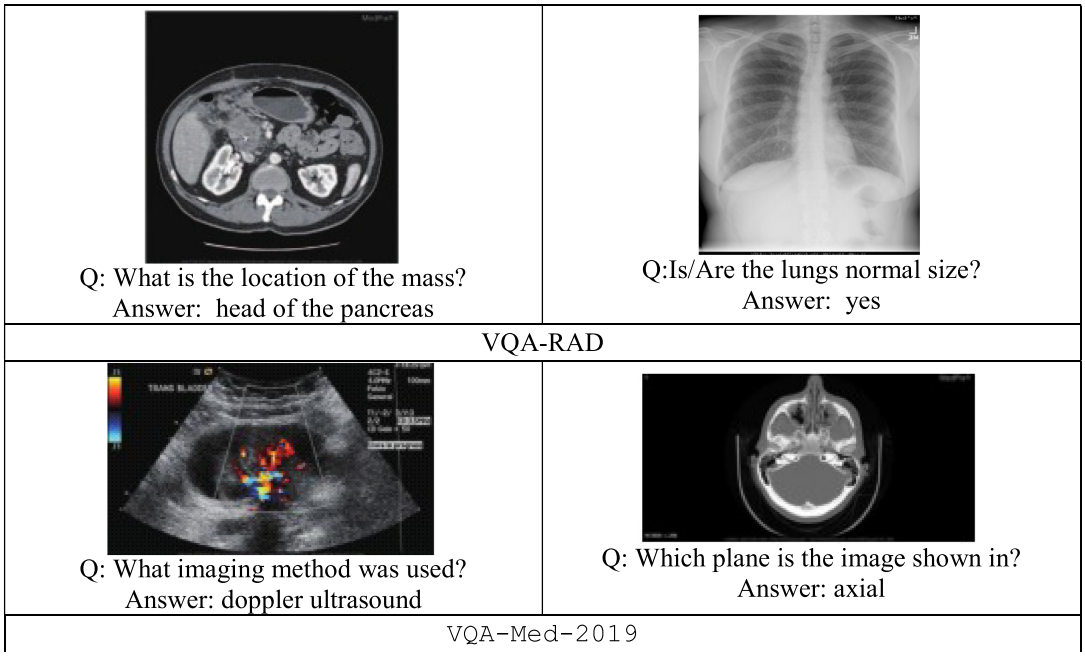


Fig. 2. Examples from the VQA-RAD dataset and VQA-Med 2019 dataset.[11,17]

on many VQA benchmarks. Also, the lack of labeled medical images makes the training process inefficient. Autoencoder helps to extract high-level features without labeled data, DAE. The DAE model can be used to extract features since it can extract robust features from a noisy image.[18] The encoder maps a noisy input image I', which is the noisy copy of the original image I, to a representation of latent w that keeps a useful number of features of the image. The training of the model aims to reduce the error of the noisy and the original images. The DAE trained on medical images to remove the added noise, as in paper.[19]

Image feature extraction describes an image as a numerical vector, and question feature describes text as a numerical vector. Image features are obtained through the pretrained convolutional DAE. A medical dataset involves various image types and organs. The characteristic of medical images is not effective to directly apply deep models; also, the small scale of existing medical datasets for VQA, fine-tuning pre-trained large models on them can lead to overfitting. To get rid of the lack of labeled data, the DAE model is used, which can be useful in extracting image features without any labeled data. This paper uses the DAE for visual feature extraction. Two types of image features will be extracted and used:

- Global features from the whole images using the Denoising encoder model.
- Local features from different regions of images are extracted by dividing the image into overlapping patches, and the extracted features are then used to train a Denoising autoencoder model.

