BioMedBLIP: Advancing Accuracy in Multimodal Medical Tasks through Bootstrapped Language-Image Pretraining

Usman Naseem, Surendrabikram Thapa, Anum Masood

Submitted to: JMIR Medical Informatics
on: January 22, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript 5
Supplementary Files 34
Figures 35
Figure 1 36
Figure 2 37
Figure 3 38
Figure 4 39
Figure 5 40
Figure 6 41
Figure 7 42
Figure 8 43
Figure 9 44
Figure 10 45
Figure 11 46
Figure 12 47
Figure 13 48

BioMedBLIP: Advancing Accuracy in Multimodal Medical Tasks through Bootstrapped Language-Image Pretraining

Usman Naseem[1, 2]; Surendrabikram Thapa[3]; Anum Masood[4, 5, 6] PhD

[1]College of Science and Engineering James Cook University Townsville AU
[2]School of Computing Macquarie University Sydney AU
[3]Department of Computer Science Virginia Tech Blacksburg US
[4]Department of Circulation and Medical Imaging Norwegian University of Science and Technology Trondheim NO
[5]Harvard Medical School Harvard University Boston US
[6]Department of Radiology Boston Children's Hospital Boston US

Corresponding Author:
Anum Masood PhD
Department of Circulation and Medical Imaging
Norwegian University of Science and Technology
B4-135, Realfagbygget Building.
Gloshaugen Campus
Trondheim
NO

Abstract

Background: Medical image analysis, particularly in the context of Visual Question Answering (VQA) and image captioning, is crucial for accurate diagnosis and educational purposes.

Objective: Our study introduces BioMedBLIP models, fine-tuned for VQA tasks using specialized medical datasets like ROCO and MIMIC-CXR, and evaluates their performance in comparison to the state-of-the-art (SOTA) Original BLIP model.

Methods: We present nine versions of BioMedBLIP across three downstream tasks in various datasets. The models are trained on a varying number of epochs. The findings indicate the strong overall performance of our models. We proposed BioMedBLIP for VQA Generation Model, VQA Classification Model, and BioMedBLIP Image Caption Model. We conducted pre-training in BLIP using medical datasets, producing an adapted BLIP model tailored for medical applications.

Results: In VQA-Generation tasks, BioMedBLIP models outperformed the SOTA on SLAKE, VQA-RAD, and ImageCLEF datasets. In VQA-Classification, our models consistently surpassed the SOTA on SLAKE. Our models also showed competitive performance on VQA-RAD and PathVQA datasets. Similarly, for image captioning tasks, our model beats the SOTA suggesting the importance of pretraining with medical datasets. Overall, in 20 different datasets and task combinations, our BioMedBLIP excels in 15 out of 20 tasks. BioMedBLIP represents a new state-of-the-art in 15 out of 20 tasks (75%) and our responses were rated higher in all 20 tasks ($P < 0.005$) in comparison to SOTA models.

Conclusions: Our BioMedBLIP models show promising performance and suggest that incorporating medical knowledge through pretraining with domain-specific medical datasets helps models achieve higher performance. Our models thus demonstrate their potential to advance medical image analysis, impacting diagnosis, medical education, and research. However, data quality, task-specific variability, computational resources, and ethical considerations should be carefully addressed. In conclusion, our models represent a contribution towards the synergy of AI and medicine. We have made BioMedBLIP freely available which will help in further advancing research in multimodal medical tasks.

(JMIR Preprints 22/01/2024:56627)
DOI: https://doi.org/10.2196/preprints.56627

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

✓ **Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all u**

Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Type of Paper
Title
Keywords
Introduction
Methods
Results
Discussion
Acknowledgements
Conflicts of Interest
Abbreviations
References

Original Paper

BioMedBLIP: Advancing Accuracy in Multimodal Medical Tasks through Bootstrapped Language-Image Pretraining

Abstract

Background

Medical image analysis, particularly in the context of Visual Question Answering (VQA) and image captioning, is crucial for accurate diagnosis and educational purposes.

Objective

Our study introduces BioMedBLIP models, fine-tuned for VQA tasks using specialized medical datasets like ROCO and MIMIC-CXR, and evaluates their performance in comparison to the state-of-the-art (SOTA) Original BLIP model.

Methods

We present nine versions of BioMedBLIP across three downstream tasks in various datasets. The models are trained on a varying number of epochs. The findings indicate the strong overall performance of our models. We proposed BioMedBLIP for VQA Generation Model, VQA Classification Model, and BioMedBLIP Image Caption Model. We conducted pre-training in BLIP using medical datasets, producing an adapted BLIP model tailored for medical applications.

Results

In VQA-Generation tasks, BioMedBLIP models outperformed the SOTA on SLAKE, VQA-RAD, and ImageCLEF datasets. In VQA-Classification, our models consistently surpassed the SOTA on SLAKE. Our models also showed competitive performance on VQA-RAD and PathVQA datasets. Similarly, for image captioning tasks, our model beats the SOTA suggesting the importance of pretraining with medical datasets. Overall, in 20 different datasets and task combinations, our BioMedBLIP excels in 15 out of 20 tasks. BioMedBLIP represents a new state-of-the-art in 15 out of 20 tasks (75%) and our responses were rated higher in all 20 tasks ($P< 0.005$) in comparison to SOTA models.

Conclusions

Our BioMedBLIP models show promising performance and suggest that incorporating medical knowledge through pretraining with domain-specific medical datasets helps models achieve higher performance. Our models thus demonstrate their potential to advance medical image analysis, impacting diagnosis, medical education, and research. However, data quality, task-specific variability, computational resources, and ethical considerations should be carefully addressed. In conclusion, our models represent a contribution towards the synergy of AI and medicine. We have made BioMedBLIP freely available which will help in further advancing research in multimodal medical tasks.

Keywords

Biomedical text mining, BioNLP, Vision-Language Pretraining, Multimodal Models

Introduction

In recent decades, the field of data analysis, machine learning, and deep learning has undergone remarkable advancements, with profound implications for various professional domains [1,2]. One of the most promising frontiers for these advancements is medical science, where data-driven models have the potential to bring about significant breakthroughs [3,4]. Medical data predominantly exists in the form of images and textual reports, encompassing X-ray images, medical records, and more. To harness the full potential of these data sources, a visual language model capable of extracting insights from both images and text becomes paramount. Visual language models, which are at the core of this research, represent a fusion of computer vision and natural language processing. These models possess the capability to understand and generate text-based descriptions for visual content, making them invaluable in contexts where both images and text are essential for comprehensive analysis.

This paper explores and adapts visual language models specifically for medical datasets, building upon the foundation laid by existing models. The primary objective is to enhance the performance of these models when confronted with medical data, such as the ROCO [5] and MIMIC-CXR [6] datasets. This will be achieved through a comprehensive process of pre-training on medical datasets and rigorous fine-tuning, with the ultimate goal of determining the optimal model configurations and parameters. This paper thus facilitates advancements in healthcare, contributing to more accurate diagnoses, streamlined medical reporting, and ultimately, improved patient care. We have made BioMedBLIP models freely available facilitating the progress of research in diverse medical applications involving multiple modalities [30].

Related Work

In the domain of visual language models and their applications within the medical field, several notable studies and advancements have paved the way for this research project. These works solve different problems within healthcare analytics and have played critical roles in shaping the paper's foundation.

Within the medical domain, image captioning has emerged as a valuable tool that enables healthcare professionals and researchers to enhance their diagnostic and reporting processes. Image captioning technology allows for the automatic generation of textual descriptions for medical images, such as X-rays, MRIs, and CT scans. This capability brings about several significant benefits. First, it aids clinicians in the diagnostic process by providing detailed descriptions of medical images, helping medical professionals to quickly and accurately identify abnormalities or pathologies in the images, thus improving the efficiency and accuracy of diagnoses. Second, image captions serve as a means of clear and standardized communication among healthcare professionals, reducing the potential for misinterpretation when multiple experts are involved in the diagnostic process. Third, image captions make medical images more accessible to a broader audience, including patients, promoting health literacy and patient engagement. Moreover, in a clinical setting, image captions expedite the process of creating medical reports, improving the overall quality of patient records. They also play a valuable role in medical education and training, aiding in the learning and teaching of medical imaging and diagnostics. Pavlopoulos et. al. [7] proposed that biomedical image captioning can significantly expedite clinicians' diagnostic processes and presented a comprehensive survey covering various aspects of medical image captioning, including datasets and evaluation measures. Furthermore, the task of automatic generation of medical image reports, introduced by Jing et. al. [8], aimed to streamline the reporting process for physicians, enhancing both efficiency and accuracy. To address this, Jing et. al. [8] employed a hierarchical Long Short-Term Memory (LSTM) model, which was tested on two publicly available datasets, IU X-Ray [9] and PEIR GROSS [10].

This connection between image captioning and medical report generation underscores the practical utility of visual language models in improving healthcare processes. In addition to these advancements, the field of medical Visual Question Answering (VQA) has gained increasing relevance. Medical VQA tasks involve developing models capable of answering questions related to medical images and bridging the gap between textual queries and visual data. Lin et. al. [11] introduced various medical datasets and also proposed methods to enhance model performance in medical VQA tasks. We use various datasets presented by Lin et. al. [11] in our experiments. Furthermore, Li et. al. [12] emphasized the significance of pre-training models on general images to capture meaningful representations of medical data, thus laying the groundwork for our approach. This insight served as our motivation to explore an approach of pre-training models on domain-specific medical datasets, with the aim of achieving enhanced performance for medical Visual Question Answering (MedVQA) tasks. Notably, Li et. al. [12] encountered a limitation in their work, as they did not pre-train models using medical datasets. Their decision was influenced by computational resource constraints, and they believed that domain-specific pre-training would be the key to improving model performance in MedVQA tasks. To address this gap in the research landscape, we took the initiative to pre-train our model using medical datasets, thereby bridging the gap between general and medical image understanding.

Transformer models have become instrumental in a diverse array of applications in various vision and language (V+L) tasks, including medical VQA. The Transformer, proposed by Vaswani et. al. [13], represents a departure from traditional recurrent or convolutional neural networks. Its architecture replaces recurrent layers with a multi-head self-attention encoder and decoder structure. Compared to traditional RNN models, the Transformer significantly reduces training time, making it a scalable solution capable of handling a wide range of inputs and applicable to diverse vision and language tasks, including the analysis of medical images.

Several prominent transformer-based models have had a significant impact on the landscape of natural language processing (NLP) and multimodal tasks. One of the most influential models is BERT, which stands for Bidirectional Encoder Representations from Transformers. Proposed by Devlin et. al. [14], BERT has demonstrated its efficacy in a wide variety of NLP tasks. This is achieved through a pre-training phase where 15% of input sequences are masked. These masked tokens can be replaced with random words, original words, or [MASK] tokens. Subsequently, the transformer auto-encodes these tokens, and fine-tuning is applied to adapt the pre-trained model to downstream tasks. The generalization capabilities of BERT are remarkable, making it adept at handling a wide array of semantic tasks. This is due in part to its bidirectional training, which allows the model to learn contextual information from both the left and right sides of a given word.

BERT's versatility allows it to be tailored for various applications, and one domain where it has shown great promise is the biomedical field. In the biomedical domain, text data often exhibits complex language patterns and domain-specific terminology. Lee et. al. [15] recognized the need for a model that could adapt to these linguistic intricacies and introduced BioBERT, a BERT-based model fine-tuned on biomedical text. BioBERT effectively addresses the word distribution shift from general data to biomedical data, making it a valuable tool for tasks like biomedical text mining. The model's workflow involves transferring weights from BERT, which is pre-trained on general domain data, to BioBERT. Subsequently, BioBERT is pre-trained on biomedical domain data, followed by fine-tuning and evaluation of various downstream tasks. This adaptation enhances BioBERT's performance in domain-specific tasks, such as biomedical text classification and named entity recognition.

The success of BERT and its adaptations has paved the way for exploring their application in multimodal tasks, where both text and image data are involved. For instance, VisualBERT, proposed by Li et. al. [16], is inspired by BERT and designed to capture rich semantics in vision and language tasks. It employs a stack of Transformer layers and integrates pre-trained object proposal systems for image feature extraction. In the training process, VisualBERT employs self-supervised learning with masked word tokens and performs image caption classification tasks with true and false captions. This approach enables the model to capture intricate relationships between text and image content, making it highly suitable for multimodal tasks where textual descriptions are needed for visual content.

Another notable model, LXMERT (Learning Cross-Modality Encoder Representations from Transformers), builds upon the success of BERT and its variants [17]. Tan and Bansal [17] recognized the importance of interpreting the semantic meaning of both images and text while exploring the relationships between vision and language. LXMERT's encoders, based on the Transformer architecture, are pre-trained on large volumes of image-text pairs. The pre-training process, inspired by BERT, includes techniques like adding random masks. Interestingly, LXMERT's pre-training approach has been found to outperform data augmentation, a common practice used to increase the amount of training data. Consequently, LXMERT is well-suited for tasks that involve understanding and generating textual descriptions for visual content, such as image captioning and visual question answering.

As the field of vision and language tasks continues to evolve, so do the transformer-based models designed to tackle them. The Vision Transformer (ViT), introduced by Dosovitskiy et. al. [18], represents an innovation to address the challenges of applying the Transformer architecture directly to computer vision tasks. ViT operates by dividing an image into 16x16 patches and processing them with position embeddings using a standard Transformer encoder. This approach has shown promise, but it demands substantial computational resources and extensive training data. Notably, ViT32 Vision Transformer was employed by Eslami et. al. [19] as part of fine-tuned versions of CLIP, comparing the performance of different models in the medical domain.

Similarly, UNITER, proposed by Chen et. al. [20], takes inspiration from the BERT model. This UNiversal Image-TExt Representation model has demonstrated strong performance in various vision and language tasks. The architecture of UNITER utilizes the Transformer as its core, with the image and text embedder working in tandem to encode image and text features into a common embedding space. This approach enables the generation of contextual embeddings, facilitating a better understanding of the relationships between vision and language. These transformer-based models collectively represent a spectrum of approaches and adaptations within the broader field of vision and language tasks.

The landscape of transformer-based models, ranging from BERT to ViT, demonstrates their adaptability and effectiveness in various domains, including NLP and multimodal tasks. CLIP (Contrastive Language-Image Pre-training), introduced by Radford et. al. [21], represents a significant step forward in this domain. CLIP is designed to connect images and text through a shared embedding space, enabling it to understand the relationship between the two modalities. By pre-training on a massive dataset containing images and their associated textual descriptions, CLIP can align images with natural language descriptions, making it a versatile tool for a wide range of tasks. This novel approach has significant implications for the medical field, where visual data, such as medical images, often require textual descriptions for comprehensive analysis and interpretation.

Building upon the success of CLIP, PubMedCLIP emerged as a tailored solution for MedVQA. Eslami et. al. [19] recognized the need for a model specifically adapted to the medical domain and developed PubMedCLIP, a fine-tuned version of CLIP trained on a dataset of medical image-text pairs from PubMed articles. This adaptation enables PubMedCLIP to better understand the nuances of medical images and text, resulting in improved performance on MedVQA tasks.

One of the recent works fusing medical imaging and text data is MedBLIP Chen et. al. [22]. MedBLIP uses a trainable 3D vision encoder MedQFormer which connects medical images with language models. However, MedBLIP could not significantly improve VQA performance and classification accuracy. For the experimental evaluation of MedBLIP, authors only used MRIs and text, which will limit the use of MedBLIP in case of other modalities such as positron emission tomography (PET), computed tomography (CT), X-ray images, etc.

Despite performing well in vision-language tasks, CLIP suffers from a number of limitations. Firstly, it is primarily focused on vision-language understanding tasks, such as image retrieval and visual question answering. This means that it is not well-suited for generation tasks, such as image captioning. Secondly, CLIP is trained on a large dataset of image-text pairs collected from the web. However, this data is often noisy and contains incorrect or misleading captions. This can lead to CLIP making mistakes when performing tasks that require an accurate understanding of the relationship between images and text.

