TOHOKU UNIVERSITY
Graduate School of Information Sciences

Discovering Improved Feature Spaces for Unsupervised Visual Anomaly Detection for Industrial Products (より良いマルチモーダル表現の学習を通じた医療視覚質問応答の性能向上)

A dissertation submitted for the Master's degree (Information Sciences)
Department of System Information Sciences

by

Mukhammad RAKHIMOV Shukhrat Ugli

February 6, 2025

Discovering Improved Feature Spaces for Unsupervised Visual Anomaly Detection for Industrial Products

Mukhammad RAKHIMOV Shukhrat Ugli

Abstract

Computer vision models have been instrumental in detection of anomalies in industrial manufacturing processes. Automation of the detection of anomalies plays a crucial role in reducing the production cost while maintaining quality control and ensuring the consistency and reliability of production lines. Current models are capable of categorizing the products as anomalous or nominal, moreover the models can outline the defective region with high accuracy.

Contemporary industrial anomaly detection models utilize multiple types of approaches. Three successful methods are: k-nearest neighbor method where statistical outliers are detected on the feature space of a pre-trained network, reconstruction based methods where a decoder is trained to predict the input of an encoder in the form of pre-trained network, student-teacher base methods where a student network is trained to match the output of the teacher pre-trained network.

As was mentioned, the use of pre-trained networks is prevalent in industrial anomaly detection models. The lack of large industrial datasets and the scarcity of samples with anomalies in those datasets make the use of classical learning methods impractical. Therefore, in the recent year, there has been an emergence of industrial anomaly detection models that rely on the features extracted by pre-trained neural networks. Out of all available pre-trained models for use as a backbone in industrial anomaly detection models, models trained on the ImageNet1k classification task are the dominant choice. ImageNet1k is a general-purpose dataset that consists of 1,000 different classes of 1,281,167 training images. While the usage of ImageNet1k-trained backbones is the dominant idea and has been shown to be effective in most cases, this thesis proposes that training on ImageNet1k generates feature spaces that are general purpose, and better performance of industrial anomaly detection models could be achieved with more specialized feature spaces generated by various methods. This investigation aims to explore and validate these alternative feature spaces to enhance model performance in detecting industrial anomalies.

In this thesis we explore methods of forming specialized feature spaces for industrial anomaly models with the purpose of improving their performance. The main proposal of this thesis is to investigate if a dataset can be formed such that, features extracted from such a dataset would show an improvement in the accuracy of anomaly detection models. The proposed method is to extract a subset of images from a large image dataset such as Laion4b, Laion400m, YFCC100M etc. while ensuring the relevancy of the images that are being extracted to the industrial use cases. To achieve that task, we train an "extractor" bi-class(which are industrial and non-industrial) classifier model, and apply the extractor model to the large image dataset. To train the "extractor" model we form another dataset that contains images consisting of industrial and non-industrial classes.

In addition to the generation of an industrial-specialized dataset, this thesis explores the effect of other feature spaces that have not been widely utilized in industrial anomaly detection models. Specifically, during our experiments, it was discovered that Vision Transformer-based models as backbones tend to perform considerably worse compared to CNN-based models. This observation highlights the limitations of Vision Transformer models in capturing the intricate patterns and features necessary for effective anomaly detection in industrial settings. Furthermore, out of all available CNN-based models, ResNet-based models showed superior performance, demonstrating their robustness and effectiveness in this domain. More precisely, WideResNet50 and WideResNet101 models

proved to achieve the highest accuracy when used as backbones for industrial anomaly detection models.

To evaluate the performance of different pre-trained models as backbones, we test them using with different types of industrial anomaly detection models. Throughout all the experiments we use MVTechAD dataset, which contains over 5000 images of industrial objects consisting of 15 different categories. For the experiments of this thesis we use models RealNet and PatchCore to represent a model per each type of industrial anomaly detection models. RealNet for reconstruction based methods and PatchCore for feature-embedding based methods.

To train the "extractor" models we set up a dataset consisting of 2 classes, industrial and non-industrial. Main part of the dataset consist of ImageNet1k classes manually separated into two categories. We train our extractor ResNet50 on the constructed dataset with the classification task. The resulting extractor model then used to extract industrial class images from a large image dataset, in our case YFCC100M. This results in a unlabeled dataset consisting of more that 6 million images. Due to the unlabeled nature of the generated dataset, we use self-supervised learning methods to achieve feature extraction.

Although it is common for Vision Transformer models to be used self-supervised learning, experiments showed that, using ViT models as backbones are ineffective compared to CNN models. Therefore, we train WideResNet50 using DINO self-supervised learning pipeline on the dataset formed from applying the extractor model on the YFCC100M dataset. We perform self-supervised learning with the same conditions on the randomly extracted dataset. Three WideResNet50 models are compared using as backbones to two industrial anomaly detection models, namely PatchCore and RealNet.

In conclusion, this thesis demonstrates that the choice of feature space and backbone architecture significantly impacts the performance of industrial anomaly detection models. By generating a specialized dataset and exploring alternative feature extraction methods, we have shown that more tailored feature spaces can enhance detection accuracy. Our experiments confirmed the superiority of CNN-based models, particularly WideResNet50 and WideResNet101, over Vision Transformer models. The findings highlight the potential of self-supervised learning on industrial-specific datasets to further improve anomaly detection capabilities. Future research should continue to refine these approaches, potentially leading to more robust and efficient industrial anomaly detection systems.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Medical Visual Question Answering (Meidcal VQA) [1, 8, 9, 10] has emerged as a critical task in the healthcare domain, leveraging the power of deep learning and multimodal representations to provide insightful answers to medical-related questions based on visual information, as shown in Figure 1.1. The ability to develop an effective medical VQA system holds immense potential in aiding medical professionals in their decision-making processes and improving patient care outcomes. In recent years, state-of-the-art (SOTA) [11, 12] models have heavily relied on the pretrain and finetune technique [4, 13, 14, 15], which involves pre-training on large-scale image-caption datasets followed by fine-tuning on specific tasks; see Figure 1.2.

We can compare a model to a medical student and follow a similar learning approach. To understand Vision-Language in the medical field, the model can imitate how medical students learn. Similar to how students study from textbooks, the model can learn basic medical knowledge from image and caption data (refer to Figure 1.3). This initial learning phase is called pre-training. Once the model has gained this foundational knowledge, it can move on to practical applications, just like a medical

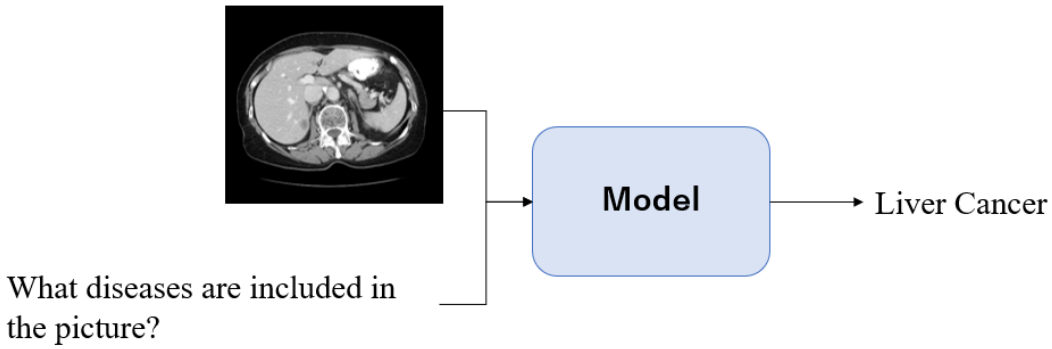


Figure 1.1: Medical Visual Question Answering (VQA) [1] is a task where the model receives medical images and corresponding questions as input and produces accurate answers as the output.

student transitioning to a real-world hospital setting. For example, the model can answer questions from patients based on X-ray images. This practical phase is known as fine-tuning, where the model refines its skills. By using pre-training and fine-tuning techniques, the model acquires both basic medical understanding and the ability to perform specific medical tasks.

However, directly applying vision-language pre-training technique in the medical domain, two significant problems arise. (1) The availability of medical image and caption datasets poses a significant challenge. Unlike other domains [16], medical datasets are often scarce and require specific expertise for collection and annotation, while vision-language pre-training requires a large-scale dataset. (2) The models lack fundamental knowledge about the medical organs and structures. Understanding the unique characteristics and context of medical imagery, such as organs, is crucial for accurate analysis and interpretation. It also helps facilitate the learning process in

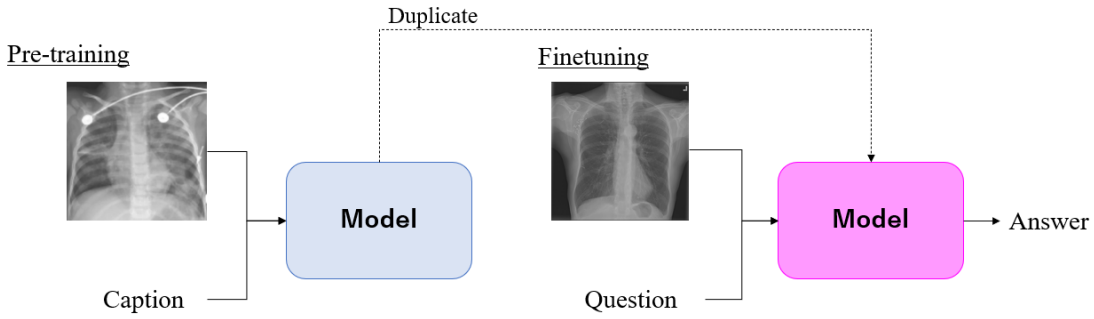


Figure 1.2: Pre-training followed by finetuning: The approach involves pre-training the model on a large-scale medical image-caption dataset [2], and subsequently fine-tuning it specifically for the task of medical visual question answering [1].

the vision-language pre-training.

Addressing the first problem, which involves the limited availability of image-caption datasets in the medical domain, poses a significant challenge. Generating a comprehensive image-caption dataset from scratch is a complex and labor-intensive effort, involving collaboration with medical professionals and accurate annotation processes [17]. As a result, our focus shifts towards tackling the second problem by exploring alternative approaches to equip models with essential organ-related knowledge.

This thesis aims to address the second problem by proposing a novel two-stage pre-training technique that focuses on equipping the model with essential organ knowledge prior the original vision-language pre-training. The first stage of pre-training involves utilizing medical image segmentation [18] as a task to equip the model with a foundational understanding of organ structures. By training the model on this task, it becomes capable at identifying and delineating different organs in medical images,

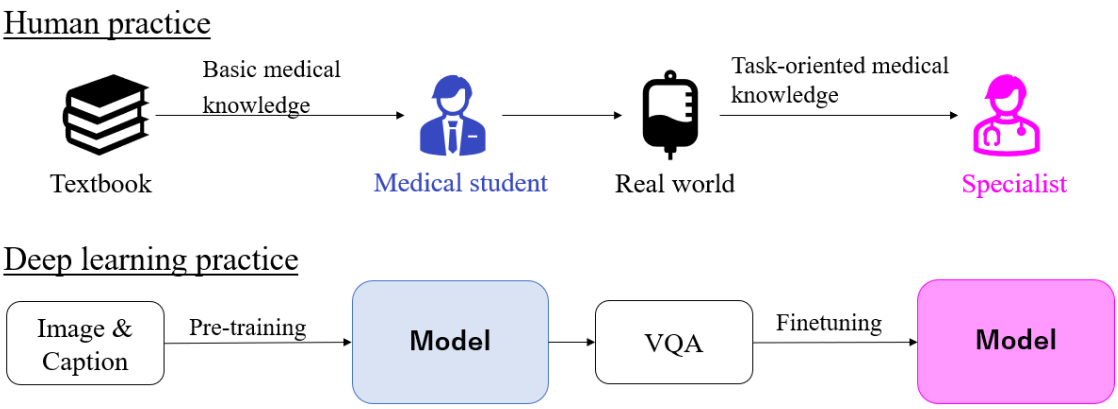


Figure 1.3: Comparing model learning to medical student training

acquiring the necessary organ-specific knowledge. The second stage of pre-training involves the original vision-language pre-training [19], where the model learns general knowledge in the medical domain by jointly processing images and their corresponding captions. This two-stage approach allows the model to build upon its foundational organ knowledge then develop a comprehensive understanding of the vision-language relationship in the medical context.

