Universidade Estadual de Campinas
Instituto de Computação

Talles Viana Vargas

Chest X-Ray Description Based on Lightweight Language Models for Accessible Medical Imaging Analysis

Descrição de Radiografias de Tórax Baseada em Modelos Leves de Linguagem para uma Análise Acessível de Imagens Médicas

CAMPINAS
2025

Talles Viana Vargas

Chest X-Ray Description Based on Lightweight Language Models for Accessible Medical Imaging Analysis

Descrição de Radiografias de Tórax Baseada em Modelos Leves de Linguagem para uma Análise Acessível de Imagens Médicas

Dissertação apresentada ao Instituto de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Mestre em Ciência da Computação.

Dissertation presented to the Institute of Computing of the University of Campinas in partial fulfillment of the requirements for the degree of Master in Computer Science.

Supervisor/Orientador: Prof. Dr. André Santanchè
Co-supervisor/Coorientador: Prof. Dr. Hélio Pedrini

Este exemplar corresponde à versão final da Dissertação defendida por Talles Viana Vargas e orientada pelo Prof. Dr. André Santanchè.

CAMPINAS
2025

Ficha catalográfica
Universidade Estadual de Campinas (UNICAMP)
Biblioteca do Instituto de Matemática, Estatística e Computação Científica
Ana Regina Machado - CRB 8/5467

Vargas, Talles Viana, 1993-
V426c Chest X-ray description based on lightweight language models for accessible medical imaging analysis / Talles Viana Vargas. – Campinas, SP : [s.n.], 2025.

Orientador: André Santanchè.
Coorientador: Hélio Pedrini.
Dissertação (mestrado) – Universidade Estadual de Campinas (UNICAMP), Instituto de Computação.

1. Grandes modelos de linguagem. 2. Tórax - Radiografia. I. Santanchè, André, 1968-. II. Pedrini, Hélio, 1963-. III. Universidade Estadual de Campinas (UNICAMP). Instituto de Computação. IV. Título.

Informações complementares

Título em outro idioma: Descrição de radiografias de tórax baseada em modelos leves de linguagem para uma análise acessível de imagens médicas
Palavras-chave em inglês:
Large language models
Chest - Radiography
Área de concentração: Ciência da Computação
Titulação: Mestre em Ciência da Computação
Banca examinadora:
André Santanchè [Orientador]
Marcelo Schweller
Rodrigo Frassetto Nogueira
Data de defesa: 08-05-2025
Programa de Pós-Graduação: Ciência da Computação

Objetivos de Desenvolvimento Sustentável (ODS)
ODS: 9. Inovação e infraestrutura

Identificação e informações acadêmicas do(a) aluno(a)
- ORCID do autor: https://orcid.org/0000-0001-5962-123X
- Currículo Lattes do autor: http://lattes.cnpq.br/7337520432327778

- Prof. Dr. André Santanchè
  Instituto de Computação, Universidade Estadual de Campinas (UNICAMP)
- Dr. Marcelo Schweller
  Pesquisador Independente
- Dr. Rodrigo Frassetto Nogueira
  Maritaca AI

A ata da defesa, assinada pelos membros da Comissão Examinadora, consta no SIGA/Sistema de Fluxo de Dissertação/Tese e na Secretaria do Programa da Unidade.

Acknowledgements

We would like to express our sincere gratitude to Dr. Ligia Cayres Ribeiro, MD, whose extensive experience in emergency medicine provided invaluable clinical insights throughout this project. This study was partially financed by CNPq (The Brazilian National Council for Scientific and Technological Development) grant numbers 428459/20188 and 400062/2023-2. This work was supported by Amazon Web Services, Inc.

Resumo

Grandes modelos de linguagem (LLMs) têm sido amplamente empregados em diversas tarefas de processamento de texto, incluindo a geração de conteúdo, tradução ou correção textual. Na visão computacional, esses modelos possuem aplicações na geração de legendas a partir de imagens e em sistemas de perguntas e respostas visuais (VQA). Na área de imagens médicas, embora existam estudos propondo diagnósticos automatizados de diferentes modalidades (raios-X, ressonância magnética, tomografia computadorizada), poucos trabalhos exploram o potencial dos LLMs nesse contexto. Os estudos existentes frequentemente priorizam a melhora de desempenho utilizando modelos com dezenas de bilhões de parâmetros, resultando em soluções computacionalmente custosas. Dessa forma, este trabalho avaliou a utilização de LLMs pré-treinados com um número relativamente menor de parâmetros para gerar descrições de imagens de raio-X torácico no contexto médico. O principal objetivo deste trabalho foi desenvolver uma arquitetura leve adotando LLM, buscando utilizar modelos pré-treinados para reduzir custos computacionais sem comprometer significativamente a qualidade dos resultados. Avaliamos múltiplas escolhas para a arquitetura, incluindo a seleção de um modelo de visão computacional ideal (U-Net vs. PSPNet) e a inicialização do módulo Q-Former com pesos do domínio (BiomedBERT). Nossos experimentos foram conduzidos utilizando métricas de eficácia clínica (CE) - precisão, revocação e medida F1 - e métricas de geração de linguagem natural (NLG), como BLEU e ROUGE. A análise qualitativa das amostras de texto geradas revelou que o modelo produz descrições detalhadas e clinicamente relevantes, frequentemente superando a brevidade dos laudos de referência. No entanto, alucinações ocasionais - frases sem sentido ou irrelevantes - foram observadas, particularmente em casos onde os achados não ficam claros. Comparações quantitativas com métodos estado da arte mostraram que nossa arquitetura, com apenas 347 milhões de parâmetros no gerador de texto, alcançou desempenho competitivo, particularmente em precisão (0,5142) e medida F1 (0,4564), mantendo um custo computacional significativamente menor em comparação com modelos como XRayGPT (7 bilhões de parâmetros) e Med-PaLM (540 bilhões de parâmetros). Os resultados obtidos demonstraram o potencial dessa abordagem para beneficiar médicos emergencistas e estudantes de medicina durante o processo de análise de raios-X torácicos, potencialmente fornecendo suporte através de pré-escrita, pré-análise e auxílio na elaboração de laudos. O sistema proposto permite a execução em computadores com configuração padrão, tornando-o acessível para ambientes com recursos limitados. A implementação desta tecnologia pode tornar as análises mais rápidas e precisas, aumentando a eficiência desses profissionais na prática clínica.

Abstract

Large Language Models (LLMs) have been widely employed in various text processing tasks, including content generation, translation, and textual correction. In computer vision, these models have applications in generating captions from images and in Visual Question-Answering (VQA) systems. However, in the field of medical imaging, while studies exist proposing automated diagnoses for different modalities (X-rays, magnetic resonance imaging, computed tomography), few works explore the potential of LLMs in this context. Existing studies frequently prioritize performance improvement using models with tens of billions of parameters, resulting in computationally expensive solutions. Addressing this gap, this work evaluates the use of pre-trained LLMs with a relatively smaller number of parameters to generate descriptions of thoracic X-ray images in the medical context. The main objective of this work was to develop a LLM-driven lightweight architecture, prioritizing pre-trained models to reduce computational costs without significantly compromising result quality. We evaluated multiple design choices, including the selection of an optimal image encoder (U-Net vs. PSPNet) and the initialization of the Q-Former module with domain-specific weights (BiomedBERT). Our experiments were conducted using both clinical efficacy (CE) metrics—precision, recall, and F1-score—and natural language generation (NLG) metrics such as BLEU and ROUGE scores. Qualitative analysis of generated text samples revealed that the model produces detailed and clinically relevant descriptions, often surpassing the brevity of reference reports. However, occasional hallucinations – meaningless or irrelevant phrases – were observed, particularly in cases of subtle findings. Quantitative comparisons against state-of-the-art methods showed that our architecture, with only 347 million parameters in its text decoder, achieved competitive performance, particularly in precision (0.5142) and F-Score (0.4564) while maintaining significantly lower computational demands compared to models such as XRayGPT (7 billion parameters) and Med-PaLM (540 billion parameters). The results obtained demonstrated the potential of this approach to benefit emergency physicians and medical students during the process of analyzing chest X-rays, potentially providing support through pre-writing, pre-analysis, and assistance in report preparation. The proposed system can be run on standard-configured computers, making it accessible to environments with limited resources. Its implementation enables faster and more accurate analyses, enhancing the efficiency of professionals in clinical practice.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

This chapter outlines the problem to be investigated, presents the main motivations of the research topic, describes the main objectives, the expected contributions of the work, and the research questions. Additionally, it provides an overview of the text organization and a publication carried out during the execution of the research.

1.1 Contextualization

Chest X-rays are a fundamental diagnostic tool in healthcare, containing valuable information that can assist healthcare professionals in diagnosing various pulmonary, cardiac, or traumatic conditions. This imaging modality provides a snapshot of a patient's internal chest structure, revealing critical details that can guide medical practitioners in their decision-making processes. Currently, X-rays are the cheapest and most widely used imaging modality. For example in Brazil, more than half of the imaging equipment in the public health system are X-ray machines [53].

Radiologists are the primary professionals responsible for interpreting chest X-rays, extracting the most relevant information, and transcribing their findings into textual reports. This manual process is not only labor-intensive but is also prone to variations in interpretation and reporting. These professionals must write their reports in a thorough manner, as they are the primary source of information for other healthcare professionals, commonly following a structured format to describe the findings, the impression, and the recommendations [39].

However, emergency physicians also need to interpret chest X-rays [58], especially in emergency situations where a radiologist is not available. In these cases, the physician must quickly interpret the image, find the traumas and make a decision based on the findings, in order to help the patient as fast as possible. In this context, an automated system that generates descriptive captions for chest X-ray images can be a valuable tool, providing immediate support to healthcare professionals in their decision-making processes. Complementary, this automated system could also help them write the reports, saving time and reducing the workload. Hence, this was the primary focus of this research.

Furthermore, in critical conditions such as pneumothorax, which can be life-threatening if missed, may present with subtle radiographic signs that could be overlooked

during busy emergency shifts. An AI system that systematically evaluates chest X-rays could serve as a "second set of eyes," helping physicians identify findings they might not have initially considered, particularly when cognitive load is high or when dealing with cases outside their immediate area of suspicion.

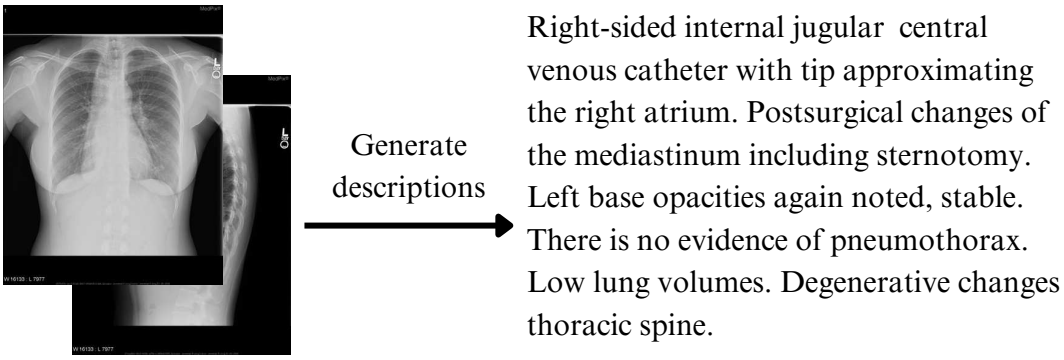


Figure 1.1: The Chest X-Ray captioning task involves extracting the findings from the X-Ray image and transcribing them to a report. Source: elaborated by this author.

In recent years, the advancements in Artificial Intelligence (AI), and specially in the Natural Language Processing (NLP) field, have revolutionized the way humans interact with machines. Large Language Models (LLMs), such as those powering ChatGPT, Claude, DeepSeek, and other emerging tools, have demonstrated exceptional text generation capabilities. These models, trained on vast textual datasets from the internet, can produce text with impressive syntactic and semantic quality [15, 26, 54, 72].

An impressive characteristic of LLMs is their ability to perform tasks for which they were not specifically trained, exhibiting properties of “few-shot” or “zero-shot learning”, as demonstrated by their performance in medical exams and potential in clinical, educational, and research contexts [71]. Furthermore, the landscape of LLMs is rapidly evolving towards multimodal capabilities [6], with models such as GPT-4 now accepting multimodal input, including text and image data [71]. This advancement indicates a step towards AI systems that can process and understand diverse data formats, exemplified by some Large Multimodal Models (LMMs) such as Gemini models and the medically specialized Med-Gemini family [79], developed to leverage various medical data including imaging. The increasing proficiency in handling multiple modalities underscores the potential of these advanced models to address complex challenges across various domains, including medicine.

Despite these promising capabilities, a critical barrier to widespread adoption of AI in healthcare settings, particularly in emergency departments and resource-constrained environments, is the substantial computational infrastructure required to run state-of-the-art LLMs. Several studies have harnessed the developments in AI and NLP to automate and streamline the process of transforming visual information from chest X-rays into informative textual descriptions. Many of them rely on small language models [9, 10], which are not well-suited for effective text generation.

Models such as BERT [20] (with only 110 million parameters in its largest variant) were designed primarily for understanding rather than generating text, lacking the capacity for complex, coherent text generation that larger models possess. In contrast, modern LLMs

such as GPT-3 [8] (with 175 billion parameters, roughly 1,600 times larger than BERT) excel at generating fluent, contextually appropriate text. Furthermore, to increase the general capacity of these models, researchers are continually expanding their parameter count, further intensifying computational requirements and widening the accessibility gap for clinical implementation [6].

Few initiatives seek to apply and harness the potential of LLMs in medical text generation, and those that do use computationally expensive models with tens of billions of parameters, demanding significant computational resources, energy consumption, and specialized hardware that may be inaccessible to many healthcare facilities [62]. This accessibility gap is particularly problematic in emergency medicine, where rapid analysis and decision-making are essential [1], and in developing regions where technological infrastructure is limited due to underfinancing, as in Brazil, for example [53].

Therefore, this research work aims to evaluate the use of more accessible LLMs with fewer parameters for generating medical descriptions from chest X-ray images, with a focus on the emergency scenario. This focus is particularly relevant because all currently available datasets originate from emergency departments. The study builds on the training approach outlined in the BLIP-2 work [46], which uses a single module to align image features extracted by a vision encoder (a Convolutional Neural Network) with the input text features of the LLM. This single-module approach is computationally efficient, reducing training costs. Smaller, more efficient models that maintain adequate performance while requiring fewer resources represent a promising direction for practical implementation in clinical settings.

The results of this research have the potential to benefit emergency physicians and students during the X-ray analysis process and training, enabling automatic preliminary diagnoses that can serve as additional support, or as a writing assistant for the reports. Furthermore, this research was refined with the guidance of a medical professional researcher with extensive experience in emergency departments in Brazil.

1.2 Problem Definition

In emergency departments, where quick decisions are crucial, physicians must frequently interpret chest X-ray images to make immediate clinical decisions, often without readily available radiologist support. This task requires careful analysis and precise documentation, following a series of cognitive steps: observing the image features, identifying potential abnormalities, mentally organizing the findings, and finally documenting them in a structured report. This process is not only time-consuming but also challenging in high-pressure emergency settings where physicians must manage multiple patients simultaneously.

The development of an automated system to assist emergency physicians in this process must mirror these cognitive steps:

1. **Chest X-ray Visual Feature Extraction**: Initially, it extracts crucial image features, mirroring the physician's visual analysis. A computer vision model, also

called vision encoder, is employed for systematically identifying relevant patterns and structures in the image.

2. **Feature Transformation and Alignment**: These extracted visual features must be transformed and aligned into a format that bridges the gap between image patterns and medical terminology. This step parallels how physicians mentally connect visual findings to their clinical descriptions.

3. **Text Generation**: Finally, an NLP model processes these aligned features to generate descriptive reports, similar to how physicians formulate their findings into structured documentation. This text aims to serve as an initial draft or supportive reference for the physician's final report.

Recent approaches to this problem have explored various technical solutions, ranging from simple neural networks to sophisticated LLMs. While LLM-based systems have shown promising results in generating detailed and coherent medical reports, they often require substantial computational resources due to their billions of parameters. This resource intensity creates a practical barrier to deployment in many healthcare settings, particularly in emergency departments where computing infrastructure may be limited.

Our research addresses this challenge by developing a lightweight chest X-ray description system that leverages smaller, more efficient LLMs while maintaining report quality. The goal is to create a practical tool that can be readily deployed in emergency settings, providing immediate support for physicians in their X-ray interpretation and documentation tasks. This system aims to enhance workflow efficiency while remaining accessible to healthcare facilities regardless of their computational resources.

1.3 Challenges

The development of an automated system for chest X-ray captioning, particularly one that leverages smaller LLMs, involves significant challenges that need to be addressed:

- **Model Architecture and Efficiency** The right balance between model performance and computational efficiency must be found. While LLMs with billions of parameters may deliver excellent results, they are often impractical in resource-limited environments. The challenge lies in developing an architecture that can generate high-quality reports using smaller, more efficient models without significantly compromising accuracy or clinical utility.

- **Quality Assessment** Evaluating the quality of generated reports presents a unique challenge, as it requires metrics that align with clinical needs. Traditional natural language metrics may not adequately capture medical accuracy or clinical relevance. This requires:

  - Development of evaluation methods that reflect clinical utility
  - Metrics that can assess both linguistic quality and medical accuracy

  - Methods to compare performance with larger, more resource-intensive models

- **Clinical Accuracy** Maintaining high clinical accuracy while using smaller models presents several challenges:
  - Processing multiple X-ray views effectively
  - Capturing subtle radiological findings
  - Maintaining consistency in medical terminology
  - Balancing between completeness and conciseness in reports

- **Practical Implementation** While our research focuses on developing a lightweight system, ensuring its practical deployment presents additional challenges:
  - Validating performance on diverse hardware configurations
  - Maintaining rapid inference times for emergency use
  - Possible integration with existing healthcare workflows and systems

Addressing these challenges is fundamental to the success of our research project. The first two challenges – model architecture and quality assessment – are primary focuses of this work. The clinical accuracy and practical implementation challenges represent important considerations that guided our design decisions and evaluation methods.

1.4 Objectives

The primary objective of our work is to design a lightweight, efficient, and accessible system for chest X-ray description that leverages smaller LLMs. This system aims to automate the generation of accurate descriptions for chest X-ray images, assisting the work of healthcare professionals and enhancing the interpretability and utility of medical imaging data.

By addressing the challenges of model efficiency, evaluation metrics, and practical implementation on standard computers, our research aims to democratize access to advanced chest X-ray analysis, promoting efficiency, accuracy, and widespread usability in healthcare settings, regardless of the computational resources available.

To reach our primary objective, some specific goals were established:

- Investigate viability of smaller in-domain pretrained LLMs – pretrained on medical text – on text generation.
- Propose and implement an architecture for chest X-ray captioning using lightweight models in both vision encoder and text decoder stages.
- Compare the proposed architecture with other available approaches.
- Evaluate the results using standard metrics, as well as obtain feedbacks from healthcare professionals.

1.5 Contributions

The primary contribution of this research is the study, design, and implementation of a lightweight LLM-driven chest X-ray description system that is not resource intensive, i.e., has fewer parameters than other techniques, in order to facilitate the application of such technology in less powerful computers. Moreover, we discuss the current most adopted metrics for medical generated text, examining whether they are a good proxy for the quality of the system or not. Beyond these technical findings, we aspire to foster discussions on accessible medical captioning systems in the research community.

1.6 Research Questions

The central research questions guiding this investigation are as follows:

Viability of Smaller, In-Domain Pre-Trained LLMs

- Can smaller in-domain pre-trained LLMs effectively generate well-written captions for chest X-ray images?
- Do these smaller LLMs demonstrate the linguistic proficiency and semantic quality required for generating coherent and informative descriptions?

Caption Accuracy

- Are smaller, in-domain pre-trained LLMs capable of generating accurate captions for chest X-ray images?
- To what extent can these models consistently capture the diagnostic content of the images in their descriptions?

Evaluation Metrics

- Are the standard commonly-used metrics for text generation capable to capture the quality of chest X-ray descriptions?
- How well do they correlate with radiologists' assessments of report accuracy?

To answer these questions, a literature review is conducted concerning the use of deep learning techniques for chest X-ray description generation. Subsequently, relevant models are studied and evaluated to address the proposed research questions.

In general, we expect that the results of this research can help advance the study of methods for chest X-ray description generation, contributing to advances in the field of machine learning and assisting healthcare professionals in daily practice.

1.7 Publication

The following scientific paper was derived from this research work:

- T.V. Vargas, H. Pedrini, A. Santanchè. *LLM-Driven Chest X-Ray Report Generation with a Modular, Reduced-Size Architecture*. Brazilian Conference on Intelligent Systems (BRACIS), Belém-PA, Brazil, pp. 199-211, November 17-21, 2024.

1.8 Text Organization

This document is organized as follows. Chapter 1 introduces the problem, describing the challenges, objectives and research questions that will guide our research. Chapter 2 describes some fundamental concepts and related work within the context of this research. Chapter 3 details the methodology, including the experimental setup, the base deep learning architecture, the datasets used for training, and the evaluation metrics. Chapter 4 presents the obtained results along with a detailed discussion. Finally, Chapter 5 concludes the work by revisiting the proposed approach, summarizing the key findings, and highlighting potential directions for future research.