To address these limitations, Li et. al. [23] proposed BLIP, a vision language pre-training framework. BLIP, as shown in Figure 1. is a unified model that can be used for both understanding and generation tasks. This is achieved by incorporating a captioning module into the model, which allows BLIP to generate captions for images. Additionally, BLIP addresses the issue of noisy web data by bootstrapping the captions. This means that a captioner generates synthetic captions and a filter removes the noisy ones. This results in a cleaner dataset that can be used to train a more robust model. As a result of these improvements, BLIP has been shown to achieve state-of-the-art results on a wide range of vision-language tasks, including image retrieval, visual question answering, image captioning, and visual grounding. Additionally, BLIP is more efficient to train and can be fine-tuned for downstream tasks with fewer data. Finally, BLIP is more interpretable than CLIP, as the captioning module allows users to understand how the model is reasoning about images.

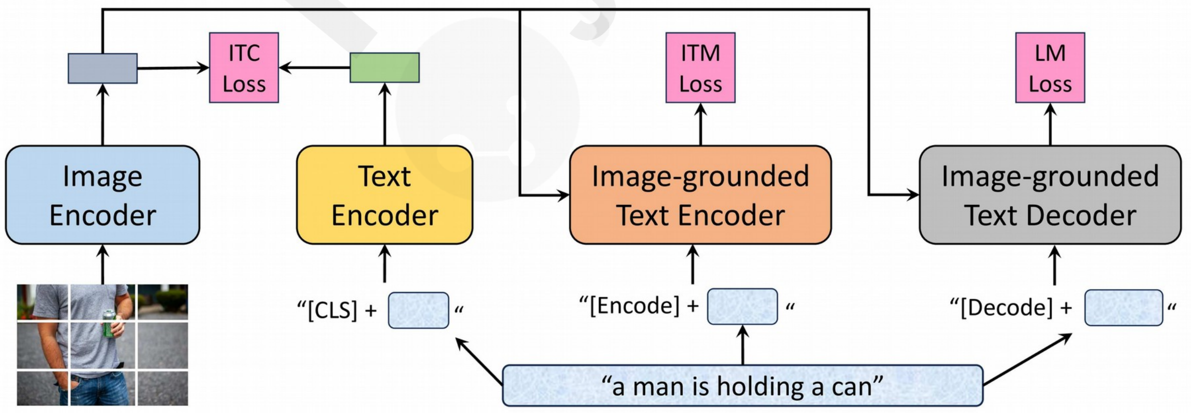


Figure 1 Pretraining architecture of BLIP.

Capitalizing on the strengths of BLIP, we propose BioMedBLIP by pre-training and fine-tuning the model with a medical dataset to achieve state-of-the-art results on medical vision-language tasks.

Specifically, BLIP's unified approach to vision-language understanding and generation makes it well-suited for tasks such as medical image classification, medical image retrieval, and medical image captioning. Additionally, BLIP's ability to handle noisy data makes it well-suited for training on medical datasets, which can often be noisy and contain incomplete or inaccurate information. We evaluate our pre-trained model using various standard task-specific performance metrics.

Methods

In this section, we describe our pretraining details along with training strategies and resources used.

BioMedBLIP

BLIP, initially pre-trained on general image datasets, possesses knowledge rooted in general image understanding. However, medical images exhibit distinct characteristics that differentiate them from general images. Many medical images are grayscale, such as X-rays and MRIs, which results in a significant divergence between the general image domain and the medical image domain. To bridge this gap, we conducted pre-training of BLIP using medical datasets, producing an adapted BLIP model tailored for medical applications.

As shown in Figure 1, BLIP is organized into four key modules.

- **Visual Transformer Block (Image Encoder):** The first module serves as an image encoder, utilizing a visual transformer to extract features from medical images.

- **BERT-Based Text Encoder:** The second module is a text encoder based on BERT. It processes textual data, ensuring a comprehensive understanding of medical texts.

- **Cross-Attention and Binary Classification:** The third module shares parameters with the text encoder, facilitating joint image-text embeddings through cross-attention. It employs a binary classifier to confirm the pairing of images and text.

- **Text Decoder:** The final module is a text decoder, which shares some components with the preceding encoders, such as feed-forward and cross-attention layers. However, it maintains its own causal self-attention layers. The text decoder generates text auto-regressively, and cross-entropy loss is applied during this process.

For BLIP, we explored various pre-training approaches. Initially, we attempted to pre-train BLIP from scratch. Subsequently, we pre-trained BLIP from a provided checkpoint using the ROCO and MIMIC datasets. Further experimentation involved extending the checkpoint with the inclusion of the ROCO dataset onto the MIMIC dataset. To apply BLIP to downstream tasks, we followed the BLIP framework's process to refactor the model's modules and assembled an adapted model tailored for specific tasks.

BioMedBLIP for VQA Generation Model

For Visual Question Answering (VQA) tasks, we adopted the structure provided by BLIP, as depicted in Figure 2. VQA tasks require the model to generate textual answers based on given images and question pairs. The process involves encoding the image to create image embeddings, producing image-question joint embeddings with the help of the question encoder, and using the answer decoder to generate the final answer.

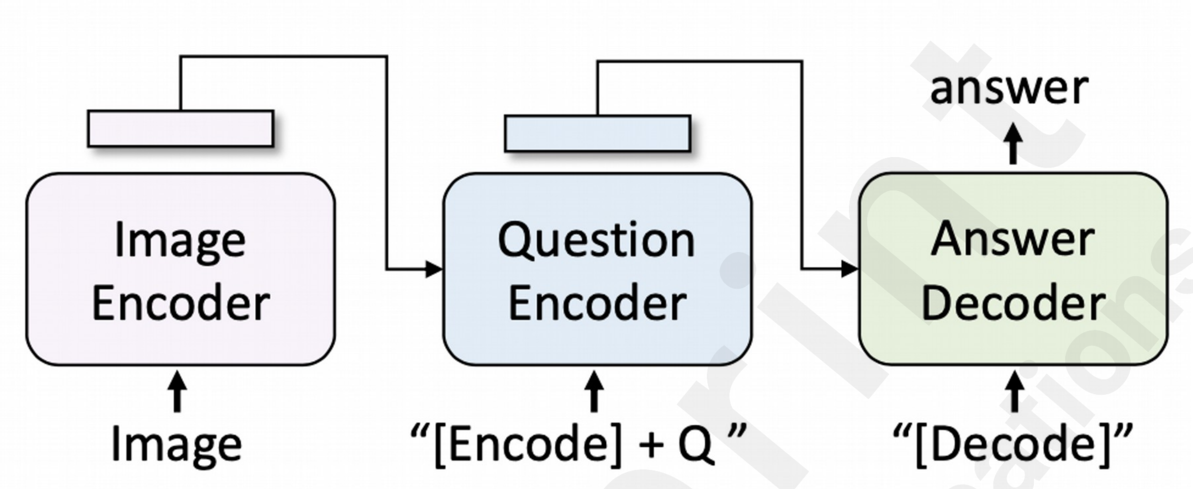


Figure 2 BioMedBLIP VQA Generation model

Modified BLIP Classification Model

The modified BLIP Classification model, illustrated in Figure 3 shares similarities with the generation model. It generates joint image-text embeddings using the image encoder and text encoder. However, instead of utilizing the answer decoder, a pooling layer is introduced to reduce the vector dimension. Subsequently, a linear classification layer is applied to produce multiple classification results.

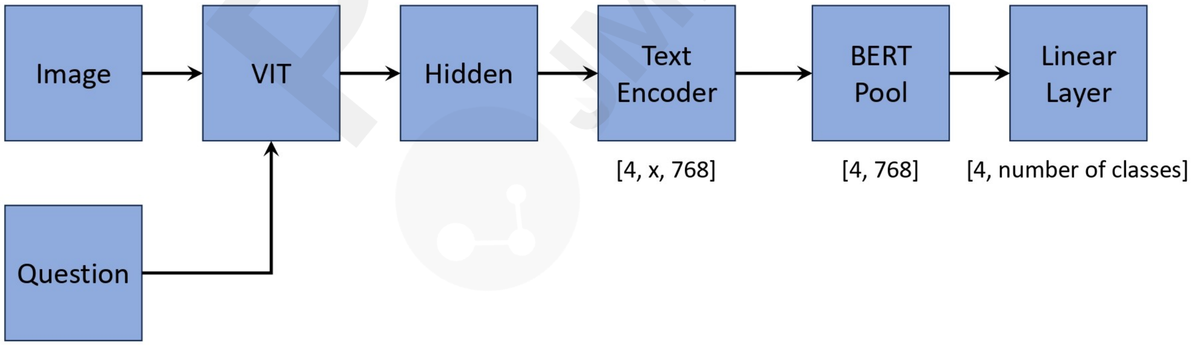


Figure 3 BioMedBLIP Classification Model

BioMedBLIP Image Caption Model

The Image Caption model, presented in Figure 4, is composed of the image encoder and text decoder, following BLIP's implementation. Unlike VQA tasks, the image caption task involves generating text based solely on images. Therefore, the text encoder is omitted, and the text decoder takes image embeddings provided by the image encoder and the `[Decode]' token as input to produce image captions.

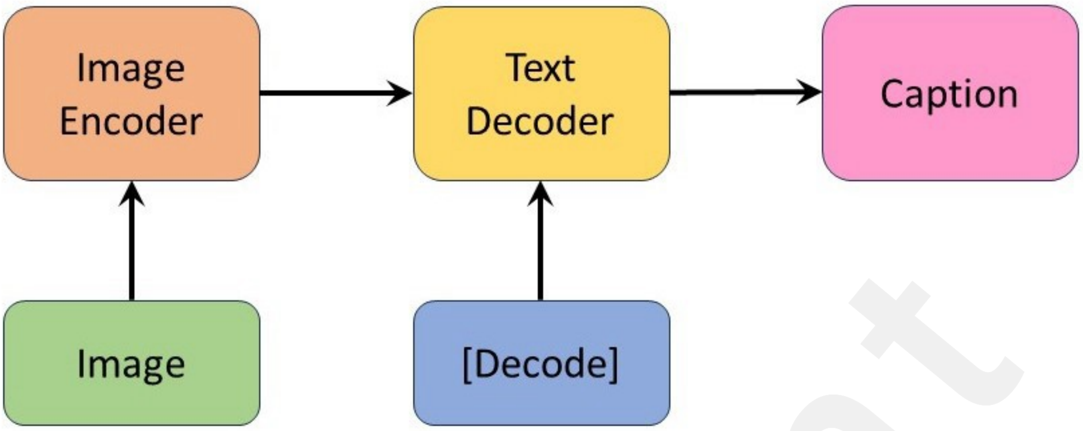


Figure 4 BioMedBLIP Image Caption Model

Datasets

Our research leverages a diverse range of medical datasets, encompassing a variety of visual and textual medical data sources. These datasets serve as the foundation for pretraining and fine-tuning our visual language model.

ROCO Dataset

The ROCO dataset [5] plays a pivotal role in pre-training BioMedBLIP. It encompasses over 81,000 radiology images representing multiple medical imaging modalities, including Computer Tomography (CT), Ultrasound, X-ray, Fluoroscopy, Positron Emission Tomography (PET), Mammography, Magnetic Resonance Imaging (MRI), and Angiography [5]. Our approach involved consolidating the training, validation, and test data into a single comprehensive JSON file, enabling the pre-training of BioMedBLIP. Notably, the captions in the ROCO dataset are sourced from peer-reviewed scientific biomedical literature and downloaded from GitHub link provided at [31]. Some examples are shown in Figure 5.

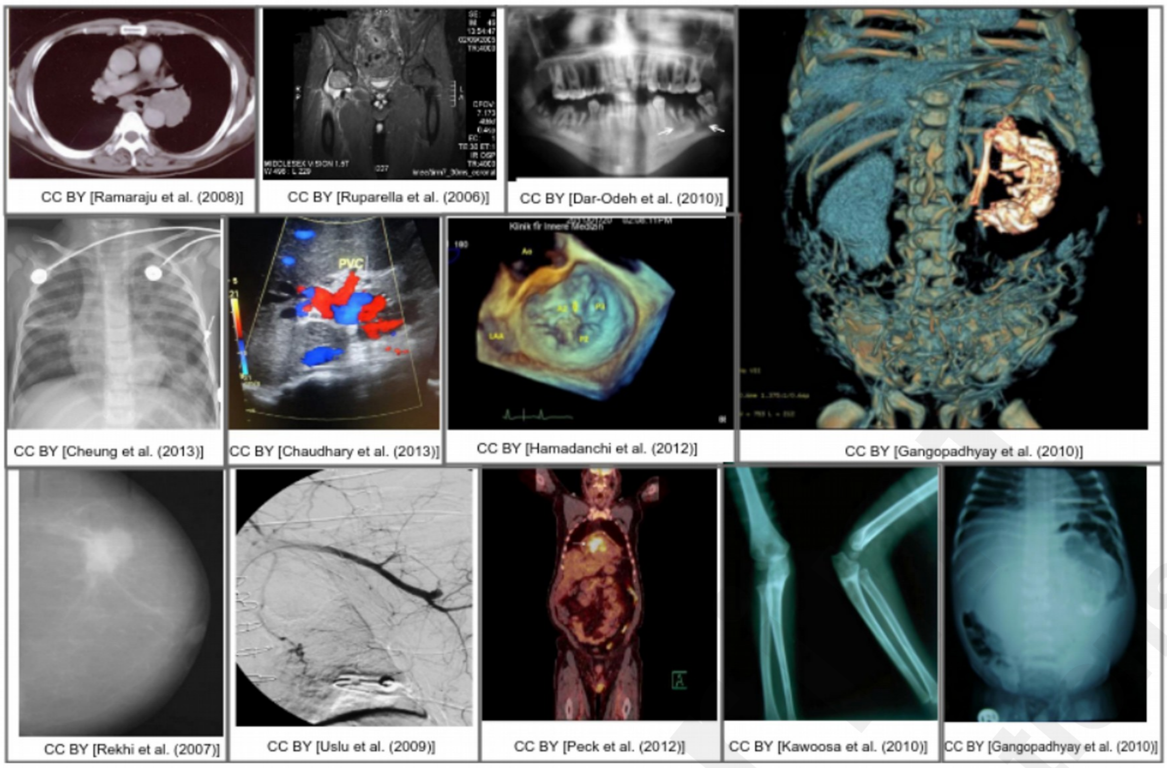

CC BY [Ramaraju et al. (2008)]
CC BY [Ruparella et al. (2006)]
CC BY [Dar-Odeh et al. (2010)]
CC BY [Cheung et al. (2013)]
CC BY [Chaudhary et al. (2013)]
CC BY [Hamadanchi et al. (2012)]
CC BY [Gangopadhyay et al. (2010)]
CC BY [Rekhi et al. (2007)]
CC BY [Uslu et al. (2009)]
CC BY [Peck et al. (2012)]
CC BY [Kawoosa et al. (2010)]
CC BY [Gangopadhyay et al. (2010)]

Figure 5 Some images in the ROCO dataset

MIMIC-CXR Dataset

The MIMIC-CXR dataset is a large dataset that consists of 377,110 chest X-rays corresponding to 227,827 imaging studies [6]. Some examples of chest X-rays from this dataset are as shown in Figure 6. In our context, we utilized this dataset for BioMedBLIP's pre-training. It's worth noting that each medical study extracted from the hospital's Electronic Health Record (EHR) system can be related to multiple chest X-rays. Our efforts focused on filtering the chest X-rays, retaining those with `AP' and `PA' positions, and ensuring that each medical report had a single associated chest X-ray. Post-processing, we obtained 218,139 image-caption pairs, which were instrumental in pre-training BioMedBLIP. The medical studies are XML files, and we extracted the `Findings' and `Impressions', as the caption for medical images.

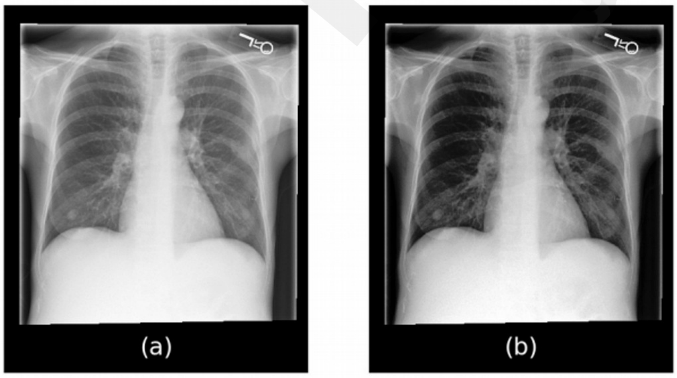


Figure 6 Chest X-rays in MIMIC-CXR dataset

ImageCLEF 2019 Dataset

The ImageCLEF 2019 dataset [24,25], provided by the ImageCLEF organization for evaluation, served as a critical component in our work. This dataset comes in three parts: training, validation, and testing sets. No preprocessing was performed on the dataset, and it was leveraged for our VQA-Generation task. It contains 12,792, 2,000, and 500 image-caption pairs for the training, validation, and testing sets, respectively. An example of a radiology image in the ImageCLEF dataset is shown in Figure 7.

