VISUAL QUESTION ANSWERING ON RADIOGRAPHIC IMAGES

Aum Sathwara (A20543213) *Prashanth VR (A20531508)*

Main paper: Self-supervised vision-language pretraining for medial visual question answering
Publisher: IEEE (Institute of Electrical and Electronics Engineer)

1. Problem Statement

Medical Visual Question Answering (VQA) systems are designed to assist radiologists by interpreting medical images and answering clinically relevant questions. However, developing effective VQA models for medical contexts is challenging due to the unique nature of medical data. Radiographic images, for instance, are complex and contain subtle details that require a deep understanding of medical imaging. Additionally, the language used in clinical settings is highly specialized, with terminology and phrasing that are often very different from everyday language. This makes it difficult for models to process both the visual and textual information accurately.

One of the biggest challenges in medical VQA is the lack of large, annotated datasets. Unlike general-purpose VQA datasets, which are more abundant and diverse, medical datasets, particularly in radiology, are limited in size and variety. As a result, medical VQA models often struggle with open-ended questions, which require a more nuanced understanding of both the image and the text. These models tend to perform better with closed-ended questions, where the answers follow more predictable patterns. This difference in performance highlights the need for methods that can handle open-ended questions with the same level of accuracy and reliability as closed-ended ones.

The task of combining visual and textual data adds another layer of complexity. The model must not only process visual information from radiographic images—such as detecting anatomical features and abnormalities—but also interpret the accompanying text, which provides context for the image. Effectively aligning and integrating these two types of information is crucial, especially in a medical context, where errors can lead to misinterpretations that could impact patient diagnosis and treatment.

In this project, we aim to address the challenges of medical Visual Question Answering (VQA) by implementing a self-supervised vision-language pretraining approach, as outlined in the paper [1] by Li et al. This approach, known as M2I2, incorporates several pre-training tasks, including Masked Image Modeling (MIM), Masked Language Modeling (MLM), Image-Text Matching (ITM), and Image-Text Contrastive Learning (ITCL). These tasks enable the model to learn both visual and textual representations by reconstructing masked inputs and aligning image-text pairs. By leveraging self-supervised learning, we aim to improve the model's ability to generalize to medical VQA tasks, even with limited annotated data, ultimately boosting its accuracy and reliability.

The goal of this project is to implement the key components of the M2I2 pre training strategy. Ultimately, we seek to create a robust VQA system that not only excels at answering closed-ended questions but also

handles open-ended ones with greater consistency. Through this work, we aim to contribute to the development of more reliable medical VQA systems that can support radiologists in making better-informed clinical decisions.

