Deep Learning-Based Visual Question Answering for Medical Imaging: Insights from the PathVQA Dataset

Esra Balık
Department of Sofware Engineering
Bandırma Onyedi Eylül University
Balıkesir / Turkey
ebalik@bandirma.edu.tr

Mehmet Kaya
Department of Computer Engineering
Fırat University
Elazığ / Turkey
kaya@firat.edu.tr

2024 International Conference on Decision Aid Sciences and Applications (DASA) | 979-8-3503-6910-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/DASA63652.2024.10836414

***Abstract*— Extracting information from medical data and making the correct diagnosis based on this information is critical in medical decision support systems. Interpretation of complex medical images requires a deep medical knowledge. Traditional methods have performance limitations in such data-intensive analyses and are insufficient in terms of obtaining accurate results. In this study, we present an effective solution to obtain fast and meaningful information from medical images and develop a Medical Visual Question Answering (MedVQA) system. With this system, it is aimed to produce correct answers by making sense of the questions directed to the contents of medical images by using deep learning-based models trained on the PathVQA dataset. In the training process of the model, transfer learning techniques and Attention mechanisms are used to better analyze the complex structure of medical terminology. With this proposed solution, it is aimed to accelerate the diagnostic processes of healthcare professionals and to provide automatic understanding of visual information. As a result of the experiments, it was observed that the MedVQA model developed as a result of the training on free-form question-answer pairs on the PathVQA dataset performed 58.63% with BERT+VGG19 and 60.13% with BioBERT+VGG19. For the question-answer pairs in the Yes-No form, 91.80% and 92.17% accuracy was obtained, respectively. Our study showed a significant performance improvement compared to existing approaches and demonstrated that MedVQA models can be effectively used for deep learning-based extraction of meaningful information from medical images and extraction of features of textual data with natural language processing models. Our study aims to contribute to healthcare professionals to obtain more accurate and faster solutions by using decision support systems with an innovative approach in the field of medical image analysis.**

Keywords*— *Medical Visual Question Answering, BERT, BioBERT, VGG19.

I. Introduction

The need to integrate and make sense of visual data and textual information is increasing day by day, and this need becomes more evident especially in complex data analysis. Visual question answering (VQA) [1] systems are artificial intelligence systems that allow you to pose a natural language question to an image and obtain a meaningful answer. VQA systems combine computer vision and natural language processing to create a multimodal system. These multimodal systems have great potential in many areas such as health, security and autonomous vehicles. Artificial intelligence techniques are increasingly being used to physically examine the human body and accelerate the diagnosis and treatment process. Medical visual question answering (MedVQA) systems are an important tool in healthcare, aiming to make clinical decisions based on certain information and data. It is known that time and accuracy are the most important issues in clinical settings, making it easier for healthcare professionals to make fast and accurate decisions based on medical images. One of the applications that MedVQA analyzes is to act like pathologists who examine body tissues and help healthcare professionals make diagnoses[2]. In addition, it is known that MedVQA can help minimize the risk of misdiagnosis by being considered as a "second opinion"[3]. To summarize the tasks of MedVQA, it focuses on answering natural language questions posed to specific medical image data. However, existing methods have some limitations in terms of visual feature extraction and their capacity to capture the right context for the questions.

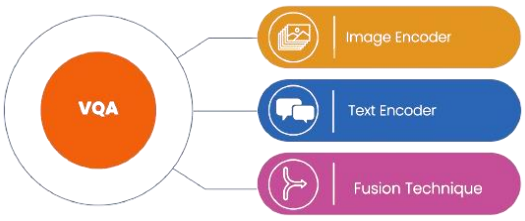

Fig. 1 VQA frameworks main components

The need for MedVQA systems is increasing day by day. It can provide efficiency in clinical decision support systems and in issues such as reducing the workload of doctors, reducing the diagnosis and treatment process for patients. The inadequacies of existing MedVQA methods can be categorized under several headings;

1) Weakness in feature extraction

2) Adapt to diversity in data sets

3) Calculation cost

4) Lack of contextual meaning and generalization

979-8-3503-6910-6/24/$31.00 ©2024 IEEE

The main objective of this study is to comparatively demonstrate the performance evaluation of VGG19[4], BERT[5] and BioBERT[6] models for deeper understanding of medical data and more effective processing of medical context. For visual features acquisition, the VGG19 model is used to understand the fine details of medical images. For textual features, BERT and BioBERT, the biomedically trained version of BERT, are compared. In this way, the impact of a general language model and a language model with a command of medical terminology on medical visual question answering systems will be evaluated. In order for visual and textual information to produce meaningful and effective results, the Attention mechanism will be used to perform the fusion process by taking into account the content and context of the question. The main contributions of this study can be listed as follows;

1) Analyzing the effects of BERT and BioBERT architectures for comparing language models in medical context,
2) Performance evaluation of the VGG19 model for analyzing complex details in images,
3) Use the Attention mechanism to help identify which information is more important according to the context of the problem,
4) It is aimed to provide a reliable and effective MedVQA solution in clinical decision support systems.