The conducted experiments that aimed to evaluate the proposed two-stage pre-training method have demonstrated its superiority over the original pre-training approach. We employed Medical Visual Question Answering [1, 10] as the evaluation task to assess the final performance. The results also clearly indicated that the order of the pre-training stages played a crucial role in achieving optimal performance. Specifically, the pre-training stage one, which focuses on equipping the model with basic organ knowledge through image segmentation, was found to be essential and should precede the second stage of pre-training. Furthermore, the proposed method shows superior performance compared to several previous works [20, 4], indicating its effectiveness and potential in the field of medical vision and language understanding.

This chapter serves as the introduction to the thesis, offering a comprehensive overview of the research topic and highlighting its significance within the field. Chapter 2 will focus on the Related Work, delving into the concept of visual question answering, Transformer attention models, and examining recent advancements in vision-language pre-training techniques that have demonstrated effectiveness in addressing various medical vision-language tasks. Chapter 3 will address the limitations of previous work and introduce the proposed method as a solution. It will highlight the key improvements and contributions of the proposed two-stage pre-training method. In Chapter 4, we will conduct a series of comprehensive experiments to evaluate the proposed method. This chapter will provide detailed information about the experimental setup and methodology used. Additionally, it will include an in-depth analysis of the obtained results, demonstrating the performance of the proposed method and its effectiveness in addressing the medical visual question answering. In Chapter 5, we will summarize the key findings obtained from our research and provide a comprehensive conclusion. We will also discuss potential areas for further improvement and future research directions in the field, aiming to contribute to the advancement of knowledge and advancements in the topic.

Chapter 2

Related Work

In Chapter 2, an in-depth exploration of the relevant literature. We begin by exploring the concept of Transformer Attention, a fundamental mechanism extensively utilized in various vision and language models (Sec. 2.1). We also explore its variants, such as the Vision Transformer (Sec. 2.1.1) and the Language Transformer (Sec. 2.1.2), which have revolutionized the field of deep learning by effectively capturing complex relationships within visual and textual data. Furthermore, we provide an in-depth analysis of the recent advancements in vision-language pre-training techniques (Sec. 2.2). These methodologies involve extensive pre-training on large-scale datasets incorporating both visual and textual data, aiming to acquire highly robust multimodal representations. Next, a comprehensive exploration is conducted on the concept of Visual Question Answering (Sec. 2.3), which involves answering questions about an image using both visual and textual information. Finally, we then narrow our focus to Medical Visual Question Answering (Sec. 2.4), a specialized domain where VQA techniques are applied to medical images and associated medical knowledge. By conducting a thorough analysis of the previous literature, this chapter offers a com-

prehensive insight into the challenges and recent advancements within the domain of vision-language pre-training in the medical domain, thus laying the foundation for the formulation of our proposed approach.

2.1 Transformer Attention

The transformer [3] is a revolutionary deep learning architecture that has had a profound impact on various domains, including natural language processing and computer vision. Unlike traditional sequential models, such as recurrent neural networks (RNNs) [21], transformers process entire input sequences in parallel, making them highly efficient for capturing long-range dependencies. As shown in Figure 2.1, the transformer model consists of an encoder and a decoder, with each component containing multiple layers of self-attention and feed-forward neural networks. A fundamental element of the transformer is its attention mechanism, which enables the model to assign varying degrees of importance to different components within the input sequence, facilitating more accurate predictions.

The attention mechanism operates from the perspective of queries Q, keys K, and values V; see formula 2.1. In this mechanism, queries represent the current input element that needs to attend to other elements in the sequence. Keys and values correspond to the other elements in the sequence. The attention mechanism computes the similarity between the query and each key using a dot product or a learned similarity function. These similarities determine the relevance or importance of each key with respect to the query. Then, the attention scores are normalized to obtain attention weights, which indicate the relative importance of each value. The final step involves calculating a weighted sum of the values, where the weights

are determined by the attention weights; see Figure 2.2 (a). This process allows the model to give more attention to relevant parts of the input sequence while suppressing irrelevant or noisy information.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

Multi-head attention is an essential component of the transformer architecture that further enhances its modeling capabilities. As shown in Figure 2.2 (b), the attention mechanism is applied multiple times in parallel, with each attention head learning different aspects of the input sequence. Each attention head operates on its own learned linear projections of the input, allowing for the simultaneous extraction of different features and representations. After computing attention independently for each head, the outputs are concatenated and linearly transformed to generate the final attention output. Using multiple attention heads allows the model to capture different patterns and relationships in the data at different levels of detail

2.1.1 Vision Transformer

The Vision Transformer (ViT) [22] is a popular architecture for visual tasks that adapts the transformer model from natural language processing. It divides the input image into fixed-size patches and maps them to token embeddings, which are then processed by the transformer encoder to capture global relationships among the patches. This allows the ViT to learn representations for various vision tasks. The model efficiently converts the input image with dimensions (C, H, W) into an image representation with dimensions $(C, H/16, W/16)$, where C is the number of channels, H is the height of the image, and W is the width of the image.

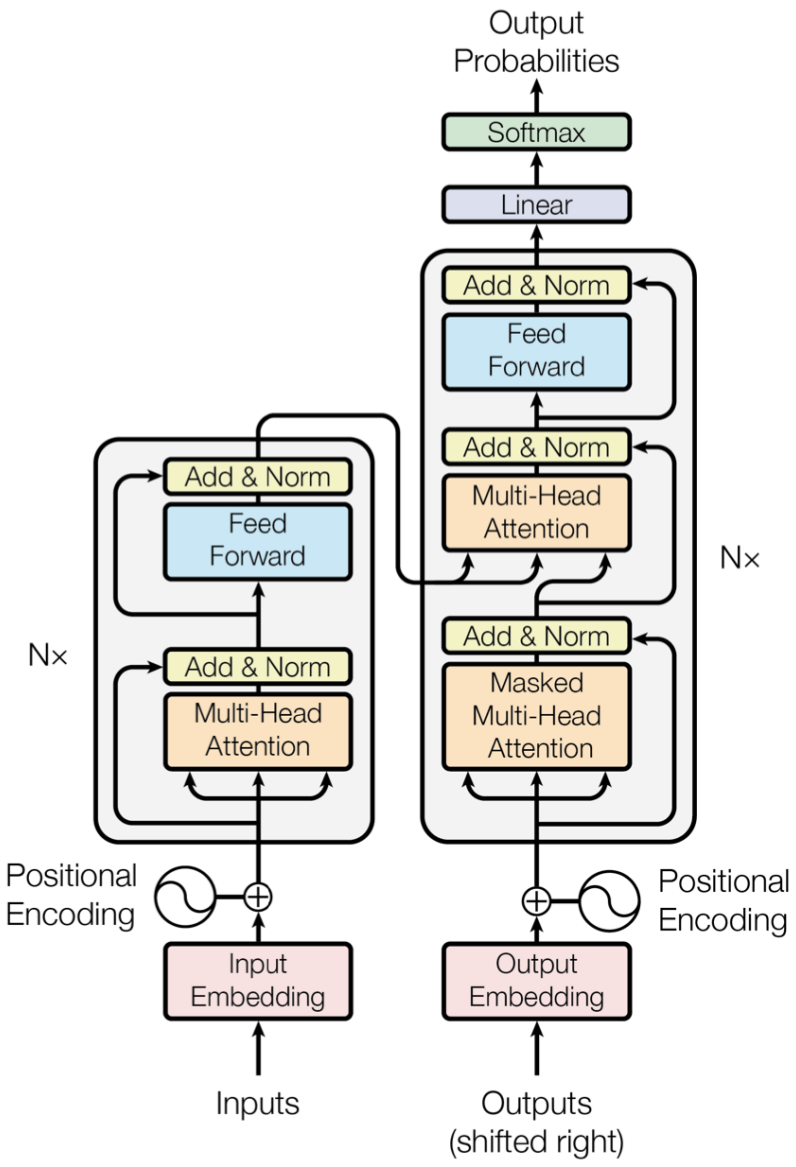


Figure 2.1: Transformer architecture [3]

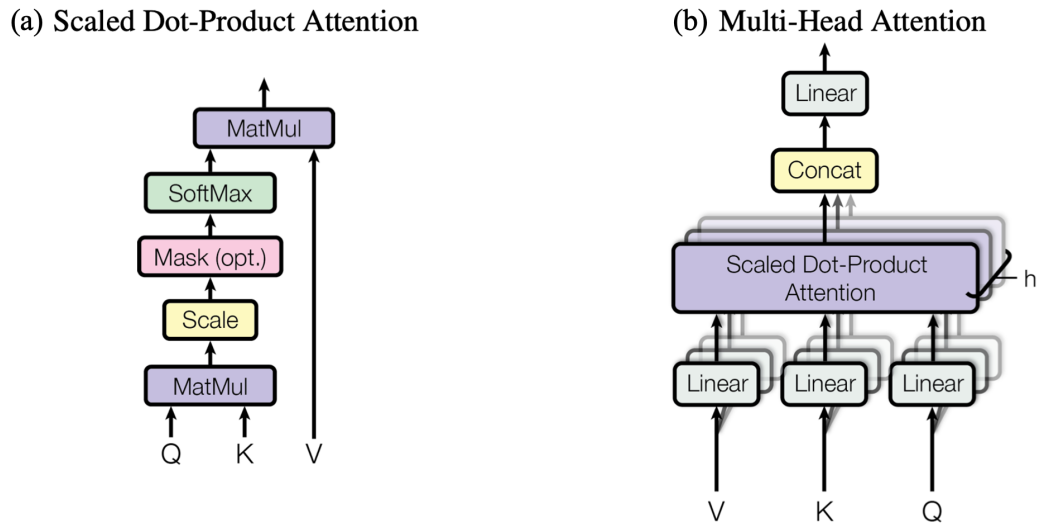


Figure 2.2: Attention mechanisms [3]: (a) Scaled dot-product attention, (b) Multi-head attention

Swin Transformer [23] is one of the variants that introduces hierarchical representations by stacking multiple stages of transformers. It divides the image into non-overlapping patches at the initial stage and applies transformers at different scales, capturing both local and global information. Swin Transformer leverages shifted windows and window partitioning to efficiently model long-range dependencies while reducing computational complexity. This variant has shown improved performance and scalability, making it an appealing choice for visual recognition tasks, especially when dealing with high-resolution images or large-scale datasets.

2.1.2 Language Transformer

The Language Transformer is a powerful architecture for natural language processing tasks. It is based on the transformer attention model. The language transformer can be classified into two architectural categories: encoder-only and encoder-decoder.

An example of the encoder-only architecture is the BERT model [24]. BERT employs a masked language modeling objective, where it randomly masks certain tokens in the input and predicts them based on the surrounding context. This allows BERT to learn deep bidirectional representations that capture both the left and right context of each token.

In contrast, The encoder-decoder architecture shares the same foundational structure as the architecture described in [3]. This setup is commonly used in generative models, such as the GPT model [25]. The encoder-decoder architecture is widely used for tasks like machine translation and text generation. The encoder processes the input sequence, while the decoder generates the output sequence autoregressively, attending to the encoder's representations to produce contextually informed predictions at each step.