2. Proposed Solution and Implementation details:

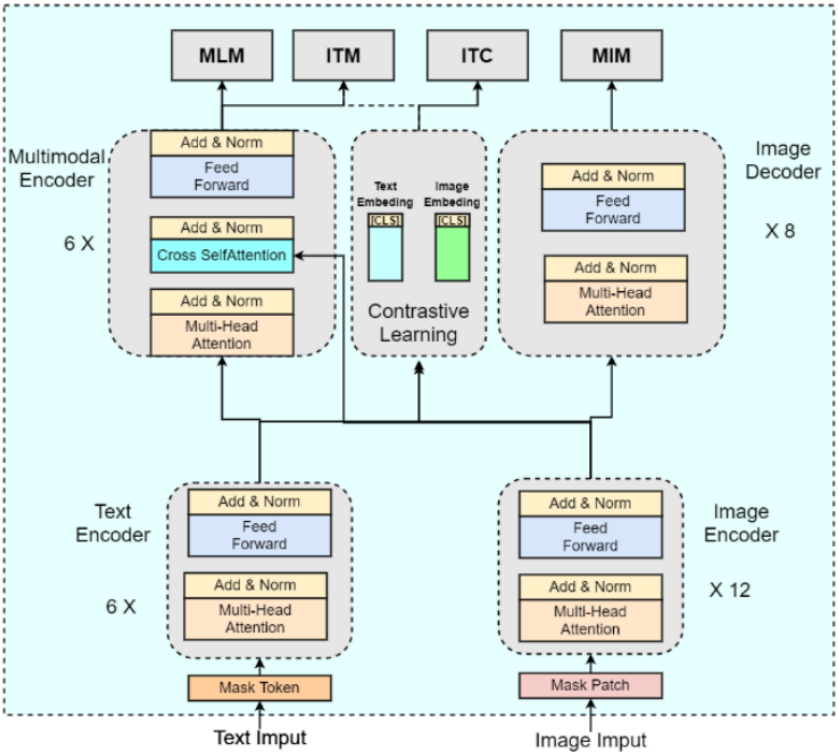


Fig. 1) VQA model Architecture

The approach focuses on four pre training tasks - Masked Image Modeling (MIM), Masked Language Modeling (MLM), Image-Text Matching (ITM), and Image-Text Contrastive Learning (ITCL) - designed to bridge the gap between visual and textual information.

2.1 Implementation Details

This project involves both pretraining and fine-tuning a multimodal model designed to understand and answer questions based on radiographic images and clinical text. The model integrates a Vision Transformer (ViT) for image processing and a BERT model for text processing. The pretraining phase is structured as a multi-task self-supervised learning pipeline, followed by fine-tuning to specialize the model on a medical Visual Question Answering (VQA) task.

Instructions for Running the Program:

1. **Environment Setup:**
    a. This program is implemented in Python and requires several dependencies, including PyTorch and the Hugging Face Transformers library.
    b. To set up the environment and install the necessary dependencies, you can use the

following command
pip install -r requirements.txt

2. **Data Preparation:**
    a. The program requires two datasets:
        i. Indiana University Chest X-rays for pretraining the multimodal model.
        ii. VQA-RAD (or another dataset for fine-tuning, depending on your setup).
    b. The data loading and preprocessing are handled by the ChestXrayDataset class, which expects the following directory structure:
        i. img_dir: The directory containing the chest X-ray images.
        ii. csv_path: A CSV file containing the captions corresponding to the images.
    c. Download the VQA_RAD Image Folder.zip and VQA_RAD Dataset Public.json from *VQA_RAD Dataset link. [3]*
        i. Extract the image folder into the directory specified by img_dir and the JSON file into the directory specified by json_dir.

3. **Running the PreTraining:**
    a. Once the environment is set up and data is ready, you can initiate the pre-training process by running the main.py script.
        i. The script requires two arguments to specify the paths to the image directory and the CSV file containing captions:
        python main.py "/path/to/images" "/path/to/captions.csv"
    b. Replace /path/to/images with the actual path to the directory containing the chest X-ray images, and /path/to/captions.csv with the path to the CSV file that contains the captions for the images.
    c. The program will:
        i. Load the data.
        ii. Set up the multimodal model (using a combination of Vision Transformer and BERT).
        iii. Start the training loop, periodically saving model checkpoints for recovery or further fine-tuning.

4. **Model Checkpoints:**
    a. During training, model checkpoints will be saved periodically. These checkpoints include:
        i. Model parameters for each individual task (MIM, MLM, ITM, ITCL).
        ii. The optimizer state.
    b. The checkpoints will be saved in the directory specified by the model_save_path and encoder_save_path arguments.
    c. After training, you can reload these checkpoints to continue training or fine-tune the model for downstream tasks.

5. **Running the Fine-tuning (VQA) Process:**
   After completing the pretraining step, you can fine-tune the model using the VQA-RAD dataset by running the Fine Tuning_VQA.ipynb file.

   Load Pretrained Weights:

   a. The pretrained weights from the pretraining task (saved in the encoder_checkpoints directory) will be loaded into the fine-tuning script.
   b. Alternatively, you can download the pretrained weights from Google Drive.
   c. Save the weights in a directory named encoder_checkpoints.

   The fine-tuning script expects the following path for the weights:

   "encoder_checkpoints/bert_model_checkpoint_epoch_1.pth"
   "encoder_checkpoints/image_encoder_checkpoint_epoch_1.pth"

2.1.1) Pre Training Phase

1. **Model and Encoder Setup**:
   We use two primary encoders in this system to process visual and textual data:

   a. **Image Encoder (ViT):** For processing radiographic images, we use a Vision Transformer (ViT) model pre-trained on the ImageNet dataset. This helps the model extract visual features from the medical images, which are key to understanding anatomical structures.
   b. **Text Encoder (BERT):** A pre-trained BERT model is used for handling the clinical text (captions, questions, etc.). BERT is great for understanding the language used in medical contexts, which often contains domain-specific terminology.