II. Related Work

Medical visual question answering systems combine natural language processing and computer vision techniques to provide meaningful answers to natural language questions posed to images. In this field, deep learning and natural language processing models have made significant progress. The process of converting visual data into text data has created a necessary motivation in many fields of study[7-14]. Revolutionizing natural language processing tasks, the BERT architecture allows us to use pre-trained generic representations with task-specific fine-tuning. BioBERT is used to analyze the structure of medical language, complex and specialized vocabulary, and to provide the highest performance in text processing, and it provides deep bidirectional representations and broader contextualization of texts. VGG19, which has high performance in image classification, provides more efficient feature extraction with small size filters. When these two models are integrated, both text and image features are expected to produce highly accurate results. When the studies in this field are examined, a pre-trained CNN architecture is generally used as an image encoder. As a text encoder, techniques such as LSTM and GRU stand out. Shengyan et al. [8] achieved 0.376 accuracy and 0.412 BLEU using an encoder-decoder model for the MedVQA task. In this model, which they presented at the ImageCLEF VQA-Med 2020 competition, seq2seq[9] was used for answer prediction. Sharma et al. [10] used two different datasets to perform the MedVQA task at high performance in their proposed MedFuseNet architecture. In addition, an innovative fusion algorithm was proposed. Two attention modules are used, allowing the features of images and questions to interact twice. Liu et al.[11] proposed to use contrastive learning technology using unlabeled data in a self-supervised scheme. They can contribute to performance improvement, especially in diseases where medical images are inadequate. There are some limitations and shortcomings of the studies that have important contributions to improve the MedVQA task. Limitations such as the dependence of the models on large datasets and computational power, the inability to fully capture details in medical images, and the shortcomings in analyzing complex medical questions, reveal the need for further research and improvement in the field of MedVQA.

III. Materials and Methods

Image Encoder: In the fields of deep learning and computer vision, image coding is the process of converting images into vectors of meaningful information called feature vectors. High-level representation vectors of images are used for image-based studies. By extracting some meaningful features from an image and summarizing the content of these features, the generated meta-information is used as input for other tasks. The fact that it can be used with other data types such as text and audio also contributes to the development of deep learning based systems. In our proposed model, the VGG19 model, which provides high efficiency with small size filters, is used. VGG architectures, short for Visual Geometry Group, are Convolutional Neural Networks with multiple layers. They have 16 or 19 layers and are mainly used to extract high-level features from images. VGG19 follows the following structure for feature extraction of medical images;

Convolutional Layers: Extracts local features of images to capture spatial information.

Pooling Layers: Reduces image sizes to minimize the processing load of the model.

Fully Connected Layers: Condenses the extracted features into a vector.

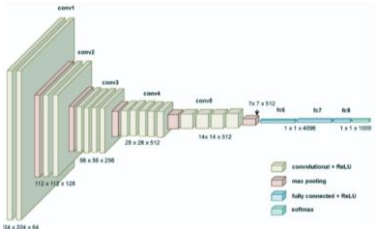

Fig. 1 VGG19 Layers

$$\upsilon = \mathrm{VGG19}(I) \in R^{\mathrm{d}}$$

The VGG19 model is applied to the system to obtain a *d-dimensional* vector *v* of the input image *I*. Low-level features of the image such as edges and textures are extracted by convolutional operations between layers. In the next layers, high-level features are extracted to recognize the prominent anomalies in the image. With this model, pathological structures are identified and abnormal

formations are detected.

Text Encoder: In the fields of natural language processing (NLP) and machine learning (ML), models that train textual data into meaningful data with low dimensionality are called text encoders. These models transform a text given in natural language into embedding vectors that capture the context of the text. Textual data can be understood at a higher level and used in machine learning applications. In addition to obtaining the properties of textual data, dimension reduction also facilitates the use of data. It can be integrated with other data types, especially in multimodal tasks.

