*IOSR Journal of Mathematics (IOSR-JM)*
*e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 22, Issue 3, Ser. 1 (May. – June. 2026), PP 38-47*
*www.iosrjournals.org*

# A Multimodal Approach To Improving Visual Question Answering Based On Extracting Attentional Features From Images And Analyzing Medical Texts By BioBert

Hussein Ali, Zainab Aqeel, Zainb Qasem
*(Wasit Directorate Of Education /The Ministry Of Education, Iraq)*
*(Wasit Directorate Of Education /The Ministry Of Education, Iraq)*
*(Wasit Directorate Of Education /The Ministry Of Education, Iraq)*

***Abstract:***
*With the advancement of technology and the expansion of the use of artificial intelligence systems in the medical field, opportunities have been provided to improve the accuracy and efficiency of diagnosing and treating diseases. One of the methods used in this field is the processing of medical images and the use of different models to improve the accuracy of diagnosing diseases and answering visual questions. Among the challenges of using the integration of different modalities, including natural language processing models and deep neural networks, is to improve the answers to visual questions in the medical field. This can be done with the methods of extracting features from images and processing medical text, which can significantly improve the accuracy and efficiency of answering visual questions in the medical field. Also, the use of BioBERT, which is an advanced natural language model for processing medical text, can play an important role in improving the answers to visual questions in the medical field. The proposed method combines image features extracted using the VGG19 network with question features extracted from the BioBERT model. A multimodal factorized binomial (MFB) layer is used to merge these features and capture interactions and dependencies between the modals. The image features are further processed through a series of convolutional (Conv) layers and ReLU activation functions, which increase the resolution and achieve higher-level representations. The final output is obtained using a softmax layer, which generates a probability distribution over the classes. Parallel processing is applied to the question features using BioBERT, followed by Conv and ReLU layers, and a softmax layer. The model was trained for 100 epochs using the Adam optimizer and achieved significant accuracy in training and validation without overfitting. The performance evaluation includes top-1 accuracy, top-5 accuracy, and BLEU benchmark, tested on the VQA-Med-2019 dataset. The results show that the proposed model has achieved a higher accuracy of 0.21 compared to the baseline papers and a BLEU score of 0.5and an efficiency of 0.398.*

***Keyword:*** *Multimodal approach, Answer optimization, Visual questions, Medical images, Medical text analysis, BioBERT*

---

Date of Submission: 01-05-2026 Date of Acceptance: 11-05-2026

---

## I. Introduction

More than forty-five percent of countries in the world had at most one doctor per 1,000 people in 2021, according to the World Health Organization . As a result, healthcare workers have to spend a lot of time, which increases the rate of human error. Artificial intelligence (AI) can help in analyzing medical images and automate this process, as medical image analysis takes up a large portion of doctors' working time (Anot et al., 2015). Today, artificial intelligence researchers are increasingly focusing on the challenges of visual question answering (VQA), which has led to the development of models capable of understanding the broader context of images and solving problems that require human-level reasoning. One of the major unresolved challenges in this field is determining the most effective way to integrate information from two different types of models: language models, typically based on recurrent neural networks, and image models, which often rely on convolutional neural network (CNN) architectures.

Visual Question Answering (VQA) is a modern and promising research area that lies at the intersection of computer vision and natural language processing (Gupta, 2017). In VQA systems, the input consists of an image and a question expressed in natural language, and the output is the correct answer. This answer may be as simple as yes/no, a multiple-choice selection, a single word, or even a complete sentence.

At first glance, the VQA task appears highly complex. Traditional computer vision techniques used to extract meaningful information from images differ significantly from natural language processing approaches applied in question answering, making their integration challenging. Moreover, generating a plausible answer

from such diverse and multimodal data further increases the difficulty. Fortunately, recent advances in deep learning have paved the way for the development of more robust and efficient VQA algorithms.

To generate answers to specific questions, VQA technology must first interpret the semantic content of images and incorporate prior knowledge, which requires the combined use of natural language processing and computer vision techniques. With the growing interest in healthcare applications, integrating VQA into the medical domain has become particularly attractive. This integration can support physicians in diagnosis and treatment planning while also enabling patients to access medical information directly, thereby improving the effectiveness of healthcare services. Early intelligent medical systems such as MYCIN demonstrated the potential of simulating diagnosis and recommending treatments using medical knowledge and predefined rules (Shortliffe, 2012).