Chapter 2

Literature Review and Concepts

This chapter describes some relevant concepts and approaches related to the topics investigated in this work.

2.1 Fundamental Concepts

This section aims to elucidate important concepts that enhance the understanding of this project. These concepts serve as building blocks for comprehending the subsequent discussions about the related work.

2.1.1 Medical Imaging and Chest X-Rays

Medical imaging encompasses a range of non-invasive techniques for visualizing internal anatomy without opening up the body [37]. The modalities include X-ray radiography, computed tomography (CT), magnetic resonance imaging (MRI), ultrasonography, and more. In the emergency department, radiographs serve as the first-line imaging tool that can quickly reveal life-threatening conditions such as tension pneumothorax, allowing medical teams to make rapid diagnostic decisions and begin appropriate interventions when needed [1].

X-ray imaging, discovered by Wilhelm Röntgen in 1895, operates by passing high-energy electromagnetic radiation through the body [37]. Tissues of varying densities absorb radiation differently: dense structures such as bones appear white, while air-filled spaces such as lungs appear dark [38]. More specifically, chest X-ray is the most commonly performed imaging test, specially in emergency settings, given its widespread availability, low cost and ability to be conducted at the patient's bedside. It offers valuable insights into lung parenchyma and associated pathologies, as well as cardiovascular and pleural abnormalities [32].

A chest X-ray can be obtained in multiple positions depending on the clinical hypothesis. The most common projections include Posterior-Anterior (PA), where the X-ray beam passes from back to front with the patient facing the detector; Anterior-Posterior (AP), where the beam travels from front to back, typically used for bedside examinations in critically ill patients; and lateral views, which are valuable for localizing lesions and

evaluating structures that may be obscured in frontal projections [38]. Figure 2.1 depicts the possible positions for a chest X-ray.

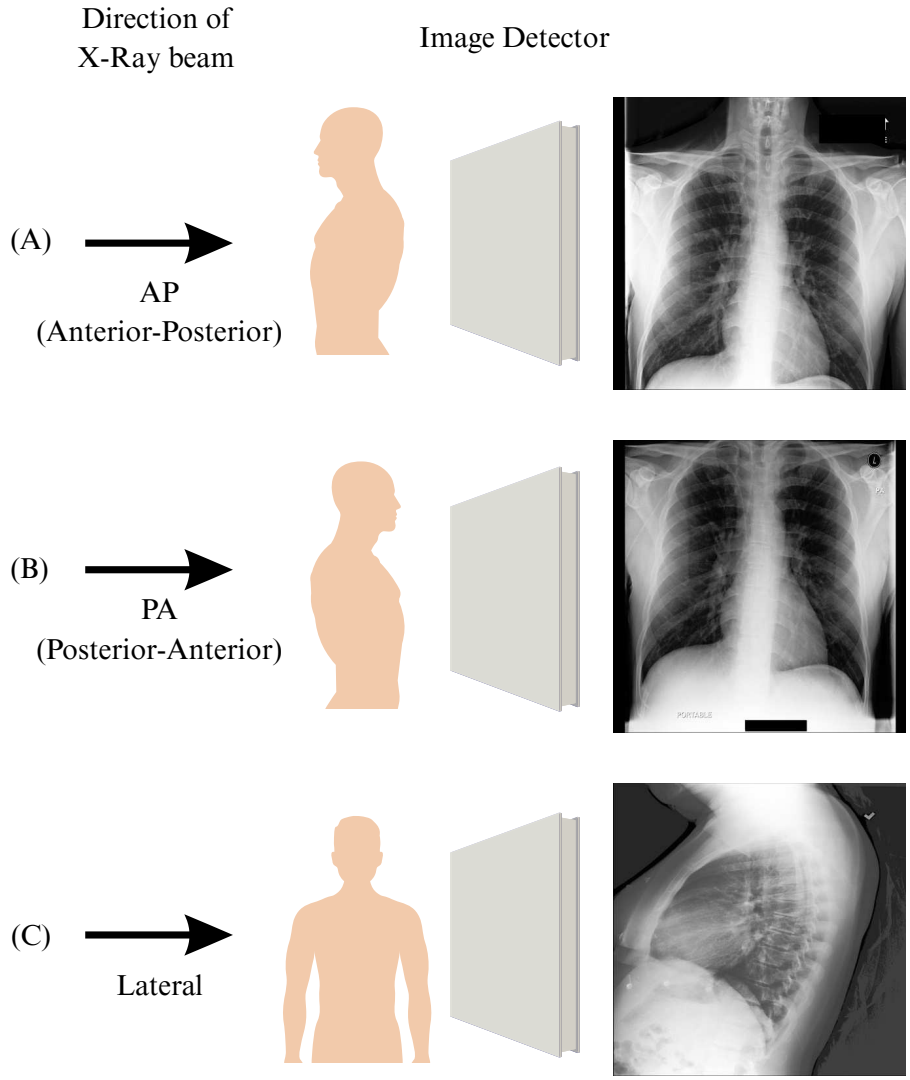


Figure 2.1: Chest X-ray imaging positions. (A) Anterior-Posterior (AP), (B) Posterior-Anterior (PA), and (C) Lateral projections, illustrating beam direction (black arrow) and image detector (gray rectangle). Each example demonstrates the resulting radiographic appearances. AP and PA views differ in beam trajectory (AP: beam enters anteriorly, exits posteriorly; PA: beam enters posteriorly, exits anteriorly), while the lateral view provides a side perspective. Source: adapted from [38].

Chest radiographs represent one of the most frequently requested examinations in radiology departments, generating substantial datasets that facilitate the development of artificial intelligence (AI) algorithms to enhance healthcare imaging analysis. Several key applications have emerged where chest X-ray AI algorithms demonstrate particular utility: (a) patient triage optimization, where AI systems analyze images to prioritize cases requiring urgent attention; (b) automated pneumothorax detection, enabling rapid identification of this potentially life-threatening condition characterized by air accumulation in the pleural cavity; and (c) COVID-19 diagnosis, where AI has shown promise in assist-

ing radiologists to differentiate positive cases [32]. These applications share a common foundation: the AI algorithm must effectively process chest X-ray images to perform the intended task.

2.1.2 Vision Encoder

A Vision Encoder, also referred to as an Image Encoder in this work, is a computational component capable of transforming an image into a latent vector representation in a different dimensional space, as illustrated in Figure 2.2. It performs the mathematical operations necessary to convert a multi-dimensional image into a compact vector representation that can be processed by machine learning architectures, conceptually similar to how the human visual system converts visual stimuli into neural signals for brain processing.

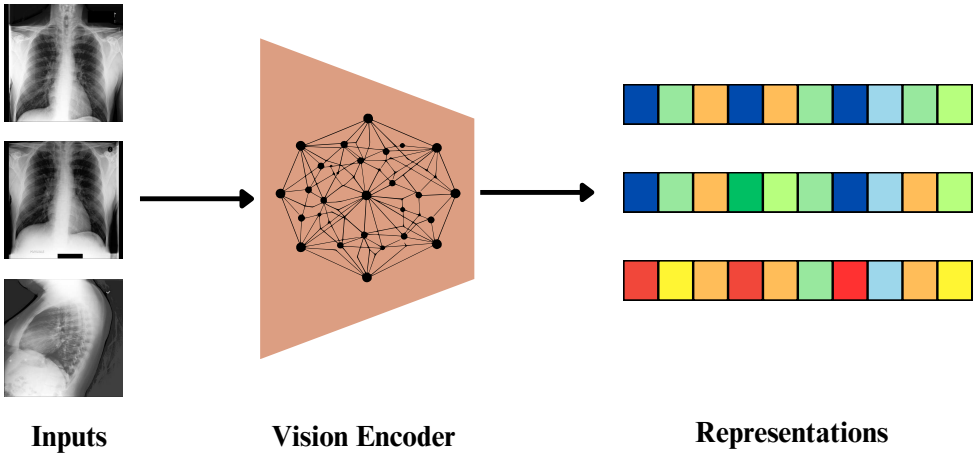


Figure 2.2: Illustration of a vision encoder's feature extraction process. The diagram shows three chest X-ray images being transformed into corresponding representation vectors. Source: elaborated by this author.

For similar X-rays, the vision encoder must produce feature vectors with similar color patterns, indicating the encoder's ability to map visual similarities to proximity in the feature space. Each vector may encode anatomical structures and potential abnormalities present in the respective X-ray, creating a valuable mathematical representation.

This transformation is crucial for extracting the most relevant features from the image, enabling further processing for computer vision tasks, such as classifying or detecting objects, and furthermore, enabling other domains applications, such as image caption generation, or, in our case, chest X-ray description generation. The extracted features must capture essential information about the image content while discarding redundant or irrelevant details [5].

Currently, neural networks, particularly Convolutional Neural Networks (CNNs), are the most common type of vision encoders in the literature. CNNs perform a sequence of matrix transformations in the input image to extract latent feature vectors, which are then used to solve specific tasks such as image classification, object detection, and semantic segmentation [43]. The hierarchical structure of CNNs allows them to capture

both low-level and high-level features, making them effective at representing complex visual patterns.

The compressed latent features extracted by the vision encoder can be integrated into more complex architectures, such as those used in multimodal contexts [28], that is, architectures where the input lies in a vision domain (e.g., X-rays) and the output lies in the text domain (e.g., X-ray description). In a typical setup, the latent features serve as input to a language model, which generates text based on the visual information provided by the encoder. This integration of vision and language modalities has been the focus of numerous research efforts, including image captioning, visual question answering, and cross-modal retrieval [3, 47].

As vision encoders continue to evolve, there is a growing interest in developing lightweight and efficient models that can be deployed on resource-constrained devices, which is particularly relevant for medical imaging applications in clinical settings with limited computational resources. Techniques such as model compression, quantization, and knowledge distillation are being explored to reduce the computational requirements of vision encoders while maintaining their diagnostic performance [11].

In the context of this work, the vision encoder plays a vital role in extracting meaningful representations from chest X-ray images. These representations serve as the foundation for generating accurate and clinically relevant captions. We aim to capture the most informative features from the images, further enabling the language model to generate descriptions that accurately reflect the visual content.

2.1.3 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a specialized type of neural network designed for image processing. They use convolutional layers to automatically learn and detect features from the images and have been highly successful in a wide range of applications in computer vision such as image recognition, object detection, and other vision tasks.

As illustrated in Figure 2.3, CNNs are composed of multiple layers, and the core of a CNN is the convolutional layer. This layer is responsible for scanning the input image with small filters (also called kernels) to detect patterns or features, such as edges, corners, and textures. The result of this operation is called a feature map. Furthermore, non-linear functions, such as ReLU (Rectified Linear Unit), are applied to the outputs of the convolutional layers and fully connected layers, introducing non-linearity into the network, allowing it to learn complex relationships in the data.

After each convolution layer, there are typically pooling layers responsible for reducing the spatial dimensions of the feature maps, which helps to reduce the computational load and makes the network more robust to variations in the position of the features.

Towards the end of the network, the feature maps begin to contain more and more abstract and latent features, i.e., compressed information. Generally, at the very end of the network, fully connected layers are used to make predictions based on the learned features. These layers learn how to map the extracted features to the desired output, such as class labels for image classification tasks, for example.

The training process of a CNN involves adjusting the weights of its layers to minimize

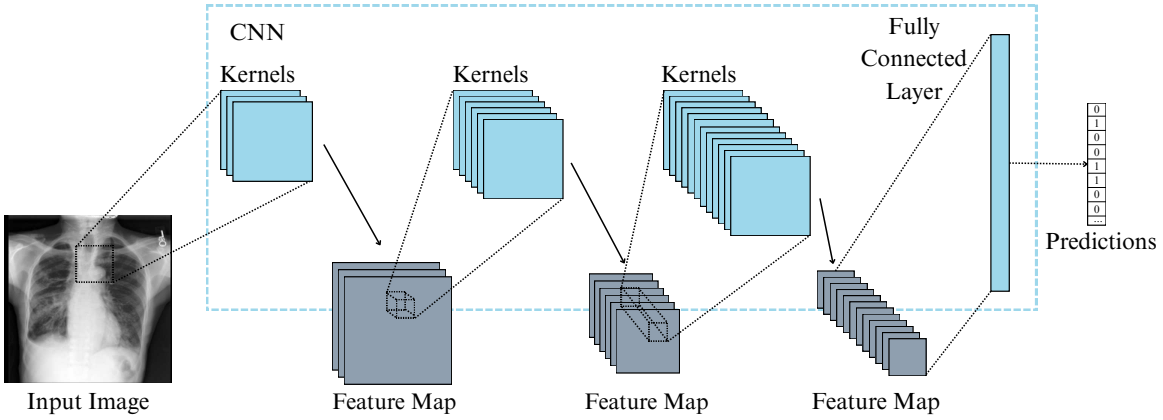


Figure 2.3: Example of a Convolutional Neural Network on a classification task context. The figure depicts the convolutional layers (also called kernels), the sequential features maps extracted from the input image, and a fully connected layer with outputs the predictions. Source: elaborated by this author.

the difference between its output predictions and ground truth. Through this process, the network gradually learns to recognize simple, abstract and complex features in images, such as shapes, textures, colors, and objects.

In the literature, several landmark CNN architectures have made significant contributions to the field of computer vision:

- LeNet (1998) [42] was one of the first successful CNNs, used for handwritten digit recognition.
- AlexNet (2012) [41] was a breakthrough model that significantly outperformed traditional computer vision techniques on the ImageNet challenge. It had a deeper architecture compared to previous models.
- VGGNet (2014) [65] further increased the depth of CNNs, showing that network depth is a critical component for good performance.
- GoogLeNet (2014) [68] introduced the Inception module, which performs convolutions with multiple filter sizes in parallel and concatenates the results, allowing the network to capture details at various scales.
- ResNet (2015) [29] introduced residual connections, allowing training of extremely deep networks (hundreds of layers) without suffering from vanishing gradients. This enabled even higher performance on various benchmarks.

These architectures form the foundation of modern CNNs and many state-of-the-art models used today build upon these ideas. One of the primary uses of CNNs is as image features extractor, working as a bridge between image domain and other domains. These extracted features can be applied in diverse scenarios, ranging from tasks such as autonomous driving to image captioning.

In our work, we leverage pretrained CNNs as feature extractors in our vision encoder. By using models pretrained on large datasets such as ImageNet or even domain-specific

medical imaging datasets, we can benefit from their learned feature hierarchies without having to train from scratch. The CNN takes the chest X-ray as input and outputs a compact feature representation that captures the salient information. This feature vector is then passed to the next stages of our captioning pipeline, namely the alignment module and language model, for further processing to generate the textual description. Employing powerful CNN feature extractors allows our model to build upon state-of-the-art computer vision techniques and focus on the captioning task.

2.1.4 Transformers

The Transformer architecture [74] is a deep learning model designed for sequence-to-sequence tasks, i.e., tasks where the input is a sequence of data (such as a sentence or a time series) and the output is also a sequence (such as a translated sentence or a predicted series). Its core innovation lies in the self-attention mechanism that allows each element in a sequence to focus on other elements. This enables the model to capture complex dependencies, even among distant elements, making it exceptionally well-suited for a wide range of tasks.

At the heart of the Transformer is the self-attention mechanism. In self-attention, each element in the input sequence computes attention scores with respect to all other elements. These scores determine how much each element should be addressed or focused on every other element. This is achieved through a compatibility function that computes the dot product between linearly projected versions of the elements. The attention scores are then normalized using a softmax function, ensuring they sum to one, and used to compute weighted averages of the value projections of the elements. This allows each element to incorporate information from the entire sequence, weighted by relevance.

Multi-head attention extends this idea by performing multiple self-attention operations in parallel, each with different linear projections. This allows the model to jointly attend to information from different representation subspaces, enhancing its expressiveness.

The Transformer architecture typically consists of an encoder and a decoder, each being a stack of self-attention and feed-forward layers. The encoder processes the input sequence and generates a contextualized representation for each element. The decoder then takes these representations, along with the output sequence generated so far, to predict the next element in an autoregressive manner. Positional encodings are added to the input embeddings to inject information about the relative or absolute position of the elements in the sequence.

The initial breakthrough of the Transformer architecture occurred in natural language processing (NLP). It has since become the foundation for various NLP applications, including machine translation, text generation, question-answering, and text classification. Pre-trained models such as BERT [20] and GPT-3 [8], have redefined the state of the art (SOTA) in NLP. These models are often fine-tuned for specific tasks, making them versatile and powerful tools for a wide array of NLP challenges.

In recent years, transformer-based architectures have emerged as state-of-the-art (SOTA) models for various computer vision tasks [27]. Vision transformers, such as the Vision Transformer (ViT) [21] and the Swin Transformer [48], have demonstrated remarkable

performance in image classification, object detection, and semantic segmentation. The Transformer [74] architecture revolutionized the way visual information is processed by shifting from convolutional operations to the introduced self-attention mechanism. The self-attention mechanism enables models to weight the importance of input features regardless of spatial distance, allowing for global feature relationships that were difficult to capture with traditional convolutional approaches. Furthermore, this mechanism allows for more efficient computation and scalability to large-scale datasets [21].

The Vision Transformer (ViT) framework, shown in Figure 2.4, considers an image as a grid of patches, each functioning as an 'image-word' token. The self-attention mechanism is applied to these patches, allowing the model to capture intricate relationships between different parts of an image. This adaptability has led to remarkable success in image classification and object detection tasks.

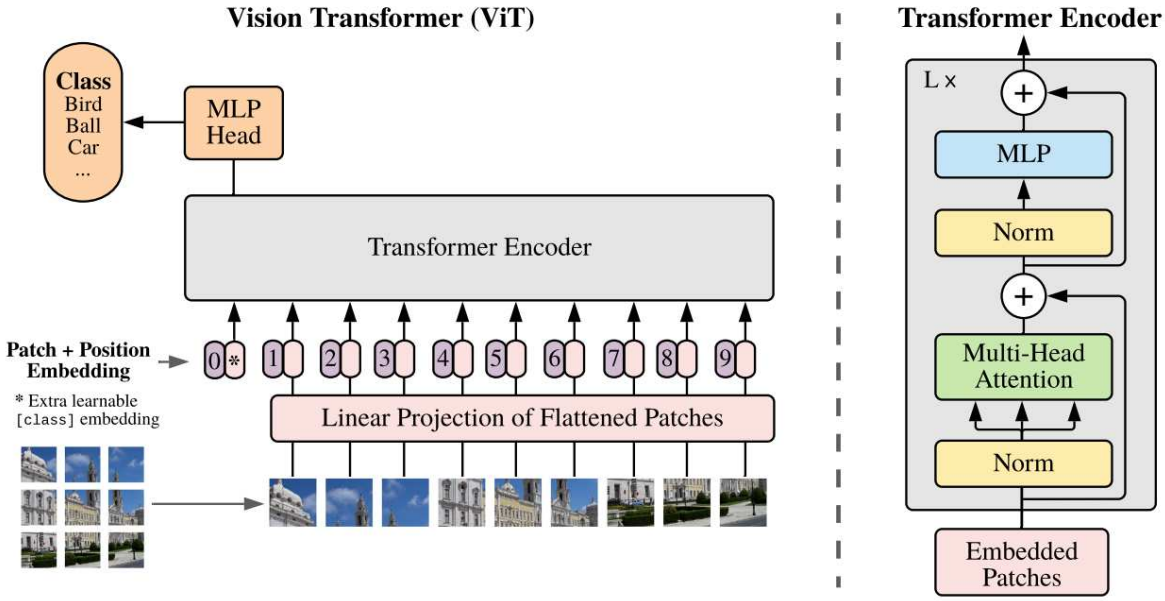


Figure 2.4: Overview of ViT model architecture. Source: [21].

Beyond the initial Transformer model, numerous variants and extensions have been proposed. For example, the Bidirectional Encoder Representations from Transformers (BERT) model introduced a masked language modeling pretraining objective, allowing the model to incorporate bidirectional context. The Generative Pre-trained Transformer (GPT) models focused on language generation and demonstrated remarkable few-shot learning capabilities. In the vision domain, the Swin Transformer introduced a hierarchical architecture with shifted windows for more efficient image processing.

In our work, we explore the use of Transformers in both alignment and language model components of our chest X-ray captioning pipeline. For the alignment component, the Q-Former module uses cross-attention to determine which visual features are most relevant for each word in the caption. This allows the model to ground the generated text in the visual content of the image.

For the language model, we leverage pretrained Transformer-based models such as GPT. These models, trained in vast amounts of text data, have the ability to generate

coherent and fluent descriptions. By conditioning the language model on the features extracted by the vision encoder, we aim to generate clinically accurate and relevant captions for the chest X-rays.

Overall, the Transformer architecture and its attention mechanisms have revolutionized both natural language processing and computer vision. By leveraging their power in our chest X-ray captioning model, we aim to capture complex dependencies within and between the visual and textual modalities, enabling the generation of high-quality, clinically relevant captions.