2. **Pre Training Task:**
   The implementation is designed as a multi-task self-supervised learning pipeline. The core pre-training tasks, as outlined in the M2I2 model, are:
   a. **Masked Image Modeling (MIM):** This task involves masking a portion of the image and training the model to reconstruct the missing patches. The loss is calculated based on how well the model can predict the masked portions of the image.
   b. **Masked Language Modeling (MLM):** In this task, certain words in the clinical text (captions) are masked, and the model is trained to predict the missing words.
   c. **Image-Text Matching (ITM):** The model is trained to determine whether a given image-text pair is correctly matched. The task involves both positive and negative image-text pairs, where negative pairs are randomly selected.
   d. **Image-Text Contrastive Learning (ITCL):** This task focuses on optimizing the similarity between matching image-text pairs and minimizing similarity between

mismatched pairs. This is achieved through a contrastive loss function.

3. **Data Pipeline**

The dataset used for pretraining is the *Indiana University Chest X-ray* dataset, containing X-ray images and corresponding clinical captions. The dataset is loaded using a custom "ChestXrayDataset" class, which handles:

a. **Image Preprocessing:** The images are resized to 224x224 pixels and transformed into tensor format.

b. **Text Tokenization:** Clinical text (captions) is tokenized using the BERT tokenizer. The column containing *impressions* is considered as the captions for their corresponding images in *IU Chest X-ray* dataset.

4. **Training Procedure**

The training procedure involves several key steps:

Batch Processing: Each batch contains paired image-text samples. Images are processed by the ViT model, while text is processed by the BERT model.

Loss Calculation: The total loss for each training step is computed by summing the losses from each task:

a. **MIM Loss:** Calculated using the reconstruction error between the original and predicted image patches.

b. **MLM Loss:** Computed using the cross-entropy loss for predicting the masked words in the text.

c. **ITM Loss:** Calculated using cross-entropy loss between predicted logits and actual labels for image-text matching.

d. **ITCL Loss:** A contrastive loss that compares matching and mismatched image-text pairs.

e. **Gradient Update:** The optimizer updates the model parameters after each batch using the computed combined loss, which includes the losses from all tasks.

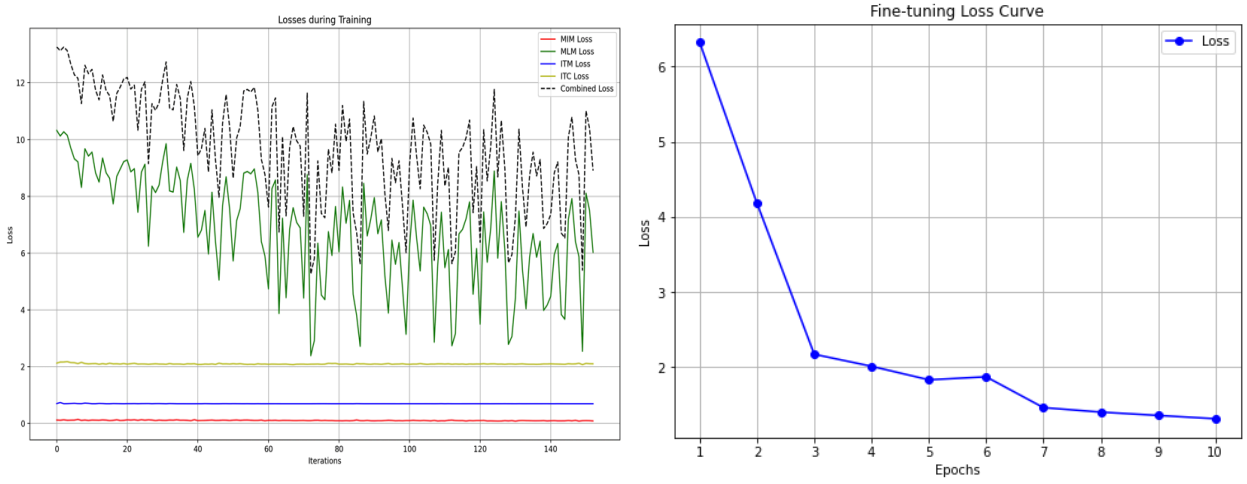


Fig. 2 a) Losses during Pre-Training VAQ Model b) Losses during Fine-Tuning of VAQ Model