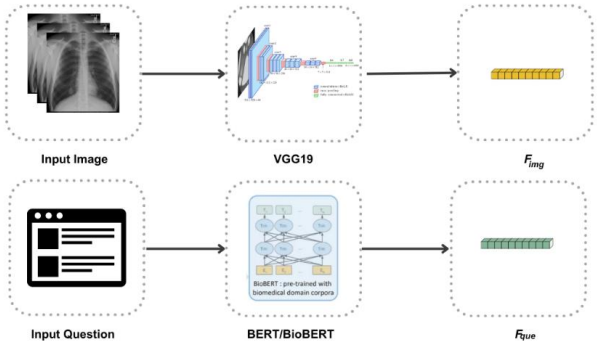

Fig. 3 Process of feature extraction from data

In MedVQA tasks, the text encoder is used to process and extract the meaning of natural language questions posed to images. BERT (Bidirectional Encoder Representations from Transformers), developed by Google in 2018 and exhibiting high performances in natural language processing, processes text data in a bidirectional manner. Each word is evaluated together with the context to its left and right. BioBERT has improved its context understanding ability in the medical field by using BERT's bidirectional context understanding ability and training on medical databases such as PMC (PubMed Central)[12] specific to the biomedical field. In this regard, it provides higher performance than other NLP models and is more effective in understanding medical terminology.

Fusion Model: In multimodal artificial intelligence applications, it is aimed to produce more accurate results by combining the extracted features of different data types (audio, text, video, image, etc.). For this, the fusion model combines the features of different data types. The use of different data types increases the generalization capability of the model.

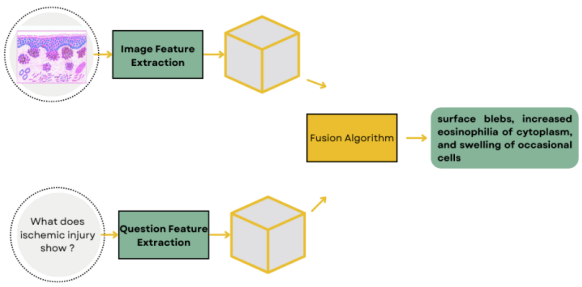

Fig. 4 Step of merging features

In order to integrate visual and textual features and extract meaningful information, it is important to combine them correctly. The features extracted with the VGG19 model used for visual feature extraction are in the form of a vector. This vector represents the content of the image. On the other hand, the vector obtained with the BERT and BioBERT architectures used to extract textual features contains features that represent the contextual content of the text. These two features need to be combined with a fusion algorithm and the extracted features need to be integrated in such a way that they complement each other [13]. The Attention mechanism, which automatically determines which features are more important, will ensure that critical information is not lost in the fusion process. In addition, it will have a direct impact on model performance in terms of its easy adaptability to different data types and computational efficiency. The correct fusion of visual and textual features is an important step that directly affects the effectiveness and reliability of a medical visual question answering system. Figure 5 shows the general steps of the basic Attention-based fusion mechanism.

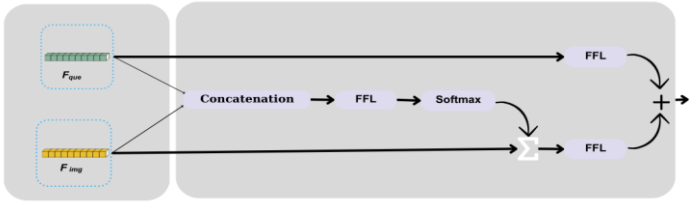

Fig. 5 Attention based fusion model

The question feature vector is compared with the image feature vector and an attention weight (importance score) is determined. The importance score represents the relevance of regions in the image to the question.

$$score(q,vi) = qTWvi,$$

where q is the question vector, vi is the feature vector of region i of the image and W is the weight matrix.

IV. Experiments and Results

A. Dataset Preparation

In this study, the PathVQA [2] dataset of pathological images was used. This publicly available dataset contains 4998 images and 32,799 question-answer pairs. The questions are categorized into 7 groups: one group contains closed-ended answers (yes-no) and 6 groups contain open-ended questions (what-where-when-how-how-who-what). The dataset is segmented into training and test. The author has developed a semi-automatic pipeline for transferring subtitles to QA pairs, so that the QA pairs can be manually controlled[10].

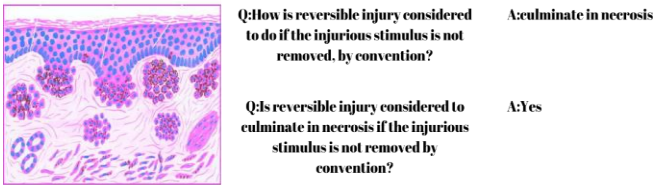

Fig. 6 An example from the PathVQA dataset

TABLE I. PATHVQA DATA SIZE

PathVQA	Data Types		
	Image	**Yes-No Questions**	**Open-Ended Questions**
Size	4998	16K	16K

B. Implementation

In this study, 3 main components were used to integrate and analyze visual and textual information. VGG19 was used as an image encoder and the features of the images were extracted. BERT and BioBERT architectures were used comparatively in the textual processing of question-answer pairs. These two feature matrices were then fused with the Attention mechanism to produce results.