2.1.5 Large Language Models

Large Language Models (LLMs), are a class of deep learning models designed for processing and generating human language text. They have gained immense prominence due to their ability to understand and generate natural language text with remarkable fluency and context-awareness. LLMs are being used mainly for chatbots as ChatGPT, Claude [15] and DeepSeek [26].

The key innovation within LLMs is their massive scale, characterized by a substantial number of model parameters. For instance, models such as GPT-3 [8], or PaLM [14], consist of hundreds of billions of parameters, enabling them to capture intricate linguistic patterns.

These models are pre-trained on large text corpora, such as books, articles, and websites. During this phase, they learn to predict the next word in a sentence. Subsequently, LLMs are fine-tuned on specific tasks or domains, allowing them to excel in a wide range of applications.

LLMs possess both generative and discriminative capabilities. They can generate human-like text, making them suitable for tasks such as text generation, dialogue systems, and creative writing.

A remarkable feature of LLMs is their ability to perform zero-shot and few-shot learning [8, 40]. This means they can make predictions or generate text on tasks they were not explicitly trained on, given a few examples as context. This adaptability broadens their utility across various domains. LLMs have transformed numerous domains, from natural language understanding and generation to code generation, summarization, translation, and more.

In our work, we explore the integration of LLMs into our chest X-ray captioning pipeline. Specifically, we investigate the use of BioGPT, a domain-specific LLM pretrained on biomedical text. By leveraging the knowledge captured in BioGPT, we aim to generate clinically relevant and accurate captions for chest X-rays.

The LLM serves as the language model component of our pipeline, taking the visual features extracted by the vision encoder as input and generating a textual description of the chest X-ray. The attention mechanism in the LLM allows it to dynamically focus on different parts of the visual representation as it generates each word in the caption.

However, the integration of LLMs also presents challenges. The computational cost of running LLMs, especially during inference, can be high. To mitigate this, we explore techniques for efficient inference, such as using smaller LLMs (for instance, BioGPT) and

optimizing the pipeline for faster execution. Additionally, ensuring the clinical accuracy and relevance of the generated captions is crucial. We address this by incorporating domain-specific pretraining and fine-tuning, as well as leveraging insights from medical experts during the development and evaluation phases.

2.1.6 Pretraining and Transfer Learning

Pretraining and transfer learning are two fundamental concepts in deep learning, particularly in the domains of computer vision and natural language processing. These techniques allow AI models to leverage prior knowledge gained from large-scale generic datasets and apply it to specific tasks with generally smaller amounts of specialized data to be trained on, enabling more efficient and effective learning [31].

Pretraining involves training a model on a large dataset to capture general knowledge and learn meaningful representations of the data. The goal of pretraining is to enable the model to understand the underlying patterns, structures, and relationships within the data domain [5]. For example, in the case of language models such as BioGPT, pretraining is done on vast amounts of biomedical text data (e.g., PubMed[1]), allowing the model to learn the intricacies of biomedical language, such as terminology, grammar, and context [50], as illustrated in Figure 2.5.

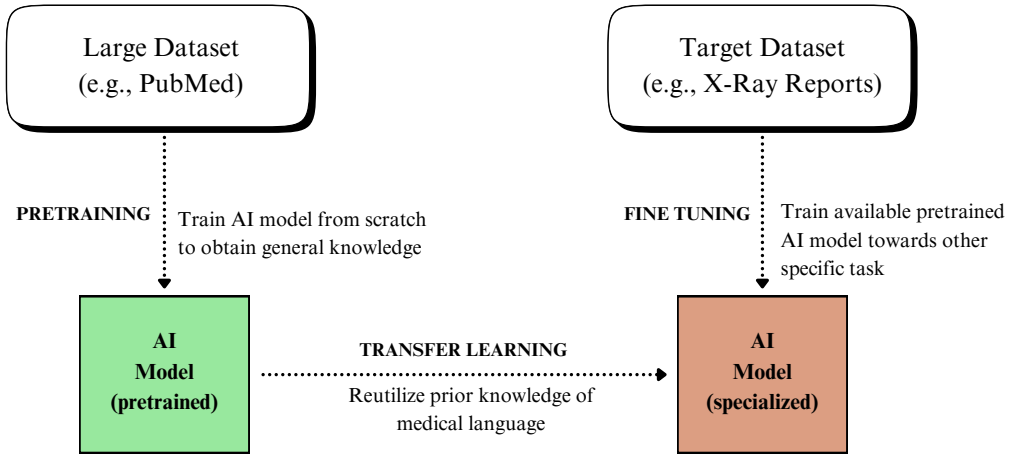


Figure 2.5: Example of pretraining and transfer learning. Initially, the model is pretrained on a large medical corpora to learn general medical language representations. Subsequently, transfer learning is applied, where the model is fine-tuned on a smaller, domain-specific dataset of X-ray reports to optimize performance for the target task. Source: elaborated by this author.

During pretraining, the model is typically trained on unsupervised or self-supervised tasks. In the case of language models, common pretraining tasks include next word prediction (predicting the next word given a sequence of words) and masked language modeling (predicting missing words in a sentence) [20]. These tasks encourage the model

[1] `https://pubmed.ncbi.nlm.nih.gov`

to learn robust representations of the language that can be used for various downstream tasks.

Similarly, in computer vision, models can be pretrained on large-scale image datasets such as ImageNet [19]. Pretraining tasks for vision models often involve classification, where the model learns to predict the object categories present in the images. By training on a diverse set of images, the model learns to extract meaningful visual features and patterns that can be generalized to other vision tasks [29].

Once a model has been pretrained, it can be fine-tuned for specific tasks (for instance, write X-ray reports) through transfer learning. Transfer learning involves taking a pretrained model and adapting it to a new task or domain by leveraging the knowledge it has already acquired [61]. Instead of training a model from scratch, which can be time-consuming and require large amounts of labeled data, transfer learning allows us to build upon the existing knowledge and refine it for the specific task at hand.

In transfer learning, the pretrained model's architecture might be modified by replacing or adding task-specific layers. For example, in the case of fine-tuning a pretrained language model for a text classification task, the final output layer is replaced with a new layer that corresponds to the number of classes in the classification problem [30]. The pretrained weights of the model are used as initialization, and the model is further trained on the task-specific dataset.

Transfer learning offers several advantages. Firstly, it reduces the need for large amounts of labeled data for the target task. Since the pretrained model has already learned meaningful representations, it can generalize well to the new task with limited training data. This is particularly beneficial in domains where labeled data is scarce or expensive to obtain, such as in medical imaging. Secondly, transfer learning speeds up the training process and reduces computational requirements by starting from informative pretrained weights rather than random initializations. Thirdly, it can lead to improved performance on the target task by capturing more robust and generalizable features.

In our work, we leverage these benefits by using pretrained models for both the vision encoder and language model components while keeping their weights frozen during training. For the vision encoder, we explore two architectures: U-Net pretrained in a chest X-ray segmentation dataset [63] and PSPNet pretrained in a different set [16]. We specifically chose models pretrained in segmentation tasks because they develop a deep understanding of spatial and structural information in medical images, including the ability to delineate different anatomical regions.

For the language component, we employ BioGPT, which has been pretrained on a large corpus of biomedical text [50], enabling it to understand medical terminology, abbreviations, and common phrases used in clinical reports. By combining these specialized pretrained models, our approach can effectively integrate visual and textual information to generate informative chest X-ray reports without requiring extensive labeled training data. It is important to note that this approach's effectiveness depends on the similarity between the pretraining tasks and our target task. In our case, the close relationship between chest X-ray segmentation and report generation provides a strong foundation for transfer learning.

In conclusion, pretraining and transfer learning are powerful techniques that enable

the development of more efficient and effective AI models. By leveraging knowledge from large-scale datasets and adapting it to specific tasks, these techniques have revolutionized the field of deep learning and opened up new possibilities for solving complex problems with limited labeled data. In our work, we mainly harness the power of pretrained models to create a model capable of generating meaningful and clinically relevant chest X-ray reports.

2.2 Related Work

Our research focused on developing a chest X-ray captioning model that incorporates LLMs while emphasizing a lightweight architecture. Therefore, this literature review focuses on three key areas: Image Captioning, Medical Captioning, and Large Language Models. In the following sections, we will explore these areas to gain insights into the state-of-the-art techniques and advancements.

We start by examining the landscape of Image Captioning, mainly focusing on different approaches and architectures designed to solve the task. Next, we make a comprehensive exploration of Medical Captioning, where we focus on finding specific techniques utilized in the medical domain that may help us better grasp the problem, find benchmarks and understand the most used metrics. Lastly, we investigate the Large Language Models, trying to identify the state-of-the-art models, the number of parameters and hardware needed to train, as well as to identify LLMs already pretrained in the medical domain.

2.2.1 Image Captioning

The image captioning task targets the automatic generation of description of natural images. In the literature, we find many methods designed to achieve this goal [46, 47, 59, 75, 81, 82]. These works vary based on: their selection of vision encoder, language model, and their approach to cross-domain alignment, which involves aligning features between the image and text domains.

Vinyals et al. [75] and Rennie et al. [59] used CNN models to extract the features from images and Long Short-Term Memory (LSTM) network to regressively generate the desired text. The CNN features provide a simple and compact representation of an image, but they can hinder further fine-grained description due to this compression. The LSTM network, due to its sequential nature, can be very slow to output the text.

Zhou et al. [81] and Li et al. [47] leveraged the potential of the Transformer [74] along region image features extracted by Fast-RCNN [22], an widely used CNN for feature extraction. Zhou et al. [81], for example, used a shared multi-layer Transformer responsible for both vision encoding and text decoding steps, hence this unique module is responsible for aligning the image features and generating the caption.

Very recent works [46, 82] take advantage of the few shot potential of LLMs [8]. Zhu et al. [82] introduced MiniGPT-4, which combines a Vision Transformer image encoder with an open-source LLM based on Llama [72], aligning both image and text domain using solely a single projection layer. The authors named it MiniGPT-4 due to its similar capability in description generation compared to GPT-4.

Li et al. [46] presented the BLIP-2 training framework, which introduces a BERT-based alignment module called Q-Former – Querying Transformer – responsible for aligning the features between a pre-trained vision encoder and a pre-trained text decoder. The Q-Former is called "querying" because it is designed around the concept of learnable query embeddings (Learned Queries) that actively extract relevant information from image features. A key aspect of its training involves splitting the process into two different stages. In their work, both Image Encoder and Text Decoder models are frozen, i.e., have their parameters fixed, and the Q-Former is the only module requiring training. Hence, this architecture allows adapting and pluging any vision encoder or text decoder into the pipeline, as only the alignment module requires training. As the Q-Former is a BERT-based model, it is able to process both image extracted features or caption text embeddings due to the different attention layers composition: the cross-attention layers, used to insert image features; and the self-attention layers, used to insert the ground-truth caption embeddings, as can be observed in the processing paths in Figure 2.6.

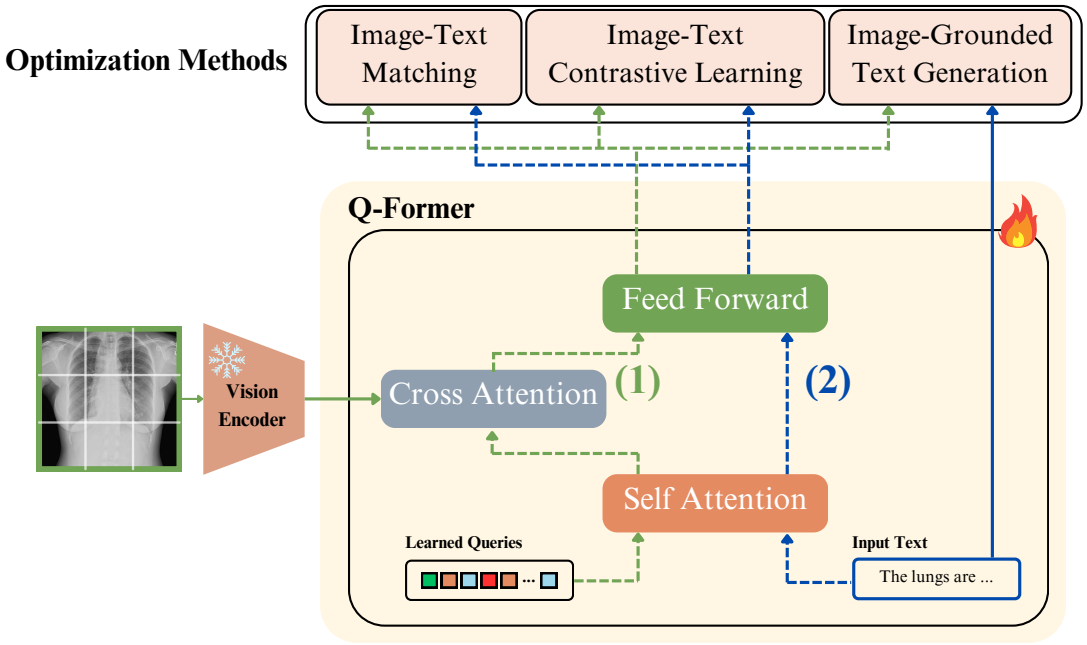


Figure 2.6: BLIP-2 first training stage - Representation Learning: optimizes Q-Former using three different optimization functions allowing the module to learn how to align features from image and text domain. Within the Q-Former architecture, data processing has 2 flows: (1) input image features path, and (2) the actual report. Each processing path will generate a feature vector in the feed forward layer output that can be used to compute the optimization methods. Source: adapted from [46].

The proposed training procedure is broken down into two stages:

1. Representation Learning - leveraging a frozen pre-trained Vision Encoder to extract features from the images, the Q-Former outputs are optimized using three different methods, as illustrated on Figure 2.6. Note that the same module is adopted for two different purposes: (1) process image features and Learned Queries; and (2) process the related text. The applied optimization methods are:

(a) Image-Text Contrastive Learning: the images and their corresponding text descriptions are processed in parallel, creating feature vectors that represent each modality (vision and text) in a shared embedding space. A contrastive loss function is then applied, which pulls related image-text pairs closer together while pushing unrelated pairs farther apart in this embedding space. This approach is intended to align visual and textual information in a common dimensional representation, enabling the model to understand relationships between images and language, and supporting multimodal applications.

(b) Image-Text Matching: it is a binary classification task to determine whether an image-text pair represents a match (positive) or a non-match (negative). Data batches are dynamically augmented to create negative pairs alongside the genuine positive pairs. This process guides the Q-Former to develop effective representations that link related visual and textual information.

(c) Image-grounded Text Generation: a BERT Language Model Head is attached to Q-Former in order to generate captions based on the input image features. A loss is computed between the generated text and the actual ground-truth caption. This process encourages the model to identify and focus on the most salient visual features that contribute to accurate and relevant caption generation.

2. Generative Learning: the prior trained Q-Former is attached to a frozen pre-trained LLM using a fully connected layer, then the whole pipeline (Figure 2.7) is trained optimizing the outputs of the LLM decoder. The fully connected layer projects the Q-Former outputs to the LLM text embedding dimension. The idea is that these projected features will serve as soft prompts to the LLM, conditioning and guiding the caption generation process.

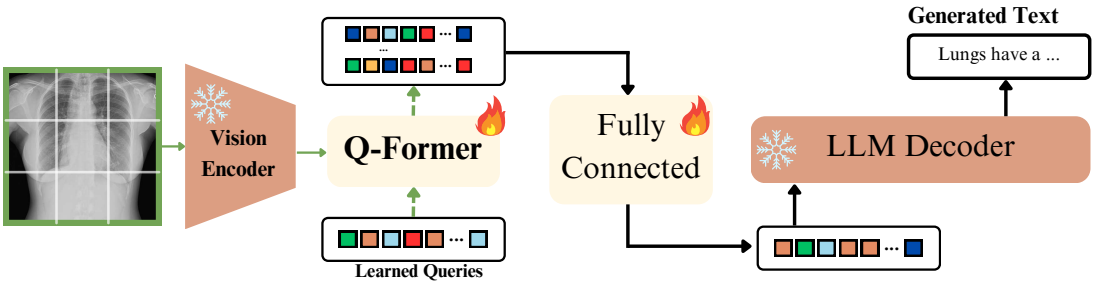


Figure 2.7: BLIP-2 second training stage - Generative Learning: optimizes Q-former and fully connected layer using the output of the LLM decoder. In this training step, the Q-Former will only process the input image features and further provide a feature vector that the LLM will use as if it was a prompt. Source: adapted from [46].

The aforementioned works have succeeded in the task of generating captions for images, however, the most recent ones adopting LLMs as text decoders, were not concerned about using lightweight models in their architecture. They have used models with billions of parameters (65 billion for MiniGPT; starting from 2.7 billion for BLIP-2) on both vision

encoder and text decoder. These models, particularly the 65-billion-parameter one, are not suitable for deployment on standard hardware due to their resource-intensive nature.

2.2.2 Medical Captioning

Medical captioning involves the automatic generation of descriptive and informative captions for medical images, such as X-rays, MRI scans, CT scans, and more. The goal is to provide accurate textual descriptions that convey relevant medical information, possibly helping medical professionals make faster and more accurate diagnoses, facilitating research and training, and ultimately improving patient care. Considerable progress has been made in this field, with numerous proposed architectures and approaches [9, 10, 17, 24, 45, 70, 77, 78].

Early research in chest X-ray captioning, including several derivative works, primarily utilized CNNs as vision encoders and Transformers as text decoders [9, 10, 24].

Chen et al. [9] introduced the use of the novel relational memory (RM) and memory-driven conditional layer normalization (MCLN) modules alongside the Transformer to enhance caption its generation power, addressing the challenge of maintaining contextual coherence across long report sequences. Building on this framework, Chen et al. [10] focused on the alignment between image and text domains, proposing a cross-modal memory network (CMN) to facilitate the interactions across these modalities. This approach barely increased the overall performance of their previous network. Furthermore, once again, these works were using Transformers as the text generator, a language model generally applied to text analysis and not text generation, as explained in Subsection 2.1.4.

Another work [24], based on the approach developed by Chen et al. [10], integrates pixel-level organ masks (e.g., bones, lungs, heart, mediastinum) extracted via the pre-trained CXAS segmentation model to enhance disease-specific region attention. The proposed Complex Organ Mask Guided (COMG) framework leverages these anatomical priors to align visual features with textual reports, addressing the common limitation of generic descriptions in radiology report generation by focusing on clinically relevant regions.

On the other hand, a more recent study [77] shifted from CNNs to Vision Transformer (ViT) as image encoders, introducing Expert Tokens designed to interact with the extracted images patches. This allowed each token to focus on different image regions for image representations and, therefore, process each patch independently to generate the output.

Recent research has pivoted toward LLMs for medical caption generation [17, 45, 70, 78], seeking to leverage their powerful text generation capabilities. These approaches can be categorized based on their architectural choices and training methodologies: general purpose LLMs with fine-tuning and specialized architectures.

In the general purpose LLMs with fine-tuning category, Yang et al. [78] adapted the BLIP-2 architecture with a large vision encoder (EVA-ViT-g) and LLM decoder (ChatGLM-6B), fine-tuning the Q-Former and LLM components for the ImageClef task. While they achieved competitive rankings, their evaluation lacked comprehensive comparison with established benchmarks, and the computational requirements of their approach

limit its accessibility for broader research and clinical applications, due to adopted models sizes. Their reliance on a general-purpose LLM, which was previously trained in Chinese and English, also raises questions about the model's ability to accurately capture medical terminology and relationships without domain-specific pretraining.

Thawkar et al. [70] combined MedCLIP's Vision Transformer [76] with Vicuna [12], a Llama-based chatbot, connecting them through a simple linear transformation layer. This approach leverages MedCLIP's medical image understanding capabilities, but employs a computationally expensive LLM without domain-specific knowledge. Furthermore, it utilizes a general-purpose chatbot rather than a model specifically designed for medical caption generation. The absence of comparative evaluations also makes it difficult to assess the method's actual contribution to the field.

In the specialized architectures category, Danu et al. [17] proposed a two-stage approach for medical caption generation. Their method first detects diseases with bounding boxes in X-rays before feeding these features to an LLM as a textual prompt. This approach introduces interpretability through localization and decouples cross-modality by separating visual detection from text generation. Additionally, it leverages an in-domain LLM, RadBloomz [36], to translate the list of abnormalities into coherent captions. However, this method relies on a very large LLM and provides limited evaluation metrics, focusing primarily on ROUGE-L scores while omitting the CE metrics commonly used in comparative studies.

Additionally, Li et al. [45] generated anatomically structured reports by decomposing the task into region-specific descriptions (e.g., heart, lungs) guided by anatomical and clinical prompts. This mimics radiologists' systematic reasoning while allowing physician input for customization. The method outperforms predecessors in clinical metrics (CE metrics) and interpretability by grounding generated text in detected regions. However, its reliance on GPT-4 [54] – a proprietary, extremely large LLM – raises concerns about computational accessibility and reproducibility. These limitations mirror those of other LLM-based approaches (e.g., [70, 78]), highlighting an ongoing trade-off between performance and practicality in medical captioning.