Global features from the whole image and local features from different regions of images, and then these two types of features are combined. Very fine information can be extracted from each part of the image by dividing the image into overlapping patches with dimensions M × N. This is very useful in cases where attention to small details is required, such as fine objects in medical images. Running DAE on each patch to extract features from the overlapping patches. These features are merged using Maxpooling, that combine the features of patches while avoiding redundancy in order to eliminate redundancy. Maxpooling ensures that only the strongest features from overlapping regions are retained. At the same time, the full-image features give a comprehensive view that helps determine the overall context. In this way, the features of the parts and the features of the whole image can be combined to improve the accuracy of the model. This is very beneficial in the localization of abnormalities and lesions in medical images for early disease diagnosis and treatment planning. Local features are used to extract features of objects that are represented by a small number of pixels in the image and lack sufficient details, making them difficult to detect using conventional detectors. The following steps are needed:

1. Use the DAE to extract features from the full image, fv1.
2. Slice the images into smaller patches. The used patch size is (80 × 80) pixels with a stride of 48 pixels (for overlapping patches), so there will be 4 × 4 patches.
3. Extract features fv2 from the smaller patches using a DAE model. Then, combine the features of patches using maxpooling, these local features of the image.
4. Combine the features from patches and the full image to get visual features, fv.

Question semantic encoder

feature extraction from text involves the conversion of specific text into features. This process yields numerical vectors, so they are commonly referred to as vectorization. These extracted features from the text are then input into the prediction model to facilitate text classification. The pre-trained models, such as GRU,[20] LSTM[20] model, and BERT[21] are used for forward embedding to represent words of text as vectors.

This paper proposes a question feature extraction based on the bidirectional encoder representation from transformers (BERT) to extract features of the questions. The version of the BERT model used in this paper is BioBERT,[22] which is a domain-specific language representation model pre-trained on large-scale biomedical corpora. BioBERT can recognize biomedical named entities that BERT cannot and can find the exact boundaries of named entities. While BERT often gives incorrect answers to simple biomedical questions, BioBERT provides correct answers to such questions.[22] The used model is a biobert-v1.1 version of BioBERT, which includes 12 layers and 768 hidden variables. Each question will be represented by a 768-d feature vector.

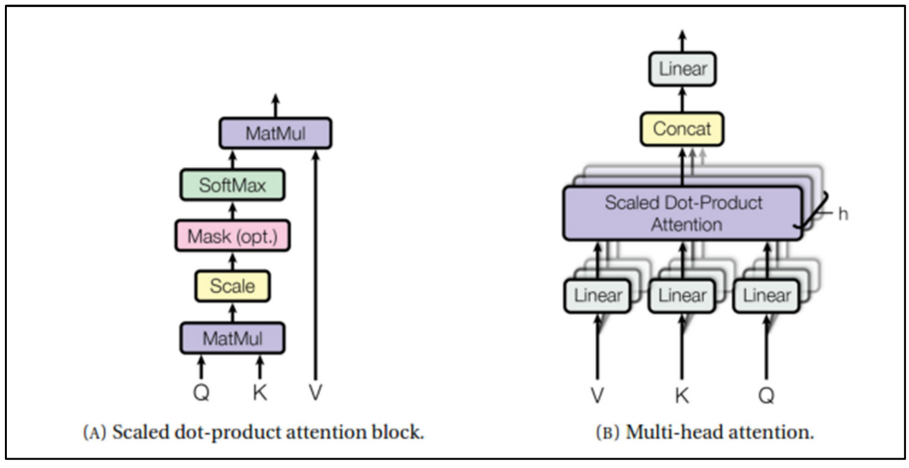
Attention mechanism

Attention models represent a recent advancement in the fields of NLP and computer vision, often referred to as "attention is all you need." This model introduces the concept of assigning varying weights to words within a sentence. Its primary objective is to comprehensively analyze a sequence, whether it's a text feature or an image feature, and condense the most informative features into a fixed-length context vector. Consequently, the model prioritizes specific words in the text while downplaying other.[21] This mechanism enables the model to concentrate on the relevant portions of both the question and the context. The key part of the transformer architecture is its self-attention mechanism. In the transformer, the attention function is expressed in terms of three elements: queries, keys, and values, all represented as vectors. The output is computed by taking the weighted sum of the values, with each value assigned a weight determined by a compatibility function between the query and its corresponding key. Self-attention mechanism is called scaled dot- product attention and is illustrated in Fig. 3.

In this process, the dot products are computed between the query and all keys, then scaled by the dimension of the keys, d_k, and finally, a softmax function assigns weights to the values. This operation can be performed for a set of queries efficiently using matrices. Denoting the matrices Q, K, and V for queries, keys, and values, respectively, the attention operation is defined by Eq. (1):[21]

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

Self-attention is performed for h "heads" $\{z_1, z_2, \ldots, z_h\}$, projecting the queries, keys, and values with learned linear functions. This enables



(A) Scaled dot-product attention block. (B) Multi-head attention.