2.2 Vision and Language Pre-Training

Vision and language pre-training techniques [4, 13, 15, 14, 26, 27] involve self-supervised learning using a combination of image and caption data. This approach has shown promising results in various vision-language downstream tasks, such as visual question answering. As illustrated in Figure 1.2, the Vision and Language pre-training consists of two phases: the pre-training phase and the fine-tuning phase. During the pre-training process, a model is trained on a large corpus of image-caption pairs without relying on explicit annotations. Subsequently, the model is fine-tuned on various downstream tasks, such as visual question answering. This allows the model to learn meaningful representations that capture the relationship between visual and textual information.

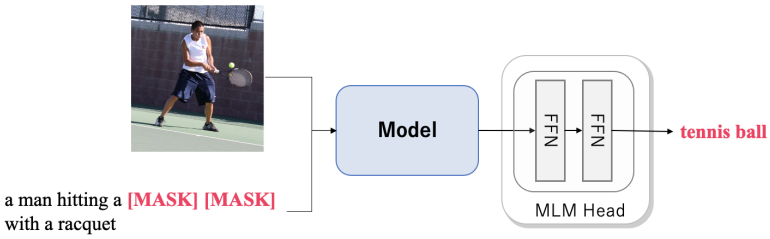


Figure 2.3: Masked language modeling [4]

In vision and language pre-training, two common objectives are employed: masked language modeling (MLM) and image-text matching (ITM).

Masked language modeling is a form of self-supervised learning objective where a portion of the input text is randomly masked, and the model is tasked with predicting the masked words based on the context provided by the surrounding words and the corresponding image; see Figure 2.3. Typically, two fully connected layers (referred to as the Masked Language Modeling Head) are appended to the model. This component allows the model to predict masked tokens within the input text, improving its ability to learn contextual representations and capture vision-language patterns.

Image-text matching aims to align the image and text representations by training the model to determine whether a given image and its corresponding caption are a correct match; see Figure 2.4. The model is extended with an image-text matching head, which incorporates two fully connected layers as depicted in Figure 2.5. By optimizing this objective, the model develops the ability to relate visual and textual information effectively.

Through self-supervised learning with MLM and ITM, the vision and language pre-training technique enables the model to acquire a rich understanding between visual and textual modalities. This knowledge can then be leveraged in downstream

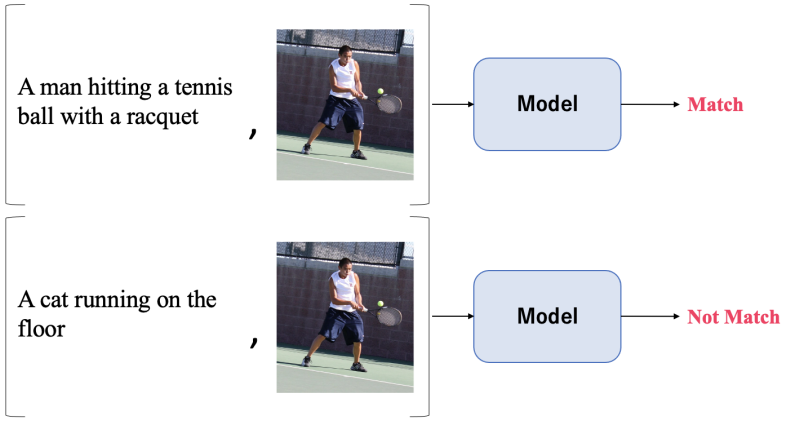


Figure 2.4: Image-text matching [4] involves determining whether an image and its corresponding caption are a match or not. The top image represents a positive pair, indicating a matching image and caption. On the other hand, the bottom image represents a negative pair, indicating a non-matching image and caption.

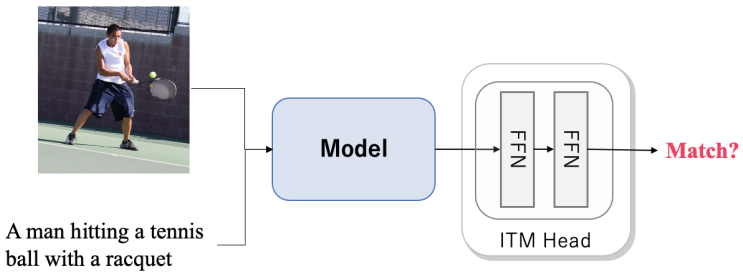


Figure 2.5: The image-text matching head [4] consists of two fully connected layers in its architecture.

tasks, where the model can generalize well and provide accurate predictions and answers based on both visual and textual inputs.

2.3 Visual Question Answering

Visual Question Answering (VQA) [28, 29, 30] is a task that aims to answer questions based on the content of an image. As shown in Figure 2.6, it involves providing a model with both an image and a corresponding question as input and expecting the model to generate the appropriate answer. VQA combines computer vision and natural language processing techniques to bridge the gap between visual information and textual understanding.

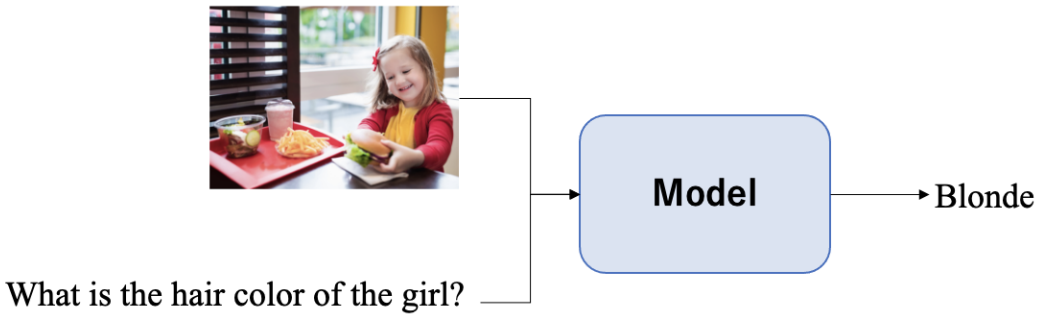


Figure 2.6: Visual question answering [5] is a task in which a model takes images and corresponding questions as input and generates accurate answers as output.

2.4 Medical Visual Question Answering

Medical Visual Question Answering (Medical VQA) [1, 8, 9, 10] is a specialized task that applies the techniques of Visual Question Answering to the medical domain. It involves answering questions related to medical images or healthcare scenarios using both visual and textual information. There are two types of previous work in this field: models that utilize pre-training [11, 31, 32, 12, 33] and models that do not rely on pre-training [1, 34, 35, 36, 37]. Generally, the approaches that employ pre-training models tend to exhibit state-of-the-art performance in this field. However, this thesis aims to illustrate that the direct application of vision-language pre-training techniques in the medical domain may not yield optimal outcomes. Instead, we propose novel techniques to enhance the performance applicability of the task.

Chapter 3

Two-Stage Pre-training Method for Medical Vision and Language Task

In this chapter, we will provide a detailed explanation of the method proposed in our research. Initially, we will discuss two problems that arise when directly applying vision-language pre-training in the medical domain (Sec. 3.1). We will then introduce our proposed method (Sec. 3.2), which aims to address these issues. The subsequent section will explore into the network architecture, layers, and components utilized in our approach (Sec. 3.3). Furthermore, we will focus on the training pipeline (Sec. 3.4) that consists of fist stage pre-training and second stage pre-training. The first stage pre-training (Sec. 3.4.1)is where the model acquires essential organ-specific knowledge. We will elaborate on the methodology employed for image segmentation and its contribution to a comprehensive understanding of organ structures. In the second stage of pre-training (Sec. 3.4.2), we will explain the original vision-language pre-training procedure that employs medical images and captions. Additionally, we will demonstrate the datasets utilized in our experiments (Sec. 3.5) and explore the

process of generating a dataset for the first-stage pre-training (Sec. 3.6).

3.1 Problems in Medical Vision-Language Pre-training

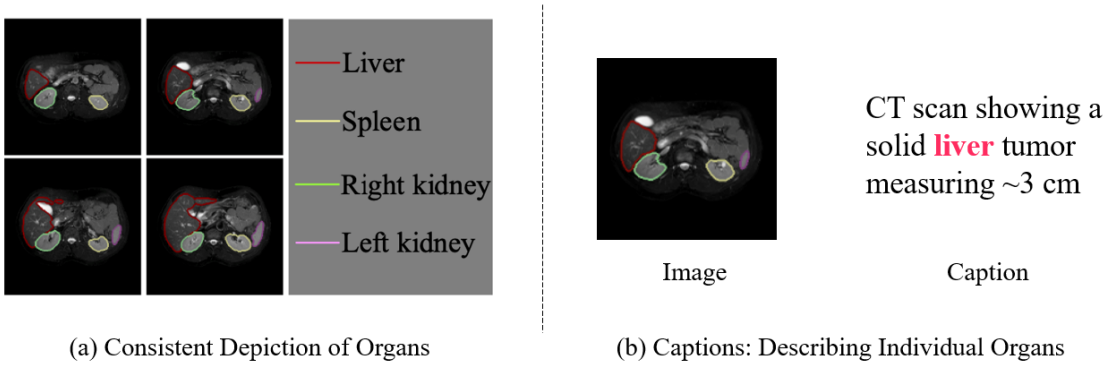
The latest advanced models in medical visual question answering use a two-step approach called pre-training and fine-tuning [11, 31, 32, 38, 12, 39]. However, when applying the vision-language pre-training technique in the medical domain, we encounter two significant challenges. The first challenge arises from the scarcity of data available for pre-training (Sec. 3.1.1). The second challenge relates to the model's lack of fundamental knowledge about medical organs (Sec. 3.1.2).

3.1.1 Shortage of Data

The first problem is the shortage of data during pre-training. Vision-language pre-training technique relies on a large-scale datasets of image-text pairs for successful learning. However, in the medical domain, the availability of such datasets is limited. Typically, in the medical domain, we have around 300K image-caption pairs [17], but pre-training techniques require approximately 4 million image-caption pairs for effective learning [14]. This scarcity of data in the medical domain can be attributed to privacy concerns and the complexity of annotating medical images, which require expert physicians to provide accurate captions. Acknowledging the difficulty and resource-intensive nature of collecting additional image-caption dataset pairs, we have chosen to set this aside and proceed with addressing the next problem. However, we believe our proposed method can work well in the low data regime.

3.1.2 Model's Lack of Organ Knowledge

The second problem is, during pre-training, the model lacks fundamental knowledge about organs, such as their size, shape, and location. The rationale behind this problem is that medical images consistently depict the same group of organs: see Figure 3.1(a), and captions primarily focus on describing individual organs; see Figure 3.1(b). Consequently, during the pre-training phase, the model faces difficulty in accurately identifying the relevant image region corresponding to the caption. For instance, in Figure 3.2, with a caption that focuses solely on one organ (liver) in an image containing multiple organs, the model struggle to identify the specific regions (such as the green, red, yellow, or purple areas) that represent the liver in the visual representation. Additionally, in Figure 3.3, the model faces challenges in tasks such as filling in masked words during masked language modeling pre-training due to its limited understanding of the visual appearance of healthy organs. Even when provided with choices such as liver, spleen, or kidneys, the model still struggles to accurately identify the organ in masked words. It is crucial to provide the model with fundamental organ knowledge prior to the original pre-training stage.



(a) Consistent Depiction of Organs

(b) Captions: Describing Individual Organs

Figure 3.1: Two reasons behind the model's lack of organ knowledge. (a) clearly illustrates four abdominal images, each showcasing the consistent presence of organs such as the liver, spleen, and kidneys [6]. In (b), the abdomen is illustrated, and the corresponding caption specifically describes the liver while disregarding the other organs present.