5. **Unified Optimizer Setup**
   A single optimizer, **Adam**, is used to update the parameters across all tasks. This is done by collecting the unique parameters from each of the four models (MIM, MLM, ITM, ITCL) and passing them to the optimizer. By using a unified optimizer, the model parameters are updated consistently, ensuring that all tasks contribute to learning the shared representation.

2.1.2 Fine-Tuning Phase

In the fine-tuning phase the model is adapted for a specific task, such as Visual Question Answering (VQA) on medical images. The pretrained weights from the pre-training phase (for both the BERT and ViT models) serve as the initialization for the fine-tuning process.

1. **Fine-Tuning Task: Visual Question Answering (VQA)**
   In this stage, the model is trained to answer specific questions related to the radiographic images. This task involves:
   - **Input**: A question (in text form) and an associated medical image.
   - **Output**: The model generates an answer (a textual response) based on both the image and the question.

2. **Fine-Tuning Procedure**
   The fine-tuning process involves the following steps:
   - **Model Setup**:
     The pre-trained ViT and BERT encoders, as well as the cross-attention-based answer decoder, are used to process the image and text inputs.
   - **Data Pipeline**:
     The VQA dataset (such as VQA-RAD for medical VQA tasks) is used. Each data sample consists of an image, a corresponding question, and the correct answer. The text is tokenized using the BERT tokenizer, while the image is processed by the ViT encoder.
   - **Training Loop**:
     - The model is trained to predict the correct answer based on the question and the image.
     - **Loss Calculation**:
       Cross-entropy loss is used to calculate the difference between the predicted and true answers. The model is trained to minimize this loss.
     - **Optimizer Update**:
       The optimizer (Adam) updates the model's weights based on the calculated loss. Fine-tuning adjusts the model's weights to better align the learned representations with the specific VQA task.

3. **Evaluation and Metrics**
   In the evaluation phase, we focus on measuring the performance of the Visual Question Answering (VQA) model using several key metrics, tailored to assess the model's ability to answer both **closed** and **open-ended** questions. The evaluation is designed to understand how

well the model generates relevant answers, particularly in the context of medical imaging and clinical text.

- **3.1 Metrics for Closed Questions**
  Closed questions, such as "Is this image a sign of pneumonia?" typically have yes/no answers. For these types of questions, the evaluation uses standard classification metrics to determine how well the model predicts the correct binary response. The following metrics are used:
  - **Accuracy**: This measures the proportion of correct predictions (yes/no) made by the model out of all predictions. It gives a straightforward sense of how well the model performs in answering yes/no questions
  - **Precision**: This evaluates how many of the model's predicted "yes" or "no" answers are correct, meaning it measures the proportion of true positives (correct yes/no answers) among all predicted positives.
  - **Recall**: This metric measures how many of the true positives (actual yes/no answers) the model correctly identified. In other words, it looks at how many of the real positive cases were detected by the model.
  - **F1 Score**: The F1 score combines both precision and recall, offering a balanced measure of performance. It's particularly useful when there is an uneven class distribution (e.g., more "yes" than "no" answers) and provides insight into the model's ability to perform across both classes.

- **3.2 Metrics for Open-Ended Questions**
  For open-ended questions, such as "What condition does the X-ray suggest?", the answers are not fixed and can vary in phrasing. For these types of questions, the evaluation focuses on comparing the generated text to the ground truth using metrics that assess the similarity between the predicted and reference answers. The following metrics are used:
  - **BLEU Score**: The BLEU (Bilingual Evaluation Understudy) score is a metric for evaluating the quality of the generated text by comparing it to the reference answer. It measures the overlap of n-grams (in this case, unigrams) between the predicted answer and the ground truth. A higher BLEU score indicates that the generated answer closely matches the reference answer.
  - **ROUGE Score**: The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) score measures the quality of generated answers based on the overlap between n-grams, longest common subsequences (LCS), and word-level precision and recall. Specifically, ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequence) are calculated. These scores provide insights into how well the generated text captures the key information in the ground truth answers.