Fig. 3. Self-attention modules of the transformer architecture.[21]

the simultaneous application of scaled dot-product attention on each head, producing output values with dimension d_v. Subsequently, these output values are concatenated and projected once again in Eqs. (2) and (3),

$$mltihead\,(Q, K, V) = con\,(z_1, z_2 \ldots, z_h)\,W^o \tag{2}$$

Where :

$$z_i = attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{3}$$

with W representing the learnable parameters of the projection layers.

In order to promote the connection between question and image features, a cross-attention mechanism is designed to capture the attentive relationships between the features. Specifically, the image features are first passed through a linear transformation layer to obtain a visual attention probability distribution over all image features. The global features of the attended regions are then extracted using this attention distribution. At the same time, the attention distribution is also utilized to reweight the visual features to focus on the crucial parts. As for the text features, similar operations are performed.

Fusion with attention mechanism for image features and question features

The attention mechanism has been shown to be a powerful technique that can capture important information from raw features in either linguistic or visual representation. The attention mechanisms are used to solve this lack by using local image features, and make the model assign varying weights to features. CNN-based approaches have achieved state-of-the-art results on many VQA benchmarks, but they are not good at reasoning about spatial relationships between different objects in an image. For example, CNN may be able to identify that there is a dog and a cat in the given image, but it may not be able to tell you which animal is in front of the other. This is because CNNs typically operate on local image patches, and they do not explicitly encode spatial information. Interpretable features can be challenging to provide explanations for answers generated by VQA systems. The semantic gap between visual and textual data is another challenge for CNNs, as VQA requires understanding the semantics of both images and questions. Traditional CNN-based approaches focus just on image features, which makes it difficult to integrate textual information effectively.

To address the mentioned limitations, attention mechanisms can be utilized for VQA to enhance context modeling, multimodal integration, and generalization. Through the attention mechanism, the system weights the visual features based on their relevance to the question, instead of using global features. Attention weights allow the output stage to focus on relevant parts of the image. Fusion involves combining the extracted image and question features, then models the hidden relationship between the language feature and visual feature to generate an answer. The main difference between several approaches is how they combine textual and image features. For example, they can simply combine them using concatenation or using pointwise (element-wise) multiplication of image and answer features.

Different attention Mechanisms are used; this paper proposes a cross-attention mechanism. According to the cross-attention mechanism, two types of attention will be learned: intramodal attention within the same modality (image attention) and intermodal attention (learn image attention according to the question as a guide and learn image attention according to the question as a guide). The intermodal attention use question features are used as a guide to compute visual features attention, and image features are used as a guide to compute textual features attention. This attention mechanism explores the interactions between features of image regions and question features, as shown in Fig. 4. We build self-attention and guided attention based on the multi-head attention mechanism. Image modality (has its own self-attention block, and cross-modal guided attention is applied from one modality to the other, as follows:

Step 1: Visual features of the input image, fv, are subjected to self-attention to learn the internal structure of the image, allowing the model to give varying importance to features in different regions, as in Eq. (4).

$$f_v = Multi - Head\ attention\,(f_v,\ f_v,\ f_v) \tag{4}$$

Step2:The textual features of question f_q *are extracted using the BioBERT model, which enables* the model to assign weights of importance to the words of the question.

Step3:The visual features, f_v, from step 1 are guided by question features to determine which parts of images are related to question words, as in Eq. (5).

$$f_v = Multi - Head\ attention\,(f_v,\ f_q, f_q) \tag{5}$$

Step 4: The Textual features, fq of question from step 2, are guided by visual features, which made the model understand which parts of the question are important for the image, as in Eq. (6).