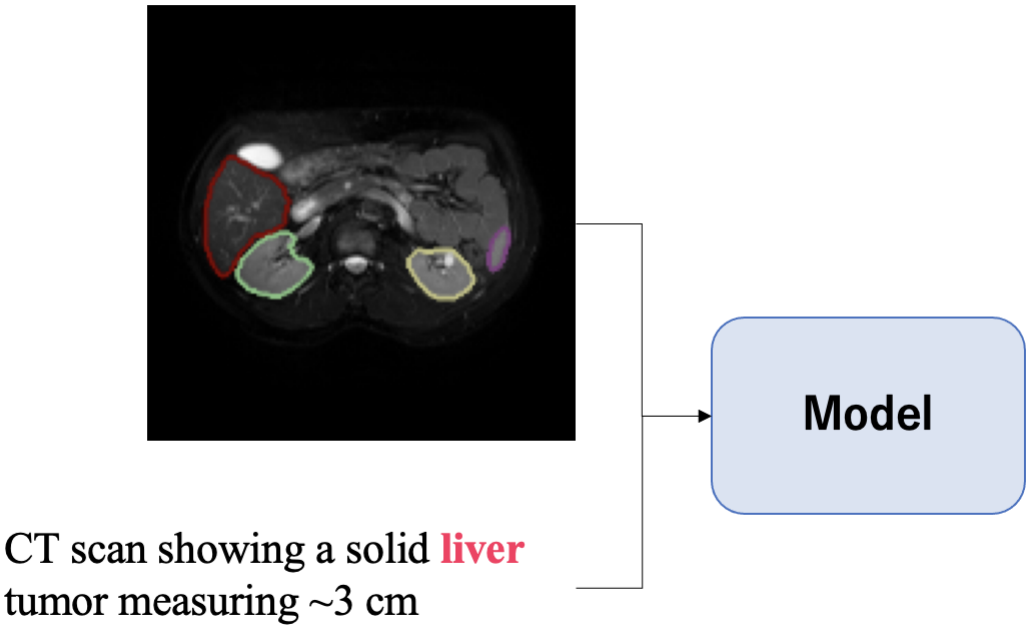


Figure 3.2: Challenge in the application of vision-language pre-training in the medical domain [6]: The model struggles to identify the specific regions (such as the green, red, yellow, or purple areas) that represent the liver in the visual representation.

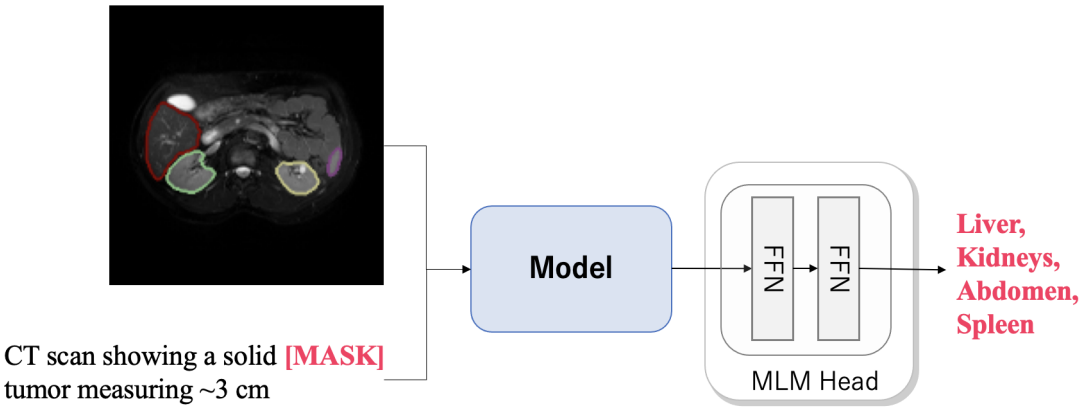


Figure 3.3: Challenge in the application of vision-language pre-training in the medical domain [6], especially in masked language modeling: The model encounters challenges in tasks such as filling masked words during pre-training due to its limited knowledge of the visual appearance of healthy organs. Despite being provided with choices such as liver, spleen, or kidneys, accurately identifying the organ in masked words remains a difficult task.

3.2 Proposed Method

In order to address the aforementioned challenge, this thesis proposes a two-stage pre-training approach aimed at equipping the model with fundamental organ knowledge prior to the original pre-training stage. More specifically, we undertake the pre-training process in two separate stages, in contrast to previous approaches [12, 11] that utilized only one stage. Figure 3.4 provides a visualization of the pre-training steps involved in our proposed method.

The first stage involves the use of an medical image segmentation [18] task to acquire important and fundamental organ knowledge, while utilizing captions as input to facilitate a comprehensive understanding of organ structures within medical images; see Figure 3.4(a). By incorporating this stage, the model gains a foundational understanding of medical organ concepts that is often lacking in previous work. In the second stage, the original vision-language pre-training process is performed using the medical images and captions: see Figure 3.4(b). The objective of the two-stage pre-training methodology is to enhance the model's understanding of medical concepts and improve the quality of multimodal representations.

Medical image segmentation is a crucial task in the field of medical imaging analysis. It involves identifying and outlining different structures or regions within medical images, like organs, tumors, or lesions; see Figure 3.5. Accurate segmentation is vital for various clinical applications, such as diagnosis, treatment planning, and disease monitoring. By performing medical image segmentation, the model gains an understanding of various aspects of organs, such as their shapes and positions within the image. This process enables the model to acquire essential knowledge about organs and their characteristics.

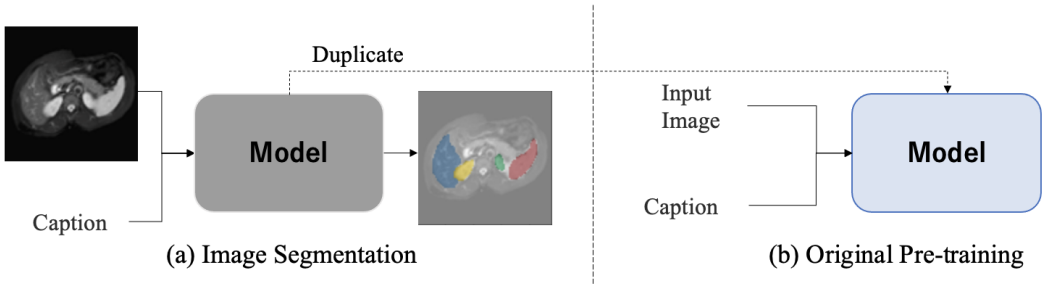


Figure 3.4: The pre-training process comprises two stages: (a) stage one focuses on a segmentation task [7], while (b) stage two adheres to the original pre-training methodology.

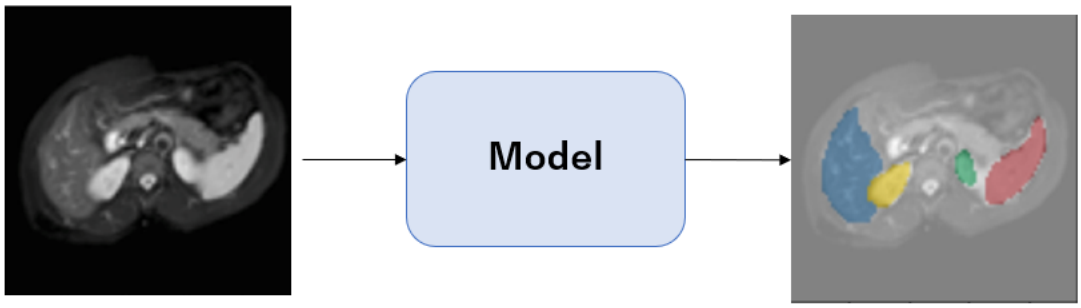


Figure 3.5: Medical image segmentation [7]

3.3 Model Architecture

The model architecture employed in this thesis builds upon previous research on vision-language pre-training models [4], incorporating three key components: the image encoder, text encoder, and fusion encoder. The image encoder focuses on mapping an image into its corresponding image representation, while the text encoder aims to map textual input into a text representation. The fusion encoder plays a crucial role in integrating the image and text representations. In the first stage of pre-training, we introduce a segmentation head to the model, specifically positioned after the fusion encoder. In the second stage of pre-training, we incorporate a Masked-Language Modeling Head and an Image-Text Matching Head into the model. During the fine-tuning phase, we incorporate a Visual Question Answering (VQA) head to perform VQA tasks. Further details regarding each architectural component will be provided in the subsequent sections. Figure 3.6 shows an overview of the architecture.

Image Encoder: The image encoder is responsible for mapping the image input to its corresponding image representation. In this thesis, we employ the ViT architecture [22] that has been pre-trained using the CLIP [40], specifically ViT/16. The input image is transformed into a sequence of embeddings, denoted as $\{v_{cls}, v_1, ..., v_N\}$, where v_{cls} represents the embedding of the $[CLS]$ token. The image encoder can be instantiated with various models capable of transforming high-resolution images into their corresponding image representations, for instance, ResNet [41] and Swin Transformer [23]. The exploration of different image encoder setups will be elaborated upon in Chapter 4.

Text Encoder: The text encoder converts the text input into its corresponding text representation. In this thesis, we use RoBERTa [42] as the chosen model for

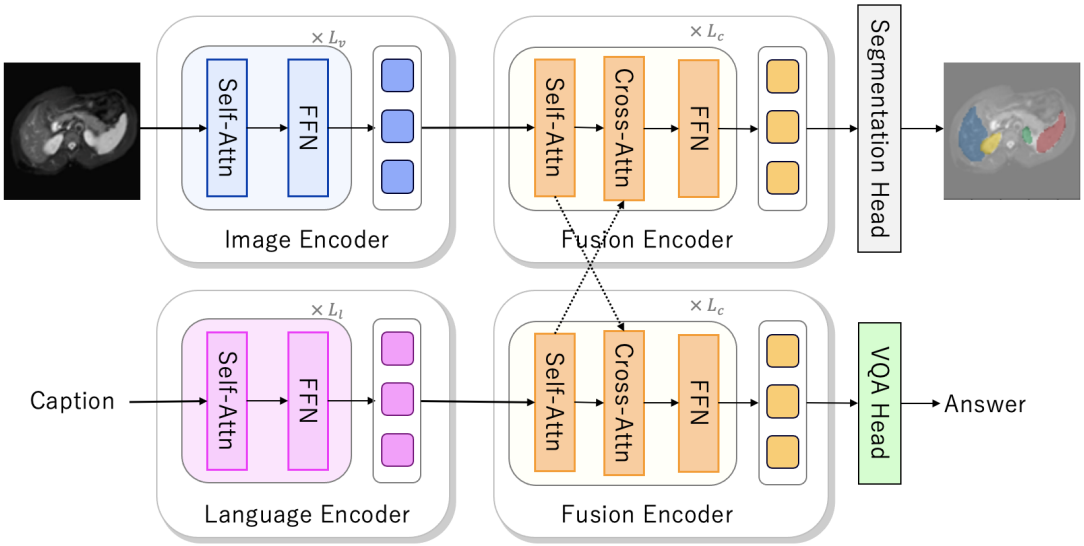


Figure 3.6: Overview of the architecture of the proposed method. The component of the image is from paper [7].

the text encoder. Specifically, the text encoder takes an input text T and transforms it into a sequence of embedding $\{w_{cls}, w_1, ..., w_N\}$. The text encoder is initialized with the weights of the RoBERTa model, which provides a strong foundation for text representation learning.