TABLE III. ANSWER PRODUCTION RESULTS

Method	**Accuracy of Free-Form Questions**	**Accuracy of Yes-No Questions**
Sharma vd.[9]	%84	
BERT+VGG19	%58,63	%91,80
BioBERT+VGG19	%60,13	%92,17

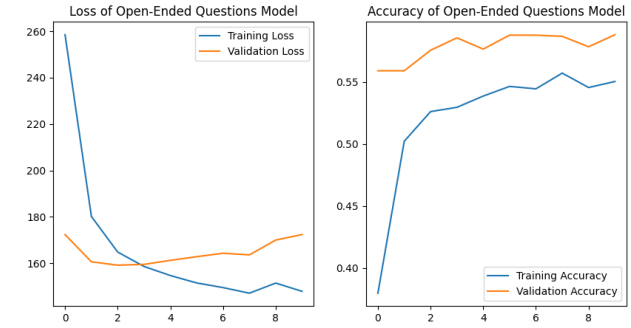

Fig. 7 Validation and Training Loss Graphs for the BERT+VGG19 model

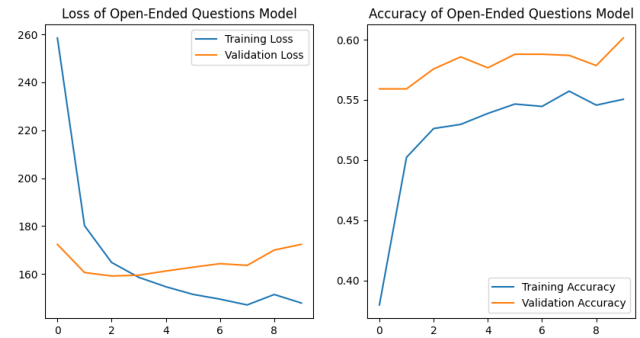

Fig. 8 Validation and Training Loss Graphs for the BioBERT+VGG19 model

V. CONCLUSIONS

This study presents a model comparing BERT and BioBERT architectures in the MedVQA domain. According to our main findings, the BioBERT model, which dominates the medical terminology, provided higher accuracy compared to the BERT architecture, which is a general language model. BioBERT is more efficient in interpreting medical terminology and extracting meaningful information. It has been shown that medical language models can produce more accurate results in the medical domain with visual data. MedVQA performance can be improved by using larger datasets and different visual feature extractor models.

VI. ACKNOWLEDGMENTS

This study was supported by the Research Projects Support Program with the number ADEP.23.11.

REFERENCES

[1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (pp. 2425-2433).

[2] He, X., Zhang, Y., Mou, L., Xing, E., & Xie, P. (2020). Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

[3] Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., ... & Kittler, H. (2020). Human–computer collaboration for skin cancer recognition. *Nature medicine*, *26*(8), 1229-1234.

[4] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[5] Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[6] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234-1240.

[7] Balık, E., Kaya, M., & Kaya, B. (2023, December). A CNN-LSTM based approach for image captioning. In *7th IET Smart Cities Symposium (SCS 2023)* (Vol. 2023, pp. 585-588). IET.

[8] Liu, S., Ding, H., & Zhou, X. (2020, September). Shengyan at VQA-Med 2020: An Encoder-Decoder Model for Medical Domain Visual Question Answering Task. In *CLEF (working notes)*.

[9] Sutskever, I. (2014). Sequence to Sequence Learning with Neural Networks. *arXiv preprint arXiv:1409.3215*.

[10] Sharma, D., Purushotham, S., & Reddy, C. K. (2021). MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, *11*(1), 19826.

[11] Liu, B., Zhan, L. M., & Wu, X. M. (2021). Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24* (pp. 210-220). Springer International Publishing.

[12] Wang, B., Xie, Q., Pei, J., Chen, Z., Tiwari, P., Li, Z., & Fu, J. (2023). Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, *56*(3), 1-52.

[13] Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., ... & Ge, Z. (2023). Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, *143*, 102611.

[14] Balık, E., & Kaya, M. (2023, October). A GAN-Based Approach to Generate Images from Text Description. In *2023 4th International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 179-183). IEEE.