$$f_q = Multi - Head\ attention\,(f_q,\ f_v,\ f_v) \tag{6}$$

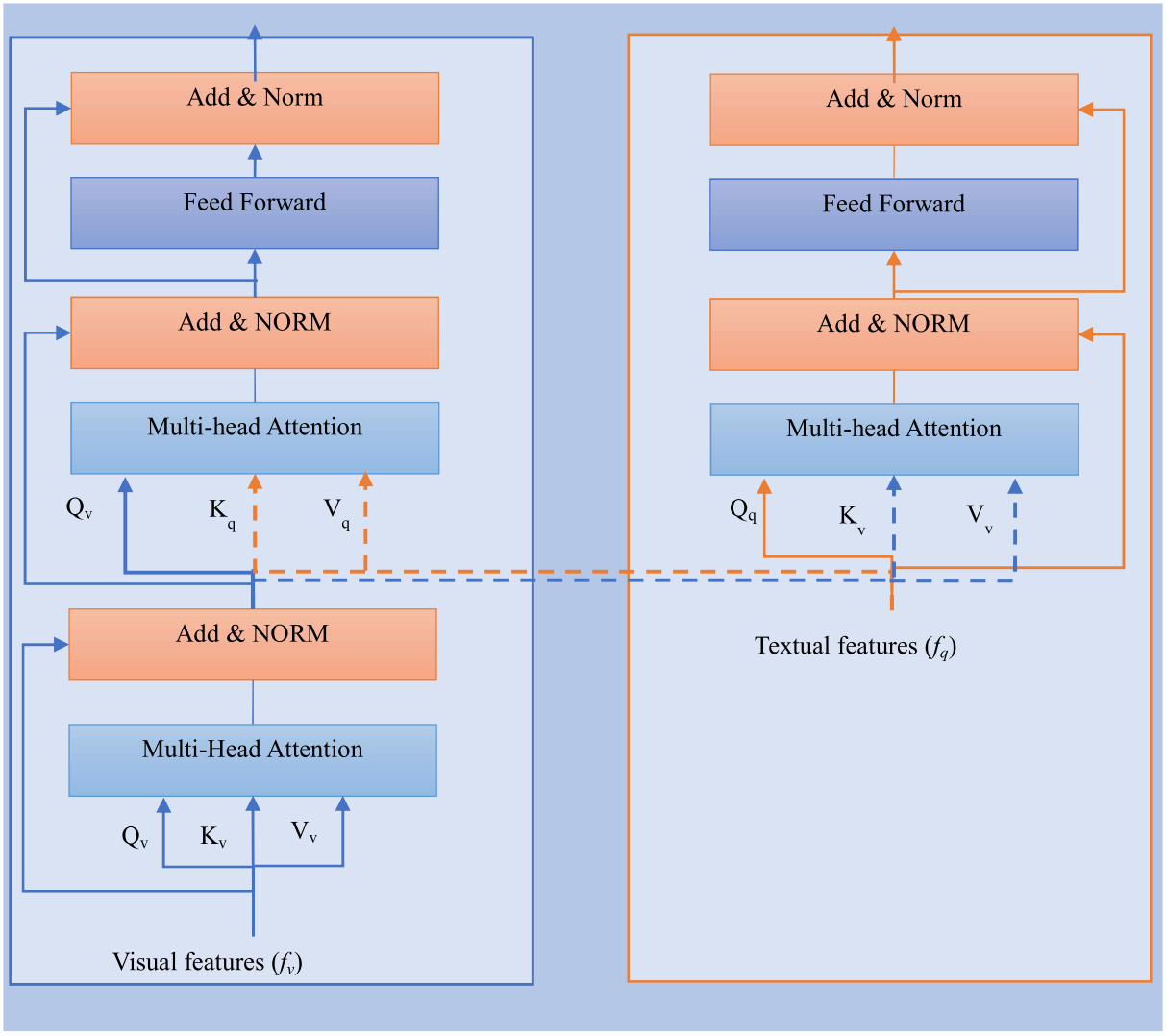


Fig. 4. Cross-attention mechanism.

Step 5: After self and cross attention are computed, the visual and textual features are fused. This fusion of features made the model integrate the information from two modalities: image and question. The fusion is performed using the elementwise addition of image and question features.

In this way, the fused feature could be obtained by carrying the relation of image question features. The fused features vector is passed to the classifier model to predict the correct answer.

Answer generation (decoder stage)

Answer prediction plays a critical role in VQA; the answer to the question about the input medical image will be predicted. VQA is usually formulated as a classification problem, where distinct answers are considered as different categories. A common choice of the classifier is a multilayer perceptron (MLP). Answer prediction is performed as a classification task. The classifier in the VQA takes the fused features from both the image and question, focusing on the relevant parts of both, processes them through fully connected layers, and finally applies a softmax layer to output the predicted answer. This classifier is designed to take the combined information from the multimodal attention network and produce a correct prediction, ensuring that both visual and textual cues are considered. The fused features are fed into a classifier that outputs the probability distribution of the N = No. answers and obtains the predicted results. The MLP, as explained in Fig. 5, is a fully connected neural network with 2 hidden layers, followed by a SoftMax layer to obtain a distribution over N answers.