Fusion Encoder: The fusion encoder plays a crucial role in combining the image and text representations, allowing the model to learn and understand both visual and textual information effectively. In this thesis, we use a fusion encoder called the dual fusion encoder, which is based on the Transformer architecture [3]. As shown in Figure 3.6, the fusion encoder is incorporated into the model following the image encoder and text encoder modules. It consists of three layers: the self-attention layer, the cross-attention layer, and the feed-forward neural networks layer (FFN). The cross-attention layer is where the fusion of the image and text representations

occurs. Within the fusion encoder, positioned subsequent to the text encoder, the image representation assumes the role of key and value inputs for cross-attention. Conversely, the self-attention output derived from the text representation serves as the query. Similarly, in the image fusion encoder, the roles are reversed, with the image representation serving as the query and the text representation acting as the key and value inputs. This allows for effective fusion and interaction between the image and text modalities, enhancing the model's multimodal understanding and representation.

Segmentation Head: In order to facilitate image segmentation during pre-training stage one, a segmentation head is integrated into the model by connecting it to the image fusion encoder. The segmentation head consists of a block of convolutional layers, the ReLU activation function, batch normalization layers, and a convolutional up-sampling layer, following the design principles outlined in the U-Net paper [18].

Masked-Language Modeling Head: To address the masked-language modeling task in the second stage of pre-training, we integrate a masked-language modeling head into our model. It is positioned after the text fusion encoder. It consists of two layers of feed-forward neural network (FFN).

Image-Text Matching Head: To facilitate the image-text matching task in the second stage of pre-training, our model is integrated with an image-text matching head. This head consists of two layers of feed-forward neural network (FFN) and is positioned after the text fusion encoder.

VQA Head: To enable visual question answering in our model, we incorporate a VQA head after the fusion encoder. The VQA head involves two Feed Forward Neural

Network layers. Following previous work [11, 12], we consider VQA as a classification task, the output layer is designed with a size corresponding to the number of candidate answers. This VQA head enables our model to generate responses to questions by leveraging its comprehensive understanding of the multimodal inputs.

3.4 Training Pipeline

As illustrated in Figure 3.7(b), the model represents the complete pipeline used in this work, consisting of three main stages. The first stage involves pre-training with an image segmentation task, where the model learns to identify and delineate organ structures in medical images. This initial pre-training stage is followed by the second stage, which is vision-language pre-training. During this stage, the model is trained to understand the relationship between medical images and corresponding textual information. Finally, the model undergoes finetuning specifically for the visual question answering task.

It is worth noting that this approach differs from previous work, as highlighted in Figure 3.7(a), where the first stage of pre-training was not incorporated. In the subsequent section, a detailed explanation of both pre-training stage one (3.4.1) and pre-training stage two (3.4.2) will be provided, exploring their respective methodologies and objectives.

3.4.1 Pre-training Stage One

During the first stage of pre-training, we employ the semantic segmentation task to convey localized knowledge to the model. The primary objective of this pre-training stage is to equip the model with fundamental organ-specific knowledge, such as the

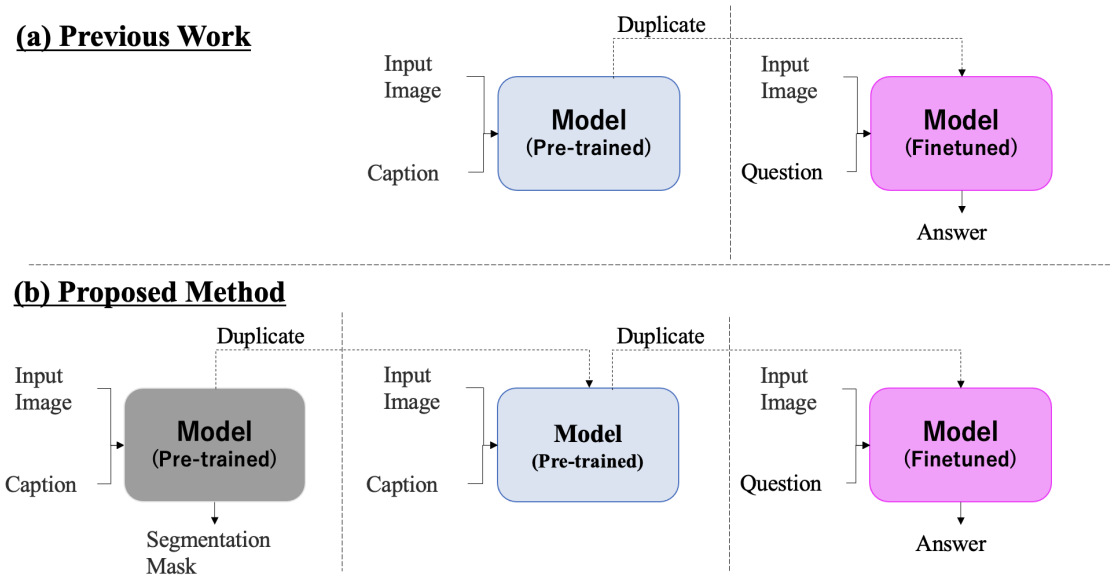


Figure 3.7: Comparative of training pipelines: Previous Work (a) versus Proposed Method (b)

spatial understanding of organs like the liver and lungs. By combining image and text inputs, the model acquires the ability to perform the segmentation task, enabling it to accurately identify the shape, size, and spatial positioning of organs within the images with the aid of textual information. The inputs to this stage consist of image-caption pairs, while the output produced is the corresponding image segmentation mask. This pre-training stage serves as a crucial foundation for subsequent stages, enhancing the model's understanding of basic organ knowledge.

Objective Function: In the pre-training stage one, the loss function includes two components: the dice loss and the binary cross-entropy loss. These loss functions are computed using the output from the model, which is the predicted image segmentation mask, and the corresponding ground truth mask.

3.4.2 Pre-training Stage Two

In the second stage of pre-training, the main objective is to train the model to acquire a comprehensive understanding of both visual and textual information. During this stage, the model receives both the image and its corresponding caption as input and learns in a self-supervised manner. The goal of this stage is to enhance the model's ability to learn meaningful representations by leveraging the rich associations between images and their captions.

Objective Function: Following previous work [11, 12], two key loss functions are utilized: the masked-language modeling (MLM) loss and the image-text matching (ITM) loss.

3.5 Datasets

Pre-training Stage One: We use the CHAOS dataset [7] as the training data. The CHAOS dataset provides a comprehensive collection of medical images with corresponding segmentation masks, which is essential for the image segmentation task. It is a collection of MRI datasets specifically focused on abdominal organs segmentation. It contains annotations for four abdominal organs: liver, right kidney, left kidney, and spleen. Figure 3.8 shows an example of the CHOAS dataset. The dataset consists of a total of 1594 images designated for training and 1537 images reserved for testing with 256x256 pixels. These masks provide pixel-level annotations that define the boundaries of the organs in the images. However, as our model relies on both text and image information as input, we face the challenge that the CHAOS dataset lacks text information. To address this, we need to generate text captions for

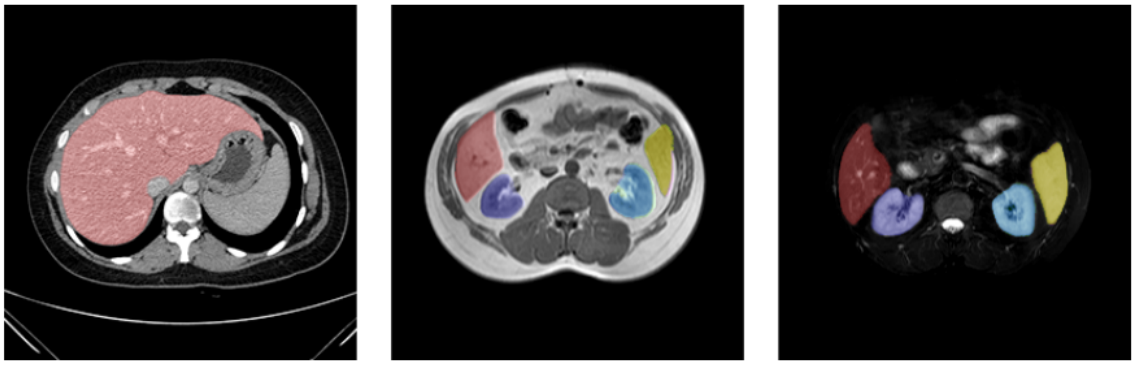


Figure 3.8: Example images from the CHAOS dataset. It shows CT and MRI images with the liver highlighted in red, right kidney in dark blue, left kidney in light blue, and spleen in yellow.

the images present in the dataset. The process of generating these captions will be explained in section 3.6.

Pre-training Stage Two: We use two datasets: the ROCO dataset [2] and the MedICat dataset [17]. These datasets offer a diverse range of medical images and corresponding text data, allowing the model to further enhance its understanding of visual and textual information.

ROCO dataset: This dataset specifically focuses on radiology images and includes over 81,000 image-caption pairs. The images encompass various medical imaging modalities such as Computer Tomography (CT), Ultrasound, X-Ray, and more. Figure 3.9 provides an illustrative example from the ROCO dataset. Figure 3.10 presents an image along with its associated textual information

MedICat dataset: This dataset comprises 217,000 image-caption pairs sourced from 131,000 open access biomedical papers. The dataset offers a diverse range of medical images and associated textual information. Figure 3.11 shows images and

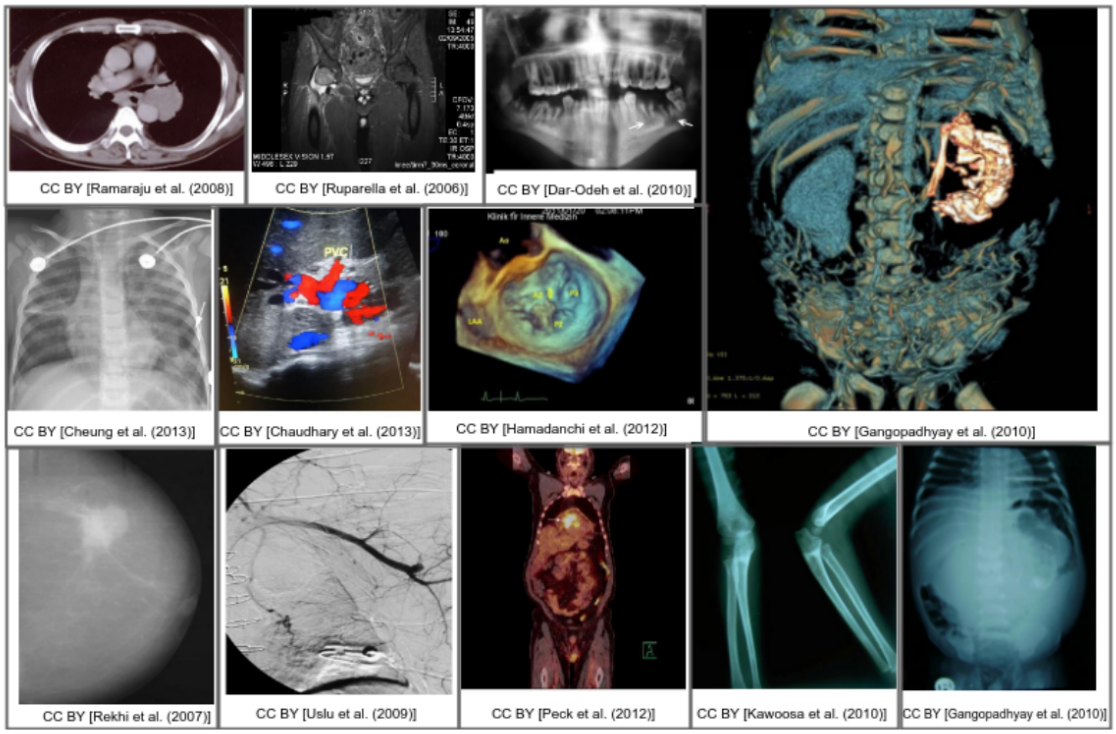


Figure 3.9: Example images from the ROCO dataset

their corresponding captions from the MedICat dataset.