It is important to note that, to the best of our knowledge, no existing architectures have leveraged the power of LLMs pre-trained on medical text for medical caption generation. Current approaches either rely on computationally intensive large-scale general LLMs or traditional encoder-decoder architectures, leaving a gap in the literature regarding efficient, domain-specific solutions.

Our work aims to address this gap by investigating the potential of domain-specific, smaller LLMs such as BioGPT [50] that have been pretrained specifically on medical literature. By combining such a model with pretrained vision encoders, we seek to develop a more computationally efficient approach that still maintains strong performance through domain relevance and effective knowledge transfer. Table 2.1 summarizes the approaches evaluated in this subsection.

Table 2.1: Systematic comparison of state-of-the-art approaches for medical captioning, highlighting the evolutionary trajectory from early transformer-based models to recent adaptations of large language models (LLMs). Prior work has predominantly focused on either using complex vision encoders or leveraging large parameter models. Our approach adopts a lightweight (0.35B parameters) LLM and utilizes available pretrained in-domain weights for both vision and text models. This strategic balance between model complexity and domain adaptation addresses a critical research gap and offers a more resource-efficient solution for clinical deployment in resource-constrained healthcare environments. * Text decoder based on Transformer [74] with some adaptations, hence we assume the number of parameters of the BERT base Transformer model.

Work	**Year**	**Text Decoder**	**Number of Parameters**	**In-Domain Weights**
Chen et al. [9]	2020	Transformer* - Encoder-Decoder	0.07 B	No
Chen et al. [10]	2022	Transformer* - Encoder-Decoder	0.07 B	No
Gu et al. [24]	2024	Transformer* - Encoder-Decoder	0.07 B	Vision
Wang et al. [77]	2023	Transformer* - Encoder-Decoder	0.07 B	No
Yang et al. [78]	2023	LLM - Decoder-only	6B	No
Thawkar et al. [70]	2023	LLM - Decoder-only	7B	Vision
Danu et al. [17]	2023	LLM - Decoder-only	7B1	Text
Li et al. [45]	2024	LLM - Decoder-only	7 B	No
Ours		LLM - Decoder-only	0.35 B	Vision + Text

Chapter 3

Materials and Methods

This chapter is intended to describe (i) the adopted datasets on training and evaluation steps, (ii) the proposed architecture for generating chest X-ray descriptions along with its training steps, (iii) the neural networks subject to evaluation on vision encoder and text decoder stages, and (iv) the metrics for analysis of the generated texts.

3.1 Datasets

This work adopts two widely-used datasets for the chest X-ray description generation task: the MIMIC-CXR-JPG [35] and the IU-RAY dataset [18].

3.1.1 MIMIC-CXR-JPG

The MIMIC-CXR-JPG dataset [35] is currently the largest chest X-ray dataset, comprising 377,100 radiographs and 227,835 associated free-text radiology reports collected from the emergency department of the Beth Israel Deaconess Medical Center (Boston, MA, USA) between 2011 and 2016. The dataset is hosted on PhysioNet[1], a repository for biomedical data, ensuring accessibility and standardization for research use. Each study in the dataset may include multiple radiographic views (e.g., anteroposterior (AP), posteroanterior (PA), and lateral projections), accompanied by a single report. The dataset provides an official split of studies into training, validation, and test subsets, facilitating reproducible benchmarking of model performance against the existing literature.

Unlike structured radiology reports, which follow a standardized format (e.g., from external to internal findings [39]), the reports in MIMIC-CXR-JPG were generated by emergency department physicians and exhibit greater variability in style and detail. A data sample is illustrated in Figure 3.1 containing a free-text report associated with a frontal and lateral X-ray. Although they are free-text emergency reports, they generally present common text sections as the Findings and Impressions. The former is intended to be a descriptive assessment of radiographic observations, and the later is a concise summary or diagnostic interpretation. In this work, we focus on automatically generating the Findings section, as it provides the most comprehensive clinical details.

[1]MIMIC-CXR-JPG webpage: `https://physionet.org/`

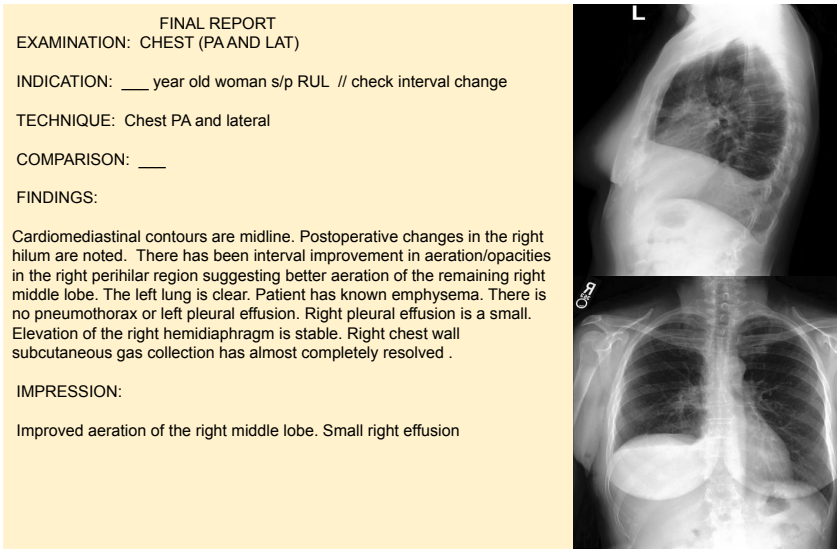


Figure 3.1: Example from the MIMIC-CXR-JPG dataset: A study comprising two chest X-ray views (AP and lateral) and the associated free-text radiology report. Emergency department reports typically include Findings (detailed observations) and Impressions (diagnostic summary). Note the presence of de-identification artifacts (e.g., "____"), which do not affect the clinical relevance of the text.

Complementary to the image data, the dataset includes two metadata files: the DICOM metadata ("mimic-cxr-2.0.0-metadata.csv.gz") and split metadata ("mimic-cxr-2.0.0-split.csv.gz"). The first contains essential metadata for each image, including view position, patient orientation, image dimensions, and identification information, while the second provides the official dataset partitioning into training, validation, and test subsets.

In this work, we conducted a comparative analysis of model performance using single-view imaging (frontal images: AP or PA) versus multi-view imaging (combining lateral and frontal views) to determine whether the inclusion of multiple perspectives enhances report generation accuracy and provides more comprehensive details in the resulting radiology reports.

3.1.2 IU X-Ray

The Indiana University Chest X-Ray Collection (IU X-Ray) [18] is a publicly accessible dataset developed for medical chest X-ray report generation research. Hosted by the National Library of Medicine[2], this collection comprises 3,996 radiology reports associated with 8,121 images, capturing both frontal and lateral radiological views. Each report is provided in a structured XML format containing metadata, associated image identifiers, and clinical text that may include findings, impressions, or both sections (as illustrated in Figure 3.2)

Since this dataset lacks an official dataset split, we implemented a random split using a 7:1:2 ratio for training, validation, and testing respectively, following established protocols in the literature [9, 10].

[2]Available at `http://openi.nlm.nih.gov/`

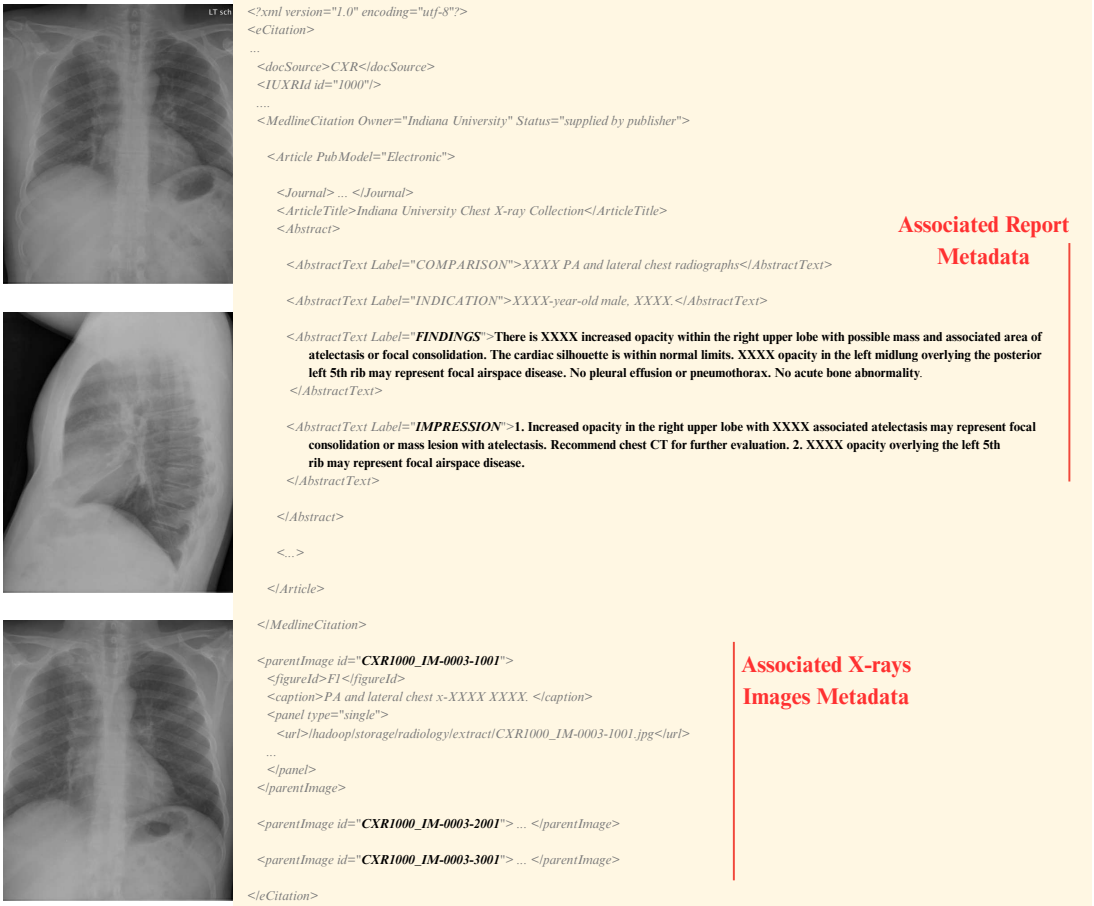


Figure 3.2: Example from the IU X-Ray dataset demonstrating data structure. The image shows three X-ray views (two frontal, one lateral) alongside their associated XML file. The XML structure reveals both metadata for the linked images and distinct sections of the actual radiological report. Note the privacy protection mechanism where confidential patient information has been systematically de-identified using 'X' characters as placeholders. Source: elaborated by this author.

Despite its relatively modest size, this dataset proved valuable during our initial development phase, allowing us to efficiently validate the architectural implementation, verify functionality, and confirm that our training methods were correctly executed before scaling to larger datasets.

3.2 Data Preprocessing

Data preprocessing plays a critical role in machine learning applications [51]. For our chest X-ray analysis system, we implemented a preprocessing pipeline to ensure data quality and consistency across both textual reports and chest X-ray images. This section details our methodical approach to preparing the data, which addresses challenges specific to medical text and imaging data while optimizing for computational efficiency during model training. Figure 3.3 provides a visual overview of our complete preprocessing workflow.

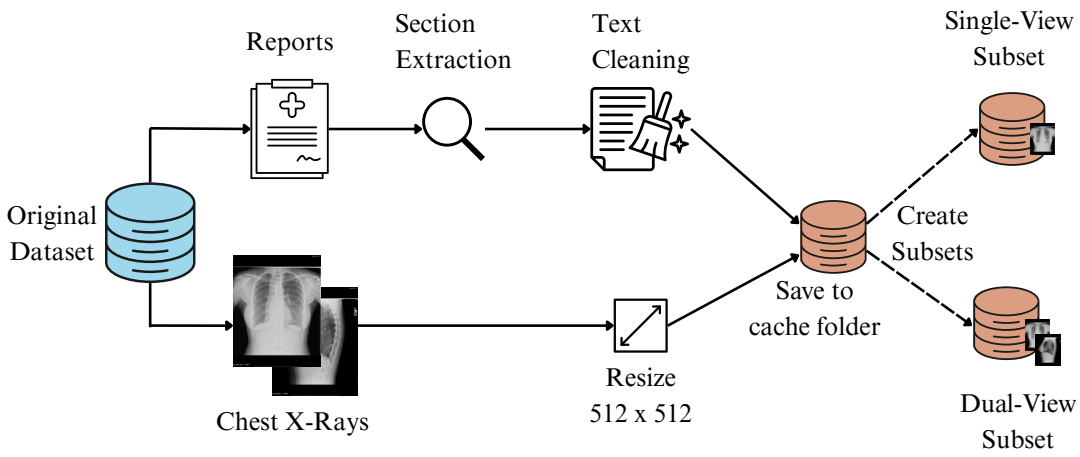


Figure 3.3: Illustration of the data preprocessing pipeline for chest X-ray datasets. The workflow shows parallel processing streams for textual reports (upper path) and radiological images (lower path). Source: elaborated by this author.

3.2.1 Text Reports Processing

Following other works in the area, we focused on the Findings section of the radiology reports, as this section contains the most detailed descriptions of the chest X-rays. When the Findings section was unavailable, we used the Impressions section as an alternative.

Our approach is focused on text generation, therefore, unlike traditional NLP tasks [69], instead of standard NLP preprocessing techniques, we implemented a customized pipeline to clean the free-text reports as much as possible, while preserving the core content of the text. The text preprocessing pipeline consisted of the following steps:

- **1. Section Extraction**: We extracted the Findings section (or Impressions section when Findings was unavailable) using regular expression pattern matching to identify section headers. For the IU X-Ray dataset, we performed a search inside the XML structured file, in order to determine and extract the right sections.
- **2. Text Cleaning**
  - Removed line breaks to create continuous text.
  - Eliminated numbers and special characters not relevant to clinical descriptions.
  - Removed multiple white spaces and standardized spacing.
  - Converted all text to lowercase to ensure consistency.
- **3. Data Caching**: The processed text was saved to local files to eliminate redundant preprocessing during training iterations.

3.2.2 Chest X-Ray Processing

For the chest X-ray images, we applied a standardized preprocessing approach:

- **1. Resizing**: All images were resized to 512x512 pixels, striking a balance between preserving clinical details and computational efficiency.

- **2. Data Caching**: Processed images were stored in cached local files to mitigate runtime overhead during repeated access operations.

3.2.3 Dataset Partition

After completing all preprocessing stages, we divided our dataset into two subsets to enable targeted analysis:

Single-view Subset: This subset of the MIMIC-CXR-JPG consists of only frontal chest X-rays (either AP or PA projections) selected from the MIMIC-CXR-JPG dataset. We created this subset to evaluate our model's performance when limited to a single viewpoint, which represents the most common clinical scenario in many healthcare settings. This process yielded 218,139 images and associated reports, representing approximately 96% of the original dataset studies.

Dual-view Subset: This subset consists of paired images for each patient, always including one frontal view (AP or PA) and one lateral view. This configuration allows us to assess how the additional perspective from the lateral view affects the quality and comprehensiveness of the generated descriptions.

These subsets were properly created using the DICOM metadata by filtering the dataset based on the 'ViewPosition' field in the metadata, filtering the images labeled as 'AP', 'PA' or 'Lateral'.

3.2.4 Benefits of Preprocessing

This preprocessing pipeline ensures that the input data is clean, consistent, and ready for training. By focusing on the Findings or Impressions sections, we prioritize the most informative parts of the reports that contain detailed descriptions of the chest X-rays. The text cleaning steps remove unnecessary characters and formatting, allowing the model to focus on the relevant content.

Resizing the images to a fixed size of 512×512 pixels standardizes the input dimensions, which is crucial for training deep learning models. This size provides a good balance between preserving important details and keeping computational requirements manageable.

Caching the preprocessed data on local files offers several benefits. It eliminates the need to perform preprocessing steps during each training iteration, reducing the overall training time. Additionally, it ensures that the data remains consistent across different runs and allows for easier data versioning and reproducibility.

By implementing this preprocessing pipeline, we streamline the data preparation process, enhance training efficiency, and ensure that our model receives high-quality input data. This lays a solid foundation for the subsequent steps in our methodology, enabling us to focus on developing and refining our architecture for generating accurate and clinically relevant chest X-ray descriptions.

3.3 Evaluation and Metrics

Metrics are quantitative measures used to assess the performance of a model or system. In the context of artificial intelligence (AI) and machine learning, well-designed metrics serve as proxies to evaluate how effectively a model achieves its intended task. The choice of metrics is crucial as they guide the development, optimization, and comparison of different models.

In order to assess the quality of the generated reports and to compare the performance of our model against other techniques, we employed two categories of metrics: Natural Language Generation (NLG) metrics and Clinical Efficacy (CE) metrics. This approach allowed us to assess both the linguistic quality of generated reports and their clinical accuracy.

3.3.1 Natural Language Generation Metrics

In the field of NLP, some of the most widely used metrics include:

- BLEU (Bilingual Evaluation Understudy): Measures overlap between machine-generated translation and reference translations by calculating precision of n-grams. Scores range from 0 to 1, with higher scores indicating better translations. BLEU has limitations, including favoring shorter translations and not accounting for semantic equivalence [55].

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Evaluates automatic summarization by measuring overlap between generated and reference summaries. ROUGE calculates recall—the percentage of n-grams from reference summaries present in the generated summary. Variants consider different n-gram types and factors such as word order and longest common subsequences [13].

- METEOR (Metric for Evaluation of Translation with Explicit ORdering): Addresses BLEU's limitations by evaluating translation quality through exact word matching, stemming, synonym matching, and paraphrase recognition. It calculates both precision and recall, emphasizing recall, and includes penalties for poor word ordering. This incorporation of linguistic features helps identify semantic equivalences that other metrics might miss [4].

We utilized standardized implementations from an established codebase[3] to ensure reproducibility of our results.

In our specific application of chest X-ray report generation, NLG metrics present several notable limitations. First, chest X-ray findings can be expressed in multiple valid ways (e.g., "cardiac enlargement" vs. "cardiomegaly"), causing NLG metrics to undervalue semantically equivalent alternatives. Second, these metrics fail to account for the clinical importance of different findings – missing a critical pathology such as pneumothorax is weighted the same as omitting a minor observation. Finally, reference reports

[3] `https://github.com/tylin/coco-caption`

themselves may contain inconsistencies or variations in reporting style, further complicating metric interpretation. Despite these problems, we include NLG metrics to maintain comparability with previous works in the field.

3.3.2 Clinical Efficacy Metrics

The Clinical Efficacy (CE) [33] metrics aim to assess the clinical accuracy and relevance of the generated text. To calculate such metrics, an external sentence labeler is adopted to categorize sentences into predefined clinical finding categories. For example, in the context of chest X-ray reports, the CheXpert [33] labeler categorizes sentences into 14 findings, such as "Cardiomegaly", "Pleural Effusion", "Consolidation", and others. The CheXpert labeler is based on a set of rules and regular expressions that match specific phrases and patterns associated with each finding.

Building upon the CheXpert labeler, the CheXbert[67] model was developed. CheXbert is a BERT-based model trained on the same dataset as CheXpert, using the outputs of its predecessor. It serves the same purpose of categorizing sentences into the 14 clinical findings but offers several advantages over the rule-based approach. CheXbert can handle variations in language, such as typos or synonyms, and can capture more complex linguistic patterns. It provides a more robust and efficient way to label sentences compared to the rule-based CheXpert labeler. The findings categories and the labeling process are illustrated in Figure 3.4.

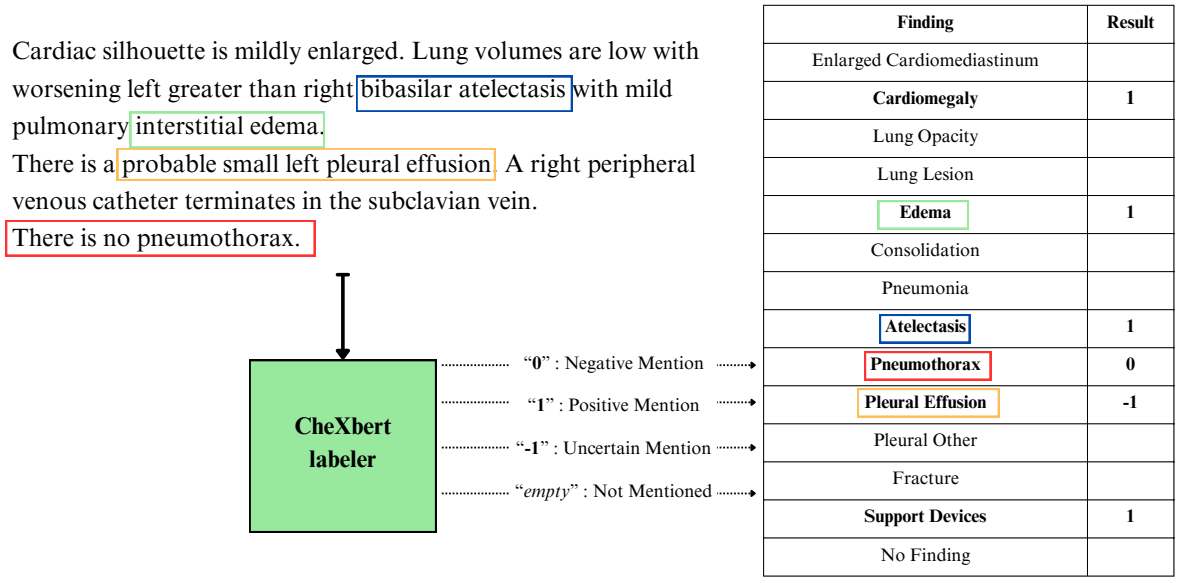


Figure 3.4: Demonstration of the CheXbert labeler's classification process. The model analyzes radiology text and outputs a 14-position vector corresponding to different clinical findings. Each position is coded as: 1 for positive mentions (Edema, highlighted in green), 0 for negative mentions (Pneumothorax, highlighted in red), -1 for uncertain mentions (Pleural Effusion), and null for findings not mentioned in the text. This example illustrates how CheXbert converts natural language descriptions into structured, machine-readable annotations, allowing further comparison between different sentences by checking accuracy, precision, recall on the labeled findings. Source: elaborated by this author.

