CS 512 - Computer Vision

Visual Question Answering on Radiographic Images

By:
Prashanth VR (A20531508) Aum Sathwara (A20543213)

Table of contents

01

Problem Statement

Problem Statement

- This project addresses the challenge of creating a Visual Question Answering (VQA) system for radiographic images. Medical VQA combines complex visual and textual data, which makes effective training challenging due to the limited annotated datasets available in this domain.
- The primary challenge lies in the combination of visual data from medical images with textual data for accurate answers. Medical VQA struggles due to the specific language in clinical contexts and limited labeled data, especially for open-ended questions.
- This project aims to create a model that can accurately answer questions regarding radiographic images, enhancing support for radiologists.

02

Related Works

Related Works

- Due to limited annotated VQA datasets, pretrained models like VGG16, ResNet (for images), and LSTM, BERT (for text) are often used, with weights from external datasets like ImageNet, despite potential domain shift issues.
- To reduce domain shift, visual encoders are pretrained on medical image datasets for tasks such as image classification before fine-tuning for VQA.
- Nguyen et al. introduced a meta-learning-based pretraining task, specifically a 3-way 6-shot classification, to improve visual representation learning on medical datasets.
- Recent methods incorporate self-supervised objectives, like masked language modeling (MLM), to jointly model image and text features, enhancing medical VQA model performance.
- Frameworks like PubMed-CLIP utilize contrastive learning to align image-text features effectively, showing promise in medical VQA by learning joint representations without relying on external datasets.

03

Dataset

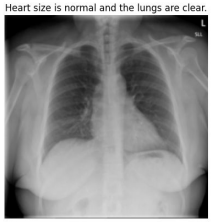
Indiana Chest X-Ray

- This dataset is a collection of radiology examinations, including images and narrative reports by radiologists and is publicly available through the Open Access Biomedical Image Search Engine (OpenI).
- The images are 7,470 frontal and lateral chest X-rays, and each radiology report consists of four sections: Comparison, Indication, Findings and Impression.
- Image captions are descriptive summaries based on radiological observations, useful for training models on clinical language generation.
- Originally intended for medical and educational purposes, this dataset is valuable for developing models that support medical diagnostics and automated report generation.

Indiana Chest X-Ray

No acute pulmonary findings.

Heart size is normal and the lungs are clear.



No acute findings, no evidence for active TB.

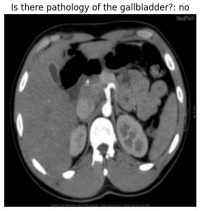
VQA-RAD

- VQA-RAD is the first ever manually created Visual Question Answering dataset in radiology where questions and answers are manually created and curated by clinicians.
- It is a high quality design with 60 hours of specialized contributions trading off quantity of automatic generation. This dataset contains 315 high quality radiology images, taken from a well balanced selection of case reports.
- These images cover key anatomical regions such as the head, chest and abdomen and were designed to assist in clinical decision making.
- There are a total of 3,515 QA pairs associated with the images. The QA pairs are split into 42% open-ended answer types and 58% close-ended. Yes/no questions represent 92% of the close-ended QA pairs.

Indiana Chest X-Ray

Are the lungs normal?: No

Is there pathology of the gallbladder?: no



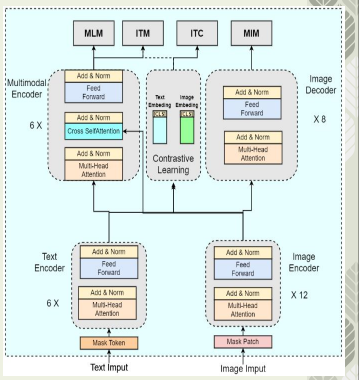
Has the midline of the mediastinum shifted?: No

04

Methodology

Model Architecture

- We first pre-train our model on various pre-training tasks such Masked Image Modeling, Masked Image Modeling, Image Text Matching and Image Text Contrastive Learning with the image-captioning dataset.
- After pre-training, the weights of the text, multimodal and image encoders along with an answer decoder are further fine-tuned for Visual question answering.



Masked Image Modeling

- Purpose: To teach the model to understand medical images by reconstructing masked patches.
- Process: A random selection of image patches is masked, and the model attempts to reconstruct these regions based on the unmasked parts.

Masked Language Modeling

- Purpose: Builds understanding of clinical language by masking words and predicting them.
- Process: Random words in captions are masked, and the model predicts the missing words based on surrounding context.