Finetuning: We use on the SLAKE dataset [1]. It offers a wide variety of medical images, accompanied by corresponding questions and answers. It consists of approximately 14,000 samples. The dataset includes a diverse collection of CT and MRI scans, providing a wide range of medical imaging data for training and evaluating our model's performance on the task of medical visual question answering. Figure 3.12 shows an example from the SLAKE dataset.

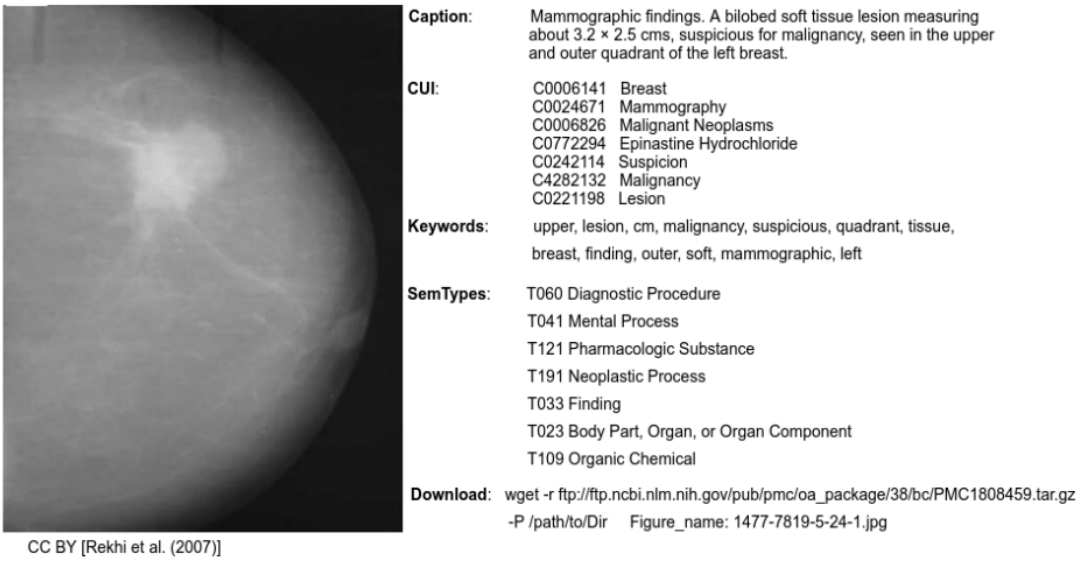


Figure 3.10: An example image and its corresponding caption are obtained from the ROCO dataset. During the pre-training phase, only the images and their corresponding captions are employed.

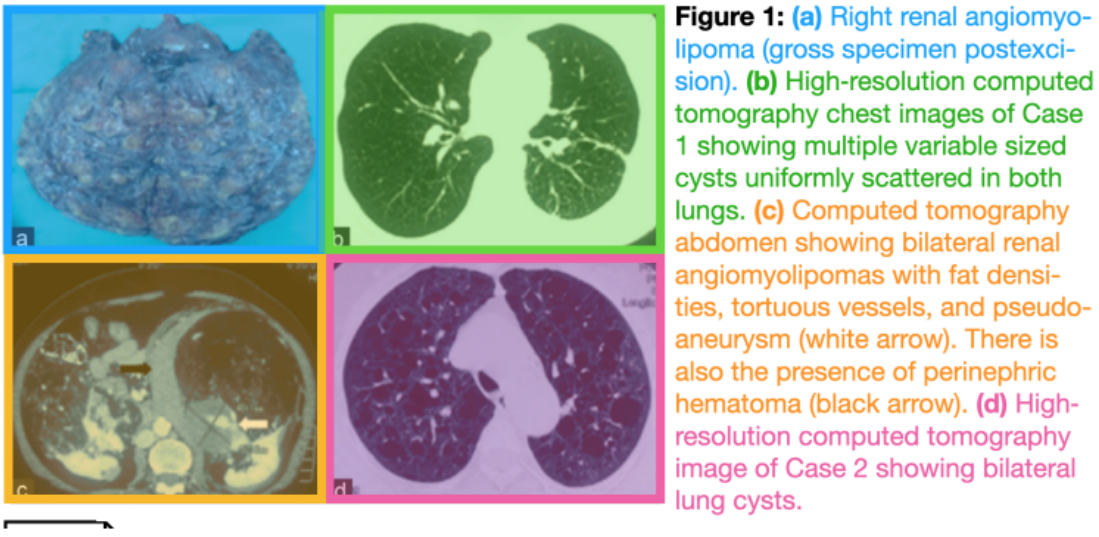
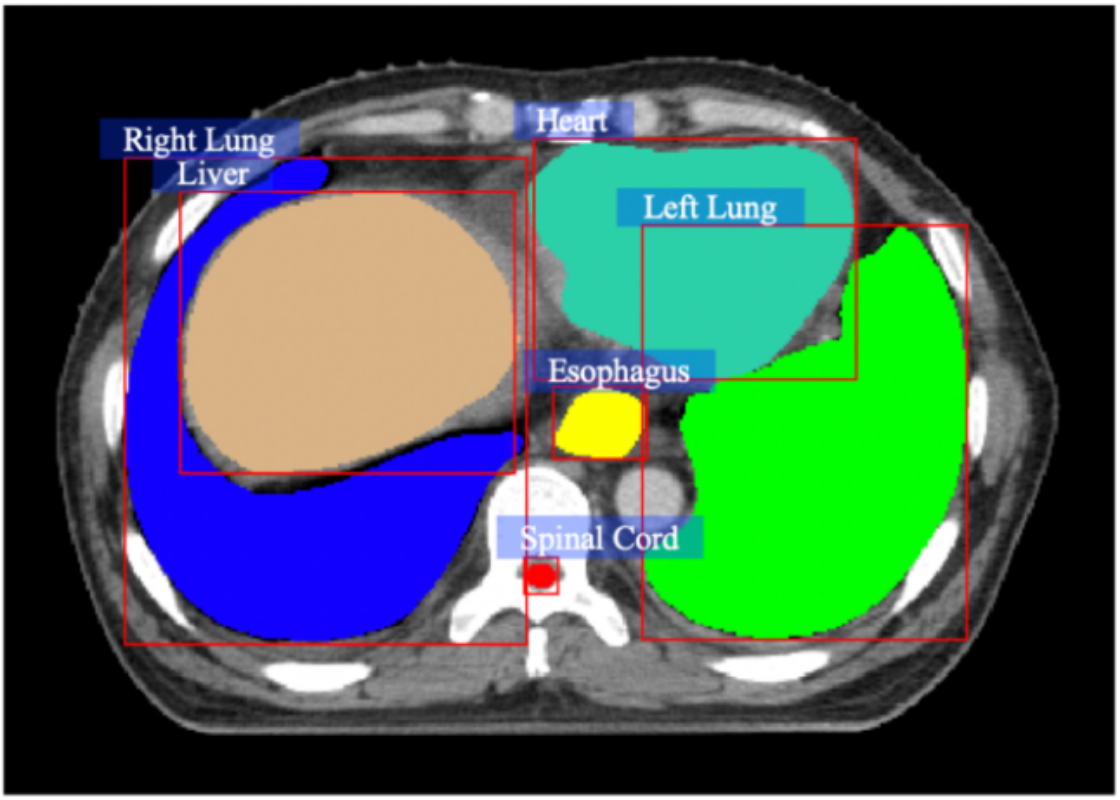


Figure 3.11: Example images and corresponding captions from the MedICat dataset



Question	➢Does the image contain left lung? ➢图片中是否包含左肺?	➢What is the function of the rightmost organ in this picture? ➢图中最右侧器官功能是什么?
Type	Vision-only	Knowledge-based
Answer Type	Closed-ended	Open-ended

Figure 3.12: An example from the SLAKE dataset is presented. It is worth noting that only English language is used during training.

3.6 Generate Caption Dataset for Pre-training Stage One

The CHAOS dataset consists of images and their segmentation masks, lacking the textual information necessary for our method. In order to address the requirement of incorporating both image and caption data in the pre-training stage one, it becomes necessary to generate the relevant textual information corresponding to the images. By synergistically integrating these two modalities, our approach effectively facilitates the pre-training process, aligning seamlessly with the core objectives of our research.

We employ the task of generating descriptive captions for the CHAOS dataset as part of our methodology. The process involves using a predefined prompt, "a magnetic resonance imaging of the $[CLASS]$," where $[CLASS]$ represents the class corresponding to the segmentation mask of the image. For instance, if the segmentation mask corresponds to the liver, the generated text would be "a magnetic resonance imaging of the Liver." see Figure 3.13. When an image contains multiple organs with their respective segmentation masks, we address this by randomly selecting one organ for each training iteration. Consequently, an image with multiple segmentation masks can yield multiple generated captions. This methodology enables the inclusion of text information that complements the image data during the pre-training stage one.

Image

A magnetic resonance imaging of the **Liver**.

Caption

Figure 3.13: Medical image segmentation with corresponding generated captions. The component of the image is from paper [7].

Chapter 4

Experiments

Chapter 4 focuses on the experiments conducted to evaluate the effectiveness of our two-stage pre-training approach (Sec. 4.1). We provide detailed insights into the experimental setup. Furthermore, we present the experimental results that demonstrate the performance of our proposed method compared to the baseline. We conduct a comprehensive analysis of the results, discussing the strengths and limitations of our approach in detail. Subsequently, we explore the impact of changing the image encoder architecture (Sec. 4.2), investigating how variations such as the use of different transformer models affect the overall performance. This analysis provides valuable insights into the generalization capabilities and adaptability of our method across different image encoders. Afterwards, we conduct a comprehensive comparison between our best-performing model and previous work methods (Sec. 4.3). By comparing our proposed method to previous approaches, we demonstrate that it brings significant advancements in integrating vision and language in the medical field. This highlights its effectiveness in addressing the challenges specific to medical applications. Finally, in Section 4.4, we present the qualitative results of our model's performance on medi-

cal VQA. This analysis includes the presentation of successful cases and failure cases, and a comprehensive examination.

4.1 Effectiveness of Two-stage Pre-training

The purpose of the proposed two-stage pre-training method is to determine if it can achieve better results compared to the original single-stage approach. We also aim to highlight the importance of acquiring basic organ knowledge before the original pre-training stage. Through the evaluation of model performance with and without the initial stage, we can demonstrate the importance of having a robust foundation in organ understanding. This analysis enables us to demonstrate the value derived from incorporating knowledge of organs into the model's learning process.

4.1.1 Experimental Setup

First Stage: The first stage involves training on the CHAOS dataset [7] for image segmentation. We utilize ViT/16 [22] pre-trained on CLIP [40] as the image encoder and RoBERTa [42] as the text encoder. The pre-training stage one takes approximately 3 hours on a single GPU A6000 with 48 GB memory. We use a batch size of 32 and an image size of 224. The epoch is set to 100, and we employ the Adam optimizer [43] with a learning rate of 1e-4.

Second Stage: In the pre-training stage two, we use the MedICat [17] and ROCO datasets [2] for the original self-supervised pre-training technique. The image encoder and text encoder maintain consistency with the configuration employed in stage one. The pre-training stage two requires around 72 hours of training on four GPU A6000 with 48 GB memory each. The batch size and image size are consistent with stage

one. We use the Adam [43] with a learning rate of 1e-5, a warm-up ratio of 10%, and linear decay of the learning rate to 0 after 10% of the total training steps. The epoch is set to 30.