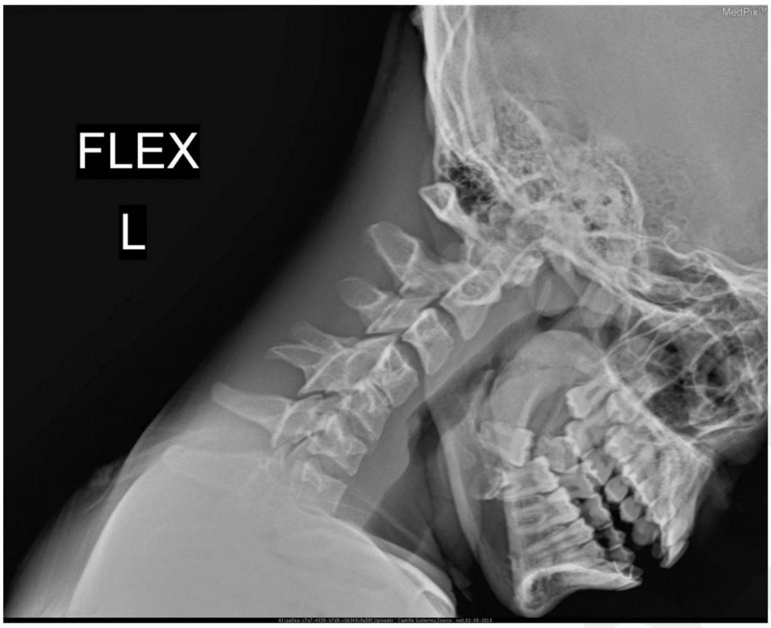


Figure 7 Radiology image in ImageCLEF 2019 dataset

SLAKE Dataset

The SLAKE dataset [32] was designed for medical visual question-answering tasks Liu et. al. [26]. We employed this dataset for VQA generation and VQA classification tasks. The SLAKE dataset has both Chinese question-answer pairs and English question-answer pairs. Our dataset preparation included filtering to retain only English question-answer pairs. After filtering, the SLAKE dataset consisted of 4,919, 1,053, and 1,061 image-caption pairs for the training, validation, and testing sets, respectively. Notably, the SLAKE dataset features two different answer types: Open and Close, allowing us to assess model performance for open-ended and close-ended questions. An example of radiology image from SLAKE dataset is shown in Figure 8.

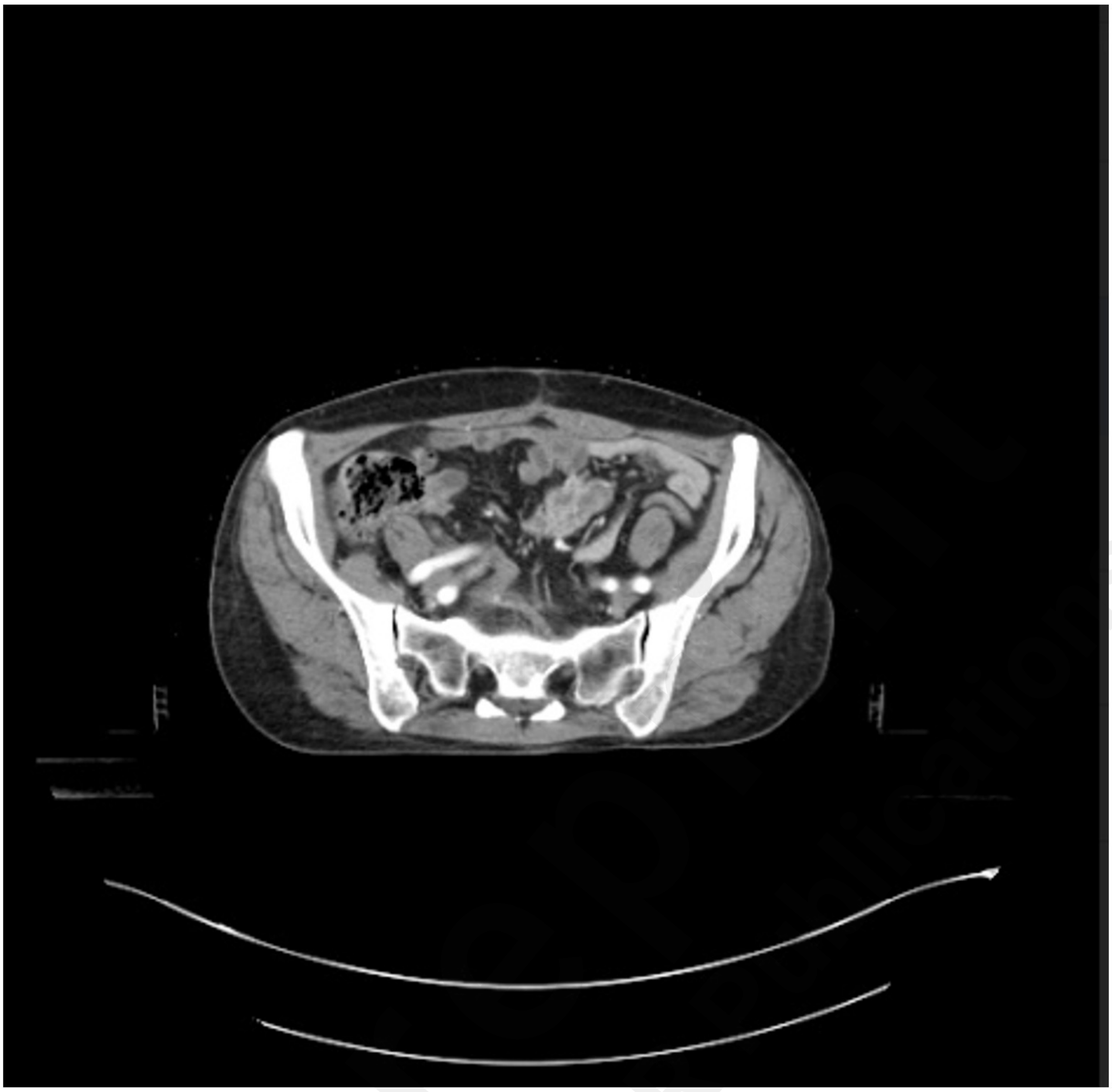


Figure 8 A radiology image from SLAKE dataset

PathVQA Dataset

The PathVQA dataset is a visual question-answering dataset for `AI Pathologist' development He et. al. [27]. It contains numerous pathology images together with questions and corresponding answers. All image-question-answer pairs are manually checked to ensure correctness. In our setting, we obtained a dataset with 32,795 image-question-answer pairs after initial preprocessing. We further categorized questions into open-ended and close-ended using our own splitting, yielding 20,968, 5,241, and 6,552 image-question-answer pairs for the training, validation, and testing sets. All of these images in the validation and testing sets are picked randomly. An example of a pathology image from the PathVQA dataset is shown in Figure 9. We use this dataset to do the VQA-Generation and VQA-Classification tasks in our project.

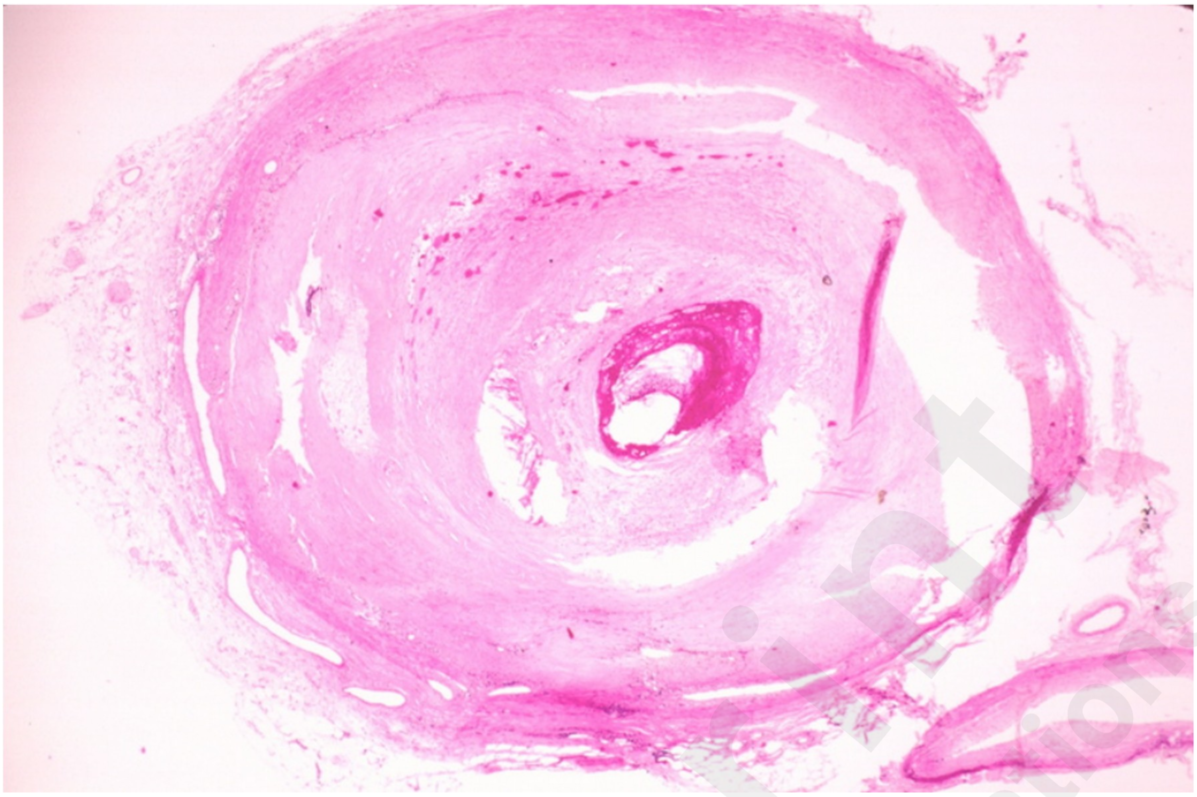


Figure 9 A pathology image from PathVQA dataset

VQA-RAD Dataset

VQA-RAD is the first dataset that was manually constructed. During the data collection process, clinicians asked natural questions about radiology images. Meanwhile, their reference answers would be provided [28]. It has radiology images together with question-answer pairs. We used the original data [33] splitting and we did not do any data preprocessing on this dataset. It contains 2452, 614, and 452 image-question-answer pairs for training, validation, and testing sets. A sample of radiology image from the VQA-RAD dataset is shown in Figure 10. We utilized the VQA-RAD dataset to implement the VQA-Generation and VQA-Classification tasks.

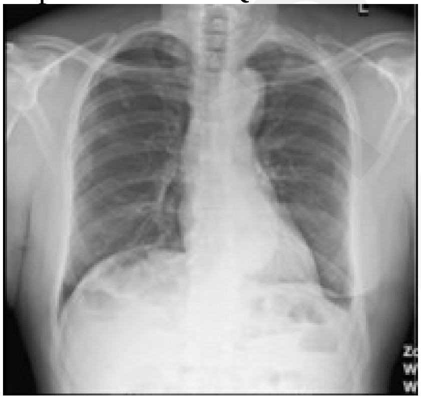


Figure 10 A radiology image from VQA-RAD dataset

Open-I Dataset

The Open-I dataset [34] is a compilation of chest X-ray images collected from open-source literature and biomedical image collections. Our focus was specifically on the chest X-ray images within this dataset. We downloaded the dataset from the official Open-I website, which comprises two parts: images and medical reports. The medical reports are stored as XML files, with the `Finding' and

`Impression' sections extracted as captions for the images. Our downloaded version contained 2,452 image-caption pairs for the training, 614 for validation, and 452 for the testing set. An example of a radiology image from Open-I dataset is shown in Figure 11.

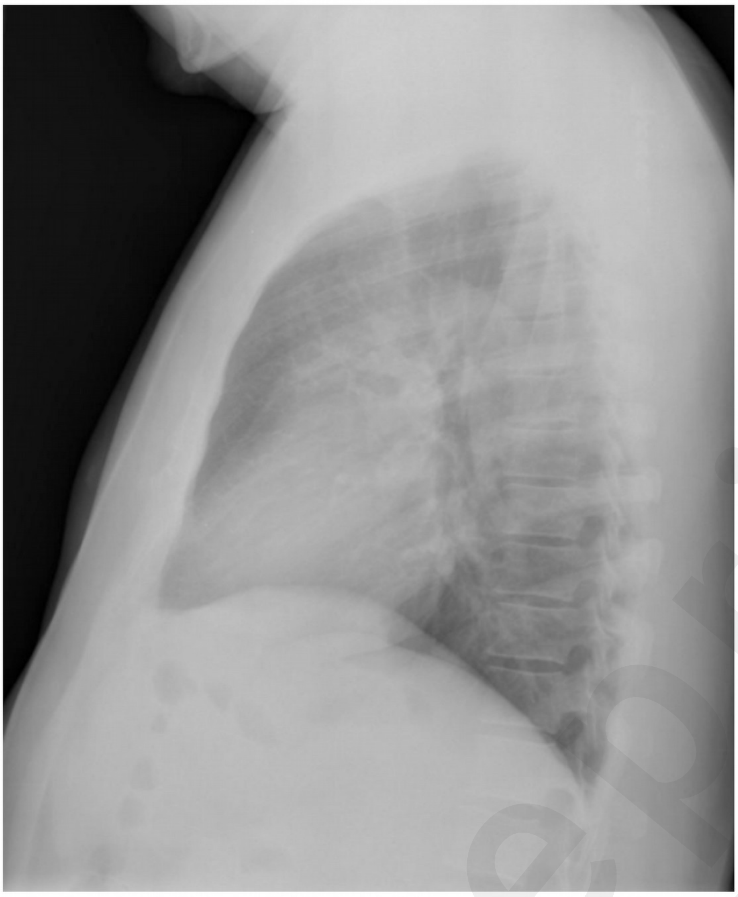


Figure 11 A radiology image from Open-I dataset

PEIR-Gross Dataset

The PEIR-Gross dataset originated from the Pathology Education Informational Resource (PEIR) and contains 7,442 image-caption pairs across 21 sub-categories [8] Our dataset preparation involved splitting it into training and testing sets. We also generated a validation set by randomly selecting 10% of the training data. After preprocessing, we had 6,029 image-caption pairs for the training, 669 for validation, and 745 for the testing set. An example of medical image from the PEIR-Gross dataset is shown in Figure 12. This dataset was employed for image captioning tasks.