As depicted in Figure 3.5, to assess clinical relevance and accuracy using CE metrics, we took the following steps:

1. Process both generated and reference reports through CheXbert, classifying each finding as positive (1), negative (0), uncertain (-1), or unmentioned (null)
2. Transform every null or uncertain classification into negative, in order to have strict binary vector for comparison.
3. Compute precision, recall, and F1-score by comparing these classifications

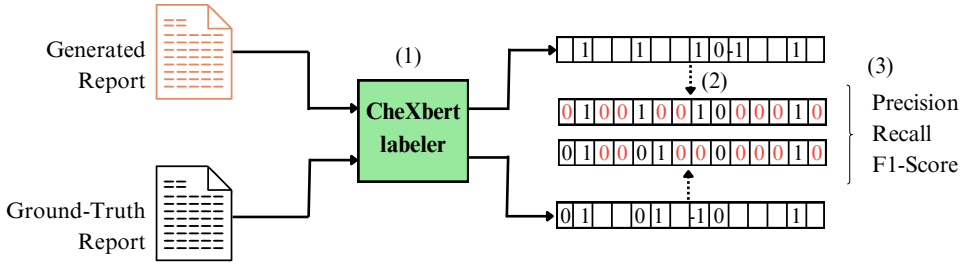


Figure 3.5: Steps to calculate CE metrics: Precision, Recall and F1-Score. (1) Classify texts into 14 findings; (2) Clean up the findings vector; (3) Calculate metrics.Source: elaborated by this author.

These CE metrics provide a more clinically relevant evaluation, as they focus on whether the model correctly identifies and reports the same findings as the reference, regardless of the specific phrasing used.

We calculated both macro-averaged metrics (equal weighting for all finding categories) and micro-averaged metrics (weighted by category frequency) to account for class imbalance in the dataset.

Despite their clinical relevance, CE metrics also have significant limitations for comprehensive evaluation. Most critically, they reduce the content of the reports to categorical classifications, losing subtle information about severity, location, and progression of findings. They cannot assess the logical flow, coherence, or professional writing quality of the generated reports, aspects that practicing radiologists value. Additionally, CE metrics depend on the quality and comprehensiveness of the CheXbert labeler itself, which may not capture emerging or rare conditions outside its 14 predefined categories.

These limitations highlight the need to develop more sophisticated evaluation frameworks that can better capture the multifaceted requirements of clinical report generation, potentially incorporating expert human evaluation to complement automated metrics. A further discussion about medical text generation metrics will be presented in Chapter 4.

3.4 Model Architecture

Our method [73] builds on the architecture and training pipeline of BLIP-2 [46], a captioning model for natural images that utilizes pretrained vision and language foundation models. BLIP-2 introduces a key innovation: the Q-Former (Querying Transformer), a

BERT-based alignment module that effectively bridges the vision encoder and the text decoder by optimizing feature representation transfer.

One of the advantages of this architecture is its modularity, allowing the integration of different vision encoders and text decoders, with only the Q-Former module requiring training. This flexibility enabled us to adapt and experiment with different vision encoders and text decoders. Complementary, as only the Q-Former must be optimized during training stages, this architecture presents fewer parameters to be fitted, making the training less computational expensive.

3.4.1 Image Encoder Stage

The architecture of our model, as illustrated in Figure 3.6, starts with the image encoder, the component responsible for extracting pertinent features from chest X-ray images. The effectiveness of this image encoder is essential to guide our architecture in generating accurate and informative captions.

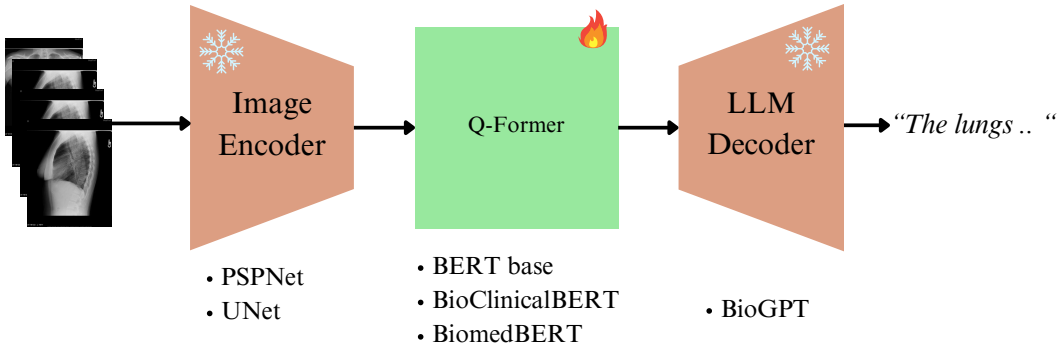


Figure 3.6: Simplified illustration of our architecture. The figure shows the three main components: the image encoder, responsible for extracting visual features from chest X-ray images; the Q-Former, responsible for transforming these visual features into a format compatible with natural language processing; and the LLM decoder, responsible for generating accurate and clinically relevant textual descriptions of the radiological findings. Each component was systematically evaluated with various neural network architectures, as indicated below their respective modules in the figure. Source: elaborated by this author.

In the original BLIP-2 study, a Vision Transformer (ViT-g/14) [57] was adopted as their vision encoder, a model containing 1.8 billion parameters. To address our objective of identifying lightweight alternatives that balance computational efficiency and performance, we deliberately selected and evaluated two significantly smaller, domain-specific pretrained vision models: (i) a PSPNet from the TorchXRayVision library [16] (66 million parameters, approximately 27 times smaller than the original ViT) and (ii) a U-Net [63] (approximately 68 million parameters, about 26 times smaller than the original ViT), implemented using code made publicly available by the authors[4].

These vision models were specifically chosen because they were previously trained on chest X-ray segmentation tasks, as illustrated in Figure 3.7, where they developed exper-

[4]https://github.com/ConstantinSeibold/ChestXRayAnatomySegmentation

tise in segmenting anatomical structures including the heart, lungs, and rib silhouettes. By adopting these pretrained vision encoders developed by other researchers, we aimed to leverage their specialized feature extraction capabilities for chest X-ray images, thereby enhancing our architecture's performance while maintaining computational efficiency.

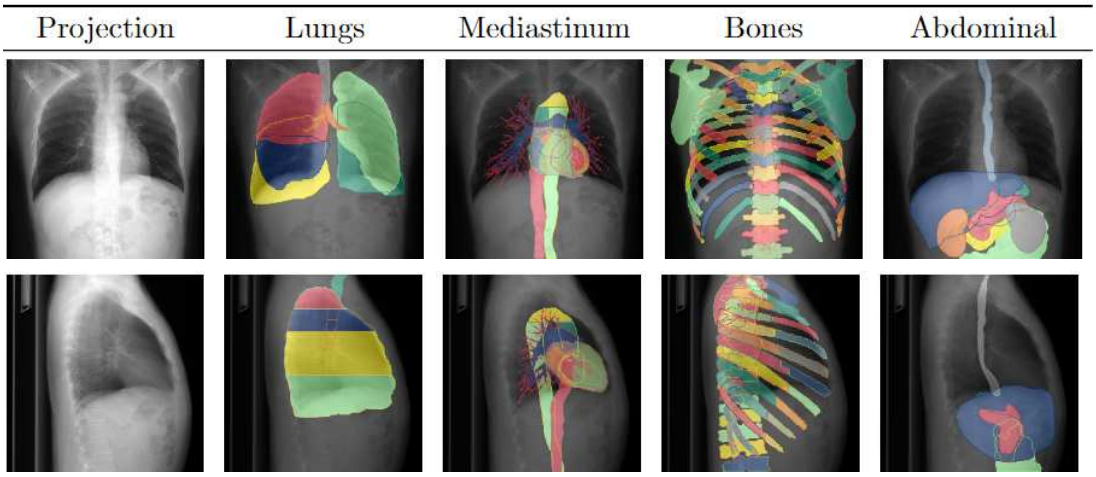


Figure 3.7: Example of a chest X-ray segmentation task for which the UNet model was previously trained. In our architecture, rather than using the segmentation output directly, we leverage the latent representations from earlier in the network to extract meaningful features for the captioning task. Source: adapted from [63].

Both vision encoder models required significant adaptations to align with our study requirements and the BLIP-2 architecture. In their original configuration, these models produce segmentation maps with dimensions matching the input image, where each pixel corresponds to a specific anatomical structure (such as lungs, ribs, heart, or other tissues). This output format is incompatible with the Q-Former module, which requires condensed feature representations. Therefore, we modified both models to extract intermediate feature maps from deeper in their architectures – before their final segmentation layers. These latent representations contain more abstract and comprehensive information about radiographic structures encoded across multiple channels, making them suitable inputs for the Q-Former attention mechanism.

For PSPNet, we utilized only the first layers up to the pyramid pooling layer, extracting a latent vector with dimensions of $128\times4\times4$. In the case of CXAS U-Net, we extracted a feature map of dimensions $64\times4\times4$ from the bottleneck layer just before the upsampling paths begin. These feature maps were then transformed using a 1×1 convolutional projection layer to achieve the dimensions required by the Q-Former attention mechanism. Figure 3.8 illustrates, using a U-Net architecture as an example, how latent features are typically extracted from the backbone of vision networks. While determining the optimal extraction point from a vision network's backbone would require further evaluation in future work, our current choices were guided by the principle of extracting latent vectors at points where information is most compressed, offering a balance between rich semantic representation and computational efficiency.

To compare both vision models and select the optimal one, we conducted a systematic evaluation. We trained the complete architecture with each vision encoder variant and

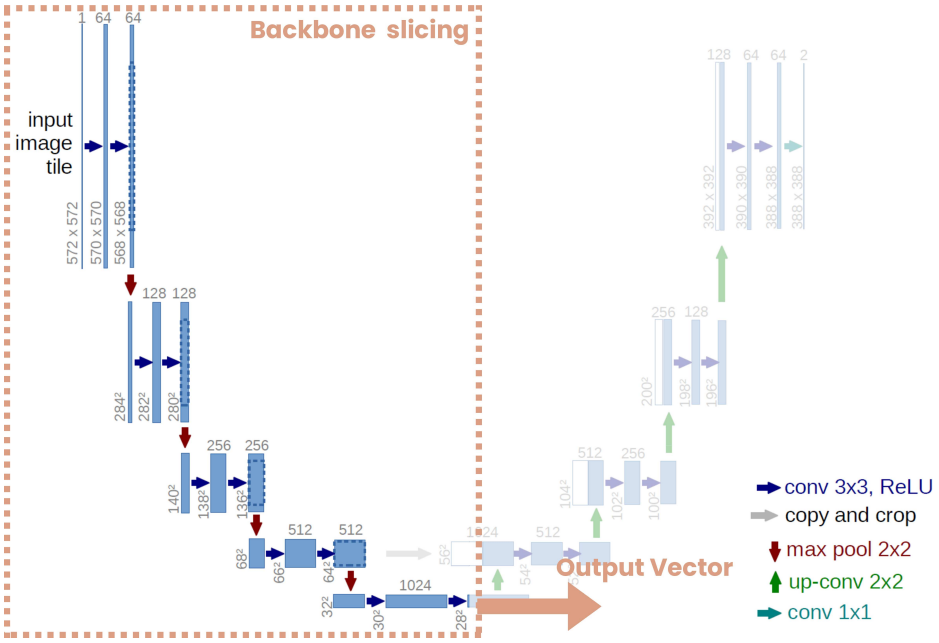


Figure 3.8: Example of re-using a vision segmentation network to extract latent feature maps. The figure demonstrates how the U-Net architecture, typically designed for image segmentation, can be repurposed to extract latent feature maps from the bottleneck of its encoder-decoder structure. The input image is processed through the encoder (contracting path), which reduces spatial dimensions while capturing increasingly abstract features. At the network's bottleneck, before the decoder begins, latent feature maps are obtained, representing a condensed yet informative embedding of the input image. These feature maps can be utilized for downstream tasks, serving as input to other models such as the Q-Former attention mechanism described in our work. The highlighted region indicates the portion of the U-Net architecture that is reused. Source: adapted from [60].

assessed their performance using standard text generation and clinical efficacy metrics. For the initial comparative experiments, we simplified our approach by using only frontal chest X-ray images as input. This choice enabled us to efficiently determine which vision encoder would be most suitable for the subsequent and more comprehensive experiments. Following this preliminary study and after selecting the pretrained vision encoder, we progressed to experiments incorporating multiple images.

For cases where two chest X-ray images were available, we adapted our feature extraction process. First, we extracted features independently from each image. Then, we aggregated these features by averaging their feature maps before feeding them to the Q-Former stage. This approach enabled us to leverage information from multiple views, resulting in more comprehensive and accurate captions. Similar strategies have proven effective in previous research [9, 10].

To assess the impact of using one versus two input images, we conducted experiments comparing the model's performance in both scenarios. By evaluating the quality of the generated captions and the model's ability to capture relevant information from single and paired chest X-rays, we aimed to determine the optimal input configuration for our architecture.

In summary, our methodology involved adapting available pretrained vision models,

more specifically a PSPNet, from TorchXRayVision Python library [16]; and a U-Net from an work [63], to serve as lightweight and efficient image encoders. By extracting latent features from these models and aggregating information from multiple views when available, we aimed to provide the Q-Former module with rich visual representations for generating accurate and informative chest X-ray captions.

3.4.2 Q-Former Alignment Module

The Q-Former module constitutes the cornerstone of our architecture, responsible for aligning extracted image features with the text decoder input. This module incorporates a BERT-based neural network and Learned Queries—trainable input parameters that guide the cross-attention mechanism over the input image features. Before the extracted image features can be processed by the Q-Former's cross-attention mechanism, they undergo dimensional alignment through a linear projection layer, responsible for performing the necessary matrix transformation to rotate the image feature representations into the appropriate dimensional space required by the transformer architecture. Figure 2.6, previously introduced in the Related Works section, illustrates the Q-Former architecture with its Learned Queries and depicts aspects of the training process that will be elaborated in a subsequent section.

Since the Q-Former is fundamentally a BERT-based model, we hypothesized that initializing it with weights from models already optimized for medical text would enhance performance. This approach has demonstrated efficacy in various medical tasks [44]. Consequently, we explored several publicly available, domain-specific pretrained weights for initializing the Q-Former:

- BERT-base[5] [20]: A English general-domain model trained on BookCorpus and Wikipedia.
- BioClinicalBERT[6] [2]: Derived from BioBERT [44] and fine-tuned on MIMIC III clinical notes [34].
- BiomedBERT[7] [25]: Pretrained from scratch on PubMed[8] abstracts and PubMedCentral[9] full-text articles.

To systematically evaluate the impact of different initializations, we trained multiple instances of our complete architecture, varying only the Q-Former initialization (BERT base, BioClinicalBERT, and BiomedBERT) while maintaining identical configurations for all other architectural components, including the image encoder and text decoder. This controlled experimental design ensured that any observed performance differences could be attributed solely to the Q-Former initialization strategy.

[5] https://huggingface.co/google-bert/bert-base-uncased
[6] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT
[7] https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext
[8] https://pubmed.ncbi.nlm.nih.gov/
[9] https://www.ncbi.nlm.nih.gov/pmc/

We evaluated each architectural variant using our established text generation and clinical efficacy metrics on the test set. Our results show that using BiomedBERT weights for initialization performed slightly better. However, these improvements were small and might need more investigation with more training iterations to determine if the results are statistically meaningful.

3.4.3 Text Decoder Module

Just as the image encoder is a critical component for feature extraction, the text decoder is the module responsible for generating the chest X-ray descriptions given the aligned features coming from the Q-Former.

Following the BLIP-2 architecture, we utilized a Large Language Model as our text decoder, due to its text generation capability. Generally, an LLM receives text prompts on its input which work as a start point, from where the model generates the text output. However, in this project, as the LLM inputs come from the Q-Former aligned features, these inputs can be considered as soft prompts, i.e., they are not explicit text embeddings, but a feature vector that help and guide the LLM decoder on its text generation task.

We adopted a model called BioGPT [50], an LLM based on GPT-2-Medium [56] architecture, since it was specifically pretrained on biomedical domain using PubMed abstracts. It is a lightweight model compared to the LLM from the original BLIP-2, containing 345 million parameters instead of 2.7 billion. This model is available at the HuggingFace website[10], allowing us to get its in-domain pretrained weights.

During the development of this work, only the BioGPT was evaluated as text decoder due to the availability of the pretrained weights.

3.5 Training Stages

The training process is divided into two distinct stages, each serving a specific purpose in enhancing the capabilities of a portion of our architecture. We conducted all model training using only the MIMIC-CXR-JPG dataset [35]. This dataset was selected for its large volume of chest X-ray images with corresponding radiological reports, and established status as a benchmark in the chest X-ray imaging studies, which enables meaningful comparisons with existing approaches.

3.5.1 Representation Learning

In this initial training stage, we focus exclusively on training the Q-Former module while utilizing features extracted from a frozen image encoder, as depicted in Figure 3.9. The primary objective of this training step is to enable the Q-Former to serve as an effective bridge between image and text domains. It prepares the module to align, understand, and generate meaningful captions, thereby enhancing the performance of the architecture as a whole.

[10] https://huggingface.co/microsoft/biogpt

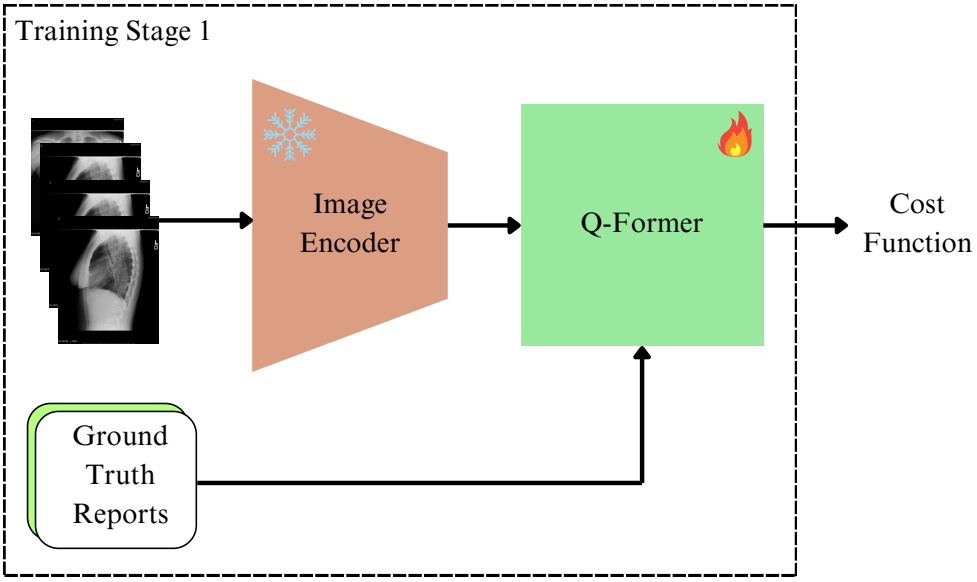


Figure 3.9: Representation Learning stage: The image encoder (orange) remains frozen while only the Q-Former (green) is trained to bridge visual and textual domains through multiple optimization objectives (cost functions). Source: elaborated by this author.

It is important to note that during this phase, the vision encoder remains completely frozen and undergoes no parameter updates. This design choice leverages the rich visual representations already captured in the pre-trained weights of the vision encoder (as detailed in Subsection 3.4.1). Only the Q-Former parameters are updated through backpropagation.