Fine-tuning: In the fine-tuning phase, we conduct experiments using the VQA SLAKE dataset [1]. The training process takes around 2 hours and is performed on a single GPU A6000 with 48 GB of memory. We utilize a batch size of 32 and set the image resolution to 384x384. The epoch is set to 15. We employ the Adam optimizer [43] with a learning rate of 1e-6. To enhance the fine-tuning process, we apply RandAugment [44], following previous work [4]. The evaluation metric used is VQA classification accuracy, as shown in Equation 4.1. We specifically analyze the performance in both closed-ended and open-ended categories. The implementation is based on the PyTorch [45] and Huggingface [46] Transformers library.

$$Accuracy = \frac{\#\ of\ correct\ answers}{\#\ of\ total\ questions} \tag{4.1}$$

4.1.2 Results

Comparing One-Stage and Two-Stage Pre-Training: Table 4.1 shows the performance comparison between one-stage pre-training and two-stage pre-training on VQA SLAKE test split. The results of the two-stage pre-training approach demonstrate its superiority over the original one-stage pre-training method by achieving a 1.6% performance improvement. This highlights the necessity of employing the two-stage pre-training technique for medical datasets.

Sequence of Pre-Training Stages: Furthermore, we investigate the impact of switching the order of pre-training stages. Table 4.2 illustrates the results, indicat-

Table 4.1: Performance of one-stage pre-training framework vs proposed two-stage pre-training.

Method	Open-ended	Closed-ended	Overall
One-Stage Pre-training [4]	79.3	85.1	81.5
Two-Stage Pre-training	80.5	87.0	83.1

ing that performing a second-stage pre-training followed by first-stage pre-training leads to a 0.5% improvement compared to the baseline of one-stage pre-training. By employing the proposed pre-training step, the model achieved a performance enhancement of 1.6%. This shows the significance of the pre-training order, indicating that conducting medical image segmentation as the first-step pre-training provides the model with organ-specific knowledge prior to the original pre-training is crucial.

Table 4.2: Comparison of performance with modified pre-training order.

Method	Open-ended	Closed-ended	Overall
One-Stage Pre-training [4]	79.3	85.1	81.5
$2nd\ Stage \Rightarrow 1st\ Stage$	79.7	85.6	82.0
$1st\ Stage \Rightarrow 2nd\ Stage$	80.5	87.0	83.1

Pre-training Stage One's Effectiveness: We also evaluate the effectiveness of pre-training stage one in isolation by excluding pre-training stage two. We conduct pre-training stage one alone and compare its performance with a model initialized with random weights. The results, presented in Table 4.3, demonstrate a substantial 3.4% improvement in performance compared to the random weight initialization approach.

This indicates that pre-training stage one alone can effectively equip the model with valuable knowledge relevant to medical vision-language tasks.

Table 4.3: Comparative performance analysis: Pre-training stage one vs. Random weights initialization

Method	Open-ended	Closed-ended	Overall
Random Weights	78.2	76.0	77.3
Pre-training Stage One	79.1	83.2	80.7

4.2 Performance of Different Image Encoder Configurations

In the two-stage pre-training experiment, we conducted an investigation to assess the model's ability to generalize across different architecture designs by changing the image encoder from Vision-Transformer (ViT/16) [22] to Swin Transformer [23]. By evaluating the performance of the two-stage pre-training method with different image encoders, we aimed to investigate the generalization capability and adaptability of the model across distinct architecture choices. The implementation details are similar to Section 4.1.1

Table 4.4 shows the downstream performance results on the SLAKE test split using different image encoders. The Swin Transformer achieved a VQA accuracy of 81.4% with original pre-training. However, when we applied the two-stage pre-training, the performance improvement was only 0.9%. On the other hand, the ViT (Vision Transformer) exhibited a more notable improvement of 1.6%. This suggests

that the ViT (Vision Transformer) is more suitable for this specific method. We hypothesize that the lack of significant improvement with Swin Transformer could be attributed to its increased model complexity. Given the limited amount of medical images for training, the Swin Transformer may have been more prone to overfitting, leading to a smaller performance gain compared to ViT.

Table 4.4: Comparative performance analysis: Different image encoder setting configurations

Method	Image Encoder	Open-ended	Closed-ended	Overall
One-stage Pre-training	Swin	79.3	84.6	81.4
Two-stage Pre-training	Swin	79.6	86.5	82.3
One-stage Pre-training	ViT	79.3	85.1	81.5
Two-stage Pre-training	ViT	80.5	87.0	83.1

4.3 Comparison with Previous Works

The purpose of our experiment was to evaluate the performance of our two-stage pre-training framework in comparison to several previous methods. By conducting this comparison, we aimed to assess the effectiveness and superiority of our proposed approach. The implementation details of our experiment were similar to those described in Section 4.1.1.

We classified the models into two groups: pre-trained models and non pre-trained models. The non pre-trained models included the baseline model from the SLAKE dataset [1], while the pre-trained models consisted of PubMedCLIP [20] , METER

[4], and ARL [11].

As shown in Table 4.5, our two-stage pre-training method demonstrated superior performance compared to the non pre-trained models (SLAKE baseline) by a significant margin. This finding highlights the crucial role of pre-training in vision-language tasks. Furthermore, our model outperformed pre-trained models that utilized a similar or smaller number of images during pre-training, such as PubMedCLIP and METER models. This showcases the effectiveness of our two-stage pre-training approach over the original one-stage pre-training method. However, our method was surpassed by ARL, which had access to a larger dataset of approximately 700K image-caption pairs during pre-training. This emphasizes the importance of the number of pre-trained images in achieving optimal results in the medical vision-language domain.

Table 4.5: Comparative performance with previous works

Method	Pre-train Images	Open-ended	Closed-ended	Overall
SLAKE baseline [1]	0	72.2	79.8	75.4
PubMedCLIP [20]	≈80K	78.4	82.5	80.1
METER [4]	≈300K	79.3	85.1	81.5
Two-stage Pre-training	≈300K	80.5	87.0	83.1
ARL [11]	≈700K	81.9	91.4	85.6

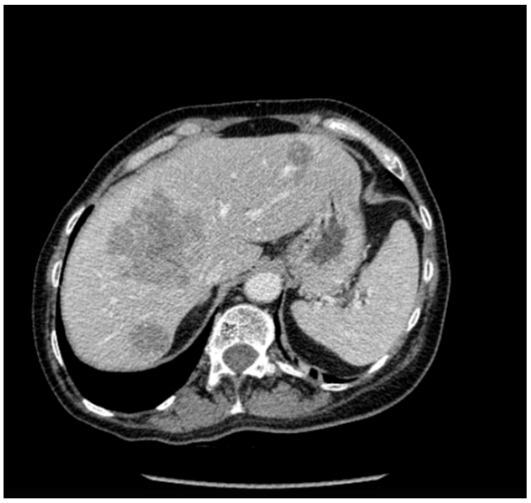
4.4 Qualitative Examples

Figure 4.1 to 4.8 show the qualitative examples from our method on the SLAKE test splits. They display both successful cases and some failure cases. The qualita-

tive results indicate that the model possesses knowledge about fundamental organ knowledge. For example, in Figure 4.1, the question is "Does the patient have liver cancer?" Our model can provide the correct answer, demonstrating its understanding of a healthy liver's appearance, which is crucial for accurate responses.

However, in Figure 4.7, we present a failure case where the question is "What diseases are included in the picture?" Our model provides an incorrect answer. Nonetheless, the model demonstrates knowledge of the abnormality's location when asked, "Where is/are the abnormality located?" and provides the correct answer. This indicates that the model possesses fundamental organ knowledge but lacks specific disease-related knowledge. This limitation can be attributed to the fact that in pre-training stage one, we focused only on semantic segmentation of organs and did not include disease-related information.

Additionally, the model exhibits some failure cases, as shown in Figure 4.8, particularly when processing abdomen images. For instance, when asked the question "Does the picture contain colon?" the model incorrectly responds. Similarly, when presented with the question "Which part of the human body is the organ located in this image?" the model mistakenly identifies it as the chest instead of the pelvic cavity. These failures can be attributed to the fact that the pre-training stage one does not include specific training on this particular body part. To address these shortcomings, one potential avenue for future improvement would involve augmenting the dataset used in pre-training stage one to encompass a wider range of organ and body part representations. By exposing the model to a more diverse set of examples, it is expected to develop a better understanding of various anatomical regions and improve its performance on related questions in the medical visual question answering task.

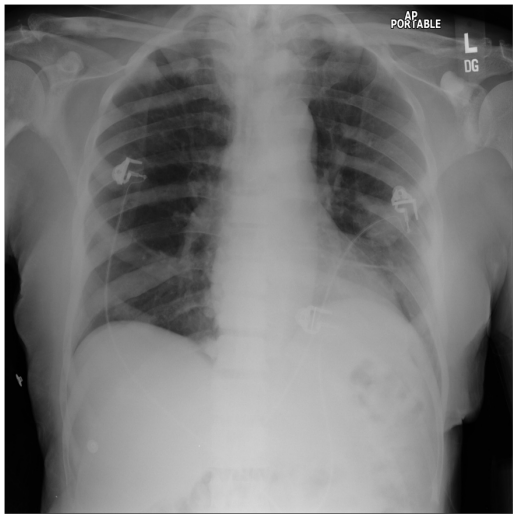


Question: Which part of the body does this image belong to?
GT: Abdomen
Ours: Abdomen

Question: Does the patient have liver cancer?
GT: Yes
Ours: Yes

Question: Where is/are the abnormality located?
GT: Liver
Ours: Liver

Figure 4.1: Qualitative examples from our method on the SLAKE test splits.

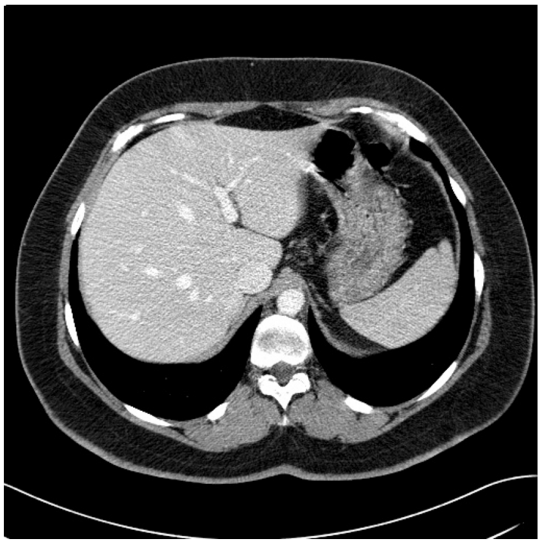


Question: Which part of the body does this image belong to?
GT: Chest
Ours: Chest

Question: Are there abnormalities in this image?
GT: Yes
Ours: Yes

Question: What is the largest organ in the picture?
GT: Lung
Ours: Lung

Figure 4.2: Qualitative examples from our method on the SLAKE test splits.

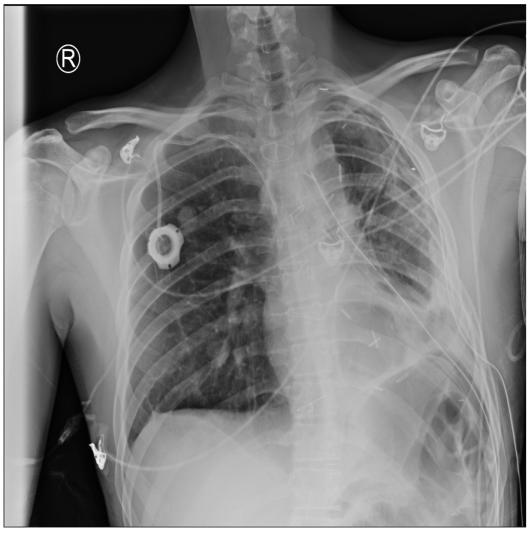


Question: Does the picture contain liver?
GT: Yes
Ours: Yes

Question: Is the liver healthy?
GT: Yes
Ours: Yes

Question: What is the organ on the left side of this image?
GT: Liver
Ours: Liver

Figure 4.3: Qualitative examples from our method on the SLAKE test splits.