3. Dataset

We use two key datasets: the Indiana University Chest X-ray Dataset for pretraining the model and the VQA-RAD Dataset for fine-tuning. Each dataset plays a unique role in building a Visual Question Answering (VQA) system, combining general medical imaging data with more specific clinical knowledge to ensure the model performs effectively in radiology-related tasks.

3.1 Pre Training Dataset: Indiana University Chest X-ray Dataset

The Indiana University Chest X-ray Dataset, accessed from the Open-i repository, serves as the primary dataset for pretraining the model's visual encoder. This dataset contains a diverse collection of radiographic images, specifically chest X-rays, which are commonly used in clinical diagnostics. The main focus during pre training is to help the model learn to recognize various visual features within medical images.

- **Data Type**: Chest X-ray images in PNG format, preprocessed from the original DICOM format for ease of use.
- **Content**: This dataset includes thousands of radiographic images, categorized into frontal and lateral views, which are crucial for developing a general understanding of radiographic image features.
- **Preprocessing**: The original DICOM images were converted to PNG format, with pixel values adjusted to remove outliers and resized to 224x224 pixels. This resizing ensures that all images are standardized and ready for input into the model. These preprocessing steps help streamline the data and reduce computational demands during training.

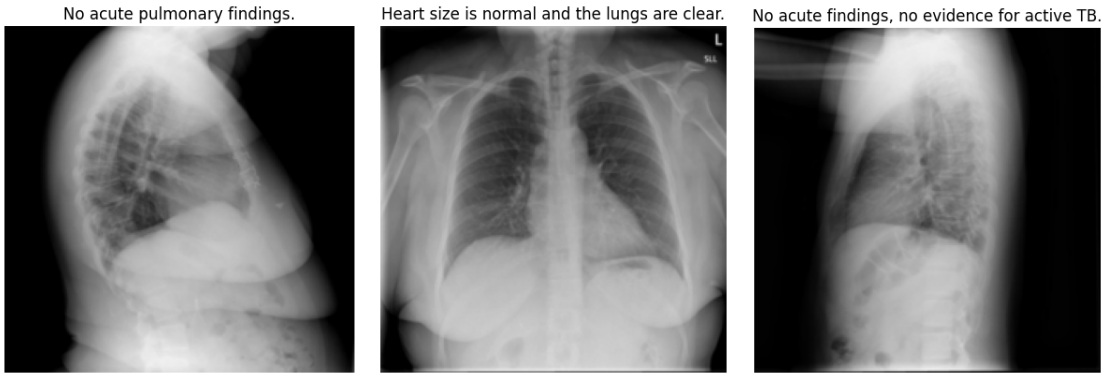


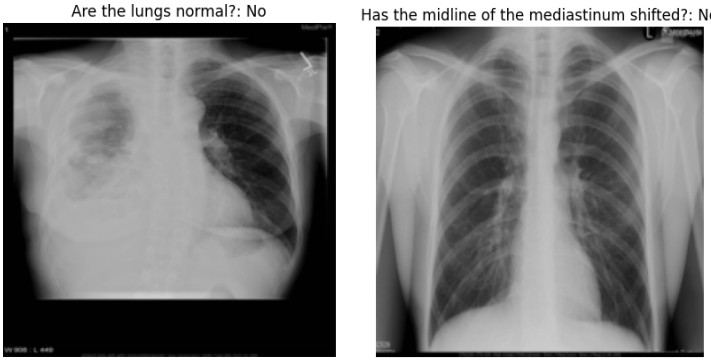
Fig 3. Samples from IU Chest X-ray dataset

3.2 Fine-tuning Dataset: VQA-RAD Dataset

The VQA-RAD Dataset is a curated dataset focused on Visual Question Answering (VQA) for radiographic images. It contains pairs of medical images and clinically relevant questions and answers, providing a rich source of annotated data that is specifically designed for training models to answer medical questions.

- **Data Type**: This dataset consists of radiographic images paired with natural-language questions and answers written by clinical experts.