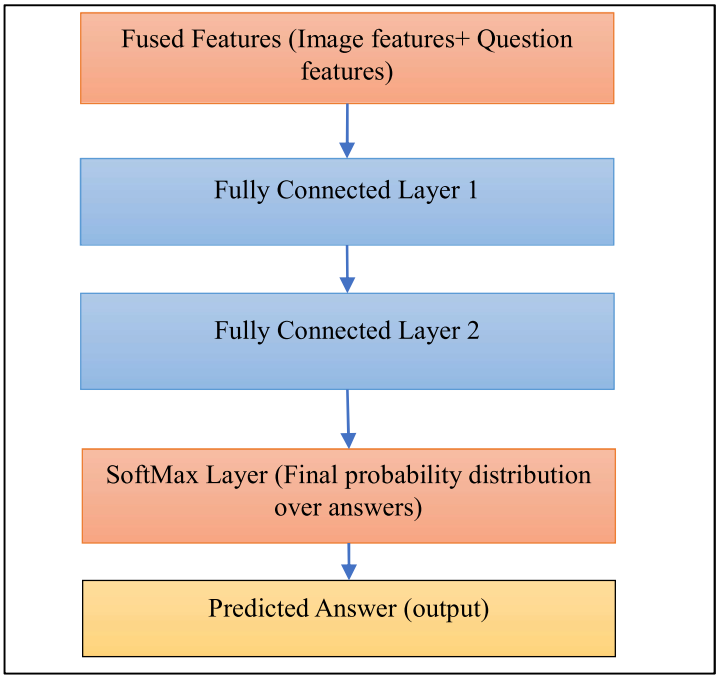


Fig. 5. Answer classification architecture.

Result and discussion

This section views and discusses the results of the proposed system. The experiments were conducted to evaluate the performance of our proposed Medical VQA framework, including its components proposed in Materials and Methods, on the published benchmark datasets VQA-RAD and Med-VQA 2019.

For the VQA-RAD dataset, exactly the same training set and test set described in 11 were used. The dataset has 458 unique answers. The test set contains 451 questions, and the rest is for training. The Med-VQA 2019 dataset has 312 unique answers, and it was divided into:

- Training set:3200 images and 12,792 questions-answers (QA).
- Validation set:500 images and 2000 QA pairs.
- Test set:500 images with 500 QA pairs.

The training set was used to train the proposed system. The data is equally distributed over four categories based on the question types, which are: “plane category, organ category, modality category, and abnormality category”.

The model uses BioBERT, DAE, and a cross-attention mechanism. A Bi-LSTM model is used for the question encoder to compare the results and effectiveness of BioBERT on the system. The accuracy metric is used to identify the convergence, and binary cross-entropy is utilized as the loss function. The BLEU score calculates the similarity of the reference (ground truth answer) and the hypothesis (predicted answer) at an n-gram level. Thus, it is a very useful metric for comparing two sequences or sentences.[23] Specifically, we use BLEU-1, BLEU-2, and BLEU-3 scores to compare the sequences at 1-gram, 2-gram, and 3-gram levels, respectively. In contrast, the Adam optimizer was used to optimize the model loss. The proposed Models were optimized by using early stopping (with patience = 10) and batch normalization. The batch size was 64, and the dropout probability was 0.2. The learning is set as 0.0001, whereas the decay rate is set as 10^{-10}. We select the accuracy and BLEU score that appear frequently in the test results.

Before running the evaluation metrics, each answer is converted to lowercase for the visual question answering. The evaluation will be conducted based on the accuracy metric. An adapted version of the accuracy metric can be used from the general domain VQA task that considers exact matching of a participant’s provided answer and the ground truth answer.