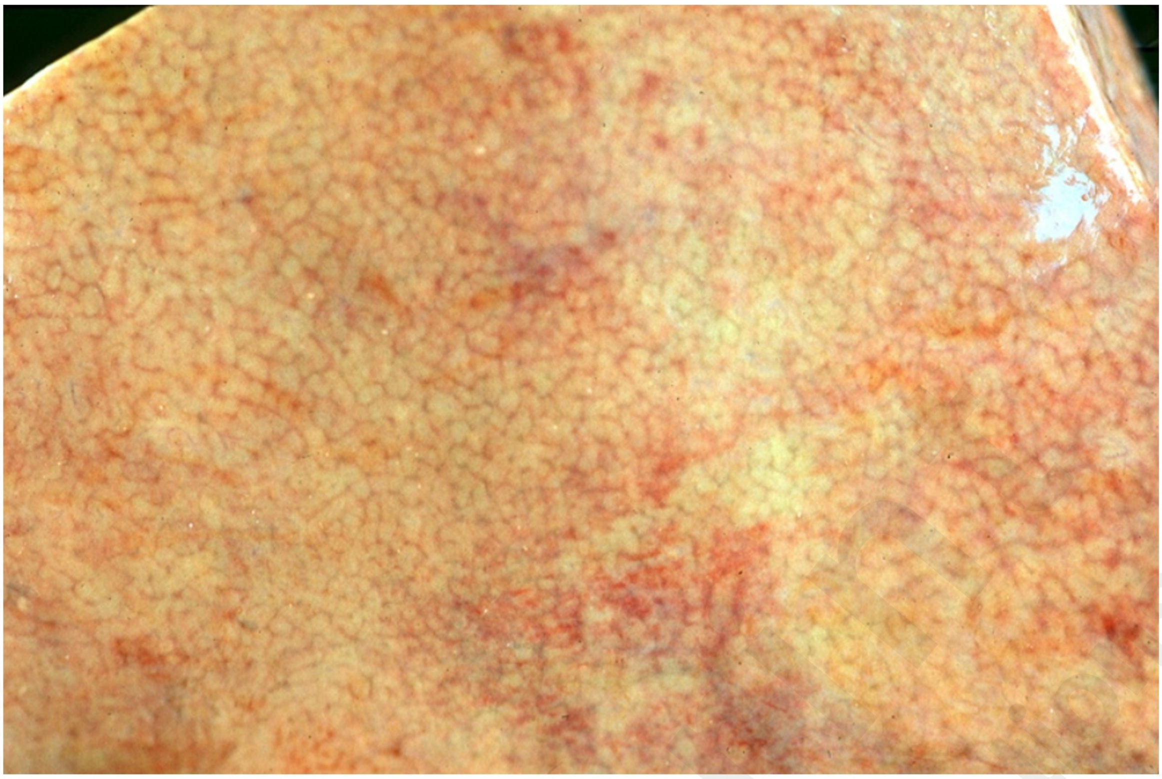


Figure 12 A medical image from PEIR-Gross dataset

Implementation

In addition to the original BLIP base model checkpoint obtained from their BLIP GitHub repository [35], we also meticulously pre-trained a series of checkpoints on various medical datasets. These checkpoints serve as critical resources that we are making available for the broader research community and our clients. Here, we provide a comprehensive list of these checkpoints:

- **Original BLIP Base Model Checkpoint (VIT-Base):** This checkpoint represents the foundational BLIP model pre-trained on the LAION115M dataset.

- **Pre-training on Combined-MED Checkpoints:** These checkpoints are derived from pre-training on a comprehensive dataset combining SLAKE, OPEN I, ImageClef, and PathVQA datasets from scratch, with versions available for both 20 and 50 epochs.

- **Pre-training on SLAKE Checkpoint:** This checkpoint is the result of pre-training on the SLAKE dataset from scratch, spanning 20 epochs.

- **Pre-training on ROCO Checkpoint**: Pre-trained on the ROCO dataset from scratch, this checkpoint encapsulates knowledge gained over 10 epochs.

- **Pre-training on MIMIC-CXR Checkpoint:** For up to 10 epochs, this checkpoint embodies the insights obtained from pre-training on the MIMIC-CXR dataset from scratch.

- **Pre-training on ROCO from the BLIP Original Checkpoint:** This checkpoint extends the pre-training on the ROCO dataset from the existing BLIP checkpoint and is extended up to 50 epochs.

- **Pre-training on MIMIC-CXR from the BLIP Original Checkpoint**: Similar to the ROCO extension, this checkpoint involves the pre-training on the MIMIC-CXR dataset from the original BLIP checkpoint and spans up to 50 epochs.

- **Pre-training on ROCO & MIMIC-CXR from the BLIP Original Checkpoint:** This checkpoint represents an amalgamation of knowledge acquired from both ROCO and MIMIC-CXR datasets, building upon the original BLIP checkpoint and extending to 50 epochs.

Pre-training Details

To undertake the pre-training of the BLIP model, our approach involved several key steps. Initially, we utilized the ROCO dataset for the pre-training of the BLIP original checkpoint, resulting in the creation of the BLIP-ROCO models. These models were uniquely identified by the number of epochs they were pre-trained for, with ``BLIP-ROCO-10", for instance, signifying the original BLIP checkpoint pre-trained on the ROCO dataset for 10 epochs. This choice was informed by the fact that the original BLIP checkpoint had been trained on millions of standard images and, therefore, already possessed the fundamental knowledge required for a Visual Language model. Our attempt to pre-train BLIP from scratch yielded unsatisfactory performance. Subsequently, we embarked on pre-training the original BLIP checkpoint with the MIMIC-CXR dataset to produce the BLIP-MIMIC models. Finally, we took the BLIP-MIMIC checkpoints and further pre-trained them using the ROCO dataset, resulting in the creation of the BLIP-ROCO& MIMIC models. In this paper, we retained checkpoints representing 10, 20, and 50 epochs for each of these models.

To optimize the pre-training process, we conducted a series of experiments aimed at identifying the most suitable hyperparameters. The selected hyperparameters for the pre-training process are as follows:

- Initial learning rate: $3e^{-5}$
- Warmup learning rate: $1e^{-6}$
- Warmup steps: 3000
- Optimizer: AdamW

Fine-tuning Details

The fine-tuning phase involved an extensive process encompassing various BLIP checkpoints and datasets, which included the BLIP original base model checkpoint, BLIP-ROCO-10, BLIP-ROCO-20, BLIP-ROCO-50, BLIP-MIMIC-10, BLIP-MIMIC-20, BLIP-MIMIC-50, BLIP-ROCO&MIMIC-10, BLIP-ROCO&MIMIC-20, and BLIP-ROCO&MIMIC-50. These checkpoints were fine-tuned on a selection of datasets, namely, ImageClef, SLAKE, VQA-RAD, PathVQA, Open I (IU X-RAY), and PEIR Gross.

For the fine-tuning process, we adhered to the common practice of splitting the datasets into training, validation, and test sets. Typically, the training set constitutes 80% of the total dataset, while the test set encompasses the remaining 20%. In certain cases, the dataset authors had already performed the necessary dataset splits, and we made no further modifications to these sets. Additionally, we created an answer list to facilitate the evaluation of predicted VQA generation sentences, VQA classification labels, and image captions.

As part of our methodology, we employed a YAML configuration file, which proved essential for adapting to different running environments. Through a series of meticulously designed experiments, we optimized the hyperparameters for each checkpoint's fine-tuning process. These hyperparameters included train batch size, test batch size, learning rates, and the number of epochs. Importantly, the optimal hyperparameters varied for each checkpoint when applied to different datasets, ensuring the fine-tuning process was meticulously tailored for each specific scenario.

Resources

This paper relies on a combination of hardware and software resources to execute efficiently. We utilize three primary platforms for code execution: Google Colab, Google Cloud Platform, and the University of Sydney's Artemis HPC super-computer. Google Colab we used had an Intel(R) Xeon(R) CPU running at 2.30GHz, an Nvidia Tesla P100-PCIE-16GB GPU, and 12.8GB of RAM. Google Cloud Platform we used had 4 CPU cores, a Tesla A100-PCIE-40GB GPU, and 26GB of RAM. On the Artemis HPC platform [36], we have access to an impressive array of resources, including 7,636 CPU cores, 45 TB of RAM, 108 NVIDIA V100 GPUs, 378 TB of storage, and 56 Gbps FDR InfiniBand networking. For our pre-training tasks, we specifically use 4 CPU cores, Tesla V100-PCIE GPUs, and 48 GB of RAM. Our code is available on HuggingFace [30].

Results

In this section, we describe our evaluation metrics along with the results and findings. The purpose of our experimentation is to evaluate how adding domain-specific information helps in improving performance of downstream tasks in biomedical domain. For this reason, we have compared the results of our models with original BLIP model.

Evaluation Metrics

In this subsection, we provide a detailed overview of the specific evaluation metrics employed to assess the performance of our model across various downstream tasks.

- **VQA Generation:** For VQA generation, we used exact match (EM) as the metric. EM assesses the model's performance by treating predictions that precisely match the ground truth as correct answers. It is particularly relevant for evaluating generative tasks.
- **VQA Classification Metrics:** For VQA classification, we use accuracy. Accuracy is a fundamental metric for classification tasks, quantifying the proportion of correctly classified instances.
- **Report Generation or Image Captioning Metrics:** BLEU, or Bilingual Evaluation Understudy, is a metric that assesses the similarity between the generated answer and the reference answer, considering n-grams. BLEU-1, in particular, focuses on 1-grams.

VQA-Generation Results

For the VQA-Generation task, we use SLAKE, VQA-RAD, PathVQA, and ImageCLEF datasets to check the performance of our model BLIP-Original, BLIP-ROCO, and BLIP-MIMIC&ROCO. These models are designed to generate answers for visual questions and have undergone different pre-training strategies. Table 1 and Table 2 present a comparison of BioMedBLIP models with the SOTA model.

Table 1 Comparison of BioMedBLIP models vs SOTA on VQA-Generation Tasks (part 1)

Dataset	Original BLIP (SOTA)	BioMedBLIP Models			
		ROCO-10	ROCO-20	ROCO-30	MIMIC-10

SLAKE-Overall	77.95	80.87	80.11	80.21	78.51
SLAKE-Open	73.80	75.81	74.57	75.04	73.80
SLAKE-Close	87.26	88.70	88.70	89.42	85.82
VQA-RAD-Overall	34.37	35.70	**37.03**	35.03	26.16
VQA-RAD-Open	39.66	43.02	**46.37**	43.02	26.82
VQA-RAD-Close	30.88	30.51	30.88	29.78	25.74
PathVQA-Overall	**66.64**	63.00	64.65	55.46	51.45
PathVQA-Open	**43.78**	38.45	40.79	23.84	18.26
PathVQA-Close	88.35	87.62	**88.57**	87.16	84.75
ImageCLEF	48.20	**58.27**	56.81	57.63	56.41

Table 2 Comparison of BioMedBLIP models vs SOTA on VQA-Generation Tasks (part 2)

Dataset	BioMedBLIP Models				
	MIMIC-20	MIMIC-50	MIMIC & ROCO-10	MIMIC & ROCO-20	MIMIC & ROCO-50
SLAKE-Overall	79.92	70.50	**82.00**	81.53	81.34
SLAKE-Open	75.50	65.12	76.28	76.28	**76.59**
SLAKE-Close	86.78	78.85	**90.87**	76.28	88.70
VQA-RAD-Overall	30.38	25.50	32.59	29.93	35.70
VQA-RAD-Open	36.87	22.35	32.96	29.05	40.78
VQA-RAD-Close	26.10	27.57	**32.35**	30.51	**32.35**
PathVQA-Overall	53.31	50.89	61.74	54.09	60.27
PathVQA-Open	20.79	17.71	36.22	22.62	33.69
PathVQA-Close	85.91	84.14	87.32	85.64	86.92
ImageCLEF	52.27	53.63	19.89	54.80	56.42

Results on SLAKE Dataset

For the overall SLAKE dataset, the BioMedBLIP MIMIC&ROCO-10 model exhibits the highest exact match accuracy, reaching an impressive 82.00%, outperforming the BLIP Original model and other variants. The results in SLAKE-Open highlight the superiority of the BioMedBLIP MIMIC&ROCO-50 model, which achieves the best performance with an exact match accuracy of 76.59%. Lastly, in the SLAKE-Close category, the BioMedBLIP MIMIC&ROCO-10 model stands out with an impressive accuracy of 90.87%, demonstrating its strong performance in generating answers that match ground truth answers exactly.

Results on VQA-RAD Dataset

The original BLIP (SOTA) model, serving as the baseline, achieved an exact match score of 34.37 in the ``VQA-RAD-Overall" category. However, it was surpassed by our BioMedBLIP model, specifically ``ROCO-20", which demonstrated exceptional performance with an exact match score of 37.03, indicating its effectiveness in generating accurate answers to visual questions. This trend continued in the ``VQA-RAD-Open" dataset, where BioMedBLIPROCO-20 outperformed the baseline with an exact match score of 46.37, highlighting its strong performance in open-ended VQA tasks. Notably, the close-ended category also saw success for the BioMedBLIP models, with ``MIMIC&ROCO-10" and ``MIMIC&ROCO-50" achieving the top exact match score of 32.35.

Results on PathVQA Dataset

In the evaluation of VQA-Generation tasks on the PathVQA dataset, the Original BLIP (SOTA) model exhibited an EM score of 66.64, setting a high standard. Among the BioMedBLIP models, ``ROCO-20" emerged as the top performer in the PathVQA-Overall and PathVQA-Open categories, achieving scores of 64.65 and 40.79, respectively. Particularly noteworthy is ``MIMIC&ROCO-10", which achieved a competitive EM score of 61.74 in the PathVQA-Overall task. In the PathVQA-Close category, ``BioMedBLIP-ROCO-20" stood out with an EM score of 88.57, surpassing the Original BLIP (SOTA) model. The BioMedBLIP models ``MIMIC&ROCO-10" and ``MIMIC&ROCO-50" also displayed strong performance. These results emphasize the effectiveness of different pre-training strategies and the potential for improved performance in VQA-Generation tasks using the PathVQA dataset with BioMedBLIP.

Results on ImageCLEF Dataset

The SOTA Original BLIP model achieves an Exact Match score of 48.20. In contrast, the BioMedBLIP models, which are pre-trained with different datasets and epochs, demonstrate notable improvements. Notably, the BioMedBLIP model pre-trained with ROCO (10 epochs) emerges as the top performer with an impressive Exact Match (EM) score of 58.27, surpassing the SOTA model. This suggests that the use of the ROCO dataset for pre-training significantly enhances the ability of the model to generate precise answers to visual questions on the ImageCLEF dataset. Other variants of the BioMedBLIP model, pre-trained with different datasets and epochs, also exhibit varying degrees of success in this task.

VQA-Classification Results

For the VQA-Classification task, we use SLAKE, VQA-RAD, and PathVQA datasets to check the performance of our model BLIP-Original, BLIP-ROCO, and BLIP-MIMIC-CXR. These models are designed to generate answers for visual questions and have undergone different pre-training strategies. Table 3 and Table 4 presents a comparison of BioMedBLIP models with the SOTA models.

Table 3 Comparison of BioMedBLIP Models vs SOTA on VQA-Classification Tasks (part 1)

Dataset	Original BLIP (SOTA)	BioMedBLIP Models			
		ROCO-10	ROCO-20	ROCO-30	MIMIC-10
SLAKE-Overall	77.85	**81.06**	80.21	80.04	78.70
SLAKE-Open	75.50	75.66	77.05	77.52	75.66
SLAKE-Close	81.49	83.31	**85.10**	84.86	83.41
VQA-RAD-Overall	**40.35**	33.70	19.96	23.95	34.36
VQA-RAD-Open	20.67	27.37	25.10	26.33	**28.49**
VQA-RAD-Close	**51.84**	39.71	33.09	39.71	38.23
PathVQA-Overall	60.09	59.25	57.77	58.65	58.85
PathVQA-Open	**37.21**	33.60	30.06	33.17	34.05
PathVQA-Close	85.15	84.96	85.54	84.20	83.71

Table 4 Comparison of BioMedBLIP Models vs SOTA on VQA-Classification Tasks (part 2)

Dataset	BioMedBLIP Models				
	MIMIC-20	MIMIC-50	MIMIC & ROCO-10	MIMIC & ROCO-20	MIMIC & ROCO-50
SLAKE-Overall	77.57	74.18	73.90	80.89	69.10
SLAKE-Open	74.88	72.25	71.62	**77.90**	71.32
SLAKE-Close	81.73	77.16	76.69	84.77	68.04
VQA-RAD-Overall	33.70	34.15	34.59	29.49	31.49
VQA-RAD-Open	**28.49**	**28.49**	**28.49**	26.26	27.93
VQA-RAD-Close	37.13	37.87	38.69	31.62	33.82

PathVQA-Overall	58.04	36.86	60.41	60.24	**61.13**
PathVQA-Open	31.98	18.96	32.61	33.32	34.09
PathVQA-Close	84.17	54.80	59.70	85.38	**86.29**

Results on SLAKE Dataset

In the context of VQA-Classification tasks using the SLAKE dataset, the evaluation results are presented in terms of accuracy, allowing for a comprehensive comparison between the Original BLIP model and several BioMedBLIP variants, each pre-trained with specific datasets and epochs. The BioMedBLIP models demonstrate their potential for significant improvements over the Original BLIP model. In the ``SLAKE-Overall" dataset category, BioMedBLIP models pre-trained with the ROCO dataset consistently outperform the Original BLIP, with ROCO-10 achieving an accuracy of 81.06. Furthermore, BioMedBLIP models pre-trained with MIMIC-CXR data, particularly MIMIC-20, showcase strong performance. For the slake open-ended dataset subtask, the superiority of the model pre-trained with both MIMIC-CXR and ROCO for 20 epochs, achieves an accuracy of 77.90, surpassing the Original BLIP. For the SLAKE close-ended dataset, the ROCO-20 variant of BioMedBLIP stands out with an accuracy of 85.10, clearly surpassing the Original BLIP's performance. These results emphasize the significance of dataset choice and pre-training duration, with BioMedBLIP models showcasing their potential for improved accuracy in classifying visual questions within the SLAKE dataset.