During this representation learning phase, the Q-Former outputs are optimized using three distinct cost functions, as in the BLIP-2 framework [46], that collectively enable effective cross-modal alignment:

Image-Text Contrastive Learning (ITC)

The contrastive learning objective aligns image and text representations in a shared embedding space. Given image features $\mathbf{q}^I \in R^{N\times K\times d}$ from query tokens and text features $\mathbf{t} \in R^{N\times d}$, we compute similarities by aggregating across query tokens:

$$sim_{i2t} = \max_k \left(\frac{\mathbf{q}_i^I \cdot \mathbf{t}_j}{\|\mathbf{q}_i^I\| \|\mathbf{t}_j\|} \right) / \tau,$$

$$sim_{t2i} = \max_k \left(\frac{\mathbf{t}_i \cdot \mathbf{q}_{j,k}^I}{\|\mathbf{t}_i\| \|\mathbf{q}_{j,k}^I\|} \right) / \tau,$$

where K is the number of query tokens, $\tau = 0.07$ is the temperature parameter, and the max operation aggregates across all query tokens. The bidirectional contrastive loss with label smoothing is:

$$\mathcal{L}_{ITC} = \frac{1}{2} \left[\mathcal{L}_{CE}(sim_{i2t}, \mathbf{y}, \alpha = 0.1) + \mathcal{L}_{CE}(\mathrm{sim}_{t2i}, \mathbf{y}, \alpha = 0.1) \right],$$

where L_{CE} denotes cross-entropy loss with label smoothing factor α, and $\mathbf{y}$ contains the correct positive pair indices.

Image-Text Matching (ITM)

The ITM loss activates the image-grounded text encoder and learns fine-grained multimodal alignment. Hard negative pairs are selected using similarity-weighted sampling from the contrastive similarities (with diagonal masking to exclude positive pairs):

$$p_{neg}(j|i) = \frac{\exp(sim_{i2t}[i,j] - \mathbb{I}_{i=j} \cdot 10^4)}{\sum_{k \neq i} \exp(sim_{i2t}[i,k])}.$$

The Q-Former processes three types of pairs: (positive image, positive text), (negative image, positive text), and (positive image, negative text). An ITM head computes matching logits from the averaged query token representations:

$$logits = \text{ITM}_{Head}\left(\frac{1}{K}\sum_{k=1}^{K} \mathbf{h}^k_{query}\right).$$

The binary classification loss is:

$$\mathcal{L}_{\text{ITM}} = \mathcal{L}_{CE}(\text{logits}, \mathbf{labels}),$$

where $\mathbf{labels} = [1, 1, ..., 1, 0, 0, ..., 0]$ with the first N entries as positive and the remaining 2N as negative pairs.

Image-Grounded Text Generation (ITG)

The ITG loss activates the image-grounded text decoder for autoregressive caption generation. Using teacher forcing with the Q-Former's past key values from the representation learning stage:

$$\mathcal{L}_{ITG} = -\sum_{t=1}^{|T|} \log P(w_t | w_{1:t-1}, \mathbf{Q}, \mathbf{I}),$$

where $\mathbf{Q}$ represents the query token embeddings, $\mathbf{I}$ denotes image features, and the probability is computed through the Q-Former's language modeling head. Labels are masked where input tokens equal the padding token ID.

Combined Loss Function

The total loss combines all three objectives with equal weighting:

$$Loss_{total} = \mathcal{L}_{ITC} + \mathcal{L}_{ITM} + \mathcal{L}_{ITG}.$$

The validation of this step focuses solely on verifying that the architecture is being optimized correctly via the defined cost functions. We monitor the learning curves, particularly checking if validation loss is decreasing appropriately, and halt training when

the validation loss plateaus or begins to increase, indicating potential overfitting.

3.5.2 Generative Learning

In the second training stage, we leverage the entire architecture, as illustrated in Figure 3.10. The previously trained Q-Former is attached to the pretrained BioGPT LLM, and then trained to optimize the outputs of the LLM decoder. This phase aims to utilize the trained Q-Former's knowledge and apply it to guide the LLM in generating accurate and coherent descriptions of input chest X-ray images.

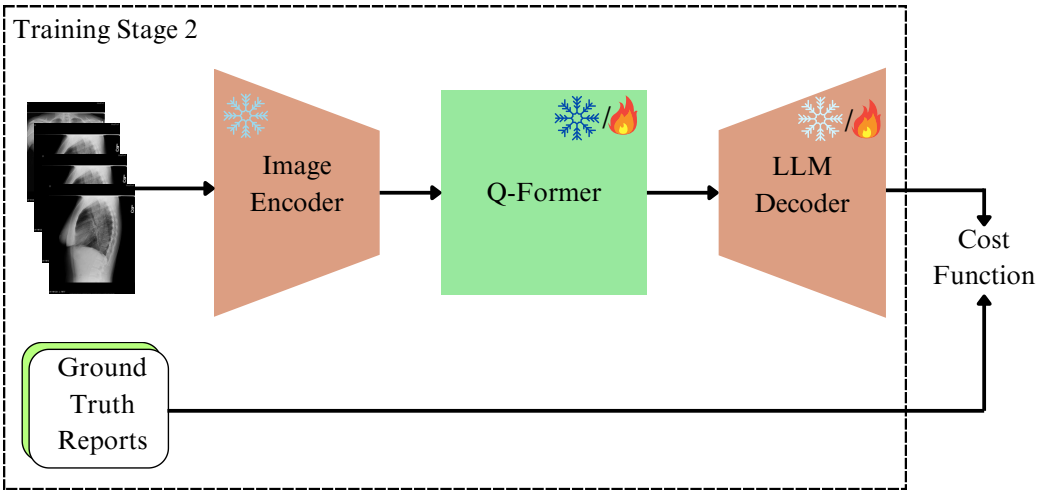


Figure 3.10: Generative Learning stage: The previously trained Q-Former is now connected to the LLM decoder. We experiment with various combinations of frozen/unfrozen components (both unfrozen, only Q-Former frozen, only LLM frozen) and optimize the trainable parameters using the LLM decoder's language modeling cost function. Source: elaborated by this author.

Our evaluation in this stage serves several purposes. Firstly, we investigate the capacity of the specialized, smaller LLM to generate precise and well-articulated descriptions based on the soft prompts provided by the Q-Former. We rely on a combination of manual qualitative analysis and established text generation metrics, which will be detailed in the metrics section. This assessment is critical in understanding the LLM's suitability for the specific domain.

To optimize performance, we systematically explored different training configurations. Specifically, we examined three distinct approaches: (1) unfreezing both the Q-Former and the LLM during training, (2) freezing the Q-Former while training only the LLM, and (3) training only the Q-Former while keeping the LLM frozen, which is the default BLIP-2 proposal. Our experimental results revealed that the second approach – freezing the Q-Former and training only the LLM—consistently yielded the best performance across our evaluation metrics. Based on these findings, all subsequent experiments presented in this thesis employ this optimal configuration in the second training stage (frozen Q-Former, trainable LLM).

3.5.3 Training Hyperparameters

To ensure reproducibility and optimize model performance, we implemented specific hyperparameter configurations for each training stage. These configurations were carefully selected based on both established practices in the literature and empirical testing in our experimental setup.

Representation Learning Hyperparameters

For the representation learning stage, where only the Q-Former module undergoes parameter updates, we employed the AdamW optimizer [49] with a learning rate of 1×10^{-4}, beta parameters of $[0.9, 0.98]$, and weight decay of 0.05. The AdamW optimizer was selected due to its effectiveness in handling the complex optimization landscape of transformer-based architectures while mitigating overfitting through weight decay regularization.

We implemented a cosine annealing learning rate scheduler to gradually reduce the learning rate throughout training, starting at 1×10^{-4} and ending at 1×10^{-5}. And also, we adopted a linear warmup phase starting from 1×10^{-6} in the first epoch, in order to maintain training stability by establishing stable gradients before accelerating the learning process [23].

Generative Learning Hyperparameters

For the generative learning stage, where the Q-Former is connected to the BioGPT LLM, we maintained the AdamW optimizer but adjusted several hyperparameters for the language model fine-tuning. Specifically, we reduced the learning rate to 1×10^{-5} and modified the beta parameters to $[0.9, 0.99]$, while keeping the weight decay constant at 0.05.

The learning rate scheduler for this stage followed a similar cosine annealing pattern but with appropriately scaled parameters: an initial learning rate of 1×10^{-5}, a minimum learning rate of 1×10^{-6}, and a warmup learning rate of 1×10^{-8} is the first epoch.

Both training stages were executed on CUDA-compatible GPU hardware, with validation performed after each epoch to monitor convergence and prevent overfitting. Early stopping was implemented based on validation loss plateaus, with model checkpoints saved at regular intervals to preserve the best-performing model states for subsequent evaluation and testing.

3.6 Computation Resources

This research utilized two distinct computational environments to support the experimental work: Amazon Web Services (AWS) SageMaker and an institutional computation server. For cloud-based computation, we leveraged AWS SageMaker service with an ml.g4dn.xlarge instance equipped with 16GiB RAM, a 125 GiB NVMe SSD, and an NVIDIA T4 GPU with 16GiB VRAM. The AWS environment operated on a PyTorch 2.10

Python 3.10 GPU Optimized container image, which included PyTorch 2.1.0, Python 3.10, and CUDA 12.1.

For on-premises computation, we utilized the Institute of Computation Server, a Docker-based environment configured for student research. This server featured an Intel(R) Xeon(R) Silver 4210R CPU running at 2.40GHz with 40 vCPUs, 126GiB RAM, and an NVIDIA RTX A5500 GPU with 24GiB VRAM. The system ran Ubuntu 22.04.4 LTS with CUDA version 12.4.

The software environment remained consistent across both platforms, employing PyTorch 2.1.2, Transformers 4.33.2, NumPy 1.26.3, Pandas 2.2.0, and OpenCV-Python-Headless 4.5.5.64. This consistent software stack ensured reproducibility of results regardless of the computational platform used for different stages of the investigation.

Chapter 4

Results and Discussion

In this chapter, we present the analysis of our proposed architecture for chest X-ray description generation. For internal evaluation of our design choices, we used CE metrics – precision, recall and F1-score. However, when comparing our approach against state-of-the-art methods, we relied on both Clinical Efficacy (CE) metrics and Natural Language Generation (NLG) metrics such as BLEU, METEOR and ROUGE scores, which are commonly used in the field.

Our experiments focused on three aspects: (1) selecting the optimal image encoder for the full architecture, (2) evaluating different Q-Former initialization weights, and (3) benchmarking our approach against existing methods using these NLG metrics. We also provided qualitative and visual examples of generated reports to demonstrate the clinical relevance and accuracy of our system. Throughout our analysis, we highlight the balance between computational efficiency and performance, showing how our approach achieved competitive results with significantly fewer parameters than current LLM-based solutions.

4.1 Qualitative Analysis

In this section, we examine the characteristics and quality of the generated chest X-ray descriptions through multiple points of view: a comparative analysis between generated reports and ground-truth references, an expert evaluation conducted by an emergency medicine specialist, and an exploration of targeted prompt strategies. This evaluation approach allowed us to better understand the clinical relevance, coherence, and potential applications of our proposed architecture.

4.1.1 Generated Text Samples

To qualitatively assess the performance of our architecture, we visually compared our system's predictions against the actual X-ray reports. This comparison allows us to better understand the quality and relevance of the generated text.

For this analysis, we utilized samples from the publicly available IU X-Ray dataset, as it provides a diverse set of chest X-ray images paired with corresponding reports. The results are presented in Figures 4.1, 4.2, and 4.3.

The main observations from the generated text samples are as follows:

Ground Truth

"heart size is normal and lungs are clear. no evidence of tuberculosis"

Our Prediction

"heart size is normal. the mediastinal and hilar contours are normal. the pulmonary vasculature is normal. lungs are clear. no pleural effusion or pneumothorax is seen. there are no acute osseous abnormalities."

Figure 4.1: Example of a chest X-ray with no abnormal findings. The model's predictions tend to be more descriptive than the actual ground-truth report, highlighting its ability to provide detailed interpretations even in cases of normal imaging. Source: elaborated by this author.

Ground Truth

"bilateral patchy pulmonary opacities noted. interval improvement in left base consolidative opacity. pulmonary vascular congestion again noted. stable enlarged cardiomediastinal silhouette. stable left. no evidence of pneumothorax. no large pleural effusions."

Our Prediction

"in comparison with the study of, there is little overall change. again there are low lung volumes with enlargement of the cardiac silhouette, pulmonary vascular congestion, and bilateral pleural effusions with compressive atelectasis at the bases. in the appropriate clinical setting, supervening pneumonia would have to be considered., md =, phd =, cta of the chest., md = telephone., paged at: a.m., one minute after the time of discovery.,"

Figure 4.2: Example of a chest X-ray with multiple abnormal findings. Once again, our model demonstrates its descriptive capabilities by identifying and detailing all relevant findings present in the X-ray. The red boxes are predictions that are not found in the ground-truth report. The other colors highlight the matching findings. Source: elaborated by this author.

1. **Descriptive Nature of Generated Text**: Since our model was trained on the MIMIC-CXR-JPG dataset, which contains highly descriptive radiology reports, it tends to generate similarly detailed text. This is evident when compared to the actual reports in the IU X-Ray dataset, as shown in Figure 4.1. The model often provides more comprehensive descriptions, even for cases where the actual report is concise or lacks detail. While our model generates comprehensive descriptions, the emergency specialist review highlighted important nuances: terminology precision matters (e.g., 'cardiomediastinal' vs 'cardiac' represent different anatomical

Ground Truth
"there are postoperative changes of sternotomy and cabg. there is stable mild cardiomegaly. there are scattered of subsegmental atelectasis, decreased from the prior chest radiograph. no focal airspace consolidation. no pleural effusion or pneumothorax. there are minimal degenerative changes of the spine."

Our Prediction
"in comparison with the study of, the left chest tube has been removed and there is no evidence of pneumothorax. atelectatic changes are again seen at the bases, especially on the left. no vascular congestion or acute focal pneumonia., md =, md =, phd =., md = telephone., paged., minutes after the discovery of the findings., md = telephone., minutes after discovery of the findings., md ="

Figure 4.3: Example of a chest X-ray with some abnormal findings. While the model generates clinically relevant text, it occasionally produces nonsensical phrases, indicating areas for further refinement. Source: elaborated by this author.

concepts), and some findings require larger images for accurate assessment (such as confirming bilateral pleural efusion conditions).

2. **Accuracy and Relevance of Findings**: The system demonstrates the ability to generate text that not only uses clinically appropriate terminology but also identifies and elaborates on findings that may be only briefly mentioned or implied in the actual report. For instance, in Figure 4.2, the model accurately identifies and describes multiple abnormalities in the X-ray, providing a more detailed account than the reference report.

3. **Occasional Hallucinations**: Despite its strengths, the model occasionally generates nonsensical or irrelevant phrases, particularly in cases where the input image is ambiguous or contains subtle findings. This phenomenon, often referred to as "hallucination", is a known challenge in text generation models and is observed in some of our predictions, as noted in Figure 4.3. These instances highlight areas for future improvement, particularly in enhancing the model's ability to generalize across diverse datasets and imaging conditions.

These samples collectively demonstrate the architecture's ability to generate coherent and clinically relevant descriptions of chest X-ray images. While the model excels in providing detailed and accurate interpretations, the occasional generation of irrelevant or nonsensical text underscores the need for continued development to improve robustness and reliability. Future work could focus on fine-tuning the model on a more diverse dataset, incorporating additional context, or employing post-processing techniques to mitigate hallucinations and enhance overall performance.

4.1.2 Emergency Specialist Analysis

To evaluate the clinical relevance of our model beyond computational metrics, we conducted a qualitative analysis with an emergency medicine specialist who assessed a random sample of 10 generated reports. The specialist independently evaluated both text quality and clinical accuracy, providing detailed feedback on each report using a structured assessment table that included the chest X-ray images alongside columns for text and clinical quality ratings (very good, good, bad, very bad) and observations.

Despite achieving competitive CE metrics on this sample (P: 0.4167, R: 0.3704, F1: 0.3922), the initial specialist assessment revealed apparent discrepancies between computational evaluation and clinical judgment. Most reports (7 out of 10) were rated as "Poor" or "Very Poor" in terms of clinical quality, highlighting several consistent issues:

1. **Technical Quality Assessment**: The specialist noted that few reports adequately addressed the technical quality of the X-rays, which is a critical first step in radiological interpretation. Factors such as patient positioning, inspiration depth, and radiation penetration significantly impact interpretation accuracy but were rarely mentioned in generated reports.

2. **Organizational Structure**: The specialist observed that ideal radiological reports follow an organized structure (from outer to inner structures: subcutaneous tissue, bones, lungs, mediastinum), which was not consistently implemented in the generated texts.

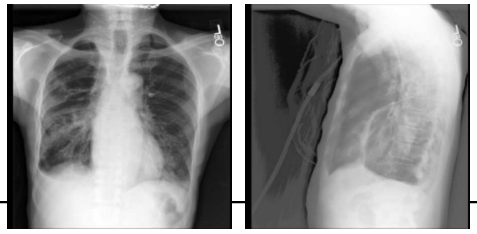
3. **Contradictory Information**: Some reports contained internally inconsistent information, such as simultaneously asserting and questioning the presence of pleural effusion. As highlighted in orange in Figure 4.4.

4. **Omission of Visible Findings**: In multiple cases, significant visible abnormalities were entirely omitted from the reports, such as unmentioned aortic calcifications or nodules. As highlighted in blue in Figure 4.4.

During our follow-up discussion with the specialist, an important contextual factor emerged—she had evaluated the reports against radiologist-level standards rather than emergency physician documentation standards. This distinction is significant because our model was trained on the MIMIC-CXR-JPG dataset, which contains reports generated by emergency department physicians rather than specialized radiologists.

Emergency department reports typically prioritize clinically urgent findings and may not follow the comprehensive, structured format of formal radiological reports. When informed of this training context, the specialist acknowledged that her evaluation criteria would have been adjusted to reflect emergency documentation standards rather than formal radiological reporting conventions.

This specialist analysis reveals important limitations in our current evaluation approach and metrics. While computational metrics suggest competitive performance, the assessment of clinical utility indicates significant room for improvement. This discrepancy underscores the need for evaluation frameworks that better align with clinical expertise



Prediction

frontal and lateral views of the chest were obtained. the heart is of normal size with normal cardiomediastinal contours. the lungs are hyperinflated, with flattening of the diaphragms, suggesting chronic obstructive pulmonary disease. there is a small right pleural effusion with overlying atelectasis. no focal consolidation is seen. there is no pneumothorax. blunting of the right costophrenic angle may be due to a trace pleural effusion. no left pleural effusion.

Actual Report

the patient is status post esophagectomy and gastric pull through. the lungs are hyperinflated. there are new patchy airspace opacities in the bilateral lung bases, concerning for aspiration. chronic medial right apex pleural thickening and triangular peripheral interstital opacities in the right mid lung field are again seen. a hazy opacity consistent with chronic scarring related to radiation treatment is again seen in the medial right upper lobe. an mm nodular opacity is again seen within the right lower lobe, unchanged from prior exam. there is atelectasis at the left lung base, a chronic right pleural effusion is again noted. there is no left pleural effusion. cardiomediastinal silhouette is stable. there is no pneumothorax. visualized osseous structures are unremarkable.

Figure 4.4: Generated report and actual report illustrating contradictory information (orange) and omission of visible findings (blue). Source: elaborated by this author.

and highlights the importance of incorporating domain experts in both the development and assessment of medical AI systems.

4.1.3 Initial Prompts

We explored the potential of using initial prompts to guide the generation of focused and structured captions. By providing prompts direct related to findings, such as “atelectasis” or “cardiomegaly”, and also, prompts related to organs, such as “the heart” or “the lungs” we aimed to steer the language model towards generating descriptions that follow a specific pattern or cover certain anatomical regions of interest, as can be seen in Figure 4.5.

The implementation of targeted prompts demonstrates clear influence over output direction. By providing condition-specific prompts or anatomical region prompts, we observe the model generating completions that address these specific elements with appropriate radiological terminology. This directed approach has both advantages and potential risks—while it ensures comprehensive coverage of key anatomical structures and potential pathologies, it may inadvertently suggest findings not present in the original image and associated ground-truth report, as we can see in the example.

These prompt-guided generations could significantly enhance emergency workflows by providing customizable reporting templates, ensuring systematic coverage of anatomical regions, standardizing reporting language, and offering preliminary assessments that physicians can rapidly verify. However, implementation would require careful balancing of guidance and potential bias introduction, with particular attention to validation against expert interpretations to prevent diagnostic errors stemming from prompt-induced sug-

Ground Truth
"the lungs are clear without infiltrate. there is no effusion or pneumothorax."

No Prompt (60 words max)
"pa and lateral chest views were obtained with patient in upright position. analysis is performed in direct comparison with the next preceding similar study of. the heart size remains normal. no configurational abnormality is present. unremarkable appearance of thoracic aorta and mediastinal structures. the pulmonary vasculature is not conge"

Prompts with max 10 words
Prompt: **cardiomegaly** ... is mild. the lungs are clear.
Prompt: **pleural effusion** ... on the right is minimal. there is
Prompt: **atelectasis** ... at the left base is mild. there is
Prompt: **edema** ... has resolved since. lungs are clear
Prompt: **lung opacity** ... in the right lower lobe is concerning for pneumon
Prompt: **pneumonia** ... in the right lower lobe. -mm right
Prompt: **the heart** ... size is normal. the hilar and mediastinal contours
Prompt: **the lungs** ... are clear. the cardiomedias
Prompt: **the mediastinum** ... is normal. the heart is not enlarged

Figure 4.5: Example of initial prompt strategies. The visualization compares different prompting approaches—including condition-specific prompts (e.g., "atelectasis", "cardiomegaly") and anatomical region prompts (e.g., "the heart", "the lungs") – demonstrating how each prompt type influences the content, structure, and focus of the resulting captions. We observe suggested findings (orange) which are not in the actual chest X-ray report. The initial prompt is in bold text and the prediction comes after the ellipsis (...). We constrained the system to generate only 10 words when given an initial word or phrase, simulating a writing assistant for emergency specialist. Source: elaborated by this author.

gestions.