The results in Tables 1 and 2 illustrate the accuracy and BLEU scores of VQA-RAD and Med-VQA 2019, respectively. These results show that our system with a cross-attention mechanism-based Medical VQA system can improve the performance of our medical VQA system for both datasets. In Table 1, the test accuracy and BLEU of VQA-RAD were shown by question category for the two models compared to the other models. There is an improvement in accuracy for the

Table 1. Accuracy and BLEU score results and comparisons using multi-models on the VQA-RAD dataset.

Models	Closed-ended Accuracy	Open-ended Accuracy	Overall- Accuracy	BLEU score
HQS-VQA [16]	—-	—-	—-	0.411
Our system(BiLSTM + DAE + Cross-attention)	82.7	64.5	73.6	0.743
Our system (BioBERT + DAE + Cross-attention)	85.8	66.1	75.9	0.781

Table 2. Accuracy and BLEU score results and comparisons using multi-models on the Med-VQA 2019 dataset.

Models	Metric	Modality	Plane	Organ	Abnormality	Overall
CGMVQA [14]	Accuracy	80.5	80.8	72.8	1.7	65.9
	BLEU	0.856	0.813	0.679	0.16	0.640
Our system (BiLSTM + DAE + Cross-attention)	Accuracy	87.1	85.7	79.6	19.2	72.7
	BLEU	0.897	0.864	0.846	0.211	0.718
Our system (BioBERT + DAE + Cross-attention)	Accuracy	88.8	87.9	80.8	22.6	73.8
	BLEU	0.879	0.884	0.816	0.23	0.745

closed-end category by (+3.1%), (+1.6%) for open-ended questions. The increase in BLEU score by 0.370 when compared with HQS-VQA. [16]

In Table 2, the test accuracy and BLEU of Med-VQA 2019 were shown by question category for the two models. There is a significant increase in accuracy in the modality category (+1.7%), (+2.2%) in plan, (+1.2%), and the increase in the abnormality category was (+3.4%), which made the BioBERT model achieve an increased overall accuracy. Also, there is an increase in the accuracy by (+7.8) and (0.105) in BLEU when compered by CGMVQA. [14]

Table 3, show the training parameters of the model.

Table 3. The summary of parameter settings.

Name	Size
Batch Size	64
Dropout Rate	0.2
Learning Rate	0.0001
Decay Rate	10^{-10}

Despite the growing interest and progress in Medical VQA, there were still limitations, such as Medical images (e.g., CT, MRI, X-rays) contain subtle features that require expert interpretation. Unlike natural images, the relevant diagnostic information may be localized, low-contrast, or multi-layered. Also, Medical VQA datasets are typically small due to privacy concerns and the high cost of expert annotation.

Conclusion

Medical VQA plays an important role in providing medical assistance to the end-users. This paper proposed a cross-attention mechanism-based Medical VQA. In the proposed system, the BioBERT model is used for textual features extraction to enable a better understanding of medical terminology, abbreviations, and phrase structures compared to general models. Adapting the DAE model for visual features extraction, which handles noise and extracts important and strong features from medical images by slicing the image into overlapping patches. The attention mechanism was applied at the outputs of BioBERT and DAE layers. According to the cross-attention mechanism, two types of attention will be learned: intramodal attention within the same modality (image attention and intermodal attention (learn image attention according to the question as a guide and learn image attention according to the question as a guide). The results illustrate that the cross-attention mechanism can significantly improve the performance of our Medical VQA system. In future investigations, it can develop the system by attention visualization, such as integrating Grad-CAM for finer-grained visual explanations. Another aspect to consider in the future is to move Medical VQA from a classification task to a generative one.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images that are not ours have been included with the necessary permission for republication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University of Babylon.

Authors' contributions statement

The idea is suggested by N F M, and I H A. N F M pays attention to the practical side of the manuscript, and I H A checked the results and edited the paper.

All authors discussed the results and contributed to the final manuscript.

Data availability

The datasets used in this study are available in the [VQA-RAD dataset: https://www.kaggle.com/datasets/shashankshekhar1205/vqa-rad-visual-question-answering-radiology. And ImageCLEF VQA-Med2019 dataset: https://www.kaggle.com/datasets/claudiopisa9884/imageclef-vqa-med-2019].