Results on VQA-RAD Dataset

In the VQA-RAD-Overall dataset, encompassing both open-ended and closed-ended questions, none of the BioMedBLIP models outperforms the Original BLIP, highlighting the challenges in achieving superior accuracy in a mixed question-type. However, in the VQA-RAD open-ended dataset, all of the BioMedBLIP models excel, surpassing the Original BLIP's accuracy, showcasing their effectiveness in open-ended question answering. Four variants of BioMedBLIP models viz. MIMIC-10, MIMIC-20, MIMIC-50, and MIMIC & ROCO-10 have the highest scores of 28.49. For the VQA-RAD-Closed Ended dataset, the BioMedBLIP models also do not surpass the Original BLIP, with the best performers being ROCO-10 and ROCO-30. This shows that further research is warranted on strategies to improve the performance on the close-ended VQA-RAD dataset.

Results on PathVQA Dataset

On the PathVQA-Overall dataset, the Original BLIP model achieved an accuracy of 60.09, while the BioMedBLIP models demonstrated varied performance. Notably, the model pre-trained with both MIMIC-CXR and ROCO datasets for 50 epochs emerged as the top performer, achieving an accuracy of 61.13. This dual pre-training approach showed significant promise in enhancing classification accuracy. For the PathVQA-Open dataset, the Original BLIP model achieved the highest accuracy of 37.21, with BioMedBLIP models pre-trained on ROCO and MIMIC-CXR data yielding slightly lower results. The highest performance among BioMedBLIP models came from the model pre-trained with both datasets for 50 epochs, reaching an accuracy of 34.09, which is very close to the Original BLIP model. In contrast, the Original BLIP model excelled on the PathVQA-Close dataset, achieving an impressive accuracy of 85.15. Nevertheless, the BioMedBLIP model pre-trained with both MIMIC-CXR and ROCO for 50 epochs outperformed the Original BLIP, achieving an accuracy of 86.29.

These results collectively indicate that the choice of pre-training strategy and the duration of training significantly influence the classification accuracy of BioMedBLIP models on the various VQA classification datasets, with the combined dataset pre-training demonstrating notable advantages in certain contexts.

Image-Caption Task Results

We used the PEIR-Gross dataset for the image-caption task to check our BioMedBLIP's performance. Figure 13 displays the results of the image-caption task when using different pre-trained BioMedBLIP models on the PEIR-Gross dataset. Specifically, it shows the BLEU-1 scores for various models. In terms of BLEU-1 scores, higher values are indicative of better performance. The results show that the BioMedBLIP-ROCO-50 model surpassed the original BLIP model on the image-captioning task with a BLEU-1 score of 25.1 (over 1.2% improvement from the original BLIP), demonstrating that our approach has the potential to enhance the model's capabilities for generating captions. On the other hand, all other models, including various pre-training strategies and epochs, exhibited slightly lower BLEU-1 scores. While some models may exhibit minor variations in performance, it's essential to emphasize that, overall, our results are quite consistent. This consistency indicates that our pre-training strategies and fine-tuning approaches are robust and capable of producing reliable outcomes.

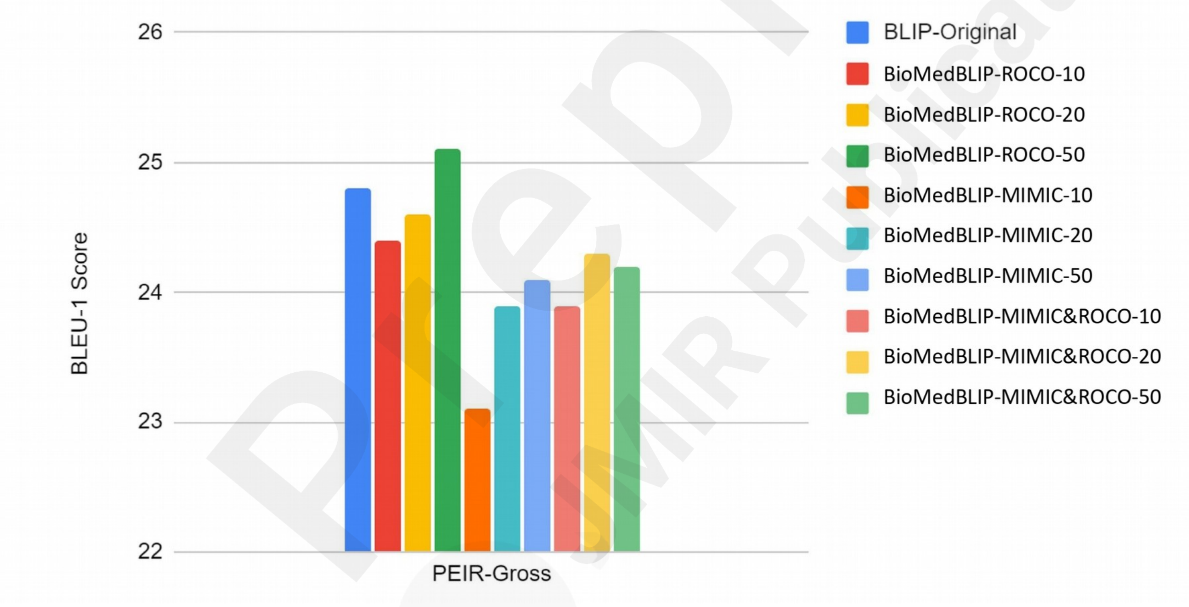


Figure 13 Visualization of the BioMedBLIP models vs SOTA (BLIP) on Image-Caption task for the PEIR-Gross dataset

Discussion

In the presented VQA-Generation, VQA-Classification, and report generation results, we aimed to assess the performance of our BioMedBLIP models against the state-of-the-art (SOTA) Original BLIP model using diverse medical image datasets and pre-training strategies. The findings indicate that our BioMedBLIP models, pre-trained with specialized medical datasets, exhibit substantial improvements in generating answers for visual questions, as well as in classifying images and questions, depending on the specific dataset and pre-training strategy.

In the discussion, it's important to emphasize the general trends and insights that can be drawn from these results:

- **Dataset Specificity:** The choice of pre-training datasets, such as ROCO and MIMIC-CXR, significantly impacts model performance in both VQA-Generation and VQA-Classification tasks. Specialized medical datasets have proven to be valuable for enhancing model capabilities in medical image analysis and question-answering.

- **Pre-Training Duration:** Longer pre-training durations, as evidenced by models like MIMIC&ROCO-50, have shown their potential to improve classification accuracy in specific categories of questions, demonstrating the importance of considering pre-training strategies tailored to the task.

- **Diverse Performance:** Our models exhibit varying levels of success across different datasets and task categories. This underscores the need for flexibility in selecting pre-training strategies, depending on the specific goals and datasets of a given application.

- **Image Captioning:** Our approach, particularly BioMedBLIP-ROCO-50, demonstrates improvements in generating captions for medical images, showcasing the model's capacity to excel in both VQA and image captioning tasks.

- **Selection of Epoch Numbers:** In our experiments, we observed that the BioMedBLIP models converge at varying numbers of epochs, depending on the dataset and the specific downstream tasks involved. This variability is typical in deep learning, where a model's loss decreases up to a certain point and then may increase, indicating that longer training periods do not necessarily yield better performance.

These results underscore the potential of our BioMedBLIP models to excel in a wide range of medical image analysis and question-answering tasks, with their performance varying depending on the specific dataset and task at hand. BioMedBLIP was tested using 20 different datasets and task combinations. our method excels in 15 out of 20 tasks. BioMedBLIP represents a new state-of-the-art in 15 out of 20 tasks (75%) and our responses were rated higher in all 20 tasks ($P < 0.005$) in comparison to SOTA models. Regression analyses showed that our model’s VQA-Generation has statistically significant predictor ($P<0.002$) on SLAKE, PathVQA, and ImageCLEF datasets. On the other hand, the VQA-Classification has a relatively lower predictor ($P<0.003$) on as SLAKE, PathVQA, and VQA-RAD datasets in accordance with the regression analyses.

VQA-Generation

For the SLAKE dataset, our model, pre-trained on a combination of general domain datasets and medical domain datasets (MIMIC-CXR and ROCO), consistently outperforms other models across various SLAKE datasets, including open-ended, close-ended, and aggregated types. In contrast to the study by Li et. al. [12], our results affirm the benefits of pre-training on both general and medical datasets, addressing limitations in their work. This underscores the advantage of a domain-specific model for specialized downstream tasks. Notably, our observations align with the findings of Eslami et. al. [29], indicating that a pre-trained Vision Transformer (ViT), such as our BLIP model, possesses a comprehensive understanding of image content and long-range dependencies, essential for interpreting the SLAKE dataset. Surprisingly, in the VQA-RAD dataset, the model pre-trained solely on the ROCO dataset (for 20 epochs) excels in open-ended and aggregated tasks, while the MIMIC-CXR and ROCO model performs better in the close-ended task. Contrary to expectations, the model pre-trained on general domain datasets and ROCO outperforms larger domain-specific datasets for PathVQA and ImageCLEF. We attribute this to superior preprocessing of the ROCO dataset, incorporating red bounding boxes that aid in learning crucial image regions. Moreover, our 50-epoch pre-training might not suffice for larger datasets, suggesting the need for further exploration with extended training. Overall, our models demonstrate superior performance on the medical datasets compared to the original BLIP model, emphasizing the efficacy of our approach in medical image analysis.

VQA-Classification

In the exploration of VQA classification tasks, diverse models undergo experimentation and fine-tuning across datasets such as SLAKE, PathVQA, and VQA-RAD, encompassing open-ended, close-ended, and aggregated question types. Notably, the SLAKE dataset consistently emerges as the dataset where BioMedBLIP models consistently exhibit superior performance. Table 3 and Table 4 highlights that, in majority of the cases, the BioMedBLIP models perform higher than the original BLIP model.

Furthermore, our experiments delve into the impact of different epoch settings on model performance. An intriguing observation emerges, indicating that the relationship between a model's performance and the number of epochs is not consistently positive. This suggests the critical importance of judiciously selecting epoch configurations during the construction of visual language models for medical datasets, challenging the notion that more epochs always lead to improved accuracy.

Image Captioning

In the context of the image-captioning task, our findings, while not entirely satisfactory, reveal promising aspects, particularly with the BLIP-ROCO 50 model. This variant surpasses the original BLIP model in BLEU measurements, hinting at potential improvements in utilizing the BLIP model for image captioning. However, the overall performance of the modified models hovers around 23% to 25%, suggesting that the BLIP model may not be inherently well-suited for image-captioning tasks.

Further Insights

The development and application of our BioMedBLIP models have far-reaching implications across the healthcare and educational sectors. First and foremost, our models can significantly contribute to improving medical diagnosis and decision support systems. By enhancing the capacity to analyze medical images and answer visual questions, they have the potential to facilitate more accurate and timely healthcare interventions, ultimately benefiting patient outcomes. In medical education and training, our models can serve as valuable tools for students and professionals alike. Automated question-answering capabilities can bridge knowledge gaps and improve learning outcomes in a field that demands continuous learning. Moreover, the automation of image analysis and question-answering tasks has the potential to reduce the workload on medical professionals, allowing them to allocate more time to complex aspects of patient care. In terms of research, our models can expedite medical investigations by streamlining the analysis of extensive datasets, potentially leading to groundbreaking discoveries and advancements in the field. Finally, on a global scale, the availability of advanced AI models like ours can improve medical services in underserved regions where access to specialized medical expertise is limited, thereby contributing to more equitable healthcare delivery. However, these opportunities are accompanied by ethical and societal responsibilities. Ensuring patient privacy, addressing biases in the data, and maintaining transparency in the development and deployment of AI models are pivotal steps to maximize their positive impact while mitigating potential risks and pitfalls in the medical field and beyond.

Our BioMedBLIP models, while showing promise in medical image analysis, come with inherent limitations. Firstly, their performance is heavily contingent on the quality and diversity of the training data. Limitations in data availability, such as smaller or less representative medical datasets, can hinder the model's ability to generalize to real-world scenarios and may introduce biases. Secondly, the variability in model performance across different datasets and task categories poses a challenge. Achieving optimal results often demands the fine-tuning of pre-training strategies for specific tasks, which may not always be straightforward in practical applications. While choosing the models appropriate for the real-world applications, there are various considerations to be made. In our work, we have presented the widely used metrics to evaluate model's performance. Different downstream tasks might have different metrics which are used for evaluation. In this case, comparison on grounds of the most relevant metrics should be made. Moreover, the computational demands for pre-training models, especially in the context of medical tasks, can be substantial, potentially limiting the accessibility of our approach to settings with limited computational resources. We recommend training models for different epochs before selecting the model for real-world applications. This is because for different datasets and tasks, the models tend to show convergence at different number of epochs. Finally, ethical and privacy concerns are paramount in the use of medical image data. Striving to ensure strict compliance with data protection regulations and maintaining patient privacy and data security is imperative in any real-world implementation.

Conclusions

In conclusion, our development and evaluation of BioMedBLIP models for medical image analysis tasks reveal both promise and practical considerations. These models have shown substantial potential in enhancing the interpretation of medical images and responding to visual questions in a healthcare context. The choice of pre-training datasets, including ROCO and MIMIC-CXR, plays a pivotal role in model performance, underscoring the importance of specialized medical data for training. Furthermore, the duration of pre-training, exemplified by the MIMIC & ROCO-50 model, has demonstrated the potential to elevate classification accuracy in specific question categories. However, our findings highlight the variability in performance across different datasets and tasks, necessitating a flexible approach to pre-training strategies. Moreover, our models have promising implications across healthcare and education. They can bolster medical diagnosis, decision support systems, and research efforts, while also streamlining medical education and reducing the workload on healthcare professionals. The global accessibility of these models can bring specialized medical expertise to underserved regions.

Acknowledgements

Authors' contributions: U.N. designed and conducted the experiments. U.N., S.T., and A.M. provided technical insights and analyzed the results. U.N., S.T., and A.M. jointly wrote the first draft and revised it. A.M. supervised the project. All authors approved the final version of this manuscript.