Several datasets are widely used for training and evaluating VQA systems, including COCO-QA, VQA-Dataset, FM-IQA, Visual Genome, Visual7W, and CLEVR. Early VQA datasets mainly focused on basic image attributes such as location, color, and object counting. Later, more complex tasks involving commonsense reasoning were introduced. In the medical domain, the most prominent VQA-Med datasets include ImageCLEF2018 VQA-Med, ImageCLEF2019 VQA-Med, and VQA-Rad (Kafle et al., 2017). These datasets primarily contain radiological images categorized into different question types, along with corresponding question-answer pairs and sets of possible answers for each question type (Li et al., 2018)

## II. Research Problem And Objectives

**Problem Statement**

Visual Question Answering (VQA) for the medical industry faces unique challenges in addition to the problems of machine vision and natural language processing. The lack of labeled data is the biggest obstacle for any supervised machine learning task in the medical domain. This is mainly due to the limitation in access to information due to privacy concerns. Furthermore, due to the lack of medical professionals, the spontaneous classification of medical data is an issue. When we examine VQA databases for practical and medical domains, there are many datasets for VQA related to real images. The number of data points in medical VQA datasets is in the thousands, while the number of data points in real VQA datasets is hundreds of times higher. Due to the inherent need for large data sets to train deep models, the application of deep learning methods for VQA in the medical field is a challenge. A natural language query and a related image are the two inputs for VQA, so it is crucial that multimodal data is managed properly to maximize the use of data collected from both modes. Furthermore, medical data is inherently complex due to the large amount of information contained in a clinical report or radiology scan. The data may have additional problems due to interferences created during scanning. Furthermore, instead of having an independent system for each body part, an ideal VQA system for medical images should be able to answer any type of query related to each organ system. Developing a solution to this challenge is also another difficult component. To generate answers, the model must generate a meaningful string of words. Therefore, developing an ideal deep learning framework that accurately integrates medical data to reduce the answer prediction error is crucial. The concerns and obstacles that medical VQA faces are different from those we face in general-domain VQA. Many of these obstacles relate to processing different types of images for different body parts and identifying regions that vary significantly with different medical conditions and diseases. The ability to understand questions and translate highly technical medical terminology in addition to frequently used non-medical terminology constitutes a second set of difficulties. There are many limitations in using these models and integrating them into a single model, and significant resources may be required to solve all of these problems. A standard deep learning solution for VQA consists of four main parts: feature extraction from the image, feature extraction from the query, fusion of two feature vectors, and an answer prediction component.

**Objectives of the study**

Our goal is to reduce the complexity of the model while maximizing the learning capacity of the model. In order to maximize the knowledge gained from the combination of input feature vectors, we propose a model for VQA in the medical domain that focuses on

1. Accurately representing the knowledge from multimodal inputs
2. The most efficient combination of feature vectors and conduct in-depth research to determine the importance of each model component and identify the essential components of the model that produce the best performance metrics.

Due to the availability and availability of patient health records, patients can now access and review their health records in relation to treatment, which helps them better understand their diseases. This raises the need for an automated process that is able to answer specific medical questions and display relevant images to verify the question and provide the correct answer. This is exactly the task addressed in this study. Given a medical image and a set of relevant clinical questions, the task aims to answer the questions using visual input images.

## III. Deep Learning

The development of artificial neural networks began with the work of Frank Rosenblatt, whose paper on the perceptron model laid the foundation for this field (Ren et al., 2018). The perceptron was inspired by biological neurons, which receive signals from multiple inputs and transmit an output signal once a certain activation threshold is reached.

A Multilayer Perceptron (MLP) consists of at least three layers: an input layer, one or more hidden layers, and an output layer. The architecture may include multiple hidden layers, each containing several neurons. Because every neuron in one layer is connected to every neuron in the next, the MLP is considered a fully connected feed-forward network in which information flows from the input layer to the output layer without forming cycles. An MLP that contains at least one hidden layer is known as a universal approximator, meaning it can approximate any finite continuous function with an arbitrarily small error (Gao et al., 2015).

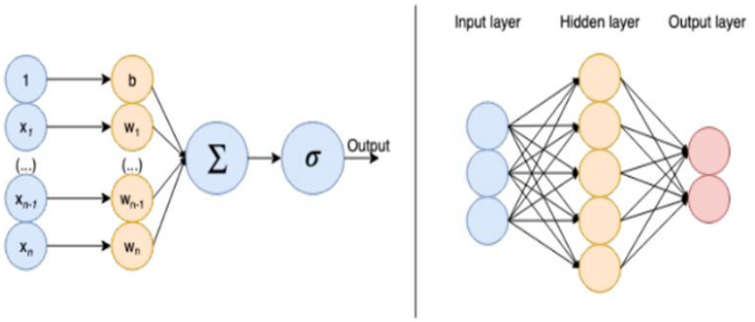



Figure (2-1) Perceptron model versus multilayer perceptron model

**Convolutional Neural Networks for Computer Vision**

A type of deep neural network widely used in image processing is the convolutional neural network (CNN). One of the problems with using MLPs for image processing is the high dimensionality of the input (Krishna et al., 2017). The number of all pixels and weights (parameters) needed to learn, so that each hidden layer has a weight for each pixel, will be very large, even in low-resolution images. CNNs combine smaller, simpler patterns into larger, more complex patterns using their understanding of the data structure. In other words, images may be combined into a format that is easier to transport, while preserving key information and spatial localization in a way that standard MLPs cannot(Zhu et al., 2016).