References

1. Lin Z, Zhang D, Tao Q, Shi D, Haffari G, Wu Q, He M, Ge Z. Medical visual question answering: A survey. Artif Intell Med. 2023 Sep 1;143:102611. https://doi.org/10.1016/j.artmed.2023.102611.
2. Sharma D, Purushotham S, Reddy CK. MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. Sci. Rep. 2021 Oct 6;11(1):1–18. https://doi.org/10.1038/s41598-021-98390-1.
3. Barra S, Bisogni C, De Marsico M, Ricciardi S. Visual question answering: Which investigated applications? Pattern Recognit. Lett. 2021 Nov 1;151:325–331. https://doi.org/10.48550/arXiv.2103.02937.
4. Ye LY, Miao XY, Cai WS, Xu WJ. Medical image diagnosis of prostate tumor based on PSP-Net+ VGG16 deep learning network. Comput. Methods Programs Biomed. 2022 Jun 1;221:106770. https://doi.org/10.1016/j.cmpb.2022.106770.
5. Asroni A, Ku-Mahamud KR, Damarjati C, Slamat HB. Arabic Speech Classification Method Based on Padding and Deep Learning Neural Network. Baghdad Sci J. 2021;18(2):925–936. https://doi.org/10.21123/bsj.2021.18.2(Suppl.).0925.
6. Bazi Y, Rahhal MM, Bashmal L, Zuair M. Vision–language model for visual question answering in medical imagery. Bioengineering. 2023 Mar 20;10(3):1–17. https://doi.org/10.3390/bioengineering10030380.
7. Ikechukwu AV, Murali S, Deepu R, Shivamurthy RC. ResNet-50 vs VGG-19 vs training from scratch: A comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. Global Transitions Proc. 2021 Nov 1;2(2):375–381. https://doi.org/10.1016/j.gltp.2021.08.027.
8. Mascarenhas S, Agarwal M. A comparison between VGG16, VGG19 and ResNet50 architecture frameworks for Image Classification. In2021 International conference on disruptive technologies for multi-disciplinary research and applications (CENTCON). IEEE. India. 2021 Nov 19;1:96–99. https://doi.org/10.1109/CENTCON52345.2021.9687944.
9. Lu S, Liu M, Yin L, Yin Z, Liu X, Zheng W. The multi-modal fusion in visual question answering: A review of attention mechanisms. PeerJ Comput. Sci. 2023 May 30;9:1–29. https://doi.org/10.7717/peerj-cs.1400.
10. Yan F, Silamu W, Li Y. Deep modular bilinear attention network for visual question answering. Sensors. 2022 Jan 28;22(3):1–15. https://doi.org/10.3390/s22031045.
11. Silva JD, Martins B, Magalhães J. Contrastive training of a multimodal encoder for medical visual question answering. Intell. Syst. Appl. 2023 May;18:1–10. https://doi.org/10.1016/j.iswa.2023.200221.
12. Gasmi K, Ltaifa IB, Lejeune G, Alshammari H, Ammar LB, Mahmood MA. Optimal deep neural network-based model for answering visual medical question. Cybern. Syst. 2022 Apr 22;53(5):403–24. https://doi.org/10.1080/01969722.2021.2018543.
13. Vu MH, Löfstedt T, Nyholm T, Sznitman R. A question-centric model for visual question answering in medical imaging. IEEE Trans. Med. Imaging. 2020 Mar 4;39(9):2856–2868. https://doi.org/10.1109/TMI.2020.2978284.
14. Ren F, Zhou Y. Cgmvqa: A new classification and generative model for medical visual question answering. IEEE Access. 2020 Mar 11;8:50626–36. https://doi.org/10.1109/ACCESS.2020.2980024.
15. Liu B, Zhan LM, Xu L, Wu XM. Medical visual question answering via conditional reasoning and contrastive learning. IEEE Trans. Med. Imaging. May 2023 26;42(5):1532–45. https://doi.org/10.1109/TMI.2022.3232411.
16. Gupta D, Suman S, Ekbal A. Hierarchical deep multi-modal network for medical visual question answering. Expert Syst. Appl. 2021 Feb 1;164:113993. https://doi.org/10.1016/j.eswa.2020.113993.
17. Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. Sci. Data. 2018 Nov 20;5(1):1–10. https://doi.org/10.1038/sdata.2018.251.
18. El-Shafai W, El-Nabi SA, El-Rabaie ES, Ali AM, Soliman NF, Algarni AD, Abd El-Samie FE. Efficient Deep-Learning-Based Autoencoder Denoising Approach for Medical Image Diagnosis. Computers, Materials & Continua. 2022 Mar 1;70(3):6107–25. https://doi.org/10.32604/cmc.2022.020698.
19. Nguyen BD, Do TT, Nguyen BX, Do T, Tjiputra E, Tran QD. Overcoming data limitation in medical visual question answering. InInternational conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing. 2019 Oct 10;522–530. https://doi.org/10.1007/978-3-030-32251-9_57.
20. Gao S, Huang Y, Zhang S, Han J, Wang G, Zhang M, Lin Q. Short-term runoff prediction with GRU and LSTM networks without requiring time step optimization during sample generation. J. Hydrol. 2020 Oct 1;589:125188. https://doi.org/10.1016/j.jhydrol.2020.125188.
21. Zhang Y, Liu C, Liu M, Liu T, Lin H, Huang CB, Ning L. Attention is all you need: Utilizing attention in AI-enabled drug discovery. Briefings Bioinf. 2024 Jan 1;25(1):1–22. https://doi.org/10.1093/bib/bbad467.
22. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020 Feb 15;36(4):1234–40. https://doi.org/10.48550/arXiv.1901.08746.
23. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002 July;311–318. http://dx.doi.org/10.3115/1073083.1073135.