Funding: This work was performed at the Norwegian University of Science and Technology with funding from the Research Council of Norway. The funding body did not play any role in the design of the study, in the collection, analysis, and interpretation of data, or in writing the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Abbreviations

BLIP: Bootstrapping Language-Image Pre-training
VQA: Visual Question Answering
SOTA: State-Of-The-Art

UNITER: UNiversal Image-TExt Representation
MedVQA: Medical Visual Question Answering
NLP: Natural Language Processing (NLP)
BERT: Bidirectional Encoder Representations from Transformers
LXMERT: Learning Cross-Modality Encoder Representations from Transformers
CT: Computed Tomography
PET: Positron Emission Tomography
MRI: Magnetic Resonance Imaging
EHR: Electronic Health Record
PEIR: Pathology Education Informational Resource
EM: Exact Match exact match (EM)
ViT: Vision Transformer

References

1. Naseem, U., Thapa, S., Zhang, Q., Hu, L., Rashid, J., Nasim, M.: Incorporating historical information by disentangling hidden representations for mental health surveillance on social media. Social Network Analysis and Mining 14(1), 9 (2023).
2. Naseem, U., Thapa, S., Zhang, Q., Hu, L., Masood, A., Nasim, M.: Reducing knowledge noise for improved semantic analysis in biomedical natural language processing applications. In: Proceedings of the 5th Clinical Natural Language Processing Workshop, pp. 272–277 (2023)
3. Li, F., Thapa, S., Bhat, S., Sarkar, A., Abbott, A.L.: A temporal encoder-decoder approach to extracting blood volume pulse signal morphology from face videos. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 5965–5974 (2023). https://doi.org/10.1109/CVPRW59228.2023.00635
4. Thapa, S., Adhikari, S.: Chatgpt, bard, and large language models for biomedical research: Opportunities and pitfalls. Annals of Biomedical Engineering 51(12), 2647–2651 (2023)
5. Pelka, O., Koitka, S., R ̈uckert, J., Nensa, F., Friedrich, C.M.: Radiology objects in context (roco): a multimodal image dataset. In: Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3, pp. 180–189 (2018). Springer
6. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.-y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data 6(1), 317(2019)
7. Pavlopoulos, J., Kougia, V., Androutsopoulos, I.: A survey on biomedical image captioning. In: Proceedings of the Second Workshop on Shortcomings in Vision and Language, pp. 26–36 (2019)
8. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2577–2586 (2018)
9. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association 23(2), 304–310 (2016)
10. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)

11. Lin, Z., Zhang, D., Tao, Q., Shi, D., Haffari, G., Wu, Q., He, M., Ge, Z.: Medical visual question answering: A survey. Artificial Intelligence in Medicine, 102611 (2023)
12. Li, Y., Wang, H., Luo, Y.: A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In: 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp. 1999–2004 (2020). IEEE
13. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)
14. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423 . https://aclanthology.org/N19-1423
15. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4), 1234–1240 (2020)
16. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.-J., Chang, K.-W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
17. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (2019). Association for Computational Linguistics
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
19. Eslami, S., Melo, G., Meinel, C.: Does clip benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint arXiv:2112.13906 (2021)
20. Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European Conference on Computer Vision, pp. 104–120 (2020). Springer
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PMLR
22. Chen, Q., Hu, X., Wang, Z., Hong, Y.: Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. arXiv preprint arXiv:2305.10799 (2023)
23. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900 (2022). PMLR
24. Ionescu, B., M ̈uller, H., P ́eteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., et al.: Imageclef 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10, pp. 358–386 (2019). Springer
25. Ben Abacha, A., Hasan, S.A., Datla, V.V., Demner-Fushman, D., M ̈uller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: Proceedings

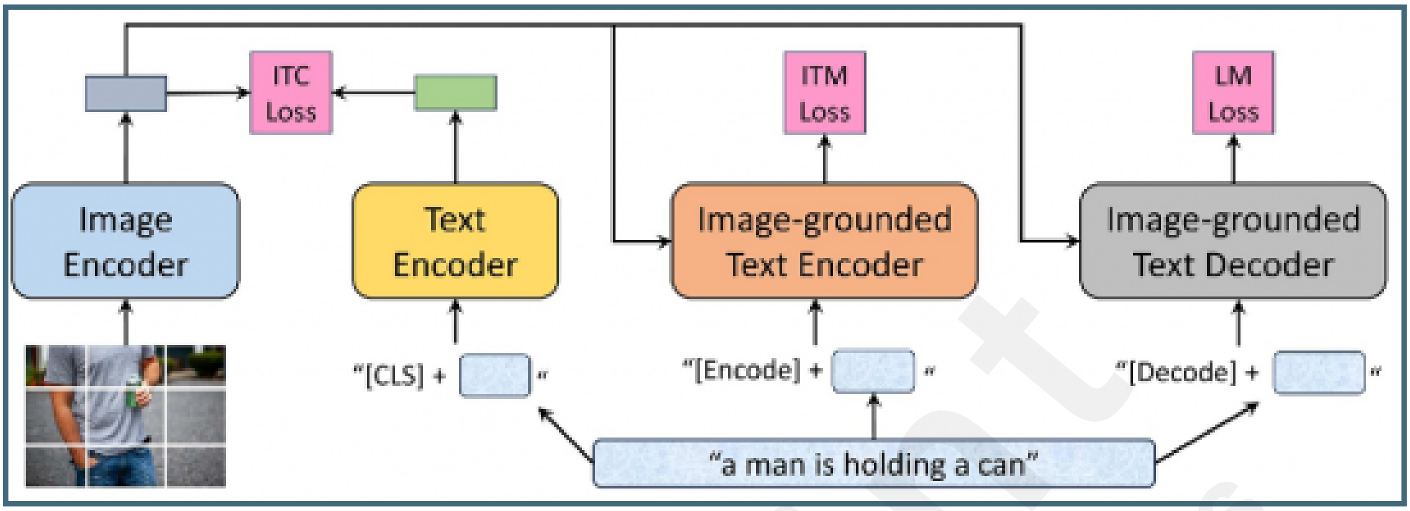
of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes (2019). 9-12 September 2019

26. Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., Wu, X.-M.: Slake: A semantically labeled knowledge-enhanced dataset for medical visual question answering. In:2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), pp. 1650–1654 (2021). IEEE
27. He, X.: Towards visual question answering on pathology images. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, vol. 2 (2021)
28. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data 5(1), 1–10 (2018)
29. Eslami, S., Meinel, C., De Melo, G.: Pubmedclip: How much does clip benefit visual question answering in the medical domain? In: Findings of the Association for Computational Linguistics: EACL 2023, pp. 1151–1163 (2023)
30. BioMedBlip Repository. Online available on HuggingFace Website. Access using the link: https://huggingface.co/biomedblip/biomedblip/tree/main
31. ROCO Dataset. Accessed on November 22nd 2023. GitHub link: https://github.com/razorx89/roco-dataset
32. SLAKE Dataset. Accessed on November 22nd 2023. Website link: https://www.med-vqa.com/slake/\#gt-Download
33. MedVQA Dataset. Accessed on November 22nd 2023. Website link: https://github.com/aioz-ai/MICCAI19-MedVQA
34. OpenAI Dataset. Accessed on November 23rd 2023. Website link: https://openi.nlm.nih.gov/
35. BLIP Dataset. Accessed on November 23rd 2023. Website link: https://github.com/salesforce/BLIP
36. University of Sydney Artemis HPC Platform. Accessed on November 24th 2023. Website Link : https://sydney-informatics-hub.github.io/training.artemis.introhpc/

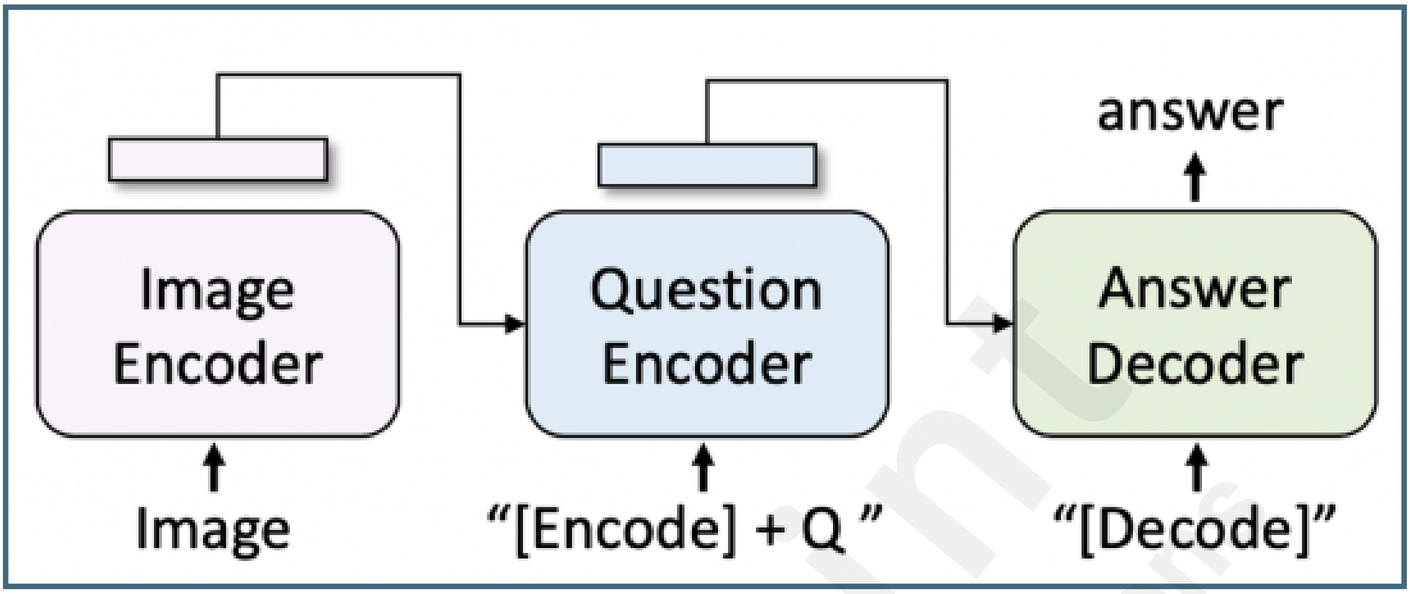
Supplementary Files

Figures

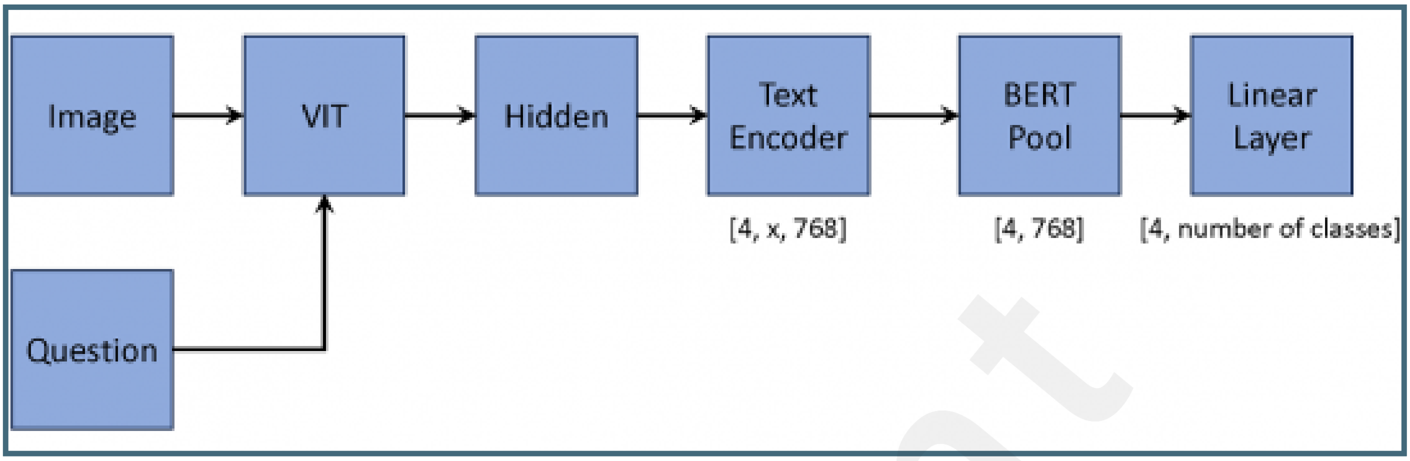
Pretraining architecture of BLIP.



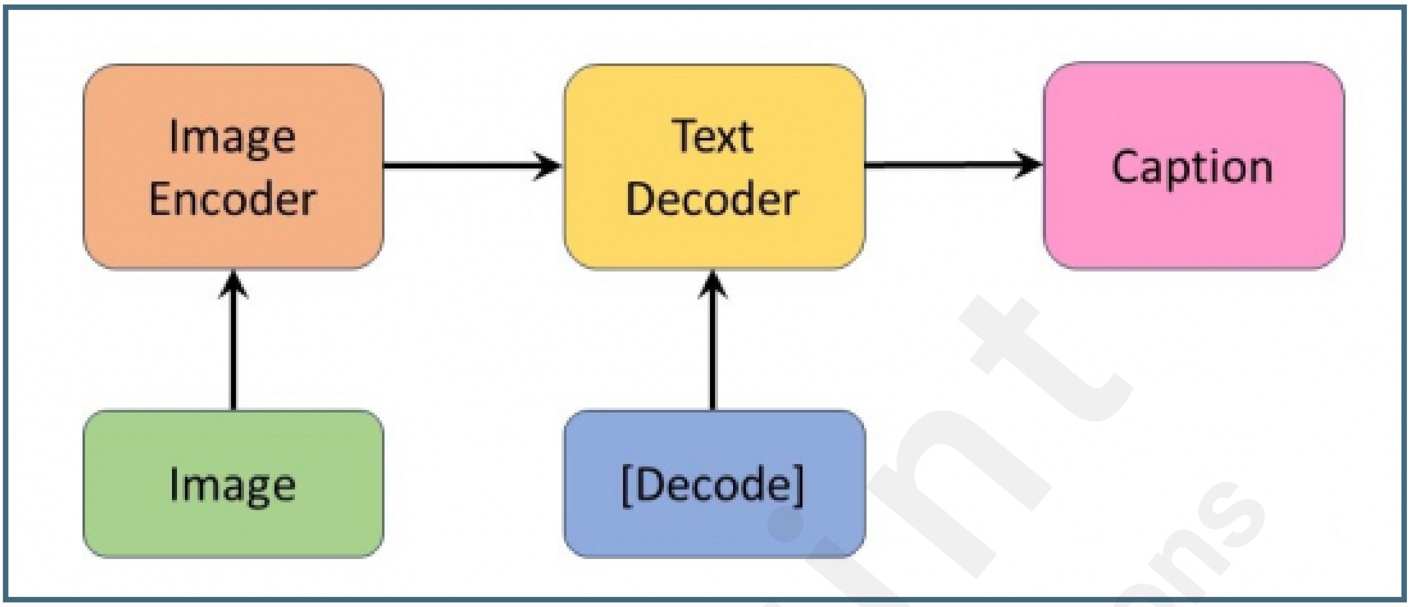
BioMedBLIP VQA Generation model.



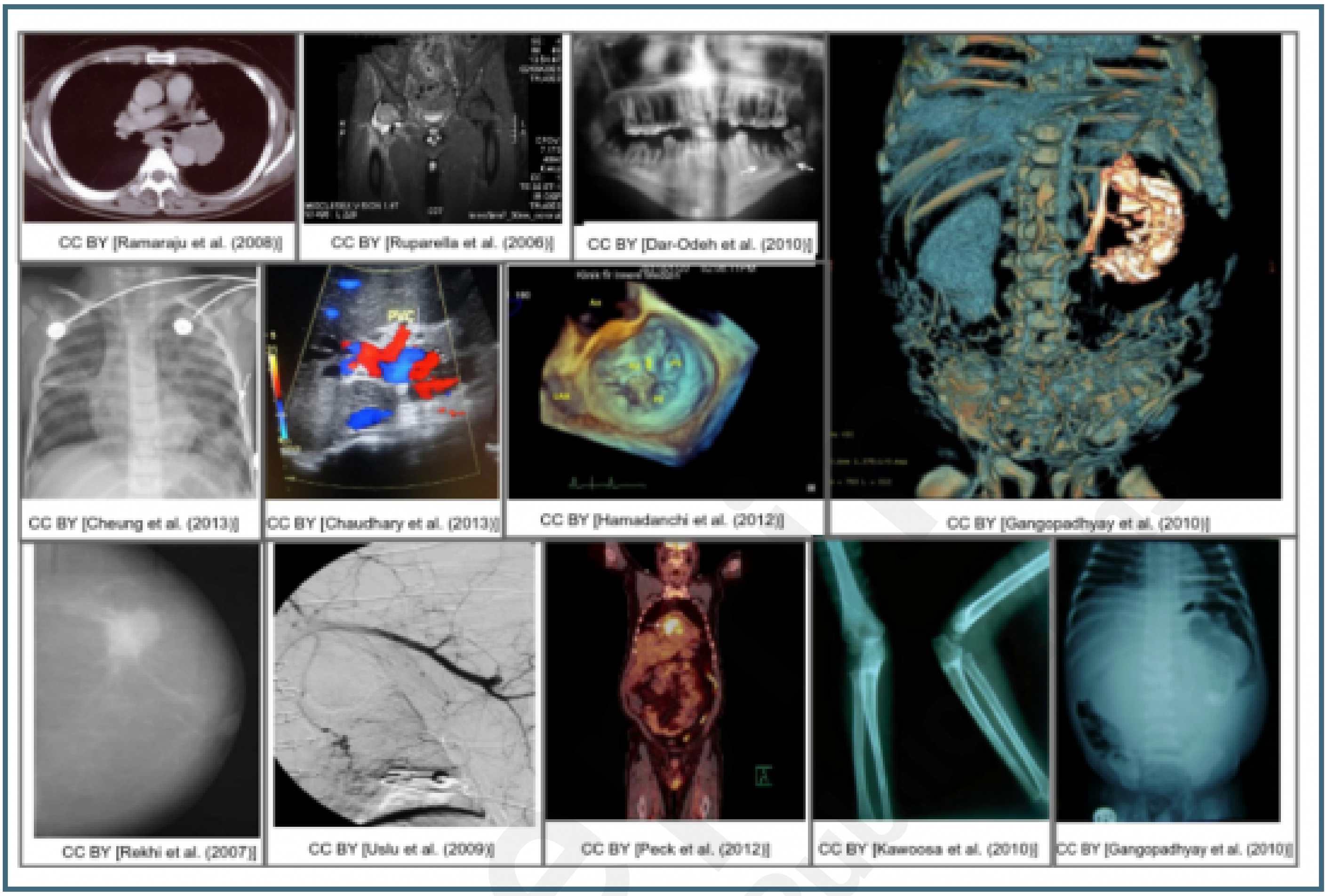
BioMedBLIP Classification model.