Future research directions should focus on optimizing prompt design to minimize confirmation bias while maintaining comprehensive coverage, developing adaptive prompting systems responsive to initial image analysis, and establishing rigorous clinical validation protocols to ensure reliability in medical decision-making contexts.

4.2 Quantitative Analysis

After performing a qualitative analysis, seeking to understand the real clinical relevance of the generated descriptions, we now turn to quantitatively comparing our models against other studies and proposed architectures. The following experiments were designed to assess both the clinical accuracy and computational efficiency of our approach, providing evidence for the viability of lightweight language models in generating meaningful chest X-ray descriptions. All experiments were conducted using the MIMIC-CXR-JPG dataset with the evaluation protocols described in Chapter 3.

4.2.1 Radiological Findings Analysis

Table 4.1 presents the system performance across different radiological findings. The model demonstrates high efficacy for some findings, such as Support Devices (0.751 F1), Cardiomegaly (0.602 F1), and Pleural Effusion (0.595 F1).

Table 4.1: Clinical Efficacy (CE) metrics for each finding within the MIMIC-CXR-JPG test set. The Positive Cases and Negative Cases columns indicate the distribution of positive and negative labels in the ground truth data, allowing for verification of class imbalance.

Observation	**Positive Cases**	**Negative Cases**	**Precision**	**Recall**	**F1**
Support Devices	1046	1599	0.713	0.793	0.751
Cardiomegaly	937	1708	0.591	0.614	0.602
Pleural Effusion	917	1728	0.703	0.516	0.595
Atelectasis	703	1942	0.395	0.333	0.361
Edema	507	2138	0.465	0.258	0.332
Lung Opacity	931	1714	0.425	0.161	0.234
No Finding	184	2461	0.148	0.495	0.228
Pneumothorax	73	2572	0.474	0.123	0.196
Consolidation	145	2500	0.242	0.103	0.145
Pneumonia	153	2492	0.137	0.065	0.088
Enlarged Cardiomediastinum	194	2451	0.079	0.026	0.039
Fracture	118	2527	0.050	0.008	0.014
Lung Lesion	154	2491	0.000	0.000	0.000
Pleural Other	88	2557	0.000	0.000	0.000
Macro-Average	-	-	0.3159	0.2497	0.2561
Micro-Average	-	-	0.5142	0.4102	0.4564

There is a strong correlation (r 0.873) between the number of positive cases and F1 scores, indicating that class imbalance significantly impacts model performance, with the system favoring majority classes during training due to their overrepresentation in the dataset. This imbalance creates a bias where the model optimizes overall accuracy at the expense of correctly identifying minority classes, leading to poor detection of rare findings. This pattern has interesting exceptions; Lung Opacity presents with high prevalence (931 positive cases) but relatively poor performance (0.234 F1), this could suggest that beyond class imbalance, the model struggles with the visual complexity or subtlety of certain radiographic features, which could be further studied in a future work.

Most findings with fewer than 200 positive examples show F1 scores below 0.1, highlighting how data unbalance limits the model's ability to learn rare conditions effectively. Furthermore, the difference between macro-average (0.2561) and micro-average (0.4564) F1 scores further confirms the influence of class imbalance on overall performance metrics.

These performance characteristics have important clinical implications, suggesting the model could serve as an effective assistive tool for common findings while requiring expert supervision for critical but rare conditions. Future work should focus on addressing the class imbalance through targeted data augmentation strategies and developing specialized architectures that can better capture the subtle visual features associated with challenging conditions such as Pneumothorax (0.196 F1) and Fracture (0.014 F1).

These results demonstrate the potential of our proposed method for generating clinically relevant and accurate text reports. The high precision ensures a low false positive

rate, while further investigation is needed to improve recall, particularly for less frequent findings.

Our architecture demonstrates competitive performance while maintaining a relatively lightweight size compared to other approaches. This highlights the effectiveness of our design choices and the benefits of leveraging pretrained models and domain-specific knowledge.

4.2.2 Choosing the Image Encoder

To determine the most suitable pretrained image encoder for our architecture, we conducted experiments with two public available models: U-Net, from the work [63], and PSPNet, from TorchXRayVision [16]. Both models were pretrained on chest X-ray segmentation tasks and adapted to fit into our proposed system, as detailed in Subsection 3.4.1

We trained the whole architecture using the same Q-Former initialization weights, BioClinicalBERT, and same LLM decoder, BioGPT, and same hyperparatemeters, detailed in previous sections.

For the training Stage 1 (Representation Learning), the whole process adopting U-Net took approximately 27 hours, while PSPNet required significantly longer at around 102 hours.

In Stage 2 (Generation Learning), training the architecture with U-Net as the image encoder took about 10 hours, compared to approximately 30 hours for PSPNet. This represents a substantial difference in computational requirements between the two encoders.

After both training stages, we evaluated the performance of both architectures using clinical efficacy (CE) metrics – precision (P), recall (R), and F1 – to check if the choice of image encoder significantly impacts the quality of generated reports. The metrics were calculated by comparing the generated reports with the ground truth reports from the test set.

Our experimental results, presented in Table 4.2, show that the U-Net encoder achieved better performance across all CE metrics. Specifically, U-Net obtained higher precision (0.5196 vs. 0.5039), recall (0.3707 vs. 0.3595), and F1 score (0.4327 vs. 0.4196) compared to PSPNet. Furthermore, the U-Net-based architecture required substantially less training time: only 37 hours compared to 132 hours for PSPNet (a reduction of approximately 72%).

Table 4.2: CE metrics comparison of the performance of full architecture adopting different vision encoder pretrained models.

Vision Encoder	**Precision**	**Recall**	**F1**	**Total Training Time (h)**
U-Net	**0.5196**	**0.3707**	**0.4327**	**37**
PSPNet	0.5039	0.3595	0.4196	132

The results suggest that U-Net model, with the available in-domain pretrained weights, originally designed for chest X-ray image segmentation, provides more effective feature extraction for chest X-ray interpretation in our multimodal framework.

Based on these findings, we selected U-Net as our image encoder for subsequent experiments, as it provided superior performance with significantly reduced computational demands for training.

4.2.3 Choosing the Q-Former Initialization

We investigated the impact of initializing the Q-Former module with different pretrained weights. Specifically, we compared the performance of the architecture when initialized with vanilla BERT-base weights and domain-specific weights from BioClinicalBERT and BiomedBERT. The CE metrics can be found in Table 4.3.

To make a fair comparison, we used the same U-Net model as vision encoder, BioGPT as LLM decoder, and identical hyperparameters during each training stage for each Q-Former initialization weight under study.

Table 4.3: CE metrics comparison of the performance of full architecture adopting different Q-Former initialization weights.

Q-Former Weights	**Precision**	**Recall**	**F1**
BERT-base	0.5154	0.3636	0.4264
BioClinicalBERT	**0.5196**	0.3707	0.4327
BiomedBERT	0.5142	**0.4102**	**0.4564**

The results demonstrated that initializing the Q-Former with BiomedBERT weights led to a slightly better performance in terms of CE metrics, with an improvement of 0.03 in F1 score compared to the baseline BERT-base initialization. Therefore, we selected BiomedBERT weights as the default Q-Former initialization weights for all subsequent experiments.

While these improvements were modest, they suggest that domain-specific pretraining can provide benefits for medical report generation tasks. To establish the statistical significance of these differences, we would need to conduct additional experiments with multiple random seeds and perform rigorous statistical testing.

4.3 Comparison Against Other Techniques

This section evaluates the effectiveness of our proposed method for generating text reports from chest X-rays. Our architecture demonstrates a notable efficiency-to-performance ratio, combining strong clinical accuracy with a significantly smaller parameter footprint compared to competing approaches. We assessed performance using both Natural Language Generation (NLG) metrics and more clinically relevant Clinical Efficacy (CE) metrics over the MIMIC-CXR-JPG test split.

4.3.1 Model Size

In order to have a fair comparison between our architecture and other techniques in terms of number of parameters, in this subsection, we have chosen studies that also utilized LLMs

as their text generator, as presented in Table 4.4.

Table 4.4: Comparison of the number of parameters of the text decoder in other studies that also adopted an LLM to generate chest X-ray descriptions.

Method	**Parameters**
XRayGPT	7 B
ChatGLM-6B	6 B
Med-PaLM	540 B
CvT21-2DistilGPT2	**82 M**
Ours	347 M

Our architecture consists of approximately 515 million parameters, where 347 million parameters are from the BioGPT, our text generator. Compared to other techniques, there is considerable variation in model sizes. XRayGPT [70] has around 7 billion parameters solely in its LLM model [12]. Similarly, Yang et al. [78] employed the ChatGLM-6B LLM model [80], which comprises approximately 6 billion parameters. Med-PaLM [66], on the other hand, incorporates 540 billion parameters in its PaLM LLM. At the other end of the spectrum, CvT21-2DistilGPT2 utilizes only about 82 million parameters. Unfortunately, the works with larger models (XRayGPT, ChatGLM-6B, and Med-PaLM) did not perform a thorough evaluation using common datasets such as MIMIC-CXR-JPG or IU X-Ray. However, they serve as works that also tried to adopt LLMs, so we can compare the size of our models with theirs.

While CvT21-2DistilGPT2 utilizes a significantly smaller parameter count than our approach, our architecture leverages BioGPT's domain-specific knowledge and achieved superior performance on CE metrics as demonstrated in Section 4.3.2. This illustrates an important trade-off in the model design space: extremely large models such as Med-PaLM (540B) may be impractical for many real-world deployments, while very small models might sacrifice critical performance. Our 347M parameter model represents a practical middle ground that balances computational efficiency with robust performance for clinical applications.

4.3.2 Text Generation Metrics

In order to quantitatively compare our approach with other studies in the area, even with the ones that do not utilize large language models, we evaluated the clinical efficacy of the reports using CheXbert labels. Precision, recall, and F1-score were calculated for each CheXbert finding. Table 4.5 summarizes the overall CE performance compared to existing methods.

Our method achieved the highest precision (0.5196) and F1 (0.4564) scores among the evaluated methods, indicating a low rate of false positive findings. While CvT21-2DistilGPT2 [52] still maintains the highest recall (0.497), our model demonstrates competitive recall (0.4102).

Table 4.6 presents a comparison of NLG metrics between other techniques and our model. Our model has low scores if compared to other techniques. This could be explained

Table 4.5: Performance comparison of medical image report generation methods using CE metrics. Our proposed approach achieves state-of-the-art precision and competitive F1 scores with only 102M trainable parameters in the Q-Former alignment model.

Method	**Precision**	**Recall**	**F1**	**Trainable Parameters**
R2Gen [9]	0.333	0.273	0.276	65M
CMN [10]	0.334	0.275	0.278	64M
METransformer [77]	0.364	0.309	0.311	152M
COMG [24]	0.424	0.291	0.345	65M
CvT21-2DistilGPT2 [52]	0.398	**0.497**	0.442	82M
Ours	**0.5142**	0.4102	**0.4564**	102M

due to the fact the those NLG metrics were developed to mainly evaluate machine translation, which is clearly not the task here. Since our core model is a pre-trained GPT-2 model, our predictions may not align as closely with the ground truth. Consequently, this misalignment could account for the observed low scores in our proposed model's metrics.

Table 4.6: Comparison of the NLG metrics between our model and other techniques.

Method	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**
R2Gen [9]	0.353	0.218	0.145	0.103	0.142	0.277
CMN [10]	0.353	0.218	0.148	0.106	0.142	0.278
METransformer [77]	0.386	0.250	0.169	0.124	0.152	0.291
COMG [24]	0.346	0.216	0.145	0.104	0.137	0.279
CvT21-2DistilGPT2 [52]	0.462	0.295	0.214	0.165	0.192	0.370
Ours	0.284	0.165	0.106	0.074	0.120	0.201

4.4 Additional Analyses

This section provides additional discussion on some results obtained with the proposed method.

4.4.1 Single Image vs. Multiple Images

We investigated the impact of using a single image versus multiple images as input to our architecture. Contrary to our initial expectations, the results demonstrated that utilizing multiple images did not lead to improved performance, as presented in Table 4.7. We hypothesized that lateral images provided less diagnostic information compared to frontal images, potentially introducing noise into the training process rather than contributing meaningful features.

For the subset containing both lateral and frontal images, we processed each image individually through the vision encoder, and then aggregated the vector representations by averaging them, an approach adopted in previous multi-view medical imaging studies,

Table 4.7: Performance comparison between multi-view and single-view approaches using CE metrics.

Approach	Precision	Recall	F1
Frontal and Lateral Views	0.5037	0.3456	0.4099
Frontal View Only	0.5142	0.4102	0.4564

as illustrated in Figure 4.6. However, this simple averaging strategy may have diluted the discriminative features present in frontal images.

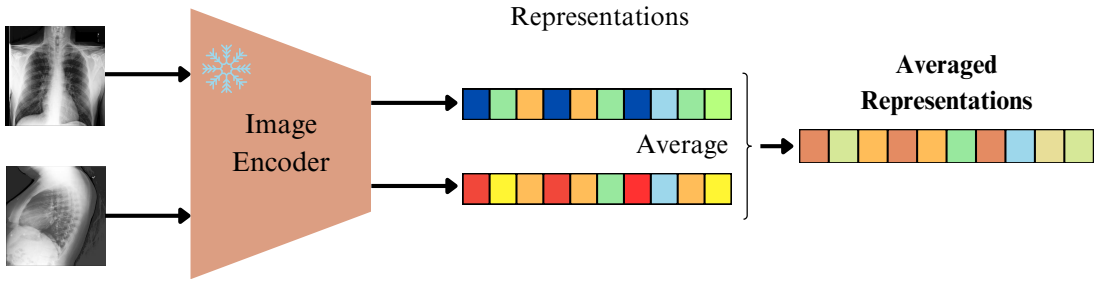


Figure 4.6: Illustration of our multi-view X-ray encoding approach. The architecture processes frontal and lateral X-ray images independently, followed by averaging the resulting vectors. Source: elaborated by this author.

Therefore, future work should explore more sophisticated methods of feature aggregation in multi-view imaging systems, such as attention mechanisms, learned weighting schemes that could better capture complementary information across different perspectives, or even simply concatenating the feature vectors instead of averaging them.

4.4.2 CE and NLG Metrics

Further analysis revealed an interesting nuance between clinical efficacy (CE) metrics and Natural Language Generation (NLG) metrics in the context of chest X-ray report generation. Specifically, within a subset of high-performing predictions based on CE metrics – where predictions perfectly conveyed observations from the X-rays (i.e., a manually selected sample with perfect Precision and Recall) – the corresponding NLG metrics once again exhibited low scores (see Table 4.8). This finding reinforces that NLG metrics alone may not reliably indicate model quality for this task, as other studies have also suggested [7].

Table 4.8: NLG metric scores (e.g., BLEU, ROUGE-L) for a subset of model predictions that achieved perfect Precision, Recall, and F1-score in the clinical efficacy (CE) evaluation.

BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
0.275	0.171	0.117	0.086	0.138	0.222

Even predictions which are clinically accurate by CE metrics might not receive high NLG scores. This highlights the limitations of NLG metrics in capturing the full spectrum

of factors that contribute to good clinical reports, such as factual accuracy or nuanced phrasing.

4.4.3 Grad-CAM Analysis

We explored the application of Gradient-weighted Class Activation Mapping (Grad-CAM) [64] to visualize the attention patterns of the vision component during word generation. This technique analyzes neural network activations during input processing—in our case, chest X-rays—and enables gradient backpropagation to identify which image regions contribute to specific outputs in the generated report.

We attempted to guide the vision model's attention by inserting specific prompts into the LLM component, but this approach proved ineffective. As noted earlier, LLM prompts are textual instructions that influence the language model's processing and output generation – in our case, serving as the initial text from which the LLM generates its response.

However, our initial analysis did not yield significant insights. In shallow layers of the vision encoder, we observed that attention was predominantly directed toward extraneous elements such as alphanumeric markers indicating the left/right sides of the radiogram or AP/PA positioning information, rather than clinically relevant findings, as can be seen in Figures 4.7 and 4.8. Even in deeper layers of the vision encoder, the attention patterns remained difficult to interpret meaningfully.

Due to time constraints, we were unable to pursue a more comprehensive refinement of our Grad-CAM methodology. Further exploration of visualization techniques remains an important direction for future work to better understand the model's attention mechanisms and improve its interpretability in clinical contexts.

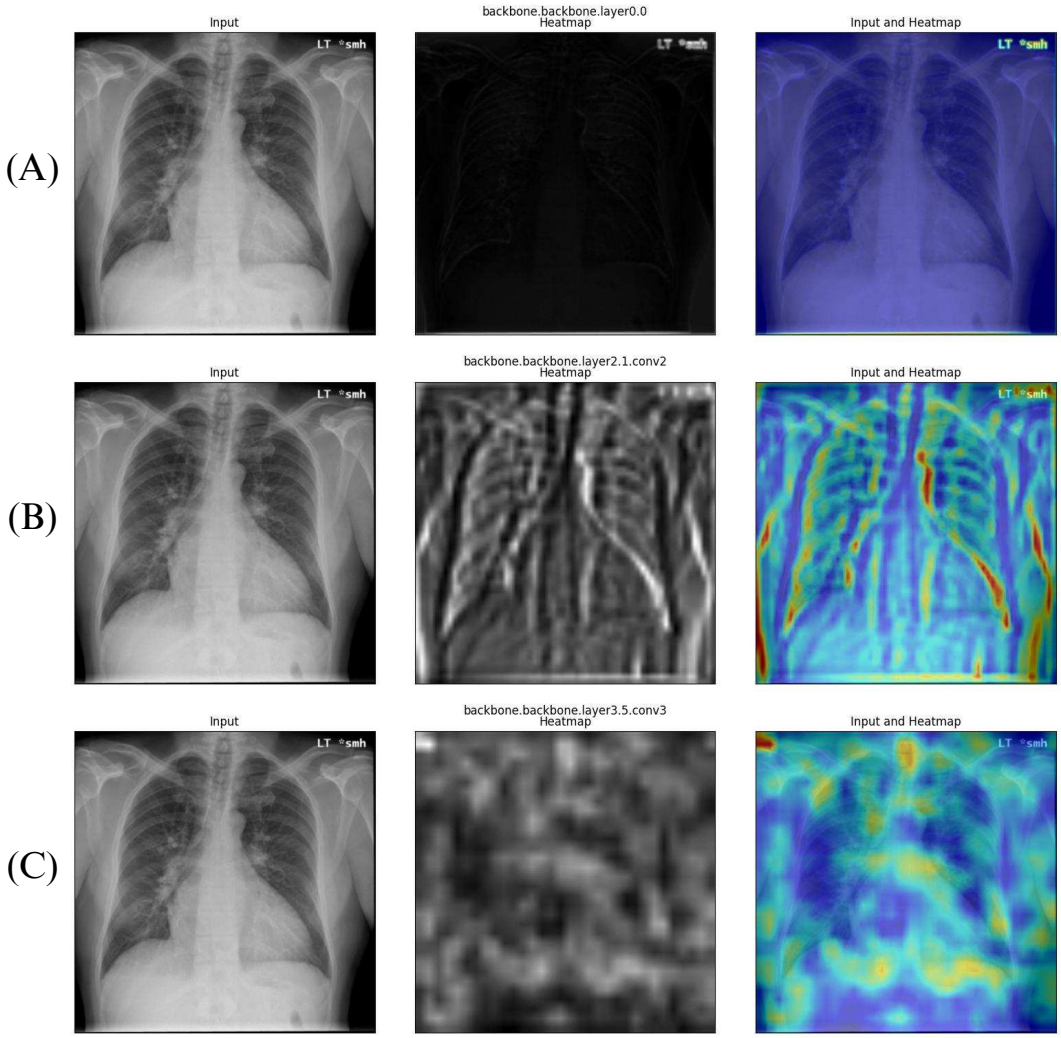


Figure 4.7: Grad-CAM visualization of vision encoder attention across three network depths (A - early, B - middle, C - late layers) when identifying cardiomegaly in chest radiograph. The given prompt was “the heart”, and the generated text was “the heart is moderately enlarged”.

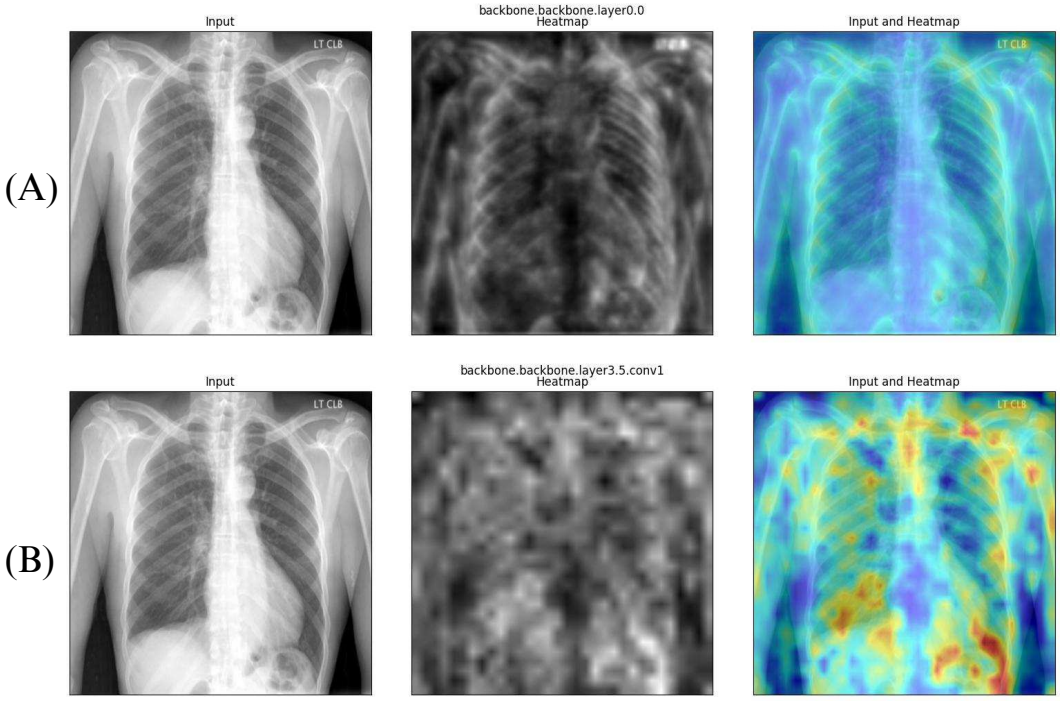


Figure 4.8: Grad-CAM visualization of vision encoder attention across three network depths (A - early, B - late layers) when identifying cardiomegaly in chest radiograph. The given prompt was “the heart”, and the generated text was “the heart size is normal”.

Chapter 5

Conclusions

In this study, we proposed a lightweight architecture for chest X-ray captioning that leverages pretrained models and domain-specific knowledge. Through extensive experiments, we demonstrated the effectiveness of our design choices, including the selection of U-Net as the image encoder and the initialization of the Q-Former module with BiomedBERT weights.

Our architecture achieved competitive performance while maintaining a compact size - with only 347 million parameters in its text decoder component (compared to 7 billion in competing approaches), attaining the highest precision (0.5142) and F1-score (0.4564) among evaluated methods while maintaining competitive recall (0.4102). This confirms that effective medical image captioning is possible without relying on computationally expensive models with billions of parameters.

The experimental results showed that our model, with only 347 million parameters in its text decoder component (compared to 6-540 billion in competing approaches), attained the highest precision (0.5142) and F1-score (0.4564) among evaluated methods while maintaining competitive recall (0.4102) score. This confirms that effective medical image captioning is possible without relying on computationally expensive models with billions of parameters.

Our qualitative analysis with an emergency medicine specialist provided critical insights beyond what computational metrics alone could reveal. Despite achieving competitive CE metrics, the specialist identified important areas for improvement in report content and structure. Importantly, this evaluation highlighted the contextual nature of report assessment; when informed that our model was trained on emergency department reports rather than formal radiological reports, the specialist acknowledged that evaluation criteria would differ. This recognition underscores the importance of aligning evaluation frameworks with the specific clinical context and intended use case of AI-generated reports.

Our work also makes an important methodological contribution through critical examination of current evaluation metrics. We demonstrated a concerning disconnect between standard metrics and clinical utility, observing that reports with perfect CE scores still received poor NLG scores. This observation questions the validity of current metrics for medical text evaluation and highlights how CE metrics reduce rich reports to categorical classifications, losing subtle information about severity or location. These findings

highlight limitations in conventional text generation evaluation methods when applied to medical reports and suggest the need for more clinically aligned evaluation frameworks.

Finally, additional analyses provided valuable insights into various aspects of the architecture. Notably, we discovered that using only frontal chest X-ray views yielded better results than incorporating multiple views, suggesting that simple image feature averaging may not be the optimal approach for aggregating information from different radiological projections.

5.1 Addressing the Research Questions

Regarding the first question of whether smaller in-domain pre-trained LLMs can effectively generate well-written captions for chest X-ray images, our findings provide a clear affirmative answer. The architecture using BioGPT (345 million parameters) as the text decoder successfully generated clinically relevant and coherent descriptions of chest X-rays. Despite being significantly smaller than competing models with 7 billion parameters, the model produced detailed and structured reports that accurately captured radiological findings. The qualitative analysis of generated reports confirms that the smaller LLM demonstrates adequate linguistic proficiency for the task, generating structurally sound reports with appropriate medical terminology and logical organization. As shown in the text samples (Figures 4.1, 4.2 and 4.3), the generated captions were often more descriptive than the original reports, providing comprehensive details about normal and abnormal findings. The model successfully maintained coherence across sentences and appropriately used domain-specific vocabulary, though occasional "hallucinations" (irrelevant or nonsensical phrases) were observed in some complex cases.

On the question of caption accuracy, the results suggest that smaller in-domain pre-trained LLMs can generate relatively accurate captions for chest X-ray images when compared to existing methods. The model achieved the highest precision (0.5142) and F1-score (0.4564) among compared methods, indicating a lower rate of false positive findings among the compared methods. While its recall (0.4102) was somewhat lower than the best-performing method, it was competitive, approaching the state-of-the-art. However, it's important to note that these metrics represent relative improvements rather than absolute measures of clinical adequacy. Furthermore, the limitations of current evaluation frameworks make it difficult to develop a comprehensive understanding of model performance in real emergency settings.

Analysis of performance across different findings categories (Table 4.1) reveals varying consistency in capturing diagnostic content. The model performed relatively well on common findings such as Support Devices (0.751 F1), Cardiomegaly (0.602 F1), and Pleural Effusion (0.595 F1), but struggled with less prevalent conditions such as Lung Lesion, Pneumothorax, and Fracture. This suggests that while the model can capture major diagnostic content in many cases, its reliability decreases significantly for rarer conditions or more subtle abnormalities. These performance disparities highlight not only room for improvement in capturing the full range of diagnostic content but also underscore the need for more nuanced evaluation metrics that better reflect clinical utility

across diverse radiological findings.

Furthermore, we demonstrated a concerning disconnect between standard metrics and clinical utility, observing that reports with perfect CE scores still received poor NLG scores. This observation questions the validity of current metrics for medical text evaluation and highlights how CE metrics may reduce rich reports to categorical classifications, losing nuanced information about severity, location, and progression.

Regarding the evaluation metrics, the research identified significant limitations in standard text generation metrics (BLEU, ROUGE, METEOR) for evaluating chest X-ray descriptions. Despite generating some clinically accurate reports (as measured by CE metrics), the model received relatively low NLG scores compared to other studies. Even in a manually selected generation subset with perfect CE scores, i.e., texts correctly describing what findings are present in the associated chest X-rays, the NLG metrics were poor because the generated text was written differently from what was expected. Therefore, the common metrics currently being utilized for analyzing medical generated text fail to capture clinical relevance and accuracy. The CE metrics (precision, recall, F1) based on the CheXbert labeler better reflect clinical relevance since they focus on the presence or absence of specific findings regardless of exact phrasing, which is more aligned with how specialists would assess report accuracy, even though these metrics still have limitations in capturing the nuanced linguistic aspects of medical reporting. Therefore, we still highlight the need for specialized evaluation frameworks that better correlate with expert medical knowledge and clinical judgment.

5.2 Limitations

Despite the promising results, this research faced several limitations that should be acknowledged. A significant challenge was the model's performance on the minority classes of findings, as Lung Lesion, Pneumothorax, and Fracture, for example. The model struggled to accurately identify such findings, achieving very low scores. This indicates a limitation in handling less prevalent chest X-ray conditions, which can be particularly problematic in emergency settings where accurately identifying critical findings can be essential for patient care.

A fundamental limitation of this work relates to dataset imbalance. The MIMIC-CXR dataset, as well as many medical datasets, exhibits significant class imbalance with certain findings being substantially more prevalent than others. For instance, common findings such as cardiomegaly and pleural effusion are well-represented, while rarer conditions such as pneumothorax and fractures appear in only a small fraction of the dataset. This imbalance directly impacted our model's performance, leading to biased predictions that favor common conditions while struggling with underrepresented ones. Therefore, in the future, techniques to address this would be essential, such as implementing advanced data augmentation method or designing different cost function what handle class imbalance.

Also, our studied approach to incorporate multiple X-rays views in the text generation process presented limitations. The simple approach to average the image features was not effective and potentially introduced noise to the diagnostic information rather than

enhancing it. This suggest that more sophisticated methods for multi-view integration should be analyzed and developed in order to leverage all complimentary information that different radiological projections could bring, which is a capability that physicians routinely employ in their diagnostic process.

Finally, the model behavior also showed limitations in consistency and reliability. The occasional generation of irrelevant or incorrect information, particularly for complex cases, indicates limitations in the model's ability to fully understand and interpret all radiological scenarios. These "hallucinations" represent a significant concern for clinical applications where factual accuracy is paramount. While our smaller model architecture demonstrated impressive capabilities overall, these instances of unreliable output highlight the ongoing challenges in developing trustworthy AI systems for healthcare.

5.3 Future Work

Our research lays the groundwork for several promising avenues of future exploration. Building upon our lightweight architecture for chest X-ray captioning, we envision a comprehensive research agenda focused on both technical advancement and clinical application.

A critical next step involves conducting comprehensive hardware cost analyses comparing our lightweight approach against off-the-shelf models like GPT-4 or available larger medical LLMs. This economic analysis would provide concrete evidence for the practical advantages of lightweight architectures in resource-constrained environments and inform deployment decisions for healthcare institutions with varying computational budgets.

The enhancement of the model's efficiency remains a priority, with opportunities to apply neural network optimization techniques such as distillation and quantization while preserving performance quality. Moreover, we see significant potential in leveraging domain-specific knowledge by pretraining both language and vision components on specialized medical datasets. By utilizing medical literature, radiology reports, and chest X-ray collections, we could develop models with deeper understanding of radiological findings and medical terminology.

Perhaps most significantly, our long-term vision extends beyond technical improvements to practical clinical applications. The system could be evolved into an assistive tool that can support emergency physicians in their reporting workflow. Beyond serving as a chest X-ray writing copilot to standardize reporting language and reduce documentation burden, the system could provide differential diagnosis support by suggesting findings that emergency physicians might not have considered based on the patient's presenting symptoms, potentially expanding the diagnostic workup when subtle or unexpected pathology is present.

Additionally, such a system holds significant promise as a training and educational tool for medical students and residents, highlighting subtle findings and demonstrating systematic radiological evaluation approaches that are fundamental to developing clinical expertise. This educational application could help bridge the gap between theoretical knowledge and practical pattern recognition skills that are essential in emergency radiol-

ogy interpretation. Such a comprehensive system would always position human experts at the center of the diagnostic process, with our technology serving in a supportive capacity to augment rather than replace clinical expertise, while providing cognitive support for both experienced physicians facing time pressure and trainees developing their diagnostic skills.

A critical area for advancement relates to evaluation methodologies for medical text generation. As discussed previously, current metrics such as BLEU, ROUGE, and even domain-specific measures such as Clinical Efficacy have significant limitations in capturing the nuances of medical language and clinical relevance. It would be beneficial to develop more sophisticated evaluation frameworks that better integrate expert medical knowledge and align with clinical judgment. This could involve creating hybrid metrics that assess not only the presence of findings but also their accurate description, location, severity, and clinical significance. Such improved metrics would provide more meaningful evaluation of model performance and better guide the development of systems that generate truly clinically useful text.

Bibliography

[1] K. Al-Dasuqi, M. H. Johnson, and J. J. Cavallo. Use of artificial intelligence in emergency radiology: An overview of current applications, challenges, and opportunities. *Clinical Imaging*, 89:61–67, 2022. 17, 22

[2] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. McDermott. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*, 2019. 49

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 25

[4] S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. 43

[5] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. 24, 30

[6] R. Bhayana. Chatbots and large language models in radiology: a practical primer for clinical and research applications. *Radiology*, 310(1):e232756, 2024. 16, 17

[7] W. Boag, H. Kané, S. Rawat, J. Wei, and A. Goehler. A pilot study in surveying clinical judgments to evaluate radiology report generation. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 458–465, 2021. 67

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 17, 27, 29, 32

[9] Z. Chen, Y. Song, T.-H. Chang, and X. Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020. 16, 35, 37, 39, 48, 66

[10] Z. Chen, Y. Shen, Y. Song, and X. Wan. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*, 2022. 16, 35, 37, 39, 48, 66

[11] Y. Cheng, D. Wang, P. Zhou, and T. Zhang. A survey of model compression and acceleration for deep neural networks. *arXiv preprint arXiv:1710.09282*, 2017. 25

[12] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, March 2023. 36, 65

[13] L. Chin-Yew. ROUGE: A Package for Automatic Evaluation of Summaries. In *Workshop on Text Summarization Branches*, Oct. 2004. 43

[14] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, and S. Gehrmann. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 29

[15] Claude Model. The Claude 3 Model Family: Opus, Sonnet, Haiku, 2025. `https://api.semanticscholar.org/CorpusID:268232499`. 16, 29

[16] J. P. Cohen, J. D. Viviano, P. Bertin, P. Morrison, P. Torabian, M. Guarrera, M. P. Lungren, A. Chaudhari, R. Brooks, and M. Hashir. TorchXRayVision: A library of chest X-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, pages 231–249. PMLR, 2022. 31, 46, 49, 63

[17] M. D. Danu, G. Marica, S. K. Karn, B. Georgescu, A. Mansoor, F. Ghesu, L. M. Itu, C. Suciu, S. Grbic, and O. Farri. Generation of Radiology Findings in Chest X-Ray by Leveraging Collaborative Knowledge. *arXiv preprint arXiv:2306.10448*, 2023. 35, 36, 37

[18] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 38, 39

[19] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 31

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 16, 27, 30, 49

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly. An image is worth 16×16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8, 27, 28

[22] R. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 32

[23] A. Gotmare, N. S. Keskar, C. Xiong, and R. Socher. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*, 2018. 54

[24] T. Gu, D. Liu, Z. Li, and W. Cai. Complex Organ Mask Guided Radiology Report Generation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7995–8004, January 2024. 35, 37, 66

[25] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2021. 49

[26] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, and X. Bi. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*, 2025. 16, 29

[27] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, and Y. Xu. A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Nachine Intelligence*, 45(1):87–110, 2022. 27

[28] I. Hartsock and G. Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence*, 7:1430984, 2024. 25

[29] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 26, 31

[30] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018. 31

[31] M. Iman, H. R. Arabnia, and K. Rasheed. A review of deep transfer learning and recent advancements. *Technologies*, 11(2):40, 2023. 30

[32] G. Irmici, M. Cè, E. Caloro, N. Khenkina, G. Della Pepa, V. Ascenti, C. Martinenghi, S. Papa, G. Oliva, and M. Cellina. Chest X-ray in emergency radiology: What artificial intelligence applications are available? *Diagnostics*, 13(2):216, 2023. 22, 24

[33] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, and K. Shpanskaya. Chexpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019. 44

[34] A. E. Johnson, T. J. Pollard, L. Shen, L.-W. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. MIMIC-III: A Freely Accessible Critical Care Database. *Scientific Data*, 3(1):1–9, 2016. 49

[35] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. MIMIC-CXR-JPG: A Large Publicly Available Database of Labeled Chest Radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 38, 50

[36] S. K. Karn, R. Ghosh, and O. Farri. SHS-NLP at RadSum23: Domain-adaptive pre-training of instruction-tuned LLMs for radiology report impression generation. *arXiv preprint arXiv:2306.03264*, 2023. 36

[37] H. Kasban, M. El-Bendary, and D. Salama. A comparative study of medical imaging techniques. *International Journal of Information Science and Intelligent System*, 4 (2):37–58, 2015. 22

[38] J. Kissane, J. A. Neutze, and H. Singh. *Radiology fundamentals: Introduction to imaging & technology*. Springer Nature, 2020. 8, 22, 23

[39] J. S. Klein and M. L. Rosado-de Christenson. A systematic approach to chest radiographic analysis. *Diseases of the Chest, Breast, Heart and Vessels 2019-2022: Diagnostic and Interventional Imaging*, pages 1–16, 2019. 15, 38

[40] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35: 22199–22213, 2022. 29

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 26

[42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 26

[43] Y. LeCun, Y. Bengio, and G. Hinton. Deep Learning. *Nature*, 521(7553):436–444, 2015. 24

[44] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 49

[45] H. Li, H. Wang, X. Sun, H. He, and J. Feng. Prompt-guided generation of structured chest X-ray report using a pre-trained LLM. In *IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2024. 35, 36, 37

[46] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 8, 17, 32, 33, 34, 45, 51

[47] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, and F. Wei. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *16th European Conference on Computer Vision*, pages 121–137, Glasgow, UK, Aug. 2020. Springer. 25, 32

[48] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 27

[49] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 54

[50] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):409–421, 2022. 30, 31, 36, 50

[51] K. Maharana, S. Mondal, and B. Nemade. A review: Data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3(1):91–99, 2022. 40

[52] A. Nicolson, J. Dowling, and B. Koopman. Improving Chest X-Ray Report Generation by Leveraging Warm Starting. *Artificial Intelligence in Medicine*, 144:102633, 2023. 65, 66

[53] G. O'Dwyer, M. T. Konder, C. V. Machado, C. P. Alves, and R. P. Alves. The current scenario of emergency care policies in Brazil. *BMC Health Services Research*, 13:1–10, 2013. 15, 17

[54] OpenAI. GPT-4 Technical Report. *arXiv 2303.08774*, 2023. 16, 36

[55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A Nethod for Automatic Evaluation of Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 43

[56] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019. 50

[57] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, and J. Clark. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 46

[58] V. M. Rao, D. C. Levin, L. Parker, A. J. Frangos, and J. H. Sunshine. Trends in utilization rates of the various imaging modalities in emergency departments: nationwide Medicare data from 2000 to 2008. *Journal of the American College of Radiology*, 8(10):706–709, 2011. 15

[59] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-Critical Sequence Training for Image Captioning. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 32

[60] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *18th International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, Munich, Germany, Oct. 2015. Springer. 10, 48

[61] A. W. Salehi, S. Khan, G. Gupta, B. I. Alabduallah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit. A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope. *Sustainability*, 15(7):5930, 2023. 31

[62] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally. From words to watts: Benchmarking the energy costs of large language model inference. In *IEEE High Performance Extreme Computing Conference*, pages 1–9. IEEE, 2023. 17

[63] C. Seibold, A. Jaus, M. A. Fink, M. Kim, S. Reiß, K. Herrmann, J. Kleesiek, and R. Stiefelhagen. Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling. *arXiv preprint arXiv:2306.03934*, 2023. 9, 31, 46, 47, 49, 63

[64] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pages 618–626, 2017. 68

[65] K. Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 26

[66] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, and S. Pfohl. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022. 65

[67] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. *arXiv preprint arXiv:2004.09167*, 2020. 44

[68] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015. 26

[69] A. Tabassum and R. R. Patil. A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, 7(06):4864–4867, 2020. 41

[70] O. Thawkar, A. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. Khan, J. Laaksonen, and F. S. Khan. XRayGPT: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023. 35, 36, 37, 65

[71] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large Language Models in Medicine. *Nature Medicine*, 29(8):1930–1940, 2023. 16

[72] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, and S. Bhosale. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 16, 32

[73] T. Vargas, H. Pedrini, and A. Santanchè. LLM-Driven Chest X-Ray Report Generation with a Modular, Reduced-Size Architecture. In *34th Brazilian Conference on Intelligent Systems*, pages 199–211, Belém-PA, Brazil, Nov. 2024. 45

[74] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017. 12, 27, 28, 32, 37

[75] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015. 32

[76] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 36

[77] Z. Wang, L. Liu, L. Wang, and L. Zhou. METransformer: Radiology Report Generation by Transformer with Multiple Learnable Expert Tokens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023. 35, 37, 66

[78] B. Yang, A. Raza, Y. Zou, and T. Zhang. Customizing General-Purpose Foundation Models for Medical Report Generation. *arXiv preprint arXiv:2306.05642*, 2023. 35, 36, 37, 65

[79] L. Yang, S. Xu, A. Sellergren, T. Kohlberger, Y. Zhou, I. Ktena, A. Kiraly, F. Ahmed, F. Hormozdiari, and T. Jaroensri. Advancing multimodal medical capabilities of Gemini. *arXiv preprint arXiv:2405.03162*, 2024. 16

[80] A. Zeng, B. Xu, B. Wang, C. Zhang, D. Yin, D. Zhang, D. Rojas, G. Feng, and H. Zhao. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*, 2024. 65

[81] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao. Unified vision-language pre-training for image captioning and VQA. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020. 32

[82] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 32