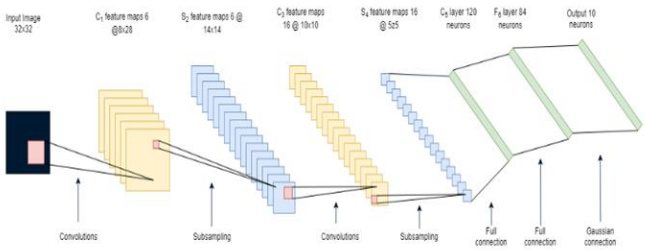



**Figure (2-2) LeNet CNN architecture from Lecun et al.**

Convolutions, nonlinear projections, and pooling operations are all performed by a CNN on the input. While the final layers identify the high-level features that are specifically at work, the initial layers identify low-level aspects of the image, such as corners and edges. These processes act as automatic feature extractors, and the output is fed into a fully connected network to produce the final output of the network. In the final layer, a softmax function is typically used to provide a probability distribution over the possible labels for classifying the input. On the input, a sliding window called a kernel performs the convolution process. This kernel is formed by the inner product between the kernel weights and the input that the kernel has chosen, and produces a feature map. Stride is the number of pixels that the filter moves per calculation. While a larger stride may ignore some input locations, reducing the dimensionality of the feature map may still be useful. Padding can be used to apply the convolution process to each input point. In other words, it places a frame of pixels around the image, usually with zero padding. The convolution operation is performed on each image channel because images are typically composed of many channels (Johnson et al., 2017).

**Residual Connections and Dense Networks**

Over the past 10 years, convolutional neural networks have consistently produced improved results on image classification challenges. The success of this approach is largely attributed to the deeper networks that are made possible by adding more layers. He and colleagues showed that adding layers to a basic model alone is not enough because deeper models often perform worse. This problem is exacerbated by vanishing gradients, although He and colleagues suspect that deeper models may perform worse than shallower models because deeper layers make it more difficult to map the identity function to zero.

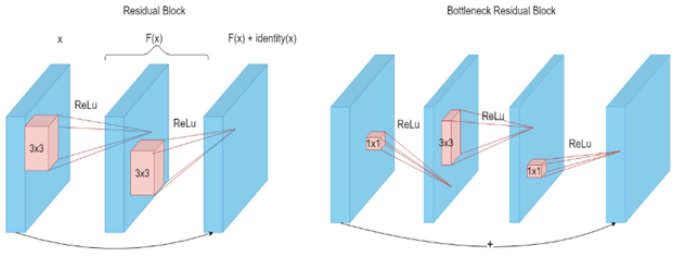



**Figure (2-3) Residual blocks and bottleneck residual blocks used in ResNets[11].**

**Transformers**

Sequence-to-sequence models that use recurrent or synchronous neural network designs, exploiting encoder and decoder architectures, are constrained by sequence operations (e.g., the computation of an RNN hidden state depends on the outcome of the previous hidden state). Vaswani et al. (2017) ] introduced a transformer design that depends only on attention mechanisms. This allows for parallelization, which significantly speeds up the training process. Dependencies may be modeled by attention mechanisms, regardless of how far apart they are in the input or output sequences (Vaswani et al., 2017).

The transformer architecture is encoder-decoder. The encoder consists of a stack of layers (the authors initially chose N = 6) with a feedforward neural network and an automatic attention layer as two sublayers in each layer. Each of the two sublayers has a residual connection that is applied to it before normalization. The decoder similarly consists of a stack of layers, in this case consisting of three sublayers. In addition to the two encoder sublayers, a third layer monitors the output of the encoder stack. Each sublayer also has a residual connection that is then normalized. Applying masking to the automatic attention layer modifies the decoder, preventing it from paying attention to input locations that are after the position being processed.