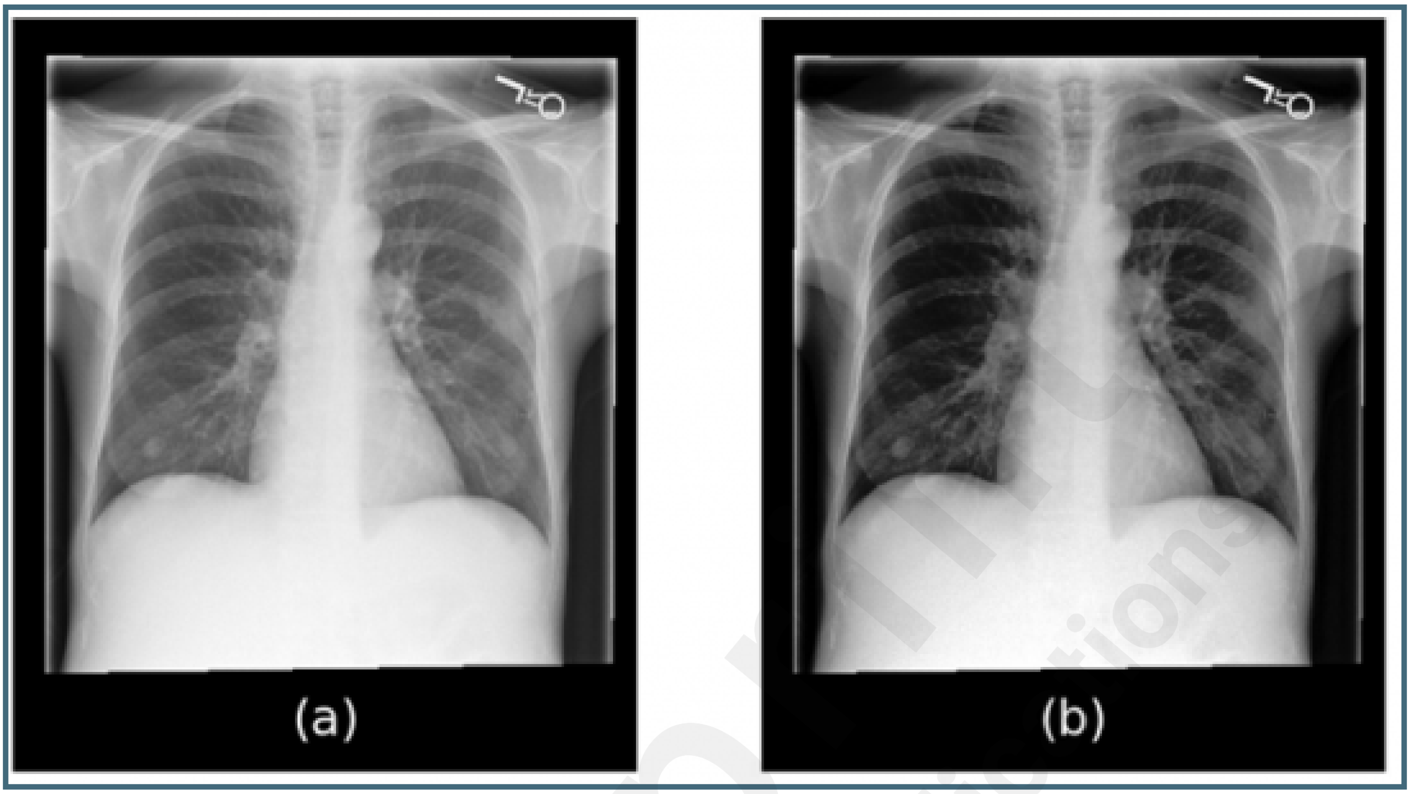
BioMedBLIP Image Caption model.



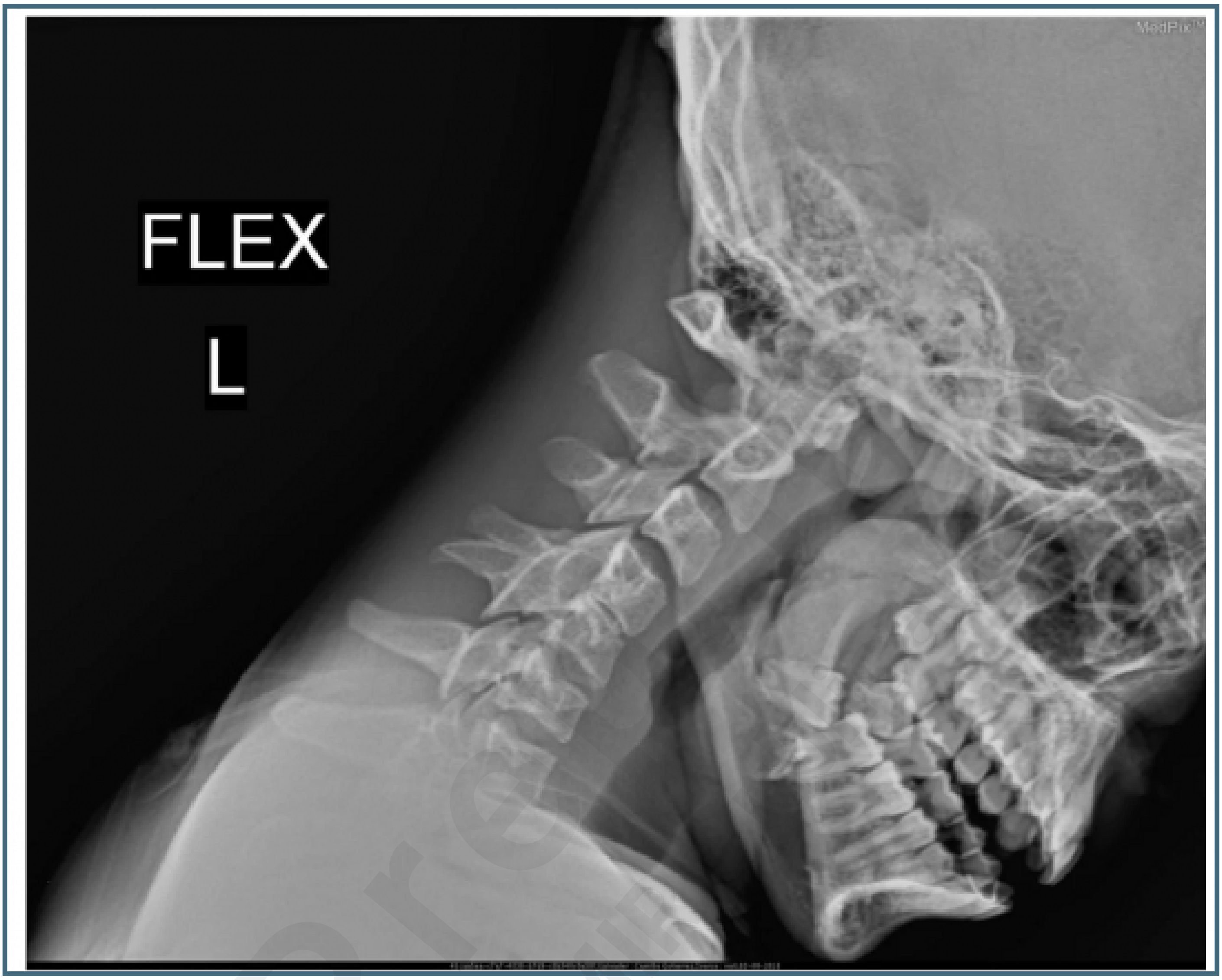
Sample images in the ROCO dataset.


CC BY [Ramaraju et al. (2008)]
CC BY [Ruparelia et al. (2006)]
CC BY [Dar-Odeh et al. (2010)]
CC BY [Cheung et al. (2013)]
CC BY [Chaudhary et al. (2013)]
PVC
CC BY [Hamadanchi et al. (2012)]
CC BY [Gangopadhyay et al. (2010)]
CC BY [Rekhi et al. (2007)]
CC BY [Uslu et al. (2009)]
CC BY [Peck et al. (2012)]
CC BY [Kawoosa et al. (2010)]
CC BY [Gangopadhyay et al. (2010)]

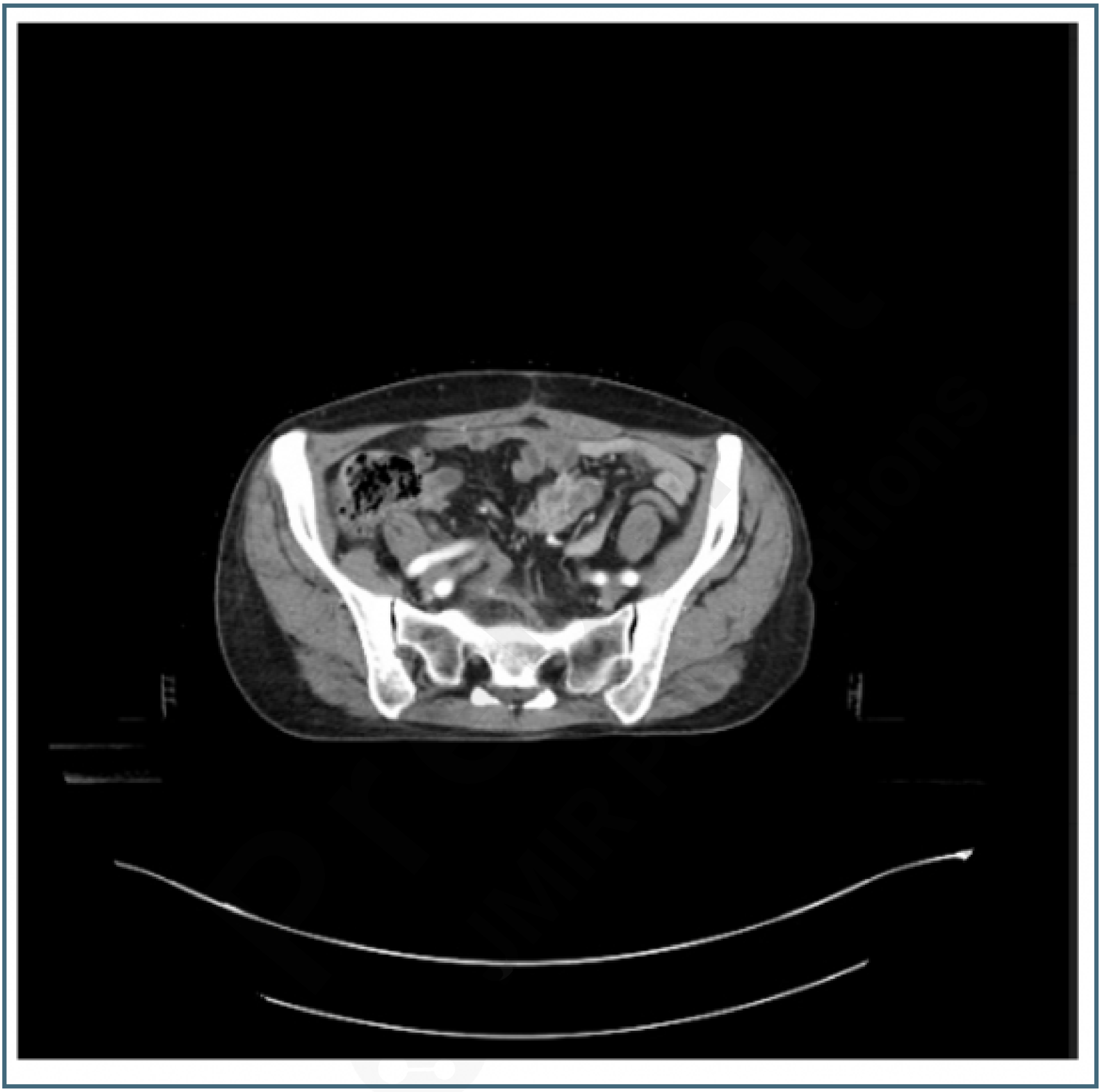
Chest X-rays in MIMIC-CXR dataset.



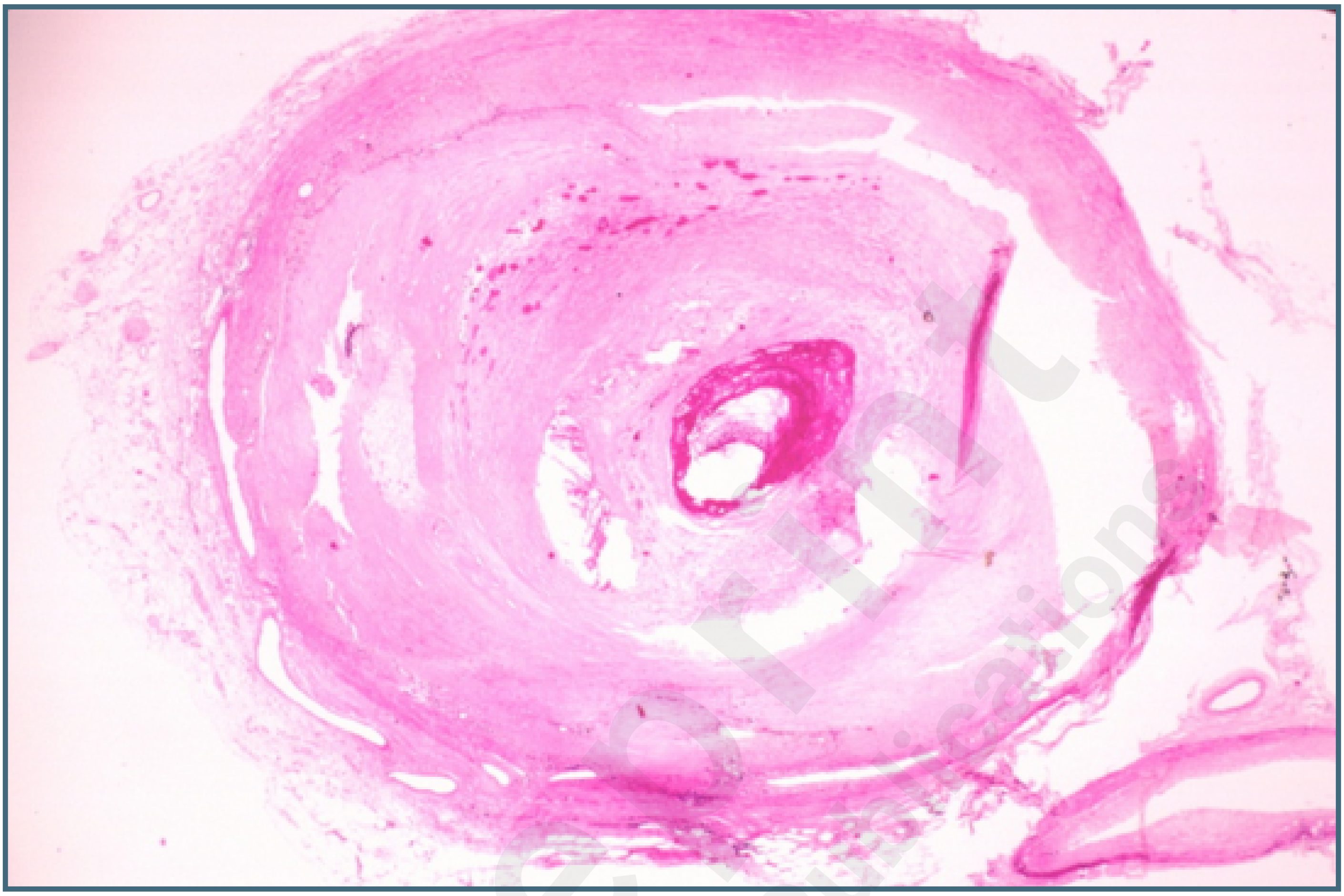
Radiology image in ImageCLEF 2019 dataset.