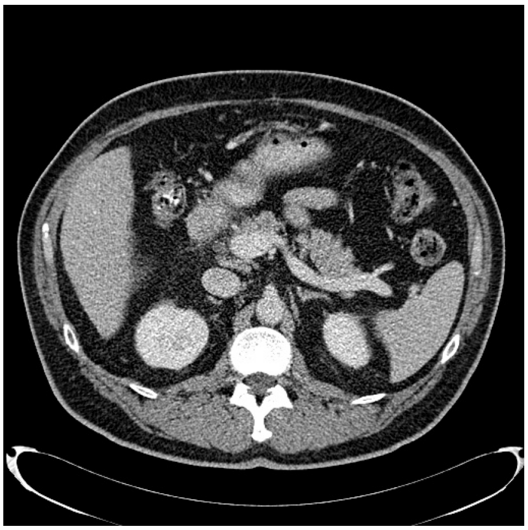


Question: Is this a transverse section?
GT: No
Ours: No

Question: Which part of the human body is the organ located in the image?
GT: Chest
Ours: Chest

Question: Is the lung healthy?
GT: No
Ours: No

Figure 4.4: Qualitative examples from our method on the SLAKE test splits.

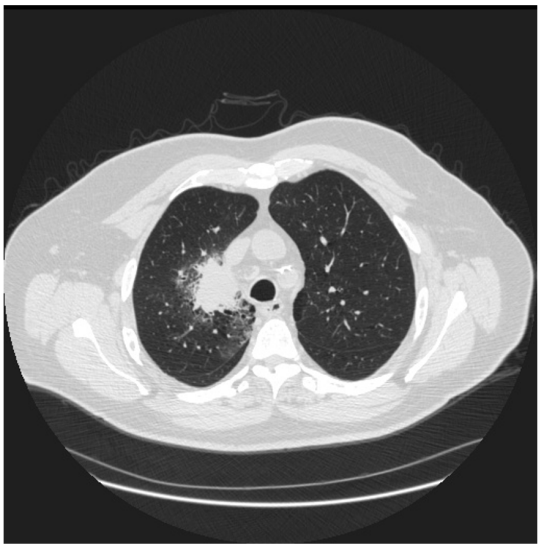


Question: Does the picture contain liver?
GT: Yes
Ours: Yes

Question: What modality is used to take this image?
GT: CT
Ours: CT

Question: What is the organ on the left side of this image?
GT: Liver
Ours: Liver

Figure 4.5: Qualitative examples from our method on the SLAKE test splits.

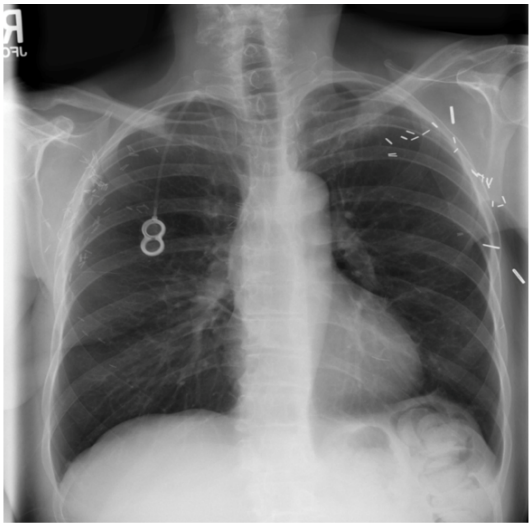


Question: Which part of the body does this image belong to?
GT: Chest
Ours: Chest

Question: What is the main organ in the image?
GT: Lung
Ours: Lung

Question: What diseases are included in the picture?
GT: Lung cancer
Ours: Lung cancer

Figure 4.6: Qualitative examples from our method on the SLAKE test splits.

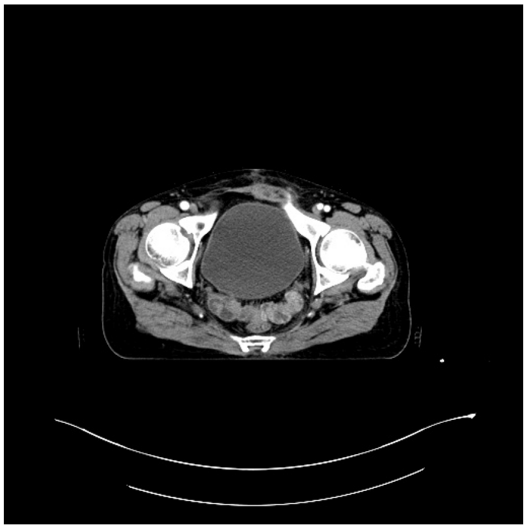


Question: Which part of the body does this image belong to?
GT: Chest
Ours: Chest

Question: Where is/are the abnormality located?
GT: Right lung, left
Ours: Right lung, left

Question: What diseases are included in the picture?
GT: Atelectasis
Ours: Mass

Figure 4.7: Qualitative examples from our method on the SLAKE test splits. The red color highlights the failure cases.



Question: Does the rectum appear in this picture?
GT: No
Ours: No

Question: Does the picture contain colon?
GT: Yes
Ours: No

Question: Which part of the human body is the organ located in the image?
GT: Pelvic cavity
Ours: Chest

Figure 4.8: Qualitative examples from our method on the SLAKE test splits. The red color highlights the failure cases.

Chapter 5

Conclusions and Future Improvements

5.1 Conclusions

In this thesis, we addressed the limitations of previous medical vision-language pre-training approaches, specifically the lack of basic organ knowledge in the models. To overcome this issue, we have proposed a two-stage pre-training method.

By incorporating an initial stage of image segmentation pre-training, our approach ensures that the model acquires fundamental organ knowledge before engaging in the original vision-language pre-training. This enables the model to acquire a better understanding of the relationship between images and textual information during the vision-language pre-training (stage two), resulting in improved performance in medical visual question answering tasks.

The results obtained from our experiments demonstrate the superiority of the two-stage pre-training method compared to the one-stage pre-training method. This high-

lights the importance of incorporating organ-specific information during pre-training to enhance the model's performance in medical vision-language tasks. Overall, our findings contribute to improving the effectiveness of vision-language models in the medical domain.

5.2 Future Improvements

In the domain of vision-language pre-training, the performance of models often shows a correlation with the scale of the pre-training dataset. Increasing the number of image-caption pairs used for pre-training can be a promising direction to further enhance the model's performance in medical downstream tasks. Increasing the training data can help the model better capture complex relationships between images and their associated language, leading to improved generalization and performance across various medical vision tasks Exploring strategies to collect and curate such large-scale datasets in the medical domain would be valuable for future research.

Another important avenue for future work is to address the challenge of catastrophic forgetting during the transition from the first-stage to the second-stage of pre-training. During the second-stage pre-training phase, the model may forget the essential organ knowledge acquired in the first-stage pre-training while focusing on learning new representations in the second-stage pre-training. To mitigate this issue, incorporating methods such as Adapters [47] can be beneficial. Adapters enable the model to fine-tune specific knowledge without overwriting previously learned information. By introducing adapters between transition stages of pre-training, the model can effectively preserve the learned organ knowledge while continuing to adapt and specialize in the second-stage pre-training. Investigating methods to optimize

the transition process would be a promising avenue for future research in medical vision-language pre-training.

Lastly, an interesting avenue to explore would be addressing the model's lack of knowledge about diseases, which becomes evident from the qualitative results. This limitation is understandable, considering that the pre-training stage one focuses primarily on organ knowledge through segmentation. To improve the model's understanding of diseases, a potential direction could involve incorporating disease segmentation as an input during training. By integrating disease-specific information into the pre-training process, we may enhance the model's ability to handle disease-related questions in the medical visual question answering task.

Bibliography

[1] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[2] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and C. Friedrich. Radiology objects in context (roco): A multimodal image dataset. In *CVII-STENT/LABELS@MICCAI*, 2018.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.

[4] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18166–18176, June 2022.

[5] Yuhang Liu, Wei Wei, Daowan Peng, and Feida Zhu. Declaration-based prompt

tuning for visual question answering. In *Proceedings of the Thirty-first International Joint Conference on Artificial Intelligence, IJCAI-22*, 2022.

[6] Zhang Chen, Zhiqiang Tian, Jihua Zhu, Ce Li, and Shaoyi Du. C-cam: Causal cam for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11676–11685, 2022.

[7] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, page 101950, 2021.

[8] Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *CoRR*, 2020.

[9] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, pages 1–10, 2018.

[10] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. 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.

[11] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5152–5161, 2022.

[12] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V*, pages 679–689. Springer, 2022.

[13] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

[14] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, pages 9694–9705, 2021.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[17] Sachin Mehta Ben Bogin Madeleine van Zuylen Sravanthi Parasa Sameer Singh Matt Gardner Sanjay Subramanian, Lucy Lu Wang and Hannaneh Hajishirzi. MedICaT: A Dataset of Medical Images, Captions, and Textual References. In *Findings of EMNLP*, 2020.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[19] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5579–5588, 2021.

[20] Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193. Association for Computational Linguistics, 2023.

[21] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1985.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg

Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, pages 1877–1901, 2020.

[26] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP (1)*, pages 5099–5110. Association for Computational Linguistics, 2019.

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 2019.

[28] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

[29] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[31] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. *arXiv preprint arXiv:2302.08958*, 2023.

[32] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, pages 6070–6080, 2022.

[33] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE, 2021.

[34] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question

answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 522–530. Springer, 2019.

[35] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 64–74. Springer, 2021.

[36] Haifan Gong, Guanqi Chen, Mingzhi Mao, Zhen Li, and Guanbin Li. Vqamix: Conditional triplet mixup for medical visual question answering. *IEEE Transactions on Medical Imaging*, pages 3332–3343, 2022.

[37] Anda Zhang, Wei Tao, Ziyan Li, Haofen Wang, and Wenqiang Zhang. Type-aware medical visual question answering. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4838–4842. IEEE, 2022.

[38] Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2982–2990, 2022.

[39] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *ArXiv*, 2022.

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[42] Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, 2021.

[43] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of International Conference on Representation Learning*, 2015.

[44] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[45] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[46] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[47] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

Acknowledgements

First and foremost, I express my deep gratitude to my supervisor Professor Takayuki OKATANI for the immense support and illuminating guidance. Truly, without his guidance and unwavering will to share knowledge, this work would not be possible in the state that it is now. I am forever grateful for having the opportunity to work under his supervision.

I am deeply grateful to Professor Hiroyuki TAKIZAWA and Associate Professor Shingo KAGAMI for their invaluable contributions as members of my thesis committee. I am sincerely thankful for their time, support, and valuable input throughout the process.

I am sincerely grateful to Assistant Professor Masanori SUGANUMA for the guidance and encouragement during times of uncertainty an overall my research.

I would like to share my thankfulness to the members of the Computer Vision lab, namely Kang Jun LIU, Nguyen Van QUANG, Bonappol LIMANOND, Kitto?? for the support and experience they shared with me throughout my research.

I also extend my appreciation for my family members, my parents BOBOJANOVA U. and BOLTABOYEV Sh. for their tremendous support throughout my life, without them I would not be here to write this work. I am also grateful to my sister RAKHIMOVA M. for always being encouraging and inspiring in all she does.

I am deeply appreciative of all the individuals mentioned above, along with everyone who, through various means, both directly and indirectly, have aided me in the completion of this work.