RadGraph: Integrating Fine-Grained and Global Features using GCNs and LLMs

by

Tejas Prakash Agrawal

August 21, 2024

A thesis submitted to the

Faculty of the Graduate School of

the University at Buffalo, The State University of New York

in partial fulfilment of the requirements for the

degree of

Master of Science

Department of Computer Science and Engineering

Acknowledgments

I wish to express my sincere gratitude to Dr. Nalini Ratha for his invaluable guidance and support over the past one and a half years. I also thank Dr. Mingchen Gao for her feedback during my thesis defense. Finally, I am grateful to my family for their unwavering support and confidence in me, which empowers me to move forward.

Abstract

The shortage of skilled radiologists and growing demand for radiology reports necessitate innovative solutions to improve diagnostic accuracy and efficiency. This thesis presents a novel method for automated radiology report generation by combining fine-grained local anatomical segmentation with global image features using Graph Convolutional Networks (GCNs) and large language models (LLMs).

Our approach segments anatomical regions in chest X-rays to produce detailed feature maps, which are then integrated with global X-ray features through GCNs. This method merges localized and global information, providing a comprehensive image representation. GCNs are particularly effective for handling variable-sized inputs and enhancing feature localization and globalization.

Empirical results indicate that integrating fine-grained anatomical features with global X-ray features significantly improves report accuracy and completeness. The use of LLMs further refines report quality, meeting clinical standards. Our evaluations show that this integrated approach achieves competitive BLEU-4, METEOR, and ROUGE-L scores compared to the current state-of-the-art models. We also review recent trends in automated report generation, such as Retrieval-Augmented Generation (RAG) and LLM-only architectures, and analyze the trade-offs between larger and domain knowledge-driven models.

This research advances medical imaging by offering a robust tool for automated radiology report generation, enhancing patient diagnosis and treatment outcomes while reducing radiologist workload. It also provides insights into current research directions and emphasizes the importance of integrating advanced AI technologies into clinical practice.

Table of Contents

List of Tables

List of Figures

Chapter1

Introduction

1.1 What is a Chest X-ray Radiology Report?

A radiology report is a critical medical document that provides an interpretation of findings from radiological images, such as chest X-rays. This report is integral to the diagnostic process, as it translates complex image data into actionable medical information. Radiologists examine these images to identify abnormalities, determine the presence of diseases, and assess the progression or resolution of conditions. The report typically includes a description of the findings, an assessment of their clinical significance, and recommendations for further diagnostic tests or treatments [1].

The report's content is structured to ensure clarity and comprehensiveness. It begins with a summary of the imaging procedure, followed by detailed observations of the anatomical structures examined. Findings are often categorized based on their severity and potential impact on patient health. For instance, a report may highlight significant findings such as tumors or fractures while noting less critical observations that may require monitoring. The radiology report is crucial not only for diagnosing and monitoring conditions but also for guiding subsequent clinical decisions and treatment plans [2]. In this work, we often use the terms Radiology report or Chest X-ray report interchangeably but they both refer to the same problem.

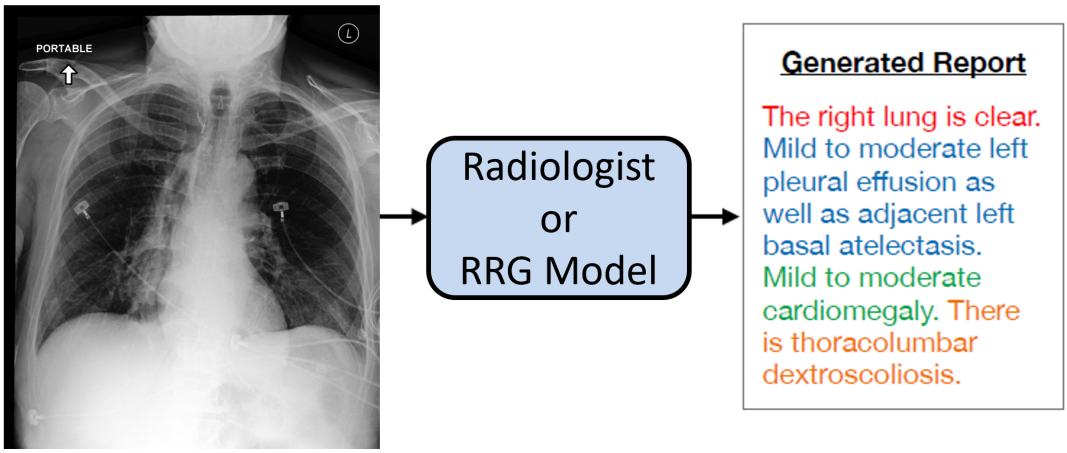


Figure 1.1: An example of a chest x-ray image and its corresponding radiology report generated by a Radiologist or a Radiology Report Generation (RRG) model.

1.2 Why Automate The Chest X-ray Report Generation ?

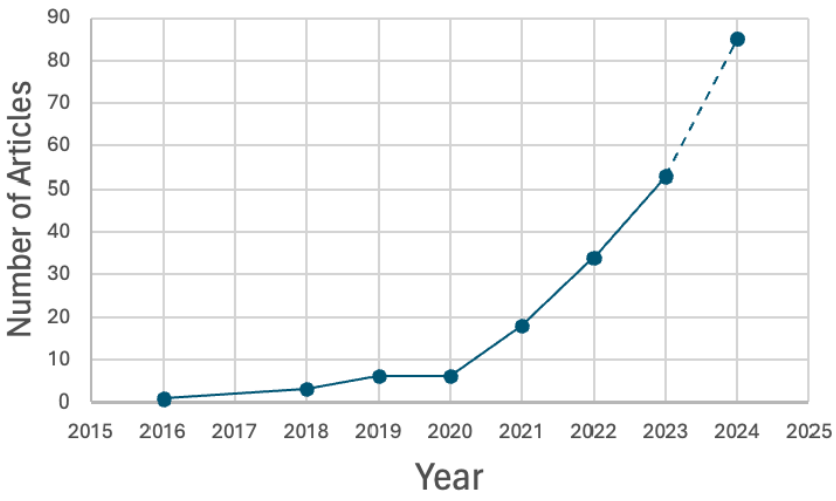


Figure 1.2: Increase in Articles related to radiology report generation. Figure taken from [3].

The automation of chest X-ray report generation is becoming increasingly urgent

due to the overwhelming demand for radiological services. In the UK, for example, 97% of imaging departments report being unable to keep up with the growing volume of imaging requests [3]. This imbalance between demand and capacity often leads to reporting delays, which can adversely affect patient care. Clinicians may be forced to make critical decisions based on preliminary or incomplete information, potentially compromising patient outcomes.

Automation has the potential to address these challenges by streamlining the report generation process. Automated systems can process large volumes of images more rapidly than human radiologists, reducing wait times and alleviating the burden on radiology departments. Moreover, by providing timely and consistent reports, automation can ensure that clinicians have access to reliable diagnostic information when making decisions, thereby improving patient management and treatment planning. Figure 1.2 shows the growing interest in automated report generation.

1.3 Challenges In Radiology Report Generation (RRG) For Chest X-ray

In addressing the complexities of chest X-ray report generation, several challenges must be overcome. This section explores the key obstacles, including data scarcity, interpretability issues, and the specialized knowledge required for effective reporting.

Data Scarcity: One of the primary challenges in automatic radiology report generation is the scarcity of high-quality, labeled datasets. Due to the sensitive nature of medical data and the complexity of image interpretation, acquiring large, annotated datasets for training AI models is difficult. The MIMIC-CXR [4] dataset, while the largest publicly available radiology dataset for chest X-rays, still has limitations in terms of data diversity and quality. This scarcity hampers the development of robust AI models capable of generating accurate and reliable reports.

Interpretability: Another significant challenge is the "black box" nature of many AI

models [5]. These models often provide outputs without transparent explanations, making it difficult for radiologists to trust and understand the generated reports. In the medical field, the rationale behind diagnostic decisions is crucial for acceptance and clinical integration. Without clear interpretability, the adoption of automated reporting systems may be limited, as radiologists need to be confident in the AI's decision-making process.

Specialized Knowledge Required for Chest X-ray Reporting: Generating accurate reports for chest X-rays involves not only the identification of anatomical structures but also an understanding of their various attributes and potential pathologies. The human chest X-ray contains numerous anatomical objects, such as the lungs, heart, and ribs, each with multiple attributes that must be considered. For example, detecting and describing a lung nodule involves assessing its size, shape, and position relative to other structures. Automated systems must be capable of accounting for these complexities to produce comprehensive and clinically useful reports.

Additionally, effective reporting requires integrating detailed anatomical knowledge with clinical context. Automated systems must be designed to recognize and interpret subtle variations in anatomical features and their implications for diagnosis and treatment. This necessitates advanced algorithms that can mimic the nuanced understanding of experienced radiologists, who combine technical expertise with clinical judgment to generate meaningful and actionable reports [6].