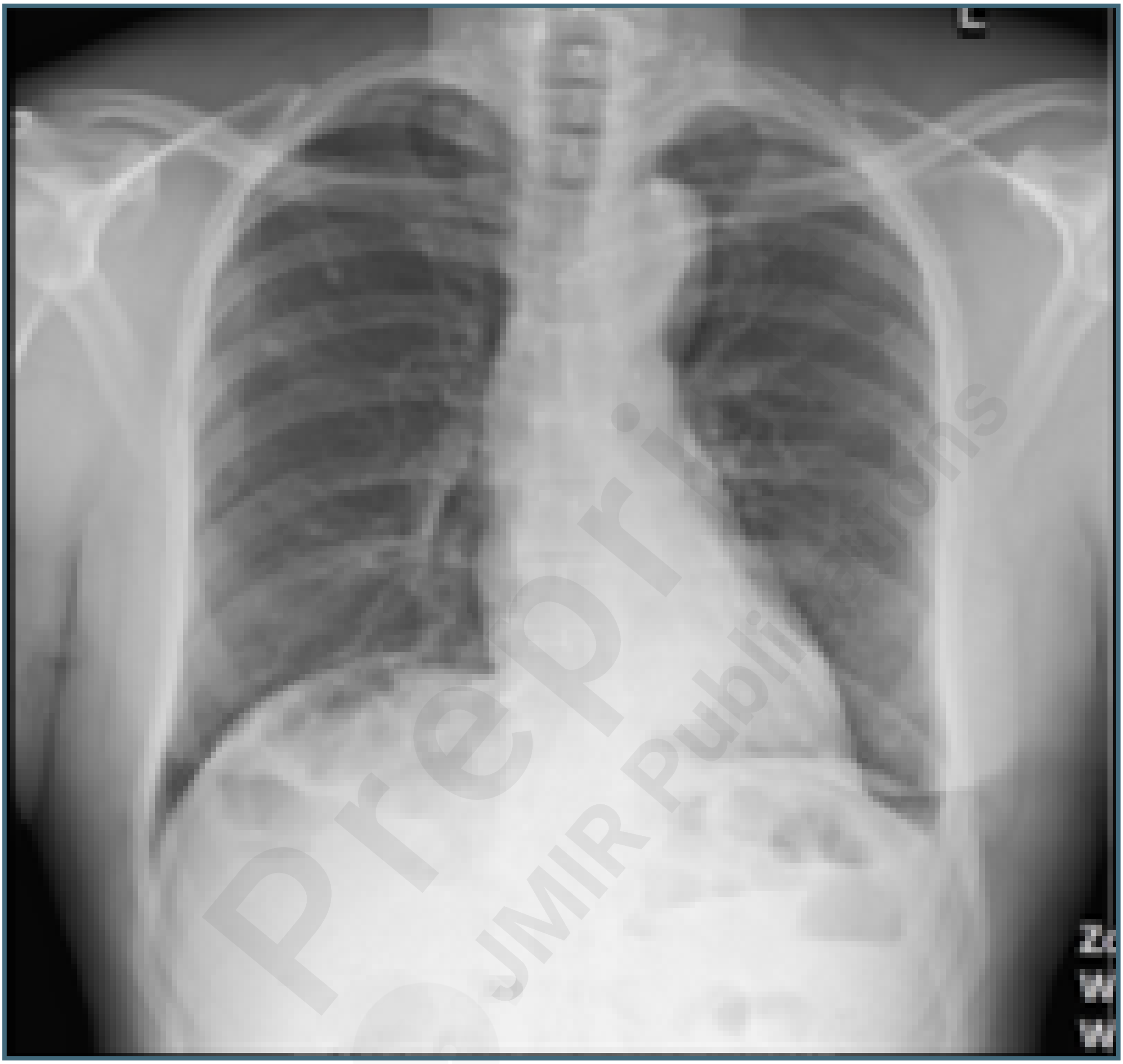
A radiology image from the SLAKE dataset.



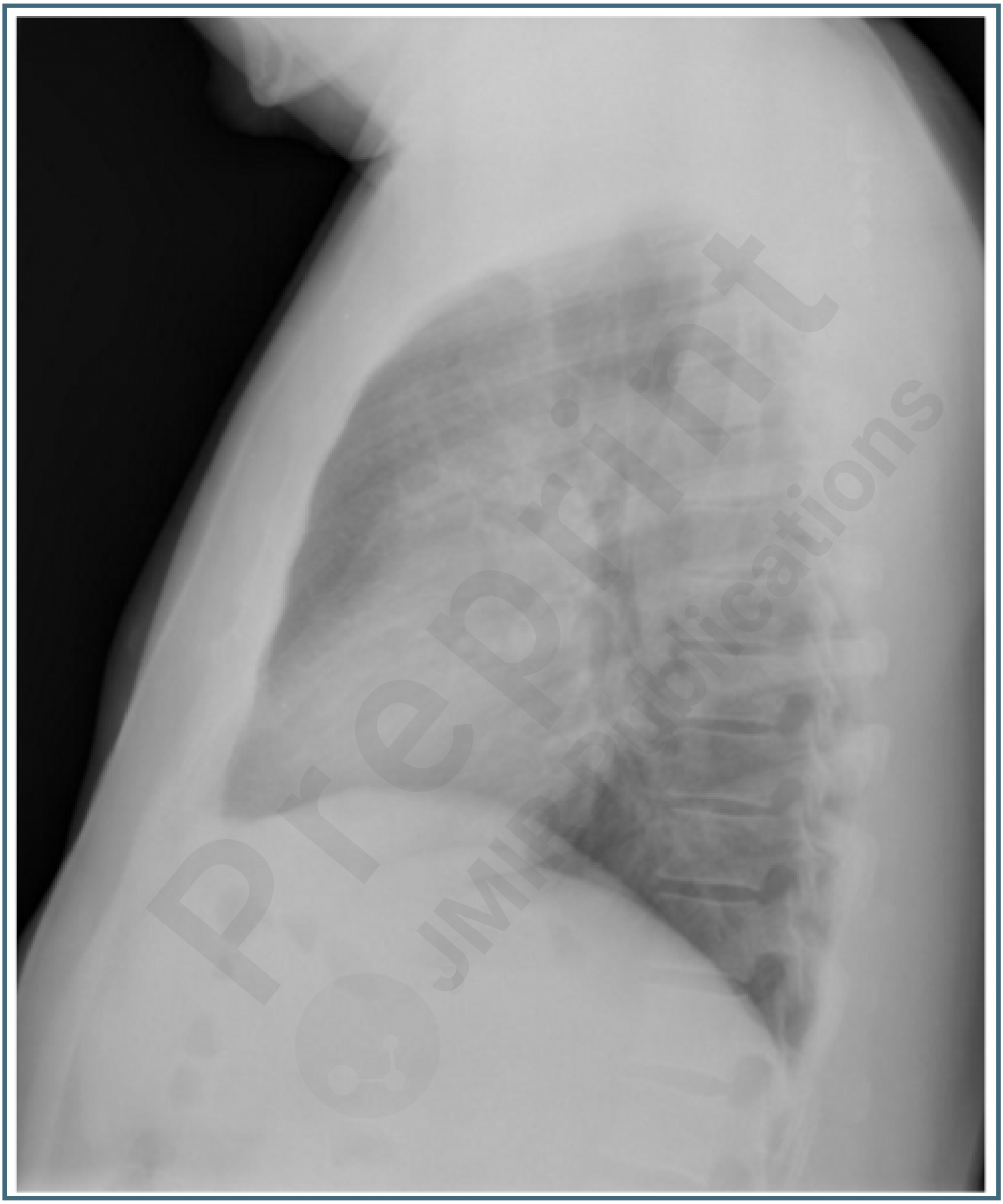
A pathology image from the PathVQA dataset.



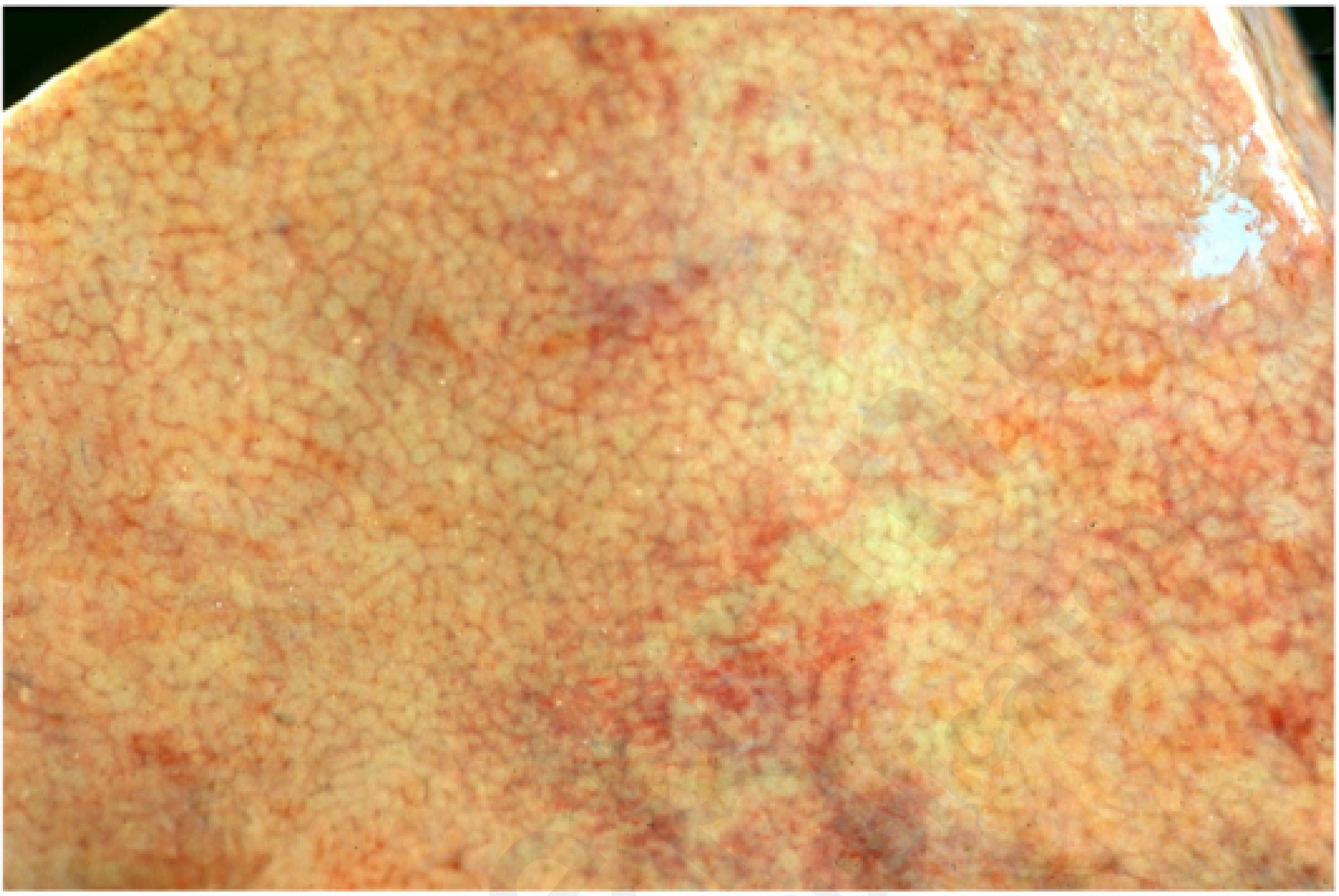
A radiology image from the VQA-RAD dataset.



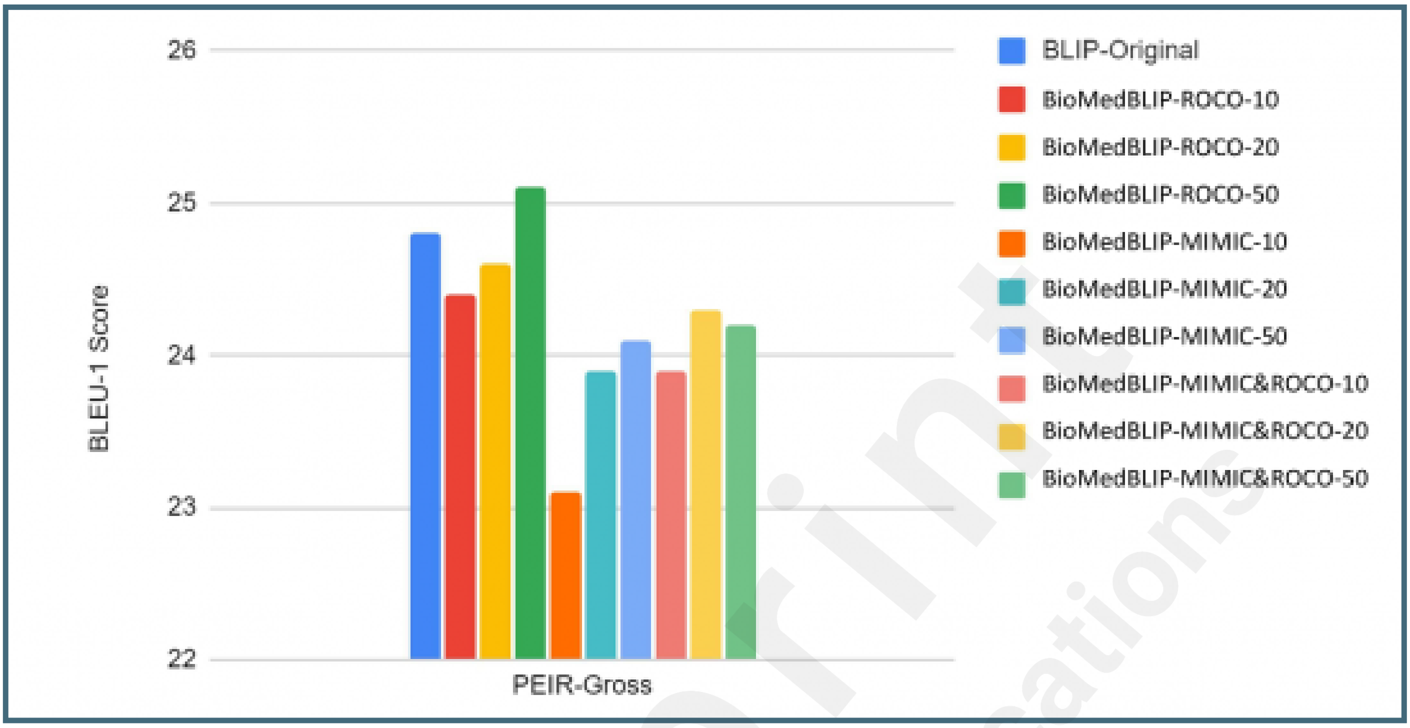
A radiology image from the Open-I dataset.



A medical image from the PEIR-Gross dataset.



Visualization of the BioMedBLIP models vs SOTA (BLIP) on the Image-Caption task for the PEIR-Gross dataset.


BLEU-1 Score
PEIR-Gross
BLIP-Original
BioMedBLIP-ROCO-10
BioMedBLIP-ROCO-20
BioMedBLIP-ROCO-50
BioMedBLIP-MIMIC-10
BioMedBLIP-MIMIC-20
BioMedBLIP-MIMIC-50
BioMedBLIP-MIMIC&ROCO-10
BioMedBLIP-MIMIC&ROCO-20
BioMedBLIP-MIMIC&ROCO-50