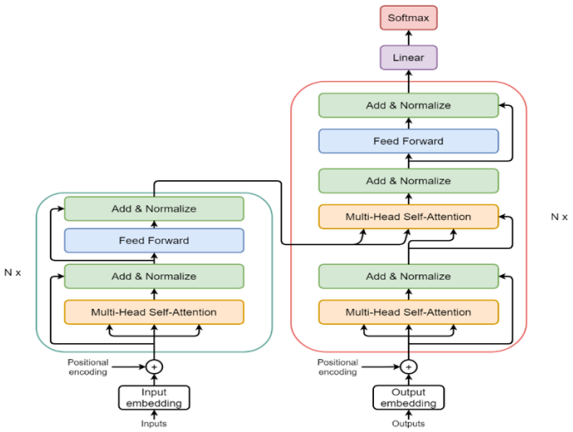



**Figure (2-4) Transformer model architecture, with the encoder on the left and the decoder on the right[14].**

**Visual Question Answering (VQA)**

The theoretical foundations of the methods used in the visual question answering (VQA) task were reviewed in the previous sections. VQA involves providing open-ended, free-form responses to questions about an image. This task formally calls for a model to learn to match (Q, I) → A a question Q and an image I to an answer A. The model must combine linguistic and visual knowledge to produce an accurate answer. Three major guidelines have been used in most VQA studies to tackle this task. The first principle is that the task is considered as a supervised-learning problem, where a model is trained using pairs of questions and images that match the desired answers. The second principle requires perceiving the task as a classification challenge rather than a production problem. This means that instead of producing an open-ended response, a probability distribution over

the most likely answer is expected. An answer as open-ended as possible would be ideal because it best represents the diversity that exists in the real world. However, this would increase the task complexity, which is already high for a neural network. Furthermore, in some of the most important datasets used in VQA research (Weu et al., 2016) , the answers often come from a relatively narrow range of words. Therefore, it is possible to reduce the possible set of answers and apply a classification strategy that requires a forced selection from a possible set of N answer words. It was a wise decision to start the study in this challenging field with this approach.

**Joint Embedding**

In the joint embedding method for VQA, a classification structure is superimposed on a joint embedding of textual and visual embeddings to arrive at an answer. There are several methods for combining textual and visual embeddings. The more popular method is to use independent fully connected layers that first map the embeddings to a common space. This is then combined using a method such as component multiplication. Different models extract textual and visual embeddings from the question and image, respectively. In most cases, a final softmax layer is placed between the two fully connected layers that form the classification structure. The joint embedding is used as input in order to predict the most likely answer. Figure (2.6) shows a typical illustration of this system(Reimers & Gurevych, 2019).

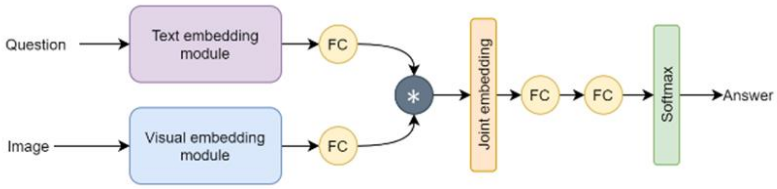



**Figure (2-6) shows the framework of the common embedding method for visual queries. The textual and visual embedding modules are two different but interchangeable concepts or techniques. A fully connected layer is represented by FC (Devlin et al., 2019).**

**VQA-Med-2021**

VQA-Med-2021 The ImageCLEF 2021 competition will release the VQA-Med-2021 dataset. The same guidelines used to generate VQA-Med-2020 will be used to create VQA-Med-2021. The VQA-Med-2020 dataset is the same as the training set. Medical experts will manually evaluate the newly collected dataset and the validation set and test set.

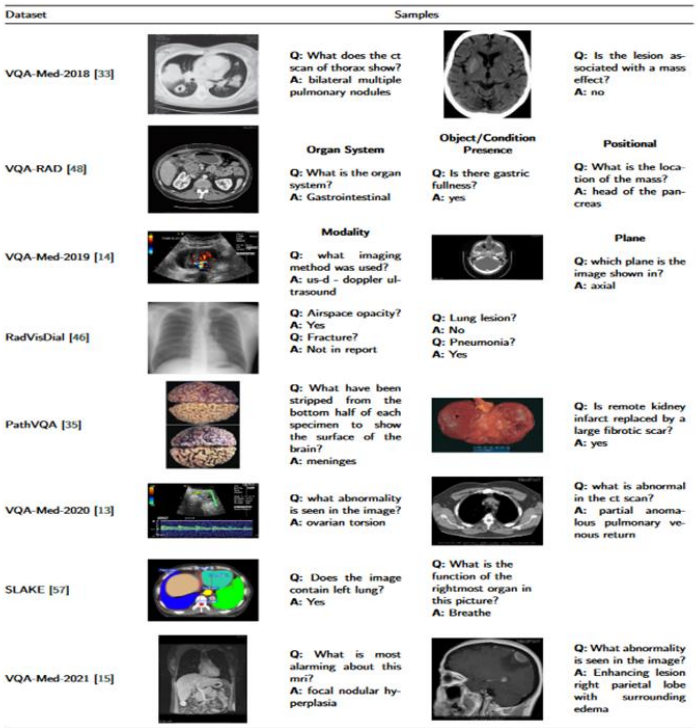


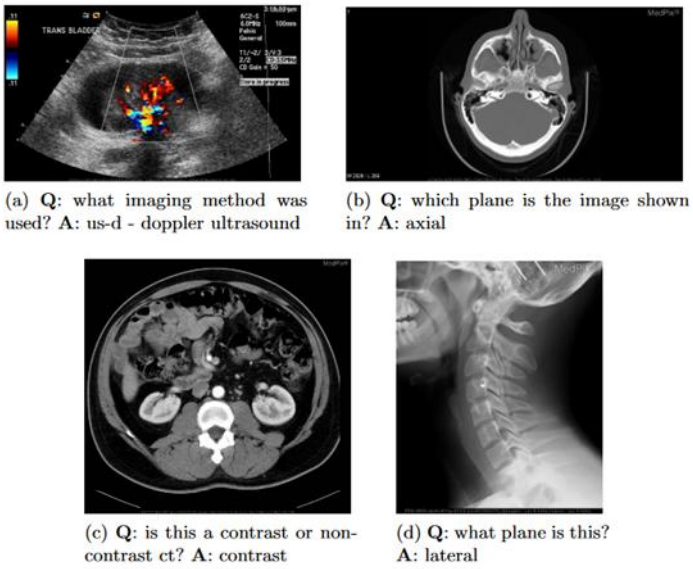

**Figure (2-7) Examples of images and question-answer pairs from the aforementioned data sets. Q = question, A = answer. The data sets are presented in chronological order .**

## IV. Proposed Method

This study discusses the selection of medical images for the VQA-Med-2019 dataset, including types, contexts, situations, and diagnostic methods. It also details the question categories and formats covering various medical imaging scenes. The dataset consists of 3,200 images for training and 500 images for validation. The test set undergoes manual binary validation by a physician and a radiologist. Finally, this chapter presents the proposed method for feature extraction and integration into the VQA-Med-2019 dataset. This chapter explains the steps for extracting image features using the VGG19 network and the multimodal bilinear (MFB) analysis technique to improve representation by detecting interactions and dependencies between image features and question features.

**Dataset**

The VQA-Med-2019 dataset was built through an automated process to create training, validation, and test sets. This process involved applying various filters to select relevant images and their associated annotations. In addition, templates were created to generate questions and their corresponding answers. To ensure the quality and accuracy of the test set, it was manually validated twice by two expert clinicians. The evaluation also noted that the dataset is publicly available, allowing researchers and practitioners to access it for their own research. Figure 1 shows examples of the VQA-Med-2019 dataset, providing visual representations of the data in the dataset.



(a) **Q**: what imaging method was used? **A**: us-d - doppler ultrasound

(b) **Q**: which plane is the image shown in? **A**: axial

(c) **Q**: is this a contrast or non-contrast ct? **A**: contrast

(d) **Q**: what plane is this? **A**: lateral

**Figure 3 - 1 Examples from the VQA-Med-2019 test suite**

**Medical Images**

For the VQA-Med 2019 dataset, relevant medical images were selected from the MedPix5 database using various filters. The selection process included considering image titles as well as criteria such as types, levels, locations, categories, and diagnostic methods. Specifically, only those cases where the diagnosis was based on the image itself were included in the dataset. The selection process ensured that the dataset included images where the diagnosis was made using a variety of diagnostic methods, including CT/MRI imaging, angiography, imaging appearance, radiography, imaging features, ultrasound, and diagnostic radiology. This broad, purposeful selection approach was intended to provide a diverse and representative collection of medical images for the VQA-Med 2019 dataset.

**Training and Validation Sets**

The training set of the VQA-Med 2019 dataset consists of 3,200 medical images. These images are accompanied by a total of 12,792 question-and-answer (QA) pairs, with an average of 3 to 4 questions per image. It may provide details about the categories (model, level, organ system, and abnormality) and lists the common answers associated with each category. In addition, the validation set of the VQA-Med 2019 dataset consists of 500 medical images. These images are accompanied by 2,000 question-and-answer (QA) pairs, representing an average of 4 question-and-answer pairs per image in the validation set. Both the training and validation sets provide abundant data for training and evaluating models in the VQA-Med 2019 challenge, allowing researchers to test and evaluate their question-and-answer systems on real medical imaging data.

**Medical Image Features**

The first step in your research involved using the VQA-Med 2019 dataset and extracting image features from the images. This was done using a VGG19 model. The first step of the research involved exploiting the VQA-Med 2019 dataset to extract relevant image features. To do this, a ResNet50 model was used, which facilitates the extraction of meaningful visual representations from input images. The process begins by passing the input image through a series of convolutional layers that perform spatial filtering to identify important image features. Then, batch normalization is applied to normalize the output of the convolutional layers, which improves the stability and efficiency of the network. An activation function is then used to introduce nonlinearity into the feature extraction process. Maximization layers are then used to downscale the feature map, reducing its spatial dimension and preserving vital information. This dimensionality reduction helps to discover more salient features while reducing computational complexity. To capture more abstract, higher-level features, the resulting feature map is reprocessed by passing through a sequence of residual layers. These residual layers allow the network to learn and represent more and more complex visual patterns and relationships. After the last residual layer, an adaptive averaging layer is applied to transform the feature map into a fixed-size representation.

This fixed-size representation is then reshaped into a one-dimensional tensor to facilitate further processing. The reshaped tensor is passed to a fully connected layer that applies linear transformations to the input features. An activation function and a dropout layer are then applied to introduce nonlinearity and reduce overfitting, respectively. The output of the fully connected layer is then reprocessed by another fully connected layer that produces the final image features. If the size of the image features is compatible with batch normalization (i.e., the batch size is greater than 1), the batchnorm operation is applied to normalize the image features, which promotes stable and efficient training. Finally, the image features are reshaped and transposed to obtain the shape (1472, 1), which is consistent with the desired format for subsequent calculations and analysis. This transformation and transposition process enables efficient management and use of image features in subsequent stages of research. Overall, this method, including the VGG19 model and the aforementioned layers, extracts informative features from the VQA-Med 2019 dataset, which allows for further investigation and analysis of the relationship between image content and relevant medical questions.

**Medical Question Features**

For question features, we use a separate process. Specifically, we use a Biobert model to encode and represent the question text. The Biobert model uses procedural text word embedding and captures the contextual meaning of the question. This step allows us to extract meaningful features from the question text that can be used for further analysis or downstream tasks.

The medical question feature extraction methodology includes the following steps:

1. Text preprocessing: Preprocess the input text by removing trailing whitespace.
2. Tokenization: Tokenize the preprocessed text using the BIOBERT tokenizer, add special tokens, limit or pad the text to the maximum possible length, and create attention masks.
3. Encoding: Pass the tokenized input through the tokenization layer of the BIOBERT model.
4. Encoding layers: Use the initial encoding layers of the BIOBERT model to obtain the hidden states in each layer.
5. Representation calculation: Calculate the representation of the question_word by averaging the outputs of encoding layers 10 and 11.

## V. Result

**Performance Measures**

In the proposed medical Q&A tasks, two types of performance measures are used: language-based measures and classification-based measures. The general measures used in classification tasks, such as accuracy and F1 score, are known as classification-based measures. They calculate the exact match accuracy, precision, and recall by considering the answer as the classification result. Classification-based measures are used as performance measures in all eight activities of the paper. For tasks that involve sentence evaluation, the general measures are known as language measures. The tasks VQA-Med-2018, VQA-Med-2019, PathVQA, VQA-Med-2020, and VQA-Med-2021 are among the tasks that use language-based measures. All four tasks use these four measures, including BLEU, which measures the phrase similarity between two sentences. However, BLEU was originally a benchmark for machine translation and is also used in medical report generation tasks.

Accuracy is a metric that measures the ratio of correct answers predicted by the model to the total number of instances. This metric indicates the overall performance of the model in providing accurate answers to the given questions. On the other hand, BLEU score is a metric that evaluates the similarity between the answers predicted by the model and the actual answers provided in the dataset. This metric calculates the n-gram overlap between the predicted and actual answers, taking into account both linguistic and retrieval. BLEU score provides

a quantitative measure of how well the answers predicted by the model match the actual answers. Both accuracy and BLEU score are commonly used as evaluation metrics in the visual Q&A field. Accuracy provides a simple measure of the correctness of the model, while BLEU score provides a more nuanced assessment by considering the linguistic similarity between the predicted and actual answers.

**Model Training**

During the training of the VQA model, the training accuracy, validation accuracy, and training loss were measured in each epoch. We trained the model for 100 epochs. An Adam optimizer was used in the training process to optimize our learning model. This optimizer adjusts the model parameters based on the calculated gradients to minimize the loss function and improve the model performance. The optimizer parameters are, for example, the learning rate is 0.0001 and the batch size is 64. Figure (4.1) shows precisely how the training and validation loss of the model are expressed over the course of 100 epochs. This clearly shows that the model did not suffer from overfitting, a phenomenon in which the model becomes too specialized to the training data and exhibits unstable performance on new and unknown data.

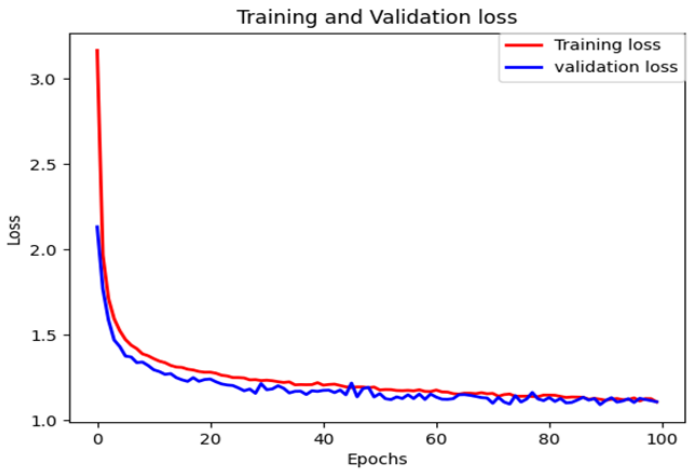



**Figure (4-1) Training / Loss of Credit**

**Superior Accuracy**

Traditional accuracy, known as p-1 accuracy, requires that the model's prediction with the highest probability be exactly the same as the expected response. This value measures the percentage of instances where the predicted label exactly matches the target singleton. Figure (4.2) shows superior accuracy for training and validation.

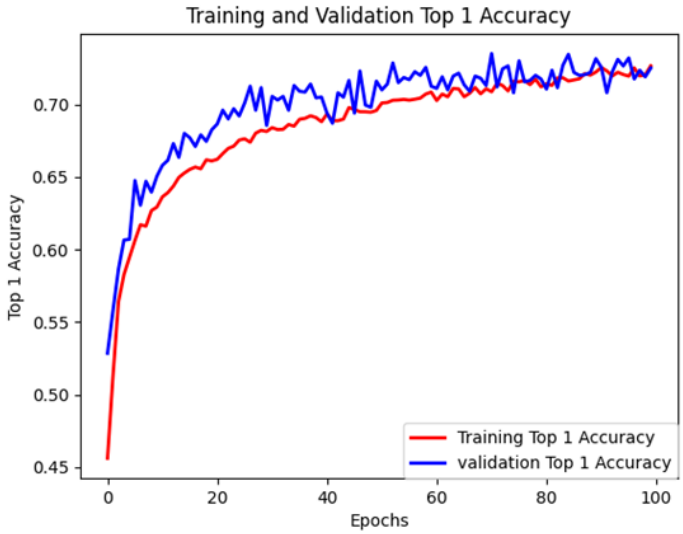



**Figure (4-2) Superior Accuracy**

Figure (4.3) shows a graph of training and evaluation scores measured using the BLEU (Bilingual Evaluation Understudy) metric for question and answer models evaluated on the VQA-Med-2019 dataset. The BLEU metric is typically used to assess the quality of machine-generated translations by comparing them to human references.

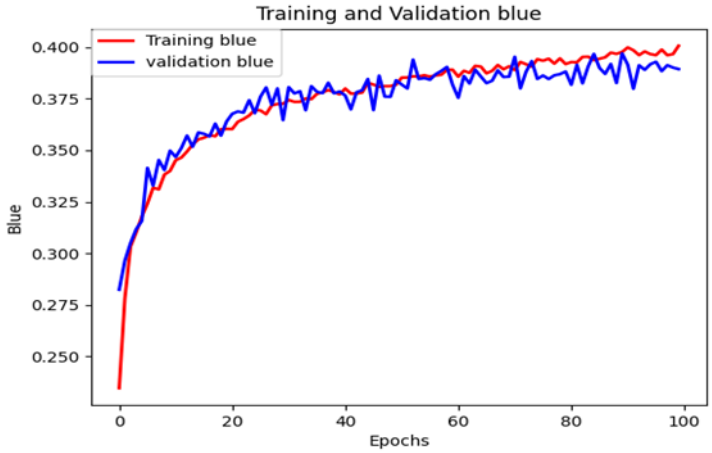



**Figure (4.3) shows a graph of training and evaluation scores measured using the BLEU**

**Results and Comparison with Baseline Studies**

Table 4.1 presents a comprehensive comparison of our proposed model with other existing models, using BLEU and accuracy metrics as performance measures. Our model shows the highest accuracy among the compared models, indicating its superior ability to provide correct answers. Furthermore, our model achieves a BLEU score of 0.4, which is considered good compared to Paper 63. However, the BLEU values of Paper 64, which include all the models, make us realize its exceptional performance in producing answers that are closely similar to human references.

**Table 1. Comparison of the Proposed Model with Baseline Methods Using BLEU Score and Accuracy**

| Method | BLEU | Accuracy |
|---|---|---|
| proposed method | 0.5 | 0.8 |
| Research Paper (Johnson et al., 2019) | 0.21 | 0.398 |
| Research Paper (Thanki & Makkithaya, 2019) | 0.460 | 0.13 |

## VI. Conclusion

In this study, we conducted experiments and presented results for the visual question-and-answer (VQA) task in the medical domain, specifically using the VQA-Med 2019 dataset. We evaluated the performance of our proposed model and compared it with existing and state-of-the-art models. The experimental results showed that our proposed model outperformed the compared models in terms of accuracy and demonstrated superior ability to provide correct answers to medical questions. Furthermore, our model achieved a good BLEU score of 0.5, indicating its efficiency in producing answers that match well with human references. Furthermore, we visualized and analyzed the model output for selected images from the VQA medical dataset. This visualization provided insights into the model performance by showing cases where the model correctly predicted the answers and cases where it was far from reality. Overall, our experimental results confirm the effectiveness and potential of our proposed model for visual question-and-answer tasks in the medical domain. These results indicate that the model has the ability to understand and answer medical questions and is effective in providing valuable insights and contributions in the medical field.

## References

[1]. Almazan, J., Gordo, A., Fornés, A., & Valveny, E. (2013). Handwritten Word Spotting With Corrected Attributes. In Proceedings Of The IEEE International Conference On Computer Vision (Pp. 1017-1024).
[2]. 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).
[3]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-Training Of Deep Bidirectional Transformers For Language Understanding. In Proceedings Of The 2019 Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies, Volume 1 (Long And Short Papers) (Pp. 4171-4186).
[4]. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., & Xu, W. (2015). Are You Talking To A Machine? Dataset And Methods For Multilingual Image Question. Advances In Neural Information Processing Systems, 28.
[5]. Gupta, A. K. (2017). Survey Of Visual Question Answering: Datasets And Techniques. Arxiv Preprint Arxiv:1705.03865.

[6]. Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., Peng, Y., ... & Horng, S. (2019). MIMIC-CXR-JPG, A Large Publicly Available Database Of Labeled Chest Radiographs. Arxiv Preprint Arxiv:1901.07042.
[7]. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A Diagnostic Dataset For Compositional Language And Elementary Visual Reasoning. In Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (Pp. 2901-2910).
[8]. Kafle, K., Yousefhussien, M., & Kanan, C. (2017, September). Data Augmentation For Visual Question Answering. In Proceedings Of The 10th International Conference On Natural Language Generation (Pp. 198-202).
[9]. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017). Visual Genome: Connecting Language And Vision Using Crowdsourced Dense Image Annotations. International Journal Of Computer Vision, 123(1), 32-73.
[10]. Li, Q., Tao, Q., Joty, S., Cai, J., & Luo, J. (2018). Vqa-E: Explaining, Elaborating, And Enhancing Your Answers For Visual Questions. In Proceedings Of The European Conference On Computer Vision (ECCV) (Pp. 552-567).
[11]. Reimers, N., & Gurevych, I. (2019, November). Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks. In Proceedings Of The 2019 Conference On Empirical Methods In Natural Language Processing And The 9th International Joint Conference On Natural Language Processing (EMNLP-IJCNLP) (Pp. 3982-3992).
[12]. Ren, M., Kiros, R., & Zemel, R. (2015). Exploring Models And Data For Image Question Answering. Advances In Neural Information Processing Systems, 28.
[13]. Shortliffe, E. (Ed.). (2012). Computer-Based Medical Consultations: MYCIN (Vol. 2). Elsevier.
[14]. Thanki, A., & Makkithaya, K. (2019, January). MIT Manipal At Imageclef 2019 Visual Question Answering In Medical Domain. In CLEF (Working Notes).
[15]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention Is All You Need. Advances In Neural Information Processing Systems, 30.
[16]. Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Dean, J. (2016). Google's Neural Machine Translation System: Bridging The Gap Between Human And Machine Translation. Arxiv Preprint Arxiv:1609.08144.
[17]. Zhu, Y., Groth, O., Bernstein, M., & Fei-Fei, L. (2016). Visual7w: Grounded Question Answering In Images. In Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition (Pp. 4995-5004).