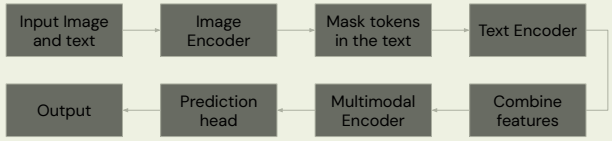


Image Text Matching

- Purpose: Helps the model align images with corresponding text, improving its ability to relate images to specific questions.
- Process: The model is trained to recognize correct image-text pairs by generating and distinguishing negative pairs.

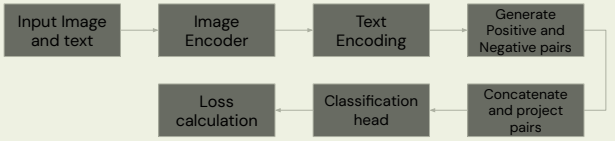
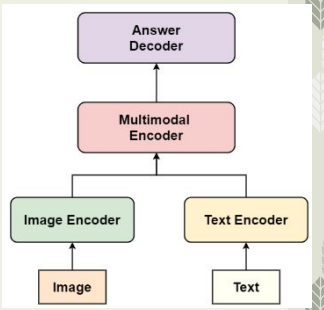


Image Text Contrastive Learning

- Purpose: Ensures robust image-text alignment, learning which features from the image and text pairs correspond closely.
- Process: The model maximizes similarity for matching image-text pairs and minimizes it for mismatched pairs, refining multimodal understanding.

Fine-Tuning on VQA

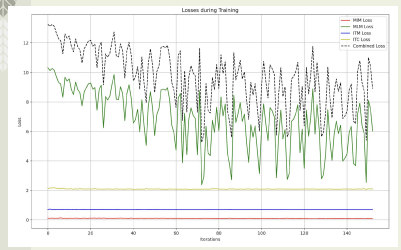
- For fine-tuning on medical VQA tasks, we initialize the image encoder, text encoder, and multimodal encoder with pretrained weights and add an answer decoder using pretrained BERT weights.
- The multimodal embeddings serve as input to the answer decoder, which generates answers through conditional language modeling, optimized by minimizing the log probability of ground-truth answers.



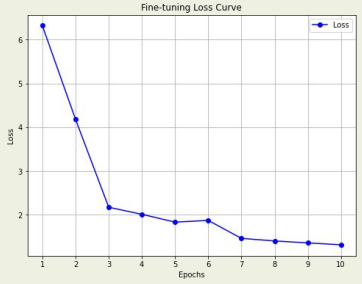
05

Results

Pre-training/ Fine-tuning Metric History



Pre-training stage



Fine-tuning stage

Inference of VQA model

Question: Is there pneumothorax?
Generated: no
Ground Truth: No

Inference on closed questions

Question: Where are the lesions?
Generated: right
Ground Truth: Paratracheal area

Inference on open ended questions

Evaluation on Closed Question

Accuracy: This indicates that the model correctly predicted the answer (either "yes" or "no") approximately 72.41% of the time.
Precision: Precision of 0.729 means that when the model predicted "yes" or "no," 72.9% of those predictions were correct.
Recall: The model correctly identified 55.2% of the actual "yes" or "no" answers, showing that it missed some correct answers.
F1 Score: The F1 score combines precision and recall into one metric. With a value of 0.6112, this reflects a moderate balance between precision and recall.

Accuracy	0.7241
Precision	0.729
Recall	0.55245
F1 Score	0.6112

Evaluation on Open Ended Question

BLEU: Measures the precision of n-grams in generated text compared to reference text.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right).$$

ROUGE: Evaluates text by comparing n-gram overlaps (precision, recall, F1) with reference text.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{\text{ReferenceSummaries}\}} \sum_{gram_n \in S} Count(gram_n)} \quad (1)$$

ROUGE-N measure

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

06

Conclusion and Future Work

Conclusion and Future Work

- In this project, we have implemented a VQA (Visual Question Answering) model specifically trained on a radiology dataset. We made use of weights of the BERT and Vision transformer models trained on various pre-training tasks and then fine-tuned on the VQA dataset, thereby letting the model learn features specific to radiographic images.
- Future works:
  - Train the model for even more epochs, making use of cloud GPU resources.
  - Fine-tune the model using other VQA datasets, such as the Slake dataset, whose size is greater than the VQA-RAD dataset containing 14,208 question-answer pairs.
  - Introduce a weighted combined loss, ensuring more weights are given for loss values that have high values.
  - Replace the vision transformer backbone with Swin transformer and ConvNeXT in the model architecture.