نظام الإجابة على الأسئلة الطبية مرئيا المعتمد على Cross-Attention

ندى فاضل محمد، اسراء هادي علي

كلية تكنولوجيا المعلومات، جامعة بابل، بابل، العراق.

الخلاصة

الاستجابة البصرية للأسئلة (VQA) هي مهمة تعلم آلي تهدف إلى إنشاء أنظمة قادرة على الإجابة على أسئلة اللغة الطبيعية بناءً على الصور المقدمة. تُعد أنظمة الاستجابة البصرية للأسئلة الطبية تطبيقًا متخصصًا في المجال الطبي لهذه التقنية، حيث تساعد في فهم المعلومات السريرية ذات الصلة بالصور الطبية. تعتمد هذه الأنظمة على تقنيات الشبكات العصبية العميقة لتوليد إجابات دقيقة للأسئلة الطبية، والتي قد تكون أسئلة مغلقة أو مفتوحة . تقدم هذه الورقة البحثية نظام استجابة بصري للأسئلة الطبية والذي يعتمد على آلية الانتباه cross-attention. حقق النظام المقترح أداءً جيدًا من خلال دمج ثلاث مكونات رئيسية، كل منها يعالج تحديات حرجة في مجال الاستجابة البصرية للأسئلة الطبية: النموذج BioBERT يتميز بالتدريب المسبق الخاص على النصوص بالمجال الطبي بتمكين فهمٍ أعلى للمصطلحات السريرية والأسئلة الطبية المعقدة مقارنة بنماذج اللغة العامة. نموذج (DAE) لاستخراج الميزات البصرية، حيث يتميز بقدرته على تقليل الضوضاء واستخراج الميزات الهرمية في معالجة تحديين أساسيين في التصوير الطبي، تحسين كشف الآفات الصغيرة من خلال معالجة(patches) والأداء القوي في للضوضاء الشديدة (مثل تشوهات التصوير بالرنين المغناطيسي) واخيرا آلية الانتباه cross-attention: لاكتشاف العلاقة بين الصورة الطبية والسؤال، حيث يمكن الانتباه عبر الوسائط من التركيز البصري الديناميكي الموجه بالسؤال. تتكون آلية الانتباه من مكونين: الانتباه ضمن نفس modality) ,الانتباه بين الوسائط (عبر أكثر من modality) مما يمكن النموذج من التركيز على الأجزاء ذات الصلة من الصورة والسؤال لتوليد الإجابة. أظهرت التجارب التي أجريت على مجموعتي البيانات VQA-RAD وMed-VQA 2019 أن النظام المقترح يحقق دقة تبلغ 76.5% و78.3% على التوالي، متفوقًا على النماذج الأساسية التي تستخدم آليات الانتباه التقليدية مثل BAN أو SAN

الكلمات المفتاحية: BioBERT، BAN، رؤية الحاسوب، آلية الانتباه المتبادل، VQA، DAE الطبي، معالجة اللغة الطبيعية.