- **Content**: The dataset includes 315 radiographic images covering anatomical regions such as the chest, abdomen, and head. It contains a total of 3,515 question-answer pairs, with questions categorized as either open-ended (43.1%) or closed-ended (yes/no).
- **Question-Answer Pairs**: The questions in this dataset focus on clinically important aspects of the images, with answers written in a concise and medically accurate manner. The questions cover a range of formats, allowing the model to learn to handle both closed and open-ended questions in a clinical context.



a) b)

Fig 4. Samples from VQA-RAD dataset: a) Closed type question; b) Open type question

4. Results and Discussion

In this section, we present the evaluation metrics and discuss the results of the Visual Question Answering (VQA) model applied to medical images, focusing on both closed-ended and open-ended questions. The model was evaluated using a combination of traditional metrics (accuracy, precision, recall, F1 score) for closed-ended questions and BLEU and ROUGE scores for open-ended questions.

4.1 Closed-Ended Question Evaluation

For closed-ended questions, which are typically yes/no type queries (e.g., "Is there evidence of a tumor?"), the model's performance was measured using the following metrics.

The model demonstrates a reasonable performance on closed-ended questions, with an accuracy of 0.7241. The precision of 0.729 indicates that when the model predicts a positive answer (e.g., "Yes, there is evidence of a tumor"), it is correct approximately 73% of the time. However, the recall score of 0.55245 reveals that the model fails to identify many positive instances, capturing only about 55% of all actual positives. This limitation is reflected in the F1 score of 0.6112, which balances precision and recall. Although the model performs adequately, the moderate F1 score suggests that it could benefit from further improvement, particularly in detecting all relevant positive cases.

Accuracy	0.7241
Precision	0.729
Recall	0.55245
F1 Score	0.6112

Table 1) Evaluation Metrics for Closed-Ended Question Classification

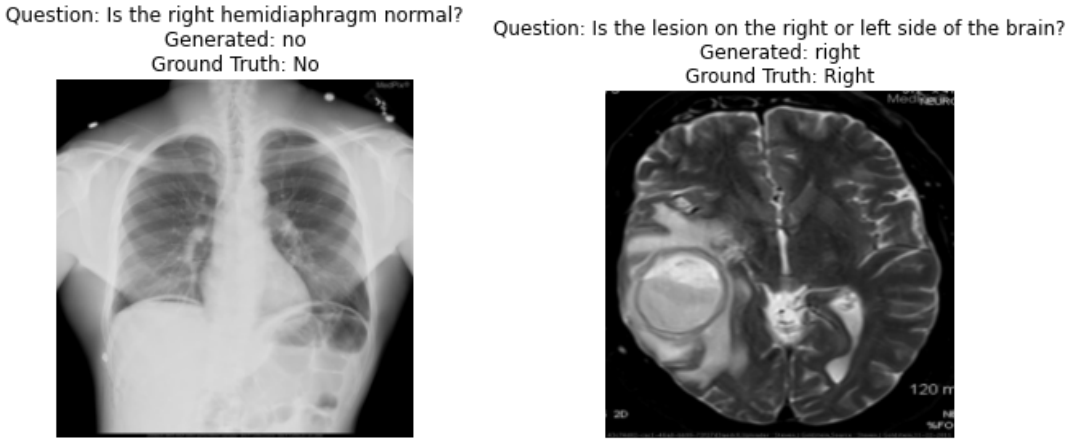


Fig 5. Outputs for closed-ended questions the VQA model

4.2 Open-Ended Question Evaluation

For open-ended questions, where the model generates natural language answers to more complex queries (e.g., "What is the size of the tumor?"), we evaluated the model's performance using BLEU and ROUGE scores.

BLEU Score:

- Measures the overlap of n-grams (unigrams, bigrams, etc.) between the generated answers and the reference answers. The model achieved a BLEU score of 0.0093, which is quite low. This suggests that while the model may generate some valid answers, there is a large gap in terms of exact matching with the reference answers.

ROUGE Scores:

- These are recall-oriented metrics that assess the overlap between the generated answers and reference answers, focusing on the matching of unigrams, bigrams, and the longest common subsequences.

The ROUGE results provide deeper insights into the model's ability to generate coherent and relevant answers. The unigram precision of 0.1880 indicates that a small proportion of the words in the generated answer match the reference. However, the recall of 0.0560 is very low, suggesting that while the model does capture some relevant words, it is missing many of the

important ones. The F1 score of 0.0836 highlights that the overall quality of generated answers is limited, with significant room for improvement in aligning the generated answers with reference answers. The bigram overlap scores were all zero, indicating that the model is unable to capture any meaningful bigram matches between the generated and reference answers. This suggests that the model struggles to produce answers that are both syntactically and semantically similar to the ground truth answers. The LCS precision and recall scores mirror the unigram scores, reinforcing the idea that while some key phrases are captured, the model's ability to produce high-quality, relevant answers is still weak.

BLEU Score	0.0093		
ROUGE Scores	**unigram overlap**	precision	0.1880
		recall	0.0560
		f1	0.0836
	bigram overlap	precision	0.0000
		recall	0.0000
		f1	0.0000
	longest common subsequence	precision	0.1880
		recall	0.0560
		f1	0.0836

Table 2) Evaluation Metrics for Open-Ended Question

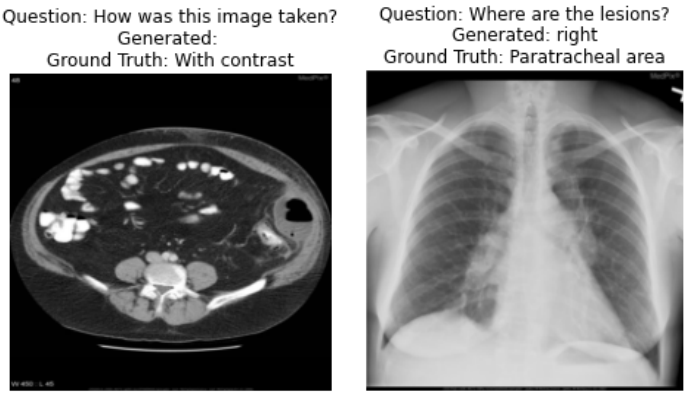


Fig 6. Outputs for open-ended questions the VQA model

4.3 Observations based on Evaluation

The evaluation reveals strengths in the model's handling of closed-ended questions and isolated keyword identification but also highlights limitations in open-ended responses and structured medical language interpretation.

1. **Strengths**:
    - The model shows reliability in answering yes/no questions, achieving a high precision in closed-ended formats.
    - Moderate ROUGE-1 and ROUGE-L scores indicate some success in identifying keywords and phrases, suggesting that the model has a partial understanding of medical terminology.
2. **Limitations**:
    - The low BLEU score and zero ROUGE-2 scores reveal a significant challenge in generating structured responses, particularly for open-ended questions requiring more complex, multi-word phrases.
    - Generated answers demonstrate a recurring issue where the model provides responses that are anatomically related but incorrect, suggesting a need for enhanced anatomical specificity and contextual understanding.
    - The model's performance on open-ended questions highlights its reliance on pretraining and limited exposure to diverse clinical queries in VQA-RAD. Given the smaller size of VQA-RAD, additional data and targeted pre-training may be necessary to address this limitation.

Conclusion:

To conclude, this project showcases our implementation of a Visual Question Answering model tailored for radiology, addressing the unique demands of medical imaging. By leveraging pre-trained weights from BERT and Vision Transformer, we were able to fine-tune on our radiology VQA dataset, enabling the model to capture critical features specific to radiographic images. While this setup demonstrated encouraging results, there's room for improvement, especially with longer training times.

References:

1. Li, Pengfei, et al. "Self-supervised vision-language pretraining for medial visual question answering." 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI). IEEE, 2023.
2. Indiana University Chest X-ray Dataset link: https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university
3. Visual Question Answering in Radiology (VQA-RAD) Dataset link: https://osf.io/89kps/
4. Lau, Jason J., et al. "A dataset of clinically generated visual questions and answers about radiology images." Scientific data 5.1 (2018): 1-10.
5. Pavlopoulos, John, Vasiliki Kougia, and Ion Androutsopoulos. "A survey on biomedical image captioning." Proceedings of the second workshop on shortcomings in vision and language. 2019.