Integration with Clinical Workflows: Even if an system generates accurate reports, integrating it into clinical workflows presents another challenge. The system needs to work seamlessly with existing healthcare infrastructure, such as electronic health records (EHRs), PACS (Picture Archiving and Communication Systems), and hospital information systems. Moreover, the system should assist rather than hinder radiologists, augmenting their capabilities without overwhelming them with false positives or irrelevant suggestions.

Ethical and Legal Concerns: The deployment of report generation systems in healthcare raises important ethical and legal concerns. Issues such as patient privacy, data security,

liability in the case of errors, and the potential for bias in AI algorithms all need to be carefully addressed. There are also concerns about over-reliance on AI systems, which could lead to deskilling of radiologists or a lack of oversight in critical cases.

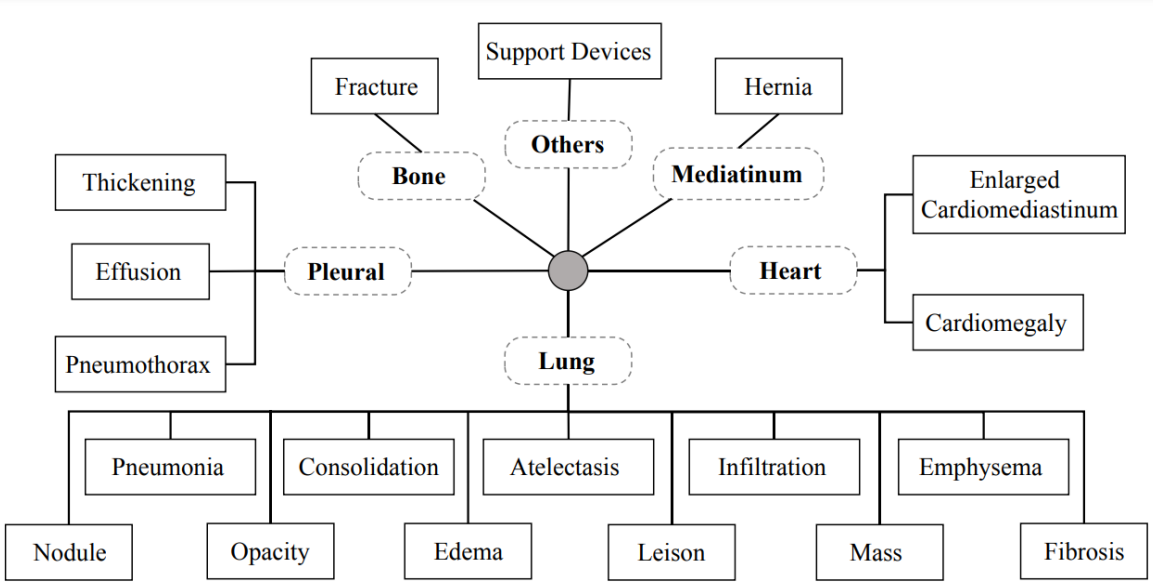


Figure 1.3: Example of the anatomical regions like lungs, heart, bone and its related attributes. Figure taken from [6].

Chapter2

Related Work

Radiology report generation (RRG) generally follows a cross-modal generation framework since it is similar to image captioning, though it differs in the length and complexity of the generated text [7] [8] [9] [10] [11] [12]. The goal is to generate a corresponding radiology report $\mathbf{R}$ from a given radiology image $\mathbf{I}$ by extracting essential semantic information from the image and producing an accurate and descriptive report. Most existing methods utilize an encoder-decoder architecture, where a visual encoder f_v extracts high-level semantics—such as latent representations, medical terms, or semantic graphs—from the image, represented as H_v. A text decoder f_t then transforms H_v into the descriptive text of the report $\mathbf{R}$. This process can be summarized as:

$$\mathbf{R} = f_t(H_v), \quad H_v = f_v(\mathbf{I}). \tag{2.1}$$

The current literature on radiology report generation encompasses five broad techniques discussed below. These techniques are not mutually exclusive but are distinguished by the core focus and methodology of each work.

2.1 Using Whole Chest X-ray Image For Global Features

One common approach in radiology report generation is treating the entire chest X-ray image as a single entity for feature extraction. Vision transformers (ViTs) [13] have emerged as a popular architecture in this domain, leveraging the ability to process global information from the full image without explicit localization of anatomical regions. Unlike traditional convolutional neural networks (CNNs), which operate on local receptive fields, vision transformers divide the image into patches and apply self-attention mechanisms to capture long-range dependencies across the entire image. This allows the model to holistically analyze the global structure of the X-ray, making it suitable for extracting complex patterns that span the whole image. However, this method may struggle to capture fine-grained details in specific anatomical regions, which could impact the accuracy of the generated report in certain clinical scenarios. Despite this limitation, treating the chest X-ray as a whole entity remains a prominent direction in the field of radiology report generation, particularly for large-scale datasets where generalization across diverse pathology is important. These studies [6] [14] [7] [15] [16] [17] [18] [19] [20] [21] harness the representational capabilities of transformers, augmented by domain-specific knowledge, to demonstrate that utilizing the entire chest X-ray image can yield strong performance in the task of radiology report generation (RRG).

2.2 Extracting Local Anatomical Visual Features

Another prominent approach in RRG focuses on extracting visual features from specific anatomical regions within the chest X-ray, such as the left lung, right lung, and heart. This method begins with segmenting these regions of interest (ROIs) to capture localized information. By isolating each anatomical structure, the model can generate region-specific

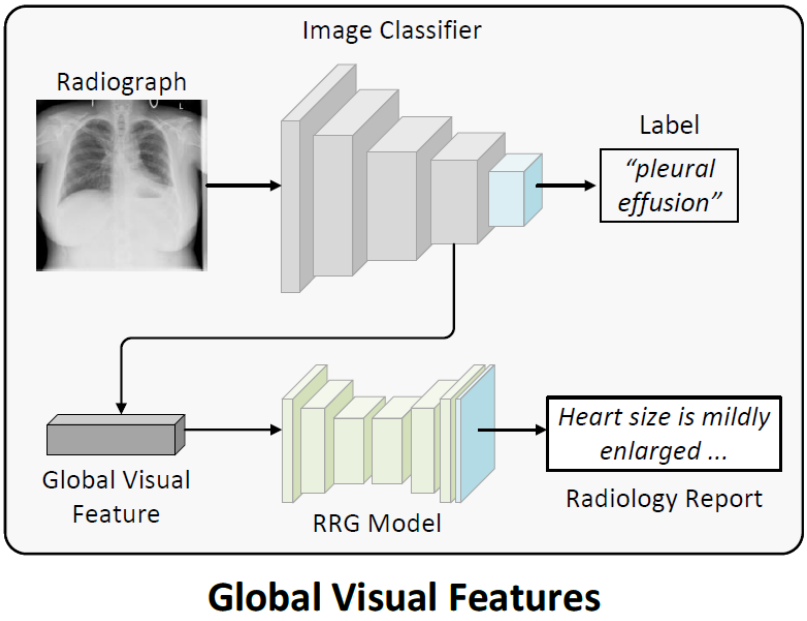


Figure 2.1: A generic architectures representing global visual features based models [22].

feature representations that more accurately reflect localized pathologies, which are often missed when treating the entire image as a single entity.

Once the features are extracted for each region, they are used to generate region-specific portions of the radiology report. This allows for more targeted descriptions of abnormalities in different parts of the chest, which is crucial in clinical practice where certain conditions are localized. After the generation of individual reports for each anatomical region, these descriptions are combined to form a cohesive and comprehensive radiology report. This modular approach improves interpretability and allows radiologists to focus on specific areas of concern, aligning the generated reports with clinical expectations and practices. Studies like [23] [24] [25] [26] [27] showcase the advantages of using these local features for RRG.

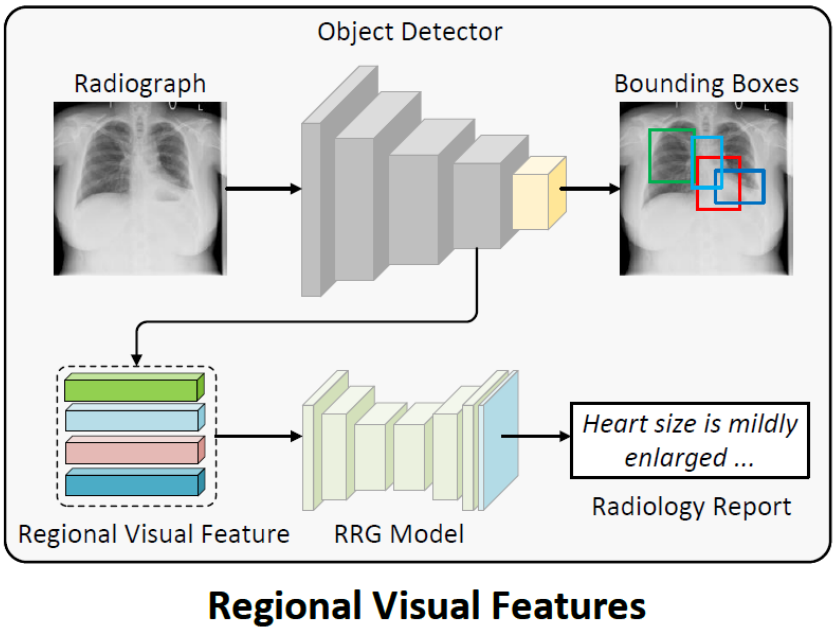


Figure 2.2: A generic architectures representing local visual features based models [22].

2.3 Fusion of Global And Local Features

A more advanced approach in Cheat X-ray radiology report generation (RRG) involves the fusion of global and local features to create a more comprehensive understanding of the chest X-ray image. This method combines the strengths of both global feature extraction, which captures overall patterns and structures across the entire image, and local feature extraction, which focuses on specific anatomical regions such as the lungs, heart, and diaphragm. By integrating these two levels of information, models can generate more accurate and detailed radiology reports.

The fusion of global and local features allows the model to balance the contextual information that spans the entire image with the precise details found in localized regions. For example, global features may help detect generalized patterns such as bilateral opacities, while local features can provide detailed descriptions of focal abnormalities like nodules or consolidations in specific regions. This holistic approach helps the model generate re-

ports that are not only more coherent but also more clinically relevant, as they take into account both the broader context and the finer details of the radiological findings. Techniques such as Graph Neural Networks (GNNs) and attention-based mechanisms are often employed to effectively integrate global and local information, enhancing the overall quality and completeness of the generated reports. Recent studies like [28] [29] [30] [31] [32] [33] have demonstrated that the fusion of global and local features effectively addresses the limitations inherent in relying solely on either approach. By integrating these two levels of feature representation, these works highlight the potential for improved diagnostic accuracy and more comprehensive radiology report generation.

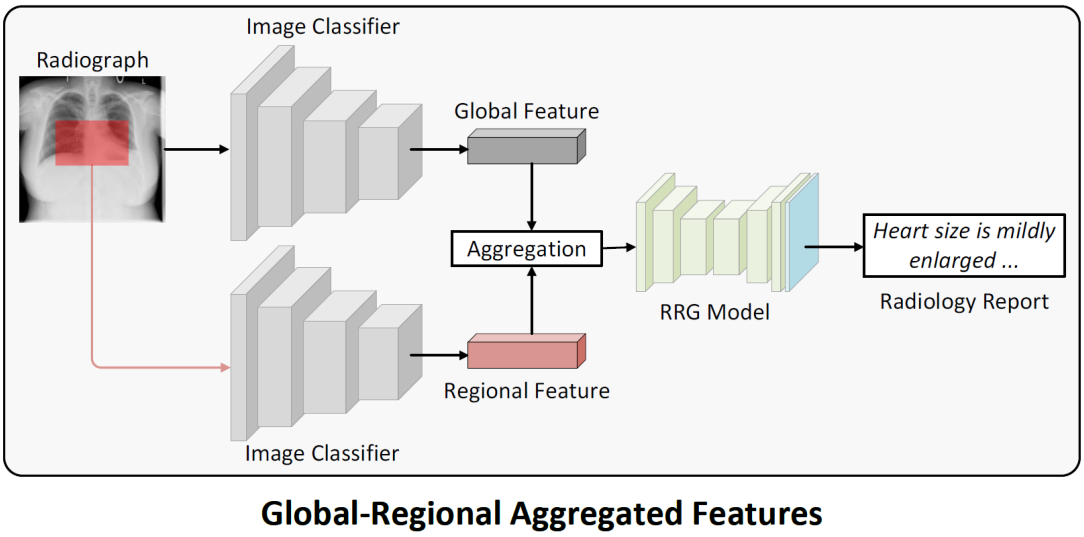


Figure 2.3: A generic architectures representing the global-local fusion based models [22].

2.4 Large Medical Domain LLMs

In recent advancements in Chest X-ray radiology report generation (RRG), large language models (LLMs) pre-trained on medical domain data have been employed with prompt engineering [34] [35] or task-specific fine tuning techniques [36] to enhance the quality and relevance of generated reports. Models such as those with up to 84 billion parameters lever-

age extensive training on medical texts to provide high expressive power and contextual understanding specific to radiology. This pre-training allows these LLMs to utilize information from diverse medical datasets, enhancing their ability to generate accurate and coherent radiology reports by drawing on a broad spectrum of medical knowledge.

The use of these large-scale LLMs offers several advantages. Their substantial capacity allows them to capture and integrate complex medical concepts, resulting in reports that are both comprehensive and contextually appropriate. Additionally, their training on extensive medical datasets enables them to provide relevant insights and details that improve the quality of the generated reports. However, there are notable challenges associated with these models. Issues such as hallucination, where the model generates plausible but incorrect or misleading information, can undermine the reliability of the generated reports. Furthermore, the substantial size of these LLMs poses significant drawbacks, including high memory footprint, reduced interpretability, and increased inference time. These factors can complicate the deployment and practical use of such models in clinical settings. Studies like [37] [38] [39] are some of the other studies that leverage LLMs for the radiology report generation and also showcase the problems of LLMs like hallucinations. The current state-of-the-art for Automated Radiology report generation [40] uses fine-tuned LLMs that can generate reports based on multiple inputs like frontal view, lateral view and previous reports.

2.5 Retrieval Augmented Generation (RAG)

In recent years, Retrieval-Augmented Generation (RAG) [41] has gained significant popularity in various generation tasks, including RRG. The core idea of RAG is to enhance generative models by incorporating external knowledge through retrieval mechanisms. Instead of relying solely on pre-trained language models, RAG systems retrieve relevant documents or information from external databases during the generation process, thereby grounding the output in real-world data and improving factual accuracy.

RAG addresses several key challenges inherent in large language models (LLMs), particularly the issue of hallucination—where the model generates content that is plausible but factually incorrect [42]. By retrieving and incorporating relevant external knowledge, RAG can mitigate this problem, ensuring that the generated reports are both accurate and reliable. This approach is especially valuable in the medical domain, where the correctness of generated information is critical. When applied to chest X-ray report generation, RAG can retrieve pertinent medical literature, case studies, or similar X-ray reports, providing a robust foundation for generating precise and contextually appropriate radiology reports.

In the context of chest X-ray report generation, RAG has the potential to significantly improve the quality and reliability of the generated reports. By leveraging a retrieval mechanism, RAG models can access relevant external data, such as previous radiology reports or clinical guidelines, which can help in generating more detailed and accurate reports. This integration of retrieval with generation offers a promising avenue for overcoming some of the limitations of traditional LLM-based approaches in RRG, particularly in ensuring that the generated content aligns with real-world medical knowledge. Recently [43] has show that RAG coupled with LLMs are a powerful tool for Radiology Report generation. Additionally studies [44] [45] has showcases the use of RAG.

Chapter3

Methodology

3.1 Overview

In this methodology, we propose a multi-stage approach for automated radiology report generation. First, 29 anatomical regions are detected in frontal chest X-rays using a Faster-RCNN model, which extracts feature representations for each region. These features are then filtered using a binary classifier, referred to as the Region Significance Classifier, to retain only **S** number of features relevant to report generation.

Simultaneously, global image features are extracted using a CLIP model. These global features are concatenated with the filtered local features, combining both levels of information. The concatenated feature set is passed through three graph convolution layers, facilitating information sharing between regions and further integrating global and local feature representations.

Finally, the enriched feature vectors are decoded using the Llama LLM. Each representation is fed into the LLM one at a time after project from 1024 dimension to 4096 dimension due to Llama's token representation vector size, which generates sentences autoregressively, one word at a time. A pseudo self-attention mechanism is employed in the first layer of the LLM to integrate region-specific features during the report generation process.

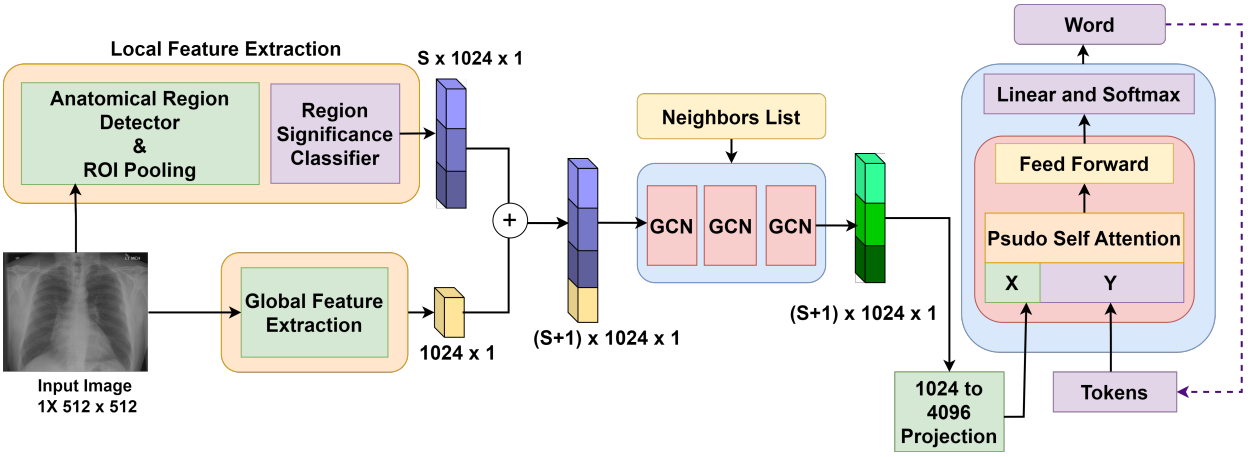


Figure 3.1: An overview of the proposed architecture.

3.2 Modules

3.2.1 Anatomical Region Detection and local feature extraction

For anatomical region detection, we employ the Faster R-CNN [46] model with a ResNet-50 backbone, pre-trained by [23] then fine-tuned by us for our objective function. The Faster R-CNN framework includes a Region Proposal Network (RPN), which generates object proposals—bounding boxes that potentially contain anatomical regions—based on feature maps extracted by the ResNet-50 backbone from the input chest X-ray image. A Region of Interest (RoI) pooling layer then maps each object proposal onto the backbone feature maps, extracting uniform-sized feature maps for each proposal. These RoI feature maps are classified into one of 30 classes, 29 anatomical region classes and 1 background in accordance with the standard Faster R-CNN procedure.

To extract the visual features of the 29 anatomical regions, we identify the "top" object proposal for each class. The top object proposal for a given region class is determined by selecting the proposal with the highest class probability score among all proposals. If a region class does not achieve the highest score in any proposal, it is considered undetected and is excluded by the region selection module. The visual features of the selected regions,

represented as a matrix of size $\mathbb{R}^{29\times1024}$, are derived from the 29 RoI pooling layer feature maps $\mathbb{R}^{29\times2048\times H\times W}$. The spatial dimensions are reduced through 2D average pooling, and the dimension of the feature maps is reduced from 2048 to 1024 via a linear transformation. Figure 3.2 is a broad overview of a Faster R-CNN workings.

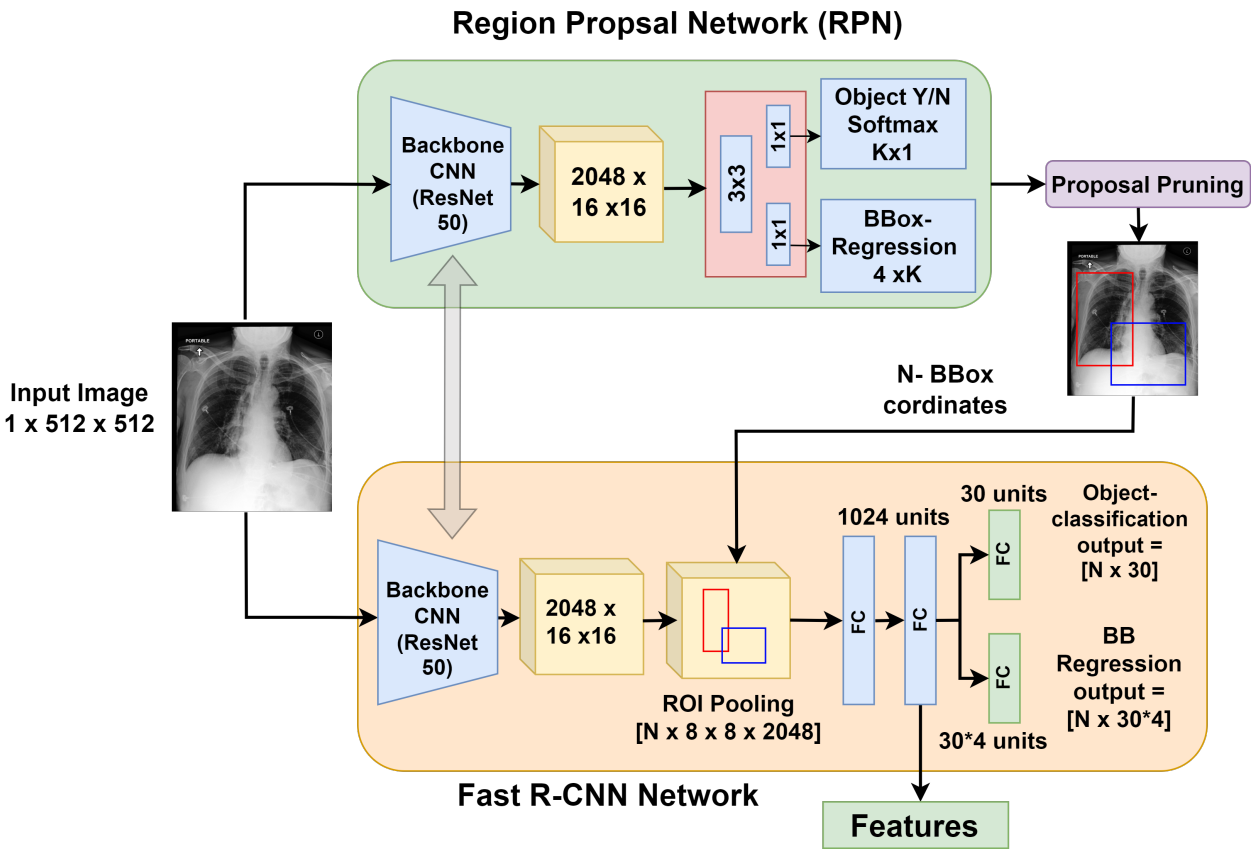


Figure 3.2: The Anatomical Region detection module that uses Faster-RCNN.

3.2.2 Global Feature Extraction

The Global Feature Extraction module processes the entire Chest X-ray image by passing it through a CLIP [47] model, which was initially pre-trained on the MIMIC-CXR dataset by [48] and subsequently fine-tuned by us for our specific objective. CLIP has gained popularity due to its effectiveness in extracting robust image features and its seamless alignment with

textual information, making it a strong candidate for image to text tasks.

The output of the global feature extraction process is a feature vector of size $\mathbb{R}^{1024}$. This global feature vector serves as a contextual representation of the entire image and is subsequently used to provide each anatomical region's local features with a broader contextual understanding of the global image. By incorporating this global context, the model can generate more coherent and accurate radiology reports.

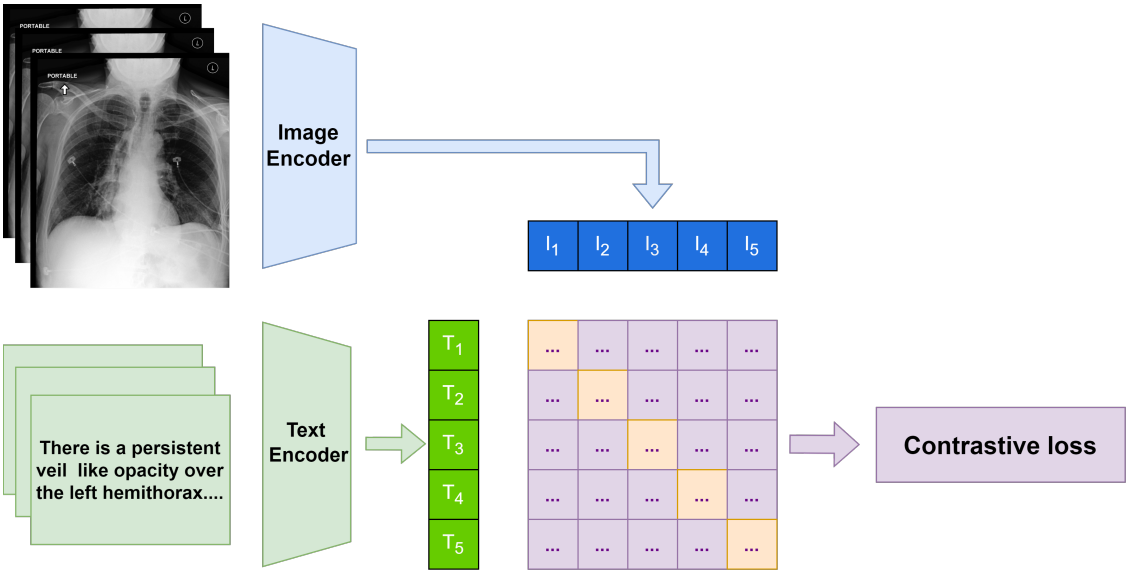


Figure 3.3: Pre-training of CLIP model for Chest X-ray images and corresponding text reports [47].

3.2.3 Graph Convolution Network

Graph Convolution Networks (GCNs) [49] extend the concept of convolution to graph structures, allowing us to operate on non-Euclidean data. Unlike traditional 2D convolution, where a weighted sum is computed over a fixed spatial neighborhood of pixels, graph convolution defines neighborhoods based on the graph structure. This flexibility allows us to define arbitrary neighborhoods, enabling GCNs to adapt to various data types, including those with irregular or complex relationships between nodes. Figure 3.4 shows the difference

between the structure of a 2D convolution and a Graph convolution. Figure 3.5 explains the graph convolution operation.

In our framework, after extracting local anatomical features, we append the global feature vector to form a set of $S + 1$ feature vectors, where S represents the number of anatomical regions. Each of these $S+1$ feature vectors is treated as a node in a graph. The connectivity of this graph is defined such that each node is connected to four other nodes: the global feature node and its three nearest neighboring nodes. The nearest neighbors are determined based on the top-3 most frequently co-occurring regions in our training dataset.

This graph is then processed through three layers of graph convolution, where each layer performs convolution operations on the graph nodes based on the defined neighborhood structure. These graph convolution operations propagate information across nodes, enabling the fusion of global and local features in a more structured manner, thereby enhancing the overall representation for generating accurate and coherent radiology reports. Figure 3.6 shows the general structure of a Graph convolution network. After the Graph Convolution, each regional feature has some information shared from other features, making them robust and context-aware.

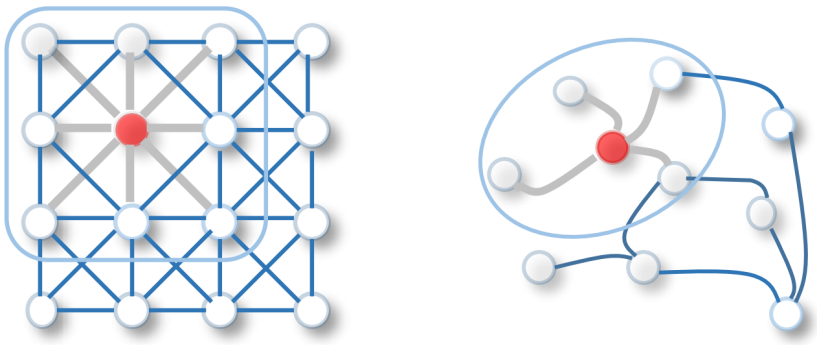


Figure 3.4: Difference between the structure of a Image with 2D convolution and a Graph convolution on a graph [50].

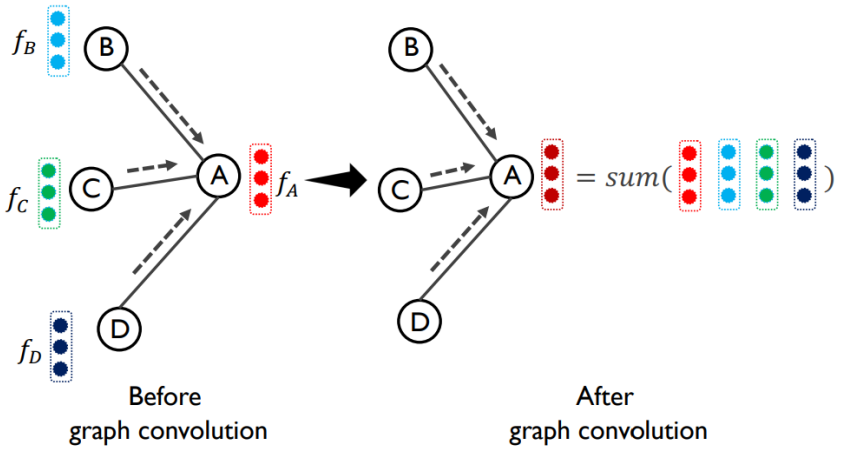


Figure 3.5: The Graph Convolution operator, **f** is the feature vector and the output of Graph convolution on node A with its neighbours B, C, D is a weighted sum of the features of A, B, C & D [51].

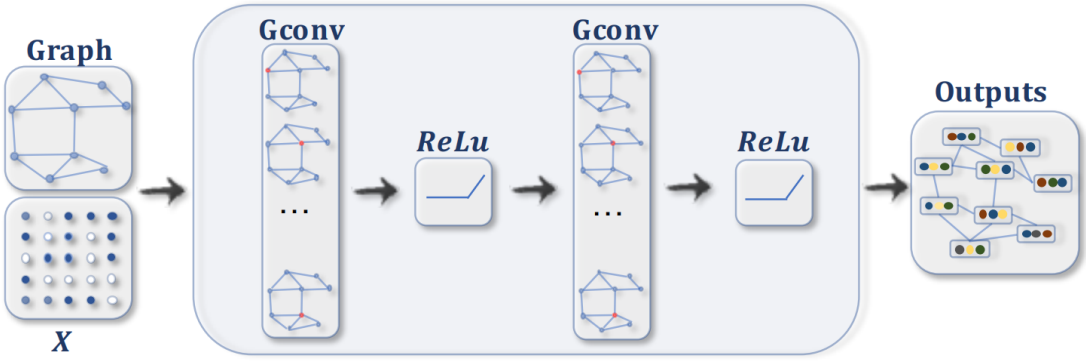


Figure 3.6: The Graph Convolution network, Each of the **S+1** feature vectors is considered a node in the **Graph** and the **X** is the neighbours list for the graph [50].

3.2.4 Large Language Model

For the language modeling component, we utilize the 8-billion-parameter Llama3 LLM model [39]. Llama3 operates as an autoregressive neural network based on self-attention mechanisms, where the generation of each token is conditioned on the preceding tokens in the sequence. We pass each of the output feature vectors from the GCNs to a Dense network for projecting to 4096 dimension representation one at a time then to the LLM for decoding till end-of-sentence token is generated. The self-attention mechanism can be formulated as:

$$\mathrm{SA}(Y) = \mathrm{softmax}\left((YW_q)(YW_k)^\top\right)(YW_v), \tag{3.1}$$

where Y denotes the token embedding, and W_q, W_k, and W_v are the projection matrices for queries, keys, and values, respectively.

To integrate region visual features into the language model, we adapt the approach of pseudo self-attention , as outlined by . This technique involves incorporating the region visual features directly into the self-attention mechanism of the model. We only modify the first attention layer of the LLM The modified pseudo self-attention can be expressed as:

$$\mathrm{PSA}(X, Y) = \mathrm{softmax}\left((YW_q)\begin{bmatrix} XU_k \\ YW_k \end{bmatrix}^\top\right)\begin{bmatrix} XU_v \\ YW_v \end{bmatrix}, \tag{3.2}$$

where X represents the region visual features, and U_k and U_v are newly initialized projection parameters for keys and values, respectively. Y represents the token from previous auto-regression step. This approach allows the model to generate text based on both the preceding tokens and the visual features of the regions.

The projection from a 1024-dimensional feature space to a 4096-dimensional space poses a significant computational bottleneck in our current model architecture. To address this

challenge, future research should explore more efficient methods for dimensional expansion. This could involve leveraging advanced projection techniques that optimize the transformation process, ensuring that essential feature information is retained while reducing computational overhead. Alternatively, a promising avenue could be the distillation of large language models (LLMs), wherein the model's representational capacity is compressed from 4096 dimensions back to 1024. Model distillation offers dual advantages: it can significantly decrease the number of parameters in the model, which in turn reduces both memory footprint and inference time. Importantly, this approach aims to maintain, if not enhance, the model's performance by retaining critical learned knowledge during the compression process. Therefore, future work should focus on integrating LLM distillation strategies to balance model efficiency with high-quality output in medical report generation tasks.

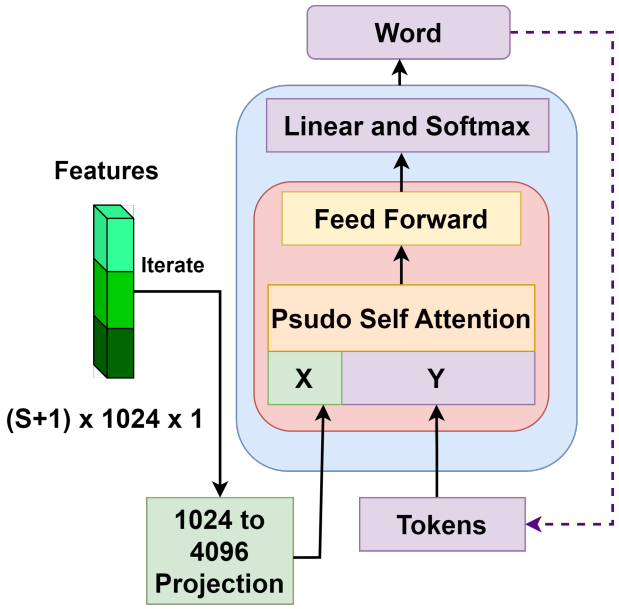


Figure 3.7: The workflow of decoding the feature vectors into natural language text using Llama3. Each feature vector is decoded one at a time.

3.3 Training

The training process for the model is executed in three distinct phases. Initially, only the Anatomical Region Detector is trained. In the second phase, this detector is integrated with the Region Significance Binary Classifiers and trained together. Finally, the complete end-to-end model, with all parameters trainable, is trained. During the training of the language model, only the region visual features associated with reference sentences are utilized, under the assumption that the region selection module will accurately identify these regions during testing. For instances where multiple sentences correspond to a region, these sentences are concatenated to enable the model to learn how to generate multiple sentences in such cases. The following is the Loss function with which the whole model is trained:

$$\mathcal{L} = \lambda_{\text{obj}} \cdot \mathcal{L}_{\text{obj}} + \lambda_{\text{sig}} \cdot \mathcal{L}_{\text{sig}} + \lambda_{\text{language}} \cdot \mathcal{L}_{\text{language}} \tag{3.3}$$

The loss function used in this model comprises three components: the anatomical region detection loss ($\mathcal{L}_{\text{obj}}$), the region significance classification loss ($\mathcal{L}_{\text{sig}}$), and the language generation loss ($\mathcal{L}_{\text{language}}$). The anatomical region detection loss focuses on regressing bounding boxes and applying Binary Cross Entropy (BCE) for detecting anatomical regions, while also incorporating dynamic class sensitivity weighting to account for varying class importance [52]. The region significance classification loss is computed using BCE, aiming to accurately identify significant regions that contribute to the final report generation. Finally, the language generation loss is handled by Cross Entropy, which ensures the model generates coherent and relevant text for the radiology reports. These three loss components are balanced by their respective weights, λ_{obj}, λ_{sig}, and $\lambda_{\text{language}}$, during the model's training process.

We use effective batch size of 64, max epochs as 40, Learning rate as 5e-05 with a learning rate scheduler, Number of beans for beam search as 4, a BERT similarity score of 0.9 for

removing similar sentences and NVIDIA H100 GPUs for training our model.

3.4 Inference

The final radiology reports are assembled by combining the generated sentences from the identified anatomical regions. In cases where certain pathologies extend across multiple regions or when anatomically similar areas (e.g., left and right lung) are free from abnormalities, the generated text may include repetitive or identical sentences. To address this redundancy, BERTScore [53] is employed to measure the similarity between sentences. When duplicates are detected, the shorter sentence is discarded in favor of the longer one, as the latter typically contains a more comprehensive and clinically valuable description.

Chapter4

Experiments and Ablation Study

4.1 Dataset and Pre-processing

We utilize the Chest ImaGenome v1.0.0 [54] dataset to train and assess our model. This dataset is derived from the MIMIC-CXR [4] collection, which includes chest X-ray images along with their associated free-text radiology reports. The Chest ImaGenome dataset offers automatically generated scene graphs for the MIMIC-CXR images. These scene graphs provide detailed descriptions of individual frontal chest X-ray images, including bounding box coordinates for 29 distinct anatomical regions in the chest. Additionally, they include sentences corresponding to these regions, extracted from the linked radiology reports when available. We adopt the official dataset split, resulting in 166,512 images for training, 23,952 for validation, and 47,389 for testing.

All images are resized to 512x512 pixels, maintaining the original aspect ratio, with padding applied if necessary. They are normalized to have zero mean and unit standard deviation. During training, data augmentation techniques such as color jitter, Gaussian noise, and affine transformations are employed to enhance the robustness of the model. For the textual data, redundant whitespaces (e.g., line breaks) are removed. In line with previous research, the findings section of the radiology reports from the MIMIC-CXR dataset is used as the reference for the report generation task. This section encompasses the radiologist's observations. Reports with empty findings sections are excluded, leaving approximately 149,000 images and reports for training, approximately 28,000 test images with corresponding

reference reports and approximately 14000 for validation. No additional processing is applied to these extracted reports, unlike in other studies.

4.2 Evaluation Metrics

We assess the model using widely recognized natural language generation (NLG) metrics, including BLEU [55], METEOR [56] and ROUGE-L [57]. These metrics evaluate the similarity between the generated and reference reports by identifying matching n-grams (i.e., overlapping words). For sentence-level evaluation, we primarily use METEOR, which is suitable for both sentence- and report-level assessments, unlike metrics such as BLEU. However, traditional NLG metrics do not effectively capture the clinical correctness of generated reports. Therefore, we also report clinical efficacy (CE) metrics as seen in previous work [23]. These CE metrics evaluate the generated and reference reports based on the presence or absence of key clinical observations, providing a measure of diagnostic accuracy.

4.3 Ablation Study

4.3.1 Global and Local Features Significance

The model's performance was evaluated using global features (GF) alone, local features (LF) alone, and a fusion of both. The use of global features yielded sub-optimal performance due to the simplicity of the CLIP model. Local features, while demonstrating good performance independently, achieved the best results when fused with global features. Consequently, the final architecture incorporates this fusion approach to optimize performance. Table 4.1 shows the performance.

4.3.2 Effect of different global feature extractors

We experimented with a basic ResNet50 as a global feature extractor and a fine-tuned ResNet50 backbone from CXR-CLIP. The CXR-CLIP variant demonstrated superior performance compared to the basic ResNet50. Table 4.2 shows that CXR-CLIP performance better hence used in the final architecture.

4.3.3 Effect of Dynamic Class Sensitivity (DCS) in loss function for local anatomical feature extraction

Certain anatomical regions are more challenging to detect than others. To address this, we incorporate dynamic class sensitivity into the loss function, penalizing the model more heavily when it fails to identify regions from classes with the lowest detection accuracy in the previous validation cycle. Class weights are determined dynamically based on their accuracy [52]. Table 4.3 shows that DCS gives better performance hence used in the final architecture.

4.3.4 Effect of different number of GCN layers

The number of GCN layers is a hyperparameter that must be empirically determined. Our experiments indicate that, for this specific problem, using three GCN layers yields the optimal performance. Table 4.4 shows the performance of each setup with different number of layers.

4.3.5 Effect of different LLMs

We evaluated the performance of different-sized language models by using a smaller GPT-2 Medium and a more recent and bigger LLaMA model with 8 billion parameters to generate natural language reports from features. As anticipated, the Llama 8B model significantly outperformed the GPT-2 Medium. Table 4.5 shows the performance of both the LLMs.

Table 4.1: Effect of using only Global features (GF), only Local Features (LF) and using both.

Experiment	**Using only GF**	**Using only LF**	**Using GF + LF**
BLUE-1	0.163	0.369	**0.381**
BLUE-2	0.128	0.249	**0.255**
BLUE-3	0.127	0.175	**0.177**
BLUE-4	0.112	0.121	**0.127**
METEOR	0.149	0.168	**0.172**
ROUGE-L	0.221	0.264	**0.280**

Table 4.2: Effect of using simple image encoder vs CLIP-CXR encoder.

Experiment	**Using simple Image encoder**	**Using CLIP-CXR**
BLUE-1	0.373	**0.381**
BLUE-2	0.249	**0.255**
BLUE-3	0.175	**0.177**
BLUE-4	0.126	**0.127**
METEOR	0.168	**0.172**
ROUGE-L	0.264	**0.280**

Table 4.3: Effect of Dynamic Class Sensitivity (DCS) for local features.

Experiment	**No DCS**	**With DCS**
BLUE-1	0.373	**0.381**
BLUE-2	0.249	**0.255**
BLUE-3	0.175	**0.177**
BLUE-4	0.126	**0.127**
METEOR	0.168	**0.172**
ROUGE-L	0.264	**0.280**

Table 4.4: Effect of using different number of GCN layers.

Experiments	**2-layers**	**3-layers**	**4-layers**
BLUE-1	0.373	**0.381**	0.380
BLUE-2	0.249	**0.255**	0.255
BLUE-3	0.175	**0.177**	0.176
BLUE-4	0.126	**0.127**	0.125
METEOR	0.168	**0.172**	0.169
ROUGE-L	0.264	**0.280**	0.270

Table 4.5: Effect of using different LLMs.

Experiment	**GPT2-Medium**	**Llama-3**
BLUE-1	0.371	**0.381**
BLUE-2	0.245	**0.255**
BLUE-3	0.176	**0.177**
BLUE-4	0.127	**0.127**
METEOR	0.170	**0.172**
ROUGE-L	0.267	**0.280**

4.4 Experiments With RAG

Retrieval-Augmented Generation (RAG) has recently gained prominence for reducing hallucinations in LLMs, prompting us to explore its application in radiology report generation, as seen in similar studies. Study [58] use a conventional RAG approach with global image features, but our experiments show that incorporating local features significantly improves report quality. We propose an advanced RAG framework that extracts and utilizes region-specific features, enhancing radiology report generation.

For our experiments, we constructed a subset of 1,000 data points from the training dataset to create a database consisting of region-wise image features and corresponding radiology report phrases. The process begins by passing an input chest X-ray image through an anatomical region detector. The segmented images are then processed using CXR-CLIP to obtain feature vectors. These feature vectors are matched based on cosine similarity scores greater than 0.95, allowing us to retrieve relevant phrases for each anatomical region from the database. The retrieved phrases are fed to a fine-tuned LLM, like ChatGPT-4 or LLaMA, which synthesizes common pathological findings and generates coherent sentences for each region, creating the final radiology report. Figure 4.1 illustrates the proof-of-concept architecture for this RAG approach. We use 5 sample input images and take 5 regions out of the 29 to calculate the scores for our proof of concept. Table 4.6 shows that out approach has a high meteor score which means the meaning of the generated report and actual report is close.

Example: Ground truth phrase : "A minimal left pleural effusion is also present. Mild cardiomegaly with mild pulmonary edema. No pneumothorax." and the findings based on similar features with cosine similarity of greater than 0.95 is "There are mild to moderate bilateral pleural effusions and bibasilar atelectasis, particularly in the retrocardiac regions. Mild to moderate pulmonary edema is noted with stable cardiomegaly. No pneumothorax is seen.". As we can see, the pathologies are correctly identified using the approach. A full

scale architecture with extensive experiments will we done in the future.

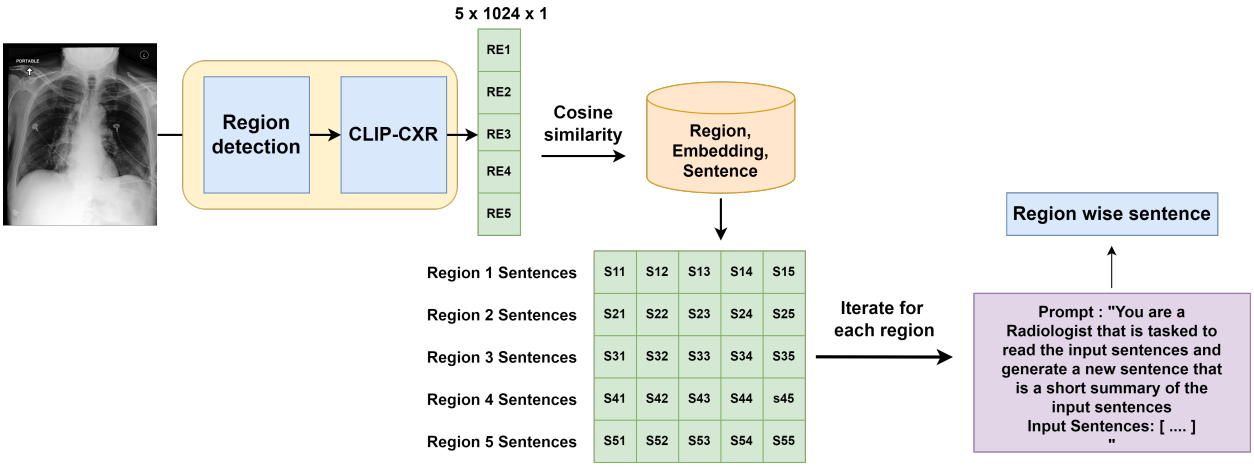


Figure 4.1: An overview of the proof of concept for RAG as a viable way for Radiology report generation.

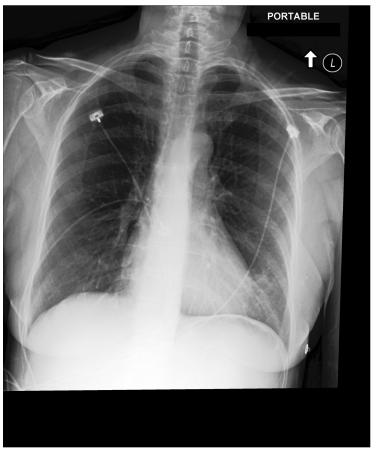
Table 4.6: NLG scores for our RAG approach over a small sample test set of 10 input images with 5 regions each.

Experiment	**Scores**
BLUE-1	0.221
BLUE-2	0.137
BLUE-3	0.089
BLUE-4	0.049
METEOR	0.329
ROUGE-L	0.242

Chapter5

Results

5.1 Sample Outputs



Generated report:
The lungs are clear. There is no pleural effusion or pneumothorax. The mediastinal and hilar contours are unremarkable. The heart size is normal. No acute cardiopulmonary process. The cardiomediastinal silhouette and hilar contours are normal. No free air below the right hemidiaphragm is seen. Left rib fractures are detected. Surgical clips on breasts noted

Reference report:
The lungs are clear of focal consolidation, pleural effusion or pneumothorax. The heart size is normal. The mediastinal contours are normal. Multiple surgical clips and old left rib fractures are noted.

Figure 5.1: Different colors font shows similar sentences in generated and reference report, abnormalities are detected and report accurately, there is one spurious sentence that does not belong.

5.2 Evaluation

Our model demonstrates competitive performance in radiology report generation relative to state-of-the-art (SOTA) methods. While it does not surpass the SOTA, it achieves comparable scores in Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics

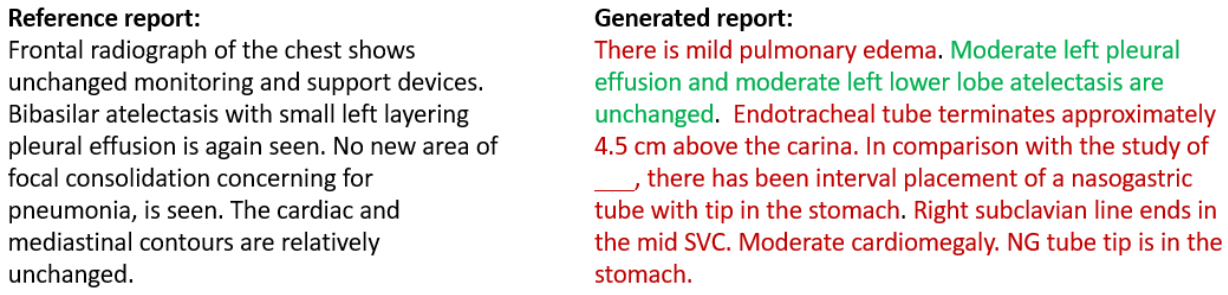


Figure 5.2: Green sentences are seen in generated as well as reference report. Red sentences not present in reference report, could be hallucinations by LLMs. This example highlights the hallucination problem of the LLMs where it generated a highly plausible but inaccurate sentence.

with significantly fewer model parameters. The METEOR score is particularly relevant for evaluating NLG, as it considers the semantic meaning of sentences. BLEU and ROUGE-L scores are utilized to assess the fluency and semantic accuracy of the generated reports. It is important to note that the current SOTA model, MAIRA-2, incorporates additional inputs such as frontal and lateral views, as well as previous reports, which contribute to its enhanced performance. Table 5.1 presents a comparison of NLG performance with other radiology report generation approaches.

Clinical Efficacy (CE) is another crucial evaluation metric for radiology report generation, emphasizing the accurate identification and reporting of abnormalities and pathologies. We compare our results with two previously established SOTA models, using similar experimental setups for fair comparison. Reports are evaluated based on 14 different pathologies, with precision, recall, and F1 score calculated against labeled ground truth, using ChestXbert for labeling. Table 5.2 indicates that our model achieves competitive CE scores.

Additional qualitative metrics to consider include data efficiency, interpretability, memory footprint, inference time, and representational power for NLG. Experiments indicate that combining medical domain-driven heuristics, such as anatomical region detection, with the expressive capabilities of large language models (LLMs) can enhance radiology report generation. Table 5.3 shows the qualitative metrics for different types of model architectures.

Table 5.1: Natural Language generation evaluation and comparison with SOTA, the scores are taken from the respective works. CMN [17], RGRG [23], Bootstrap-LLM [43], MAIRA-2 [40].

NLG Evaluation	**CMN**	**RGRG**	**Bootstrap-LLM**	**MAIRA-2**	**Proposed**
BLUE-1(↑)	0.353	0.373	0.402	**0.479**	0.381
BLUE-2 (↑)	0.218	0.249	**0.262**	-	0.255
BLUE-3 (↑)	0.148	0.175	**0.180**	-	0.177
BLUE-4 (↑)	0.106	0.126	0.128	**0.243**	0.127
METEOR(↑)	0.142	0.168	0.175	**0.430**	0.172
ROUGE-L (↑)	0.278	0.264	0.291	**0.391**	0.280
No. of parameters	300M	400M	14.2B	13B	8B

Table 5.2: Clinical Efficacy Evaluation compared to studies with similar experimental setup. Scores taken from respective studies. RGRG [23], Bootstrap-LLM [43].

CE evaluation	**P(↑)**	**R(↑)**	**F1(↑)**
RGRG	0.461	0.475	0.447
Bootstrap-LLM	**0.465**	**0.482**	**0.473**
Proposed	0.462	0.477	0.469

Table 5.3: Qualitative metrics to be taken into account when deciding the performance of a model.

Metric	**Knowledge-driven**	**LLMs based**	**Combined**
Memory Footprint	Low	High	Moderate
Inference time	Low	High	Moderate
Data efficiency	High	Low	Moderate
Interpretability	High	Low	Moderate
NLG capability	Low	High	Moderate

Chapter6

Conclusion & Future Work

6.1 Conclusion

- Radiology Report Generation (RRG) presents significant challenges due to the complexity of the task and various underlying difficulties.
- This study demonstrates that integrating domain-specific knowledge with the expressive capabilities of large language models (LLMs) yields superior outcomes compared to using either approach in isolation. This integration effectively addresses challenges such as data scarcity and interpretability.
- The fusion of multiple features, including global and local anatomical features, enhances both Natural Language Generation (NLG) and Clinical Efficacy metrics, resulting in improved overall model performance.
- Our approach achieves competitive results relative to state-of-the-art (SOTA) methods, even with the use of only frontal chest X-ray images and a comparatively smaller number of model parameters.

6.2 Future Work

- Explore integrating RAG with feature fusion techniques to mitigate the issue of hallucinations in large language models (LLMs). This could help ensure that generated reports are more accurate and grounded in relevant medical knowledge.

- Investigate the impact of incorporating lateral chest X-ray views and previous radiology reports alongside current frontal view images as additional inputs. This could provide richer representations and improve both report generation accuracy and Clinical Efficacy metrics.

- Experiment with various methods of feature fusion, such as attention-based mechanisms or weighted combinations, to identify the most effective approach for enhancing model performance and balancing global and local feature contributions.

- Research the potential of knowledge distillation techniques to transfer the knowledge of large language models (LLMs) into smaller, more efficient models. This could help reduce the model's computational requirements while maintaining competitive performance.

Bibliography

[1] Michael P. Hartung et al. "How to Create a Great Radiology Report". In: *RadioGraphics* 40.6 (2020). PMID: 33001790, pp. 1658–1670. DOI: 10.1148/rg.2020200020.

[2] John R Wilcox. "The written radiology report." In: *Applied Radiology* 35.7 (2006).

[3] Phillip Sloan et al. "Automated Radiology Report Generation: A Review of Recent Advances". In: *IEEE Reviews in Biomedical Engineering* (2024), 1–20. ISSN: 1941-1189. DOI: 10.1109/rbme.2024.3408456. URL: http://dx.doi.org/10.1109/RBME.2024.3408456.

[4] Alistair E. W. Johnson et al. *MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs*. 2019. arXiv: 1901.07042 [cs.CV]. URL: https://arxiv.org/abs/1901.07042.

[5] Mauricio Reyes et al. "On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities". In: *Radiology: Artificial Intelligence* 2.3 (2020). PMID: 32510054, e190043. DOI: 10.1148/ryai.2020190043.

[6] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. *KiUT: Knowledge-injected U-Transformer for Radiology Report Generation*. 2023. arXiv: 2306.11345 [cs.CV]. URL: https://arxiv.org/abs/2306.11345.

[7] Baoyu Jing, Zeya Wang, and Eric Xing. "Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. DOI: 10.18653/v1/p19-1657. URL: http://dx.doi.org/10.18653/v1/P19-1657.

[8] Junhua Mao et al. *Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN)*. 2015. arXiv: 1412.6632 [cs.CV]. URL: https://arxiv.org/abs/1412.6632.

[9] Steven J. Rennie et al. *Self-critical Sequence Training for Image Captioning*. 2017. arXiv: 1612.00563 [cs.LG]. URL: https://arxiv.org/abs/1612.00563.

[10] Jiasen Lu et al. *Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning*. 2017. arXiv: 1612.01887 [cs.CV]. URL: https://arxiv.org/abs/1612.01887.

[11] Peter Anderson et al. *Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering*. 2018. arXiv: 1707.07998 [cs.CV]. URL: https://arxiv.org/abs/1707.07998.

[12] Marcella Cornia et al. *Meshed-Memory Transformer for Image Captioning*. 2020. arXiv: 1912.08226 [cs.CV]. URL: https://arxiv.org/abs/1912.08226.

[13] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[14] Zhihong Chen et al. *Generating Radiology Reports via Memory-driven Transformer*. 2022. arXiv: 2010.16056 [cs.CL]. URL: https://arxiv.org/abs/2010.16056.

[15] Yixiao Zhang et al. *When Radiology Report Generation Meets Knowledge Graph*. 2020. arXiv: 2002.08277 [cs.CV]. URL: https://arxiv.org/abs/2002.08277.

[16] Fenglin Liu et al. *Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation*. 2021. arXiv: 2106.06963 [cs.CV]. URL: https://arxiv.org/abs/2106.06963.

[17] Zhihong Chen et al. *Cross-modal Memory Networks for Radiology Report Generation*. 2022. arXiv: 2204.13258 [cs.CL]. URL: https://arxiv.org/abs/2204.13258.

[18] Han Qin and Yan Song. "Reinforced Cross-modal Alignment for Radiology Report Generation". In: *Findings of the Association for Computational Linguistics: ACL 2022*. Ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 448–458. DOI: 10.18653/v1/2022.findings-acl.38. URL: https://aclanthology.org/2022.findings-acl.38.

[19] Kaveri Kale et al. "KGVL-BART: Knowledge Graph Augmented Visual Language BART for Radiology Report Generation". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 3401–3411. DOI: 10.18653/v1/2023.eacl-main.246. URL: https://aclanthology.org/2023.eacl-main.246.

[20] Kaveri Kale, pushpak Bhattacharyya, and Kshitij Jadhav. *Replace and Report: NLP Assisted Radiology Report Generation*. 2023. arXiv: 2306.17180 [cs.CL]. URL: https://arxiv.org/abs/2306.17180.

[21] An Yan et al. *Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation*. 2021. arXiv: 2109.12242 [cs.CL]. URL: https://arxiv.org/abs/2109.12242.

[22] Chang Liu, Yuanhe Tian, and Yan Song. *A Systematic Review of Deep Learning-based Research on Radiology Report Generation*. 2024. arXiv: 2311.14199 [cs.CV]. URL: https://arxiv.org/abs/2311.14199.

[23] Tim Tanida et al. “Interactive and Explainable Region-guided Radiology Report Generation”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. DOI: 10.1109/cvpr52729.2023.00718. URL: http://dx.doi.org/10.1109/CVPR52729.2023.00718.

[24] Lin Wang et al. “An Inclusive Task-Aware Framework for Radiology Report Generation”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*. Singapore, Singapore: Springer-Verlag, 2022, 568–577. ISBN: 978-3-031-16451-4. DOI: 10.1007/978-3-031-16452-1_54. URL: https://doi.org/10.1007/978-3-031-16452-1_54.

[25] Tiancheng Gu et al. “Complex Organ Mask Guided Radiology Report Generation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2024, pp. 7995–8004.

[26] Ke Zhang et al. “Attribute Prototype-guided Iterative Scene Graph for Explainable Radiology Report Generation”. In: *IEEE Transactions on Medical Imaging* (2024), pp. 1–1. DOI: 10.1109/TMI.2024.3424505.

[27] Wenjun Hou et al. *ICON: Improving Inter-Report Consistency of Radiology Report Generation via Lesion-aware Mix-up Augmentation*. 2024. arXiv: 2402.12844 [cs.CV]. URL: https://arxiv.org/abs/2402.12844.

[28] Mingjie Li et al. *Auxiliary Signal-Guided Knowledge Encoder-Decoder for Medical Report Generation*. 2020. arXiv: 2006.03744 [cs.CV]. URL: https://arxiv.org/abs/2006.03744.

[29] Yuhao Wang et al. “Self adaptive global-local feature enhancement for radiology report generation”. In: *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2023, pp. 2275–2279.

[30] Jianbo Yuan et al. “Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment”. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*. Ed. by Dinggang Shen et al. Cham: Springer International Publishing, 2019, pp. 721–729. ISBN: 978-3-030-32226-7.

[31] Liming Xu et al. “CGFTrans: Cross-Modal Global Feature Fusion Transformer for Medical Report Generation”. In: *IEEE Journal of Biomedical and Health Informatics* (2024), pp. 1–12. DOI: 10.1109/JBHI.2024.3414413.

[32] Yi Guo et al. “IFNet: An Image-Enhanced Cross-Modal Fusion Network for Radiology Report Generation”. In: *Bioinformatics Research and Applications*. Ed. by Wei Peng, Zhipeng Cai, and Pavel Skums. Singapore: Springer Nature Singapore, 2024, pp. 286–297. ISBN: 978-981-97-5128-0.

[33] Ke Zhang et al. "Semi-Supervised Medical Report Generation via Graph-Guided Hybrid Feature Consistency". In: *IEEE Transactions on Multimedia* 26 (2024), pp. 904–915. DOI: 10.1109/TMM.2023.3273390.

[34] Tao Tu et al. *Towards Generalist Biomedical AI*. 2023. arXiv: 2307.14334 [cs.CL]. URL: https://arxiv.org/abs/2307.14334.

[35] Rajesh Bhayana. "Chatbots and large language models in radiology: a practical primer for clinical and research applications". In: *Radiology* 310.1 (2024), e232756.

[36] Shaoting Zhang and Dimitris Metaxas. *On the Challenges and Perspectives of Foundation Models for Medical Image Analysis*. 2023. arXiv: 2306.05705 [eess.IV]. URL: https://arxiv.org/abs/2306.05705.

[37] Kai Zhang et al. "A generalist vision–language foundation model for diverse biomedical tasks". In: *Nature Medicine* (Aug. