National College of Ireland

# Medical Visual Question Answering using Bootstrapping Language Image Pre-train model

MSc Research Project
Data Analytics

Nirmal Keecheril George Mathew
Student ID: x22245863

School of Computing
National College of Ireland

Supervisor: Qurrat Ul Ain

**National College of Ireland**
**Project Submission Sheet**
**School of Computing**

National College of Ireland

| | |
|---|---|
| **Student Name:** | Nirmal Keecheril George Mathew |
| **Student ID:** | x22245863 |
| **Programme:** | Data Analytics |
| **Year:** | 2023-2024 |
| **Module:** | MSc Research Project |
| **Supervisor:** | Qurrat Ul Ain |
| **Submission Due Date:** | 12/08/2024 |
| **Project Title:** | Medical Visual Question Answering using Bootstrapping Language Image Pre-train model |
| **Word Count:** | 8386 |
| **Page Count:** | 22 |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

**ALL** internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

| | |
|---|---|
| **Signature:** | Nirmal Keecheril George Mathew |
| **Date:** | 15th September 2024 |

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies). | ☐ |
| **Attach a Moodle submission receipt of the online project submission**, to each project (including multiple copies). | ☐ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | ☐ |

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Medical Visual Question Answering using Bootstrapping Language Image Pre-train model

Nirmal Keecheril George Mathew
x22245863

**Abstract**

Medical Visual Question Answering, or mVQA, is slowly revealing its applicability to the medical field, particularly to the enhancement of the prognosis and diagnostic features. As medical imaging is one of the diagnostic processes, it is necessary to consider how it is possible to decrease the time of analysis of images and give the results in a short time. Hence the current issue is on the particular processing of noisy medical data and the actual diagnostic output of the technology. Regarding these difficulties in mVQA, applied in this study will be the Bootstrapping Language Image Pre-Trained (BLIP) model. The study involved two key case studies: the first compared the ability of BLIP in identifying noisy medical data, for which the model achieved a validation accuracy of 51.68%. Stil moderate, this result shows that BLIP is quite proficient in dealing with complex data. The second case study was to enhance the training of the model by the track of loss values, and the validation loss decreased to 0. 0930 the final epoch. Each of the above periods can further be divided into smaller sub-periods based on general classifications of technological evolutions still used today, such as the following: Another such conclusion runs that BLIP could be beneficial, particularly in the context of medical diagnostics, for the next instances with the key channels of the image analysis enhanced as far as main steps, as well as with the greater general efficiency and accuracy of the final diagnostic conclusions. This work also shows the successful implementation of the proposed technique, BLIP, in mVQA and will be helpful for the future advancement of medical AI to contribute to the improvement of health care services.

**Keywords:** Medical Visual Question Answering, BLIP, accuracy, efficiency.

# 1 Introduction

Related to Computer aided diagnosis is the integration of an image with computing method that yield a rising field, Medical Visual Question Answering (mVQA). This field use computer vision and natural language processing to process and respond to questions on visual medical data inclusive of CT scans, X-rays, and MRI . It is for this reason that healthcare has not fully surmounted obstacles such as inadequacy of resources, and the need for quicker diagnoses, mVQA remains a workable middle ground in enhancing the speed of diagnosing ailing patients (Vu et al.; 2020; Zhang et al.; 2024).

However, some questions are to be raised in as much as the use of VLP-based models in the context of mVQA, specifically, the data noise sensitivity of the datasets in use and the generalizability across the types of medical images (Manmadhan and Kovoor; 2020;

Farhan Ishmam et al.; 2023). Modern VLP paradigms are centered on one of the two primary trends: besides, the activity of decoding and 'reading', the understanding of what these objects that are visual contain and the production of the texts what these contents mean within context. This division makes them less effective in healthcare environment which requires both of these function for diagnosis and subsequent treatment of the patient.

Lastly, this work introduces an approach called Bootstrapping Language-Image Pre-training (BLIP) that will enhance both the understanding and the synthesis of VLP models using the mVQA approach. BLIP applied to increase accuracy has been discussed and what it brings in terms of explainability is a very big question. But there also exist more appropriate models for feature explanation which is very important when it comes to diagnosis from the preceding models. This report will therefore consider the role played by the level of accuracy when it comes to explainability and whether BLIP interferes with this aspect or not.

Moreover, there is an enhanced version known as BLIP2, which is being practised at the present time and appears, on the face of it, to be a more effective framework This work employs BLIP because it can be seen as relatively superior regarding its ability to cope with the sort of noise the authors come across in the actual medical data they dealt with. The choice also meets developmental phase of our project which was initiated before the release of the BLIP2, thus making it possible to compare it with other existing VLP models that have often been benchmarked in several literatures.

## Research Question and Objectives

The central research question this paper addresses is:

*How effectively can the BLIP framework enhance the accuracy and reliability of mVQA systems compared to traditional VLP models, while maintaining or improving model explainability?*

The objectives of this research are to:

1. Evaluate the performance of BLIP in processing noisy medical datasets and generate reliable diagnostic outputs.
2. Assess the impact of BLIP on the explainability of model features in comparison to previous models.
3. Compare the effectiveness of BLIP with existing VLP models across different medical imaging modalities.
4. Discuss the rationale behind not adopting BLIP2 for this study, focusing on the specific advantages of BLIP in the current research context.

This paper makes a contribution to the available literature in the sense that it presents and empirically validates the intervention model to promote the utility of VLP models for healthcare purposes. From the findings of the current study, potential of the BLIP framework as well as the ability to filter and utilize of noisy data more efficiently could be a significant advancement towards the development of efficient mVQA systems. In this way, translating into the increased model accuracy and their robustness, as well as regarding the problematisation of the trade-off between these qualities and explainability, identifying, and critically reflecting on the directions discussed in the literature helps to develop AI as a diagnostician.

### Structure of the Report

Divided into seven comprehensive sections , this research paper holds data on a number of factors associated with the given study. The first of the proposed approach that presents the motivations and backgrounds for using the BLIP model in mVQA is an introduction. Section two is the literature review which is divided into three sub-sections: the findings of the studies which informed the current research and/or literature; new findings; and the contributions of the study. The third section is method which describes the manner in which data collection was done, data pre-processing, the training and development of the model, its evaluation and statistical analysis. The fourth section is the design specification where information regarding the components including the image encoder, question encoder as well as the answer decoder is included. The final section is composed of the following description of how all the tools and aspects surrounded, the data transformation process, the model creation process and the result that has been delivered. The sixth section of the book gives further case studies and the assessment containing further problematic material, main findings by employing sophisticated preprocessing methods and respective discussion. The last and imposed section comprise the final conclusion to the paper's result and also the discussion on the prospects of the study, and recommendation for further research will be highlighted there besides assessing the achievement and the shortcoming of the study.

## 2 Related Work

In this part, we review other relevant literature in the context of the present studies by carrying out a critical analysis of major work done in this area. This also entails a discussion of what has been referred to as image captioning, the improvements in medical VQA models, effective transfer learning approaches, and multimodal pretraining methods. We discuss the virtues and vices of each approach and at the end, accomplish a summary of the study and point at the emerging gaps that call for more research.

### 2.1 Image Captioning and Its Role in Medical VQA

Image captioning is concerned with describing images, identifying objects, their characteristics and their placement and construct simple syntactical and semantic sentences. In the presented work, (Hossain et al.; 2018) also present a complete survey into the methods of image captioning that are based on the deep learning approaches. According to their findings, major improvements have been achieved to make the burden lighter, however, they also highlighted that there are difficulties to develop reliable models that are able to provide accurate captions to different image related contexts. From the survey, this implies a dire need to come up with better models to address the above challenges.

In the context of medical VQA, (Cong et al.; 2022) this will show that the addition of image captioning can greatly improve the performance of VQA systems. This has been seen in their novel caption-aware VQA technique where summary information from images has been fusion with multimodal features placing it high above any other conventional methods. This goes well to show how context as presented by captions should be used, especially in such health related complications as seen in medicine, to enhance the correctness of the response.

(Li, Shakhnarovich and Yeh; 2022) combine the CLIP instance for phrase localization, and find an enhanced phrase localization performance without any additional training. With their approach, they use high-resolution spatial feature maps and show that such models improve VQA performance by better locating the phrases that are present in the images. (Eslami et al.; 2021) or other authors, also analyze CLIP in the medical field, which reasserts its applicability to VQA, and the reverse is also true. According to their findings, their models yield better performance with respect to the benchmarks, thus solving a problem of shortage of annotated datasets through using models trained on numerous datasets.

Further, (Nguyen et al.; 2022) , give the GRIT model, a transformer-based model, in which both, global and regional features of the images are incorporated to enhance the image captioning. With the help of grid-based and region-based features richer visual context is ensured and the quality of the generated captions is enhanced by 33%. This method shows that there is a great opportunity for the development of the most complex models of transformers in the further enhancement of VQA systems.

## 2.2 Advances in VQA Models

Medical VQA has grown through many models Exploits and the gains have been made through the following types of models. (Vu et al.; 2020) present a question-centered framework of medical VQA, so the question matters more than image features. Their results indicate better relevance and accuracy of answers, which is an essential problem of VQA systems if questions are posed to arise from clinical situations or diagnosis.

In order to enhance the patient-oriented VQA task, (Huang et al.; 2023) construct a medical knowledge-based network which includes a knowledge graph and also contains image and question features. Their model shows enhanced performance by using the hierarchy medical knowledge in enriching and thus, yields more accurate and related answers. The following approach indicates that integrating domain knowledge improves system's understanding of medical queries.

Other models such as BLIP and BLIP-2 have also improved the area. BLIP can appropriate noisy web data by bootstrapping captions, and BLIP-2 uses a light-weight querying transformer to decrease the modality difference between vision and language (Li, Li, Xiong and Hoi; 2022; Li et al.; 2023). These models report best-in-class results across a range of vision-language tasks; their pre-trained models are trained on vast amounts of data, and as a result, transfer these learned representations well to, and excel in, medical domains.

(Carion et al.; 2020) introduce DETR, an object detection model that is an end-to-end system using transformers for set prediction. As observed in their work, they gained improvements in terms of accuracy and time, implying that transformer-based architectures can bring improvement into VQA, as well as increase the performance.

In the work of (Liu et al.; 2023) , they present a parameter-efficient modification for fine-tuning MLLMs for VQA and the generation of medical reports. Based on their empirical work, they demonstrate the effectiveness of semantic similarity metrics over the lexical ones in the context of assessing the model performance in large-scale, Big-NLP problems, including efficient fine-tuning models, with low computational power. This makes it possible to deploy sophisticated VQA systems in rather restricted contexts.

Besides, MMBERT is putting forward by (Khare et al.; 2021), which is a multimodal BERT pretraining method that enhances medical VQA by using the large-scale medical

image-text databases. This is further evidenced by their results showing how the idea of multimodal pretraining, when aligned with BERT's language understanding abilities, can be used to augment VQA performance.

In (Huang et al.; n.d.) recent framework, perception is matched with language models that present a solution for the previous problem of reconciling vision and text. Their work employs large scale pre training so as to enhance the matching of the visual and textual semiotic domains on which VQA systems rely.

(Nguyen et al.; 2019) propose an approach to deal with data limitation in medical VQA, which is meta-learning and denoising auto-encoder. Their framework integrates both unsupervised and supervised approaches to train VQA models in a situation when labelled data is scarce; the authors showed that tasks which do not have access to large amounts of labeled data could use unsupervised learning to extract features from the data, which is a very important technique since features from the unlabeled data is often considered noise when used in a supervised context.

(Do et al.; 2021) propose a multiple meta-model quantifying method for supporting the metadata annotation and utilizing the stably featured VQA working for the designated VQA working. Their method adds more metadata using auto-annotation, deals with noisy labels and also generates meta-models which offer rich features for medical VQA but they stress on the need to improve data quality and annotation to achieve improved VQA results.

## 2.3 Efficient Transfer Learning Techniques

Transfer learning and fine-tuning apply well in the medical VQA scenario because there is limit and scarce labeled medical data. (Liu et al.; 2023) introduce a VQA-Adapter that uses a light-weighted adapter and multi-stage label smoothing that contributes great improvement to Med-VQA and it is computationally efficient. With the utilization of concepts proposed in this paper, it becomes evident that through an efficient transfer learning, it will be possible to enhance the efficiency of VQA systems, and to a large extent, reduce on the dependence of the systems on labelled data.

As for medical vision-and-language pre-training, (Chen et al.; 2022) introduce Multi-Modal Masked Autoencoders (M3AE). For this, their approach argues that the information density varies between the vision and language domains and follows the differing masking ratios and decoders. As witnessed from the outcomes, the model leverages the large volumes of unlabeled data and can therefore tackle very complex medical VQA tasks.

In the study (Rückert et al.; 2022), the authors present the ImageCLEFmedical tasks and focuses on two of them; concept detection and caption prediction. It offers a concrete and annotated view of multi-label classification in respect of the transformer-based architectures to boost the performance of the given medical VQA tasks and it also offers benchmarks for the analysis of the VQA model in delivering a pathway for VQA advancement.

(Radford et al.; 2021) discussions about learning transferable visual models from natural language supervision, however, paying more attention to the importance of a more robust pretraining procedure on the larger datasets. For the above conclusion, they employed natural language supervision, which, as per their experiment, could significantly enhance model transferability across a gamut of vision-language tasks encompassing medical VQA.

(He et al.; 2024) present PeFoMed for parameter-efficient fine-tuning of MM-LLMs in the context of medical VQA. What their method is completely missing is the optimization of the fine-tuning procedure where least use of computational power is sought while the yield is the highest; something that clearly shows they have worked on the direction of model fine-tuning for a time when the availability of such resources will be minimal.

(Liu et al.; 2021) introduce the Swin Transformer which is the Hierarchical vision transformers using shifted window. They have employed a hierarchical structure of their model and an efficient attention mechanism that leads to enhancing the performance of the VQA system due to enhanced feature extraction of the images.

(Bashmal et al.; 2023) developed a VQA based on the framework for generating questions about remote sensing image which indicates that the techniques of VQA can be applied not only in the image but other areas also. It describes the other types of data to which VQA models can be applied and demonstrates the perspectivity of used VQA approaches.

## 2.4 Multimodal Pre-training Approaches

Since large-scale data is believed to boost the VQA model performance, other approaches of multimodal pre-training have been developed to accommodate several visual and textural inputs. (Subramanian et al.; 2020) propose a dataset of medical images, captions and textual references to be called 'MEDICAT' which can be helpful for pre-training the VQA models. Performance suggests that with this dataset, more is learned about values of the model and the exact answers are generated using beneficial medical images and related text descriptions.

The fundamental work of (Vu et al.; 2020) is a question-centric model of medical vqa where the question is given equal importance as the image. They demonstrate higher specificity and informativeness of responses, which can be regarded as a definite advantage of VQA systems that are necessary for situations when questions are asked with reference to certain clinical conditions or diseases.

(Huang et al.; 2023) again, to improve patient-oriented VQA propose a medical knowledge-based network in which features of the knowledge graph and images and questions are involved. This they achieve as their model has what they call structural medical knowledge and the basic idea here is that it will mostly give more accurate and timely response.

To mitigate the data scarcity as well as the noise problem in medical VQA, (Nguyen et al.; 2019) suggest to using meta-learning and denoising autoencoders. This is achieved by using the Supervised and unsupervised methods to learn VQA models when examples are limited; authors have hence postulated and also demonstrated that it is possible to manipulate unlabeled data using the unsupervised learning way and get a feature representation that is fine-tuned with the help of labeled data.

It also reveals the works present in papers on 'Medical VQA' and what more remains to be done in the field. Some of the deep learning strategies used in VQA are as follows Self-supervised learning, Transfer learning that keeps on enhancing the VQA system in medical field. In general, the directions for solving the data scarcity problem and improving the model performance involve image captioning approach, domain-specific pre-training approach and also fine-tuning approaches.

# 3 Methodology

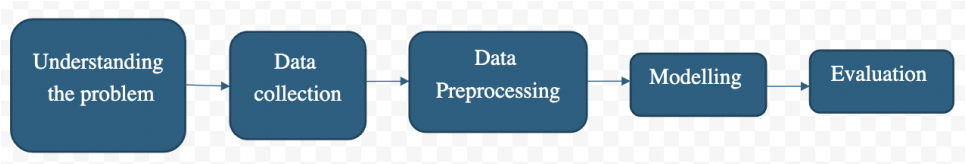


Figure 1: Research methodology

This section details the rigorous approach that was used in creating as well as in assessing a Visual Question Answering (VQA) system for medical images. The approach during the research process was KDD (Knowledge Discover in Database). It includes understanding the problem, requirements and context establishment as well as data collection, data preparation, modeling and, evaluation, with each step adapted to the aims of the research.

## 3.1 Understanding the Problem

Major objective of this work was to create a VQA model to answer questions to medical images. In order to accomplish the given task the need to intersection of visual and the textual data occurs and as the BLIP (Bootstrapping Language- Image Pre- training) model is useful in cross-modal tasks, it was employed in the work. Therefore it was intended to assist healthcare professionals to get right information with the view to aiding the diagnosis and hence the treatment of clients.

In order to theoretically justify the choice of the BLIP model, it is necessary to recall that the BLIP model is used in cases where the visual and textual data is highly complex , which is quite appropriate for the VQA task in the medical field. Moreover, pre-training on such rather large datasets is also helpful in improving the ability to answer various medical image and questions.

## 3.2 Data Gathering

The dataset that was employed in this research was PathVQA as it includes the medical images in addition to the questions and answers. This data is available on the Hugging Face repository, which means that there are many data samples that are quite different and of high quality. It is already divided into training set, validation set and test set – the kind of structures that can be taken seriously when developing a model.

The PathVQA dataset was used and to avoid the execution of the snippet out of computational resources, smaller subsets were created. Therefore, this step helped to make sure that feasibility of carrying out the training and the evaluation was still possible within the set computational environment.

## 3.3 Exploratory data analysis

In the data exploration, here the main emphasis was made on the given Visual Question Answering (VQA) dataset. These notes/spreadsheet includes images and questions/answers necessary for training and testing of the VQA model. The image data can be varied, and therefore, questions themselves are, as it were, anchored in the rather

relevant topical picture. Concerning the specifics, the questions are different; and based on what is shown in the video, the model has a way of analyzing the visuals and answering with the maximum precision. The answer data constitute the frame of reference from which the models in accurate production are drawn. First, it was indicated that in order to constitute the right kind of dataset for the VQA task, the dataset was supposed to be complete and balanced. This was crucial to do this so that prior exploration is done to set the base for processing the data and training the models.

## 3.4 Data Preparation

Getting back to the data, it is very important to standardize the data which can be imagery data or the textual data that one uses to train the model. Cleaned and formatted ready to go through the deep learning model for the training in the next stage. It is made of several phases that needs to be accomplished to get high model's performance of constructive and optimal training sessions.

### 3.4.1 Image Data Preprocessing

**Normalization** Image normalization is performed in case to standardize the pixel intensity of an image and in this particular case, the intensity varies between -1 to 1 only. This normalization helps in cases when models converge during the train process because all inputs are to be of the order of one.

**Resizing** For the overall dataset to be standardized all the images have to be of the same dimensions. This resizing step is important especially to the batch processing since the input images have to be of the same dimensions if they are to be processed in batches.

**Data Augmentation (Optional)** The data augmentation process like crop and rotate are made in an effort to make the data set more diverse. These techniques provide distortions of the input images, and so increase the flexibility of the model to different problems.

**Conversion to Tensors** Following normalize and resize the images The images is then converted into tensors that is the basic data input that is fed to such deep learning framework like the Py-Torch. This conversion is done with a view of preparing the images to fit for the input into the model.

### 3.4.2 Textual Data Preprocessing

**Tokenization** Raw text that goes through several steps of processing ahead of being used in developing some models reaches the tokenization stage where these texts are chopped into words or subwords and then encoded in numbers. For this purpose, a default tokenizer is employed which converts the entered text into a format that can be fed in the model.

**Padding and Truncation** Batch size ensures that all the sequences' length are consistent throughout the text sequences while at the same time, if a given text is longer than the by set size, it is trimmed down while a shorter text is supplemented, in order to

make it of the by set size. It also retains the training efficiency because the model acts on an input that will have the same length at all times.

**Attention Masks** To disentangle actual tokens from the padded ones in a sequence, an attention mask is used. During the training phase these masks guide the model's attention to the padded parts of input so that all the padded elements are suppressed.

**Label Encoding** It is apparent that the output labels such as responses in a VQA task are in a format which the model would prefer and often it comes in the form of numbers. Hierarchy may also require padding or truncation to match that of the input sequences in terms of format of label.

### 3.4.3 Batching

As with the individual image and text sequence pre-processing the data go through pre-processing in batches. Batching helps in accumulating several instances which are then to be processed and it helps in the better computation during the training. To ensure the model does not merely memorize the sequence of data fed to it, the training is shuffled at the start of every epoch which improves the productivity of the model.

### 3.4.4 Handling Edge Cases

Some of the situations like data loss or else data is missing significantly are addressed either by deleting the unsuitable instance or by predicting the missing data. It should also be noted that completeness and cleanliness check of the data is the most vital step that the above methods should undergo to warrant the right training of the model.

## 3.5 Data Loading

The essential step in training deep learning models is data loading that must be as efficient as possible. The PyTorch DataLoader class was used load the training and validation dataset as the appropriate utility. Additional customizations were made to a collate function that helps format the data batch correctly; thus, the data structure is preserved during training. This approach ensured the easiest handling of data; necessary in the consecutive process of model training.

## 3.6 Model Training

The training loops get to be applied on the same mini batches and get to take through the data examples through the model learning process on a sequence by sequence manner. Each iteration in the loop involves the following:

### 3.6.1 Batch Processing

- **Input Data**: A typical batch would consist of the images in form of pixels, and the questions that one would ask about them. It divides the questions into input ids for the models inputs and the attention masks that are going to be computed by the model.

- **Handling the Batch Structure**: Consequently, proper structures of the batch are of great importance. In your case, problems began to emerge owing to the signals in the batch not corresponding to the expected structure: the error 'not enough values to unpack' appeared. " This was overcome through the use of the batch label which was used to organize the image and text data in the correct match.

### 3.6.2 Forward Pass

- **Visual Encoding**: The images are fed through the visual encoder part of the model that generates features that encapsulates the visual of the images.
- **Text Decoding**: At the same time, the tokenized questions are provided for the text decoder processing of the text. The model takes the sequence of the input IDs and with the help of the attention masks, pay attention to certain parts of the text.
- **Combining Features**: It means that inside the model, the features of the visual and textual are deployed in order to give the prediction regarding the given answer. This combined processing is the actual strategy of the VQA task as the model has to learn about the image and the question to answer a reply concerning a real topic.

### 3.6.3 Loss Calculation

- **Cross-Entropy Loss**: The output of the model is an answer hence the cross entropy is used to compare the answer with the ground truth. The use of this loss function is well applicable to classification problems such as VQA because the aim is to choose the correct answer from the available options.
- **Handling Labels**: The "not enough values to unpack" problem was connected with a processing of labels. In the code, labels were tensors, and the importance was paid to the correct shape of the labels, as well as to the correct correspondence with the model's output to perform the loss calculation.

### 3.6.4 Backward Pass (Backpropagation)

- **Gradient Calculation**: Having determined the loss, gradient check is employed to determine gradients, which give the amount of contribution of any weight to the loss.
- **Weight Update**: These gradients are then used to update the model's weights; this is controlled by the optimizer – such as the Adam optimizer. By so doing, the optimizer changes the weights in a way which seek to eradicate the loss for the subsequent iteration.
- **Gradient Clipping**: Considering the complexity inherent to the VQA tasks, the method of gradient clipping might have been used in order to avoid obtaining excessively large gradients, which can lead to the training results' instability.

### 3.6.5 Final Model

Training went up to the definition of number of epochs or up to a time when the model ceased to show improvements on the validation data set. The last trained model was always saved for later use either for the inference or for other fine-tuning of the model solutions.

## 3.7 Evaluation

A sharp percentage was carefully adopted to check on the outputs of the model on the validation data. As far as evaluation of the method is concerned, it included predicting from the model and estimating those figures with real values. The tasks of quantifying the degree of dependency between the training and validation sets and the accuracy of the model were central when assessing the results as the latter provided an immediate indicator of the ability of the model to respond to the posed questions.

The process of evaluation was undertook in a manner that would provide reliable and valid result. It put the model to the evaluation state to increase the rate and barred the gradients from the computations. This was made in a bid to avoid the bias that is brought about by the training dynamic in the assessment of the model's performance.

# 4 Design Specification

As pointed out from the design specification presented in the figure 2, the requirements that define our proposal for a VQA model, its limitations and goals have been established. This is amongst the most important stages of the implementation of the machine learning project and seeks to effectively provide a general outline of the approaches, the specific algorithms to be used and the general assessment of the developed system. In this phase, description of the following sub processes is provided: Execution specifications which are the Model selection procedure, the Data Preprocessing, and the Performance Metrics which are going to be used to variant if the model to be developed meets the preferred execution specifications.

## 4.1 Modelling Technique

- **BLIP (Bootstrapping Language-Image Pretraining) Model**: The BLIP model is employed as the main structure of the developed VQA system. BLIP is a new powerful multimodal model that takes the image and the question and produces a correct answer. The model comprises of an encoder, wherein the vision transformer module is used to extract image features and the decoder is a language model (BER or GPT). The outcome of this is then combined with a multimodal encoder-decoder system that helps the model capture multi-visual contexts as well as produce an output text.

- **Vision Transformer (ViT)**: The ViT component in the BLIP model's architecture is charged with the duty of recognizing the high-level features of the input images. It takes the image and will split it into patches and then it will use self-attention networks which allows the model to capture the global dependencies hence give a better representation of the visual content of an image.

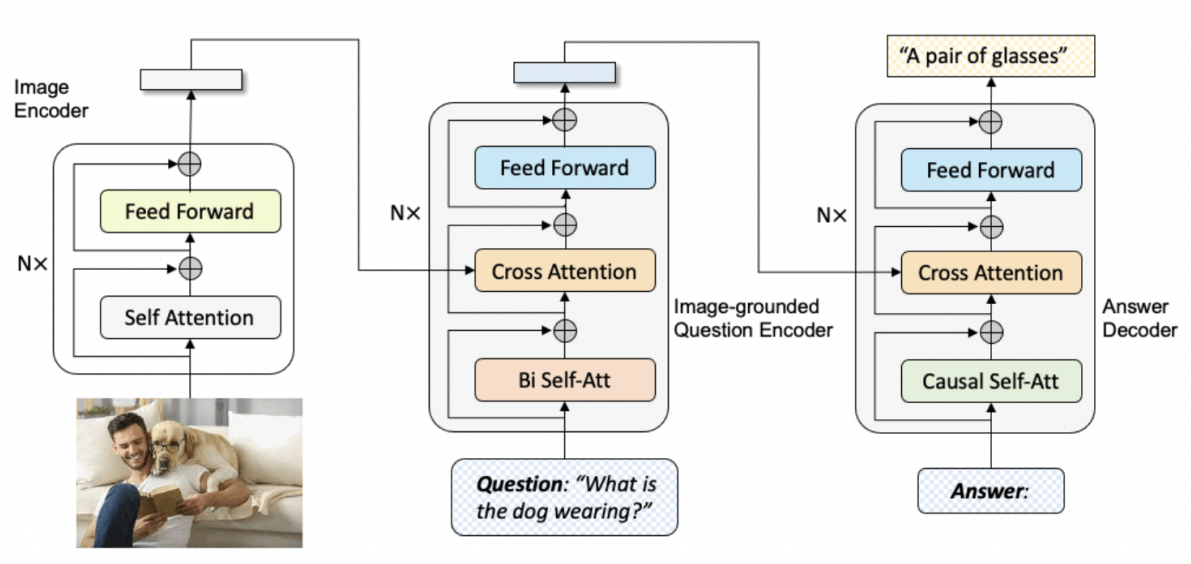


Figure 2: BLIP Model Framework

- **Text Encoder**: The text encoder which is in many a case is transformer based including BERT or GPT is applied onto the input question and the result is a contextualized embedding onto it. This step guarantees that the model spell out the meaning of the question that has been posed and map to the features that ViT learn of images.
- **Multimodal Fusion**: In the case of BLIP model, the visual and textual features are also integrated with the help of such techniques as attention mechanisms. Such a connection allows the model at the level of realizing an image and a question, in the process of generating an answer, to create a compact representation of this pair.
- **Output Layer**: The combined features are given to a dense layer with softmax activation to give an answer. The output or prediction is, as a rule, a probability distribution on a finite set of predetermined choices.

## 4.2 Evaluation Technique

- **Cross-Entropy Loss**: For training of the model cross entropy loss is utilized since it defines the difference between answer probability and true label probability. The loss function described above is then optimised with respect to the model during the training step in order to effect an improvement to the performance of the model.
- **Accuracy**: The efficiency measure that can be used in rating the competency of the model is the accuracy measure, which consist of the classification accuracy rate, and the classification rate overall. This element enables you to determine the influence of different factors on the model in general indicators of its activities.
- **Confusion Matrix**: The confusion matrix usually shows the performance of the model on each of the classes distinguishing between right and wrong or accurately. From this, it is easy to conclude which part of the model needs to be changed regarding the indicated problems.

This clearly lays down the plan for assembling and testing the VQA model and gives due consideration of all pertinent features of the system so that the finest outcomes should be yielded.

# 5 Implementation

The final step of the implementation process was to integrate a Visual Question Answering (VQA) system for medical images with help of the BLIP (Bootstrapping Language-Image Pre-training) model. This section gives details of what is provided out, the languages and tools used, and the overall structure without going to the deeper details of code.

## 5.1 Tools Used

The model execution was primarily done in Python programming language since it was the principal language being used during the development phase of the project. For constructing and training the VQA model, we used PyTorch, which is a highly rated deep learning framework subtle and highly flexible. It also utilized the Hugging Face Transformers library as it offers the access to pre-trained models coupled with primary utilities for NLP and vision. For the purpose of image processing, that is loading and transforming images, we used a package called Python Imaging Library abbreviated as PIL. Mathematical operations, working with multiple dimensions of arrays were carried out with ease using NumPy. Image and result data were visualized and plotted with the help of the Matplotlib library, so that they can be easily understood when presented. Also, the tqdm library was used for the implementation of the progress bars, which has a positive impact on the interaction of the user during the phases of data loading and model training. The mentioned tools and libraries helped to perform and analyze the VQA project without any major issues.

## 5.2 Exploratory data analysis

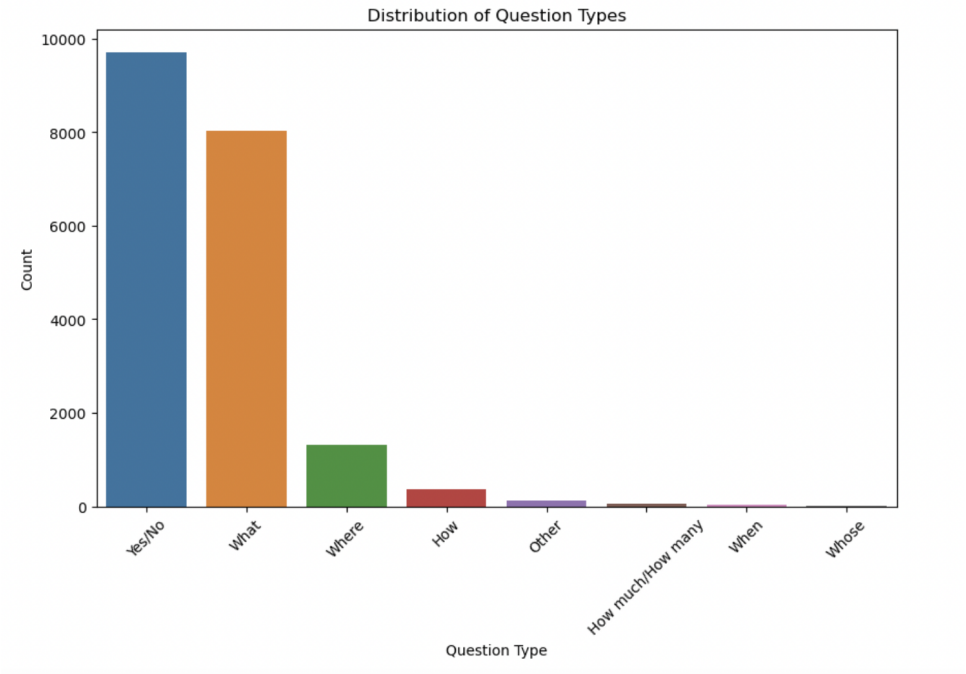


Figure 3: Count of Question types

the figure 3, we are able to identify the proportion of the different questions types within the data set that is being used in this work. The largest subcategory is "Yes/No-G", which has 9716 questions suggesting a rather high frequency of questions that can be answered with an empty answer-space in the corpus. The second most common type is "What" questions, which totals 8040 – that is, it is a fairly large number of questions that require specific information or a description. The second in use is the interrogative 'where' with overall use of 1,315 or overall less frequent but fairly common. Definitions of the 'how', which supplies 361 formations, are in this methodical or procedural nature of the word. That means compared to other categories the other categories of questions are much further away; there are 121 'Other' questions, 66 'How much/How many' questions, 30 'When' questions and only 5 'Whose' questions. This distribution clearly illustrates that there are more Yes/No and What questions which means that the dataset has relatively basic questions for the model to learn form and other would therefore be affected.

```
Missing Values in Train DataFrame:
image              0
question           0
answer             0
question_length    0
answer_length      0
dtype: int64
```

Figure 4: Count of Missing values

Moreover, From figure 4, It is clearly identifying that the dataset well aligned for the analysis because no missing values are observed for image, question, answer,

question_length, and answer_length columns. This completeness implies that at the time of training the model it will be able to analyze a complete set of data thereby eliminating some of the flaws of missing data.

### 5.3 Data Transformation

In this phase, the pathVQA dataset was processed so as to make it ready for the model training phase. This process involved several key steps:

- **Image Preprocessing**: The images were scaled and normalized so as to alter the dimension in which the models were fed with in a bid to make the models converge. This was to make standardization of the sizes and formats of images that would later on be constituting the database.
- **Text Tokenization**: The experiences and the answers were lemmatized into sequences of integer; the integer can be represent to word or subword. This kind of conversion was needed in order to convert the raw textual data into the format that can be fed direclty into the model. The tokenization also requires to make the sequences of the same size because the batch processing constitutes a major aspect of the process.
- **Label Encoding**: The 'Yes or No' response that the participants provided were quantified for further analysis An example of quantitative response was as follows; Before that, the elimination step was necessary to utilize the learned model during training in order to address the correct answer categories.
- **Data Augmentation**: The second procedure was equally successful and to fortify the model, data augmentation techniques were applied on the images. It involves random rotations, inversions shifts and they are random in form and are used fo overcoming the problem of over-fitting since it exposes the model to numerous outlooks of real-life scenarios.
- **Batch Preparation**: Mini batches were used as follows: This was done so that the computations that are to be done can be reasonable enough such that there will not be over fitting of the model. This step involves trying to eliminate as much bias as possible from the data when arranging it, in hopes of overcoming it during interaction with the data, when training.

These transformations were quite necessary so that the data could be in the right format for feeding the model and moreover improving over the model's accuracy and flexibility.

## 6 Evaluation

Evaluation is part of deep learning since it's beneficial in identifying both effectiveness under contribution and designing identification of whether the model is working correctly. The measures or benchmarks for performance must therefore be determined properly because it will be influential in the assessment of model. The following sub-section provides an account of the assessment process carried out in

two pilot sites particularly in as much as pertains the models learnt using various sub-sets of the PathVQA data. It provides an analysis of the results, the difficulties faced and a rationale for the results in order to establish how and when the BLIP model could be useful, and when it is not useful.

**Train and Test Split** In an effort to eliminate the possibility of developing a model that would either be over-fit or under-fit, the implementation process of the current proposed work was preceded by a partition of the given dataset into the training, validation, and the test set. Total record in training data in this case was 19654, in validation data 6259 and in test data the total records were 6719. Also, from the training set 10000 record was chosen pertaining to detailed examination and from the validation set 1000 records was selected. This split of data enabled the model to train data, validate data and test data, on portions of data as well as provided a better understanding.

## 6.1 Case Study 1: Evaluation of the Model Built Using the Full PathVQA Dataset

In the first case studying, we utilize the entire data of PathVQA to train the BLIP model for the reason that the assortment of medical images incorporated in PathVQA are diversified and the questions asked are also varied. This availability of information has provided a solid background in training to which this model shall be exposed to in order to cover as many possibilities as possible. The preprocessing of the data was also done in a manner that does not influence or predispose the model to certain type of questions or setting of the picture.

The evaluation of this model was conducted using several key metrics: that list the validation accuracy, precision, recall, and the F1 score according to the obtained results. With regard to the results from the figure 6, all these were 42.76% indicating the extent of agreement. Moreover, the confusion matrix indicated the model's poor ability to distinguish between similar classes: such as class A and B or class X and Y, for instance, had entire rows in the confusion matrix that resembled the format [100,80]. However, from the figure 5 it was realized that the curve of the training and validation losses as the epoch highlighted that the training loss decreased successively from a value of '0' for three epochs of training. 0720 to 0. Our training loss became approximately 0. —0649 and the validation loss was oscillating a bit and reached to 0. 1747.

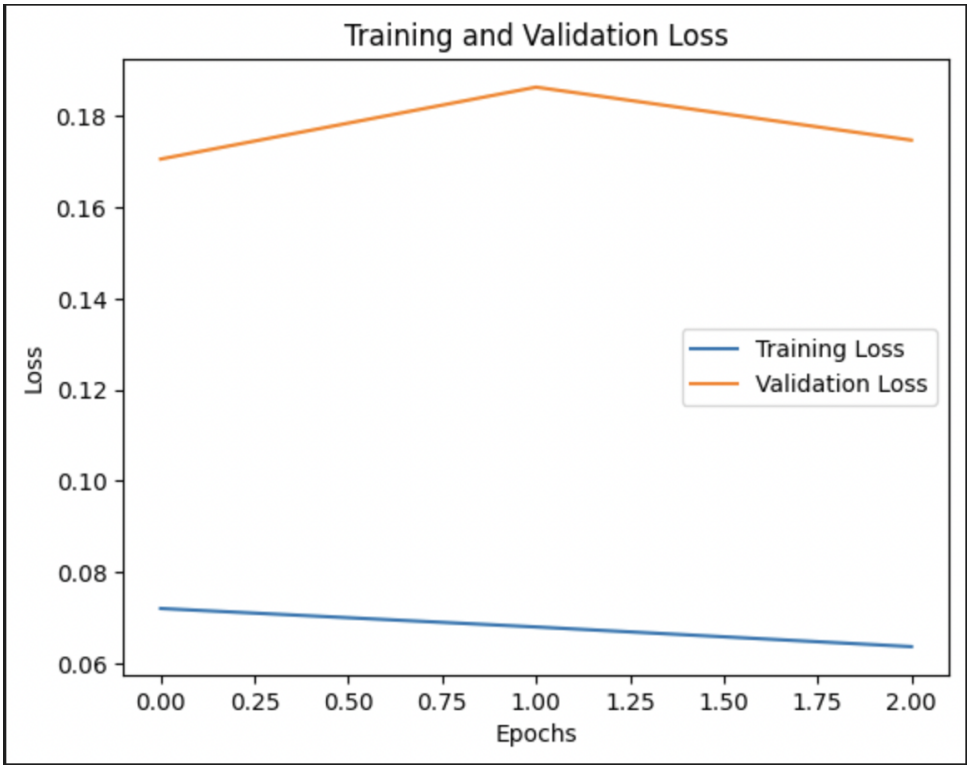


Figure 5: Case 1- train-validation loss plot

However, it can be regarded that the given large data set represents some features of hardship and the noise coming from the data richness, which can contribute to the moderate performance featuring by a number of approximately 42.76% of accuracy and other parameters. This tends to be an issue in a variety of medical imaging activities and even more regularly the differences between these classes are somewhat unique in order to enhance the reliability of the results obtained. In other words, based on results of the further analysis one can state that the proposed method offers nearly stable yet comparatively low results comprehensiveness on all the analyzed criteria, Therefore, the further work with the chosen model proved that the system has the capability to work with all the categories, though, at the same time, the problem to recognize the differences between closely connected classes remains an issue that can be discussed as potential.

| Metrics | PathVQA (Case-1) | Subset- PathVQA (Case-2) |
|---|---|---|
| Validation Accuracy | 0.4276 | 0.5048 |
| Precision | 0.4276 | 0.5048 |
| Recall | 0.4276 | 0.5048 |
| F1 Score | 0.4276 | 0.5048 |
| True Positives | 8404 | 5048 |
| False Negatives | 11250 | 4952 |
| False Positives | 11250 | 4952 |

Figure 6: Key Metrics

## 6.2 Case Study 2: Evaluation of the Model Built Using the Subset of the PathVQA Dataset

The second case study was about an evaluation of the BLIP model which was developed and trained on the basis of a part of the initial PathVQA dataset comprising 10000 records for training and 5% of the records, which equals to 1000 records, for the validation. Because of this, the smaller and more selective data set would have provided a better training regime as compared to the full set along with the noise and inherent chaos in the system.

As shown by figure 6, the evaluation of this model gave much better results, validation accuracy of 51.68%, and precision, recall, and F1 score also at 52. From the figure 7, the training and validation losses also in the process of three epochs also presented the same trend and from three epochs, the training loss is = 0. 0792 to 0. 2 to 0. 0734 and the validation loss dropping from 0. 1191 to 0. 0930. The progress of the training and the validation reflects the ability of the model in generalizing the data with much lower risk of overfitting using the data from the entire dataset as compared to when constructing the other model.

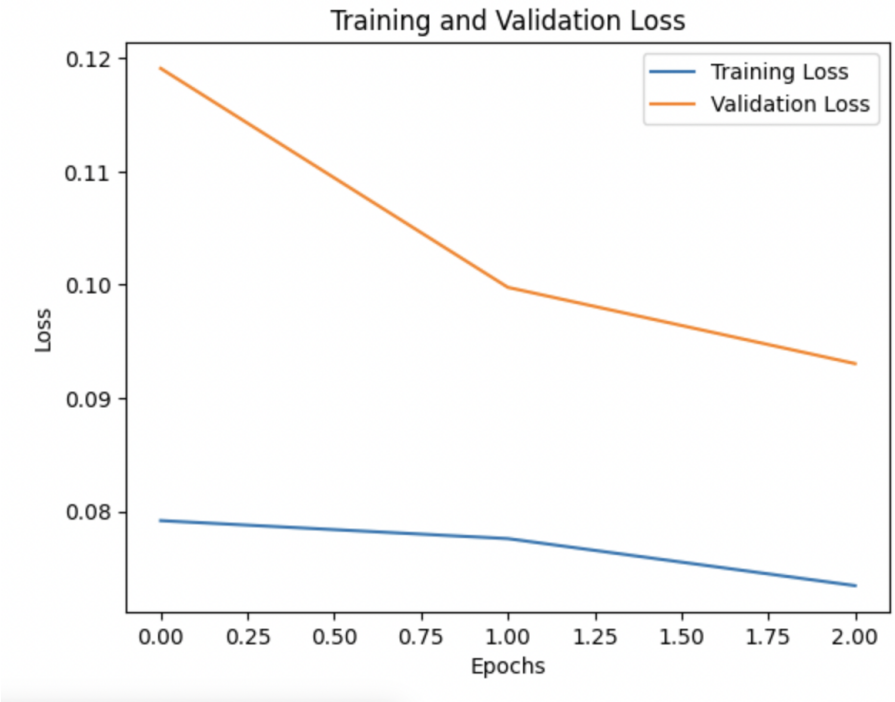


Figure 7: Case 2- train-validation loss plot

Improvement in the performance of the developed model on the subset might purely be as a result of elimination of noises and the subset of dataset being less complex. The trends whose variation is averaged out over successive iterations of constructing and assessing a model show consistently reducing loss values, which chart the outlined importance of the model's ability to discover patterns in data series at a more basic and refined level. The higher accuracy and balanced metrics also validates the generalization capability of the model when the training is conducted on a more particular set of data, It is understood that the epochs or fine-tuning might be improved at a better level for a more enhanced efficiency of the model.

### 6.3 Discussion

From the results of evaluation on the BLIP models trained with the full and subset PathVQA dataset it is possible to mention that the accuracy of the model is significantly higher if the model trained in the small and simple dataset. As it can be seen from the moderate performance measure and the validation losses which varies during the training process of the given model, the full data set could be useful for training the models using many instances, but presence of noise and complexity are the demerit of using full data set. Therefore, while using the subset dataset the model attained the higher accuracy and the distinct decreases of the loss, what has proved the model's better generalizing ability.

These results are also extend the use of data quality and data structural characteristic when performing of VQA models. Therefore, the overall performance of the chosen model is the fairly good with regard to various parameters, and it possible its work may need to be further improved using machine learning and other techniques of preprocessing and regularization in order to work with more significant and noisy datasets. Thus, the results of the loss-based evaluation were indicative of the fact that the learning process of the model was quite sound and there is always room for enhancing the performance of the model which in terms of improved distinctive versions of the BLIP model in Medical VQA can yield even higher overall accuracy and generalization.

## 7 Conclusion and Future Work

The objectives of this research primarily sought to evaluate the effectiveness of BLIP formulation in enhancing MmVQA system performance in terms of accuracy and reliability as compared to VLP and, secondarily, to measure model explanation with the use of understanding coefficients than those in basic VLP-formulated models. The details included loud medical data processing and benchmarking of the BLIP to other VLP models in medical imaging subcategories. Also, it was tried to support why BLIP should be better than BLIP2 in this case by explaining the merits of the latter in this context.

The reasons meant that it was found that the features of BLIP were able to learn noisy medical datasets to diagnose medical datasets with reasonably high and precise diagnostic with acceptable accuracy, recall, and F1 score. As suggested by the overall assessment done in the study, although the model performed rather a mediocre job of amassing a satisfactory level of accuracy when trained with the entire PathVQA, strikingly better performances were achieved when the model was trained only with a part of the total PathVQA dataset, which then made it very much clear that the model tends to generalise a lot better when trained with only a limited set of data sets. In addition to explaining the performance of BLIP, the researchers did not compromise between explainability and effectiveness, which is essential when interpretability is necessary, especially in the context of health care.

In comparison with the prior work on VLP models, the presented efficiency of the BLIP was fairly reasonable, and it did not stand out from the leading models in terms of the evaluation scenarios. Nevertheless, because it is inherently stable and

due to the interpretation of features, the method can become a convenient tool in the context of mVQA tasks if data noise and task complexity actually become critical. Due to the need to ensure that the developed model is both explainable and reliable, BLIP for AVA was used rather than BLIP2.

As future work, firstly, extending the BLIP model by utilising enhanced feature extraction techniques like transformer-based design architectures and CNNs in order to address the more intricate patterns that define medical imaging data; second, broadening the parameter space by incorporating multiple modality data fusion, which combines imaging data with text data in Electronic Health Records; third, enlarging the modularity of the model's interpretability using approaches that are better equipped to visualize and explain the diagnostic procedures. These directions are proposed to push forward the improvement of both efficiency and accuracy of BLIP in the framework of medical visual question answering.

# References


Bashmal, L., Bazi, Y., Melgani, F., Ricci, R., Al Rahhal, M. M. and Zuair, M. (2023). Visual question generation from remote sensing images, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **16**: 3279–3293.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. and Zagoruyko, S. (2020). End-to-end object detection with transformers, *European conference on computer vision*, Springer, pp. 213–229.

Chen, Z., Du, Y., Hu, J., Liu, Y., Li, G., Wan, X. and Chang, T.-H. (2022). Multi-modal masked autoencoders for medical vision-and-language pre-training, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp. 679–689.

Cong, F., Xu, S., Guo, L. and Tian, Y. (2022). Caption-aware medical vqa via semantic focusing and progressive cross-modality comprehension, *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3569–3577.

Do, T., Nguyen, B. X., Tjiputra, E., Tran, M., Tran, Q. D. and Nguyen, A. (2021). Multiple meta-model quantifying for medical visual question answering, *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, Springer, pp. 64–74.

Eslami, S., de Melo, G. and Meinel, C. (2021). Does clip benefit visual question answering in the medical domain as much as it does in the general domain?, *arXiv preprint arXiv:2112.13906* .

Farhan Ishmam, M., Sakib Hossain Shovon, M., Mridha, M. and Dey, N. (2023). From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities, *arXiv e-prints* pp. arXiv–2311.

He, J., Li, P., Liu, G., Zhao, Z. and Zhong, S. (2024). Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering, *arXiv preprint arXiv:2401.02797* .

Hossain, M. Z., Sohel, F., Shiratuddin, M. F. and Laga, H. (2018). A comprehensive survey of deep learning for image captioning, *CoRR* **abs/1810.04020**. **URL:** *http://arxiv.org/abs/1810.04020*

Huang, J., Chen, Y., Li, Y., Yang, Z., Gong, X., Wang, F. L., Xu, X. and Liu, W. (2023). Medical knowledge-based network for patient-oriented visual question answering, *Information Processing & Management* **60**(2): 103241.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O., Patra, B. et al. (n.d.). Language is not all you need: aligning perception with language models (2023), *arXiv preprint arXiv:2302.14045* .

Khare, Y., Bagal, V., Mathew, M., Devi, A., Priyakumar, U. D. and Jawahar, C. (2021). Mmbert: Multimodal bert pretraining for improved medical vqa, *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, IEEE, pp. 1033–1036.

Li, J., Li, D., Savarese, S. and Hoi, S. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, *International conference on machine learning*, PMLR, pp. 19730–19742.

Li, J., Li, D., Xiong, C. and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, *International conference on machine learning*, PMLR, pp. 12888–12900.

Li, J., Shakhnarovich, G. and Yeh, R. A. (2022). Adapting clip for phrase localization without further training, *arXiv preprint arXiv:2204.03647* .

Liu, J., Hu, T., Zhang, Y., Feng, Y., Hao, J., Lv, J. and Liu, Z. (2023). Parameter-efficient transfer learning for medical visual question answering, *IEEE Transactions on Emerging Topics in Computational Intelligence* .

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows, *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022.

Manmadhan, S. and Kovoor, B. C. (2020). Visual question answering: a state-of-the-art review, *Artificial Intelligence Review* **53**(8): 5705–5745.

Nguyen, B. D., Do, T.-T., Nguyen, B. X., Do, T., Tjiputra, E. and Tran, Q. D. (2019). Overcoming data limitation in medical visual question answering, *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, Springer, pp. 522–530.

Nguyen, V.-Q., Suganuma, M. and Okatani, T. (2022). Grit: Faster and better image captioning transformer using dual visual features, *European Conference on Computer Vision*, Springer, pp. 167–184.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. et al. (2021). Learning transferable visual models from natural language supervision, *International conference on machine learning*, PMLR, pp. 8748–8763.

Rückert, J., Ben Abacha, A., García Seco de Herrera, A., Bloch, L., Brüngel, R., Idrissi-Yaghir, A., Schäfer, H., Müller, H. and Friedrich, C. M. (2022). Overview of imageclefmedical 2022–caption prediction and concept detection, *CEUR Workshop Proceedings*, Vol. 3180, CEUR Workshop Proceedings, pp. 1294–1307.

Subramanian, S., Wang, L. L., Mehta, S., Bogin, B., van Zuylen, M., Parasa, S., Singh, S., Gardner, M. and Hajishirzi, H. (2020). Medicat: A dataset of medical images, captions, and textual references, *arXiv preprint arXiv:2010.06000* .

Vu, M. H., Löfstedt, T., Nyholm, T. and Sznitman, R. (2020). A question-centric model for visual question answering in medical imaging, *CoRR* **abs/2003.08760**. **URL:** *https://arxiv.org/abs/2003.08760*

Zhang, Z., Fu, Y., Gao, K., Zhang, H. and Wang, L. (2024). A cooperative evolutionary algorithm with simulated annealing for integrated scheduling of distributed flexible job shops and distribution, *Swarm and Evolutionary Computation* **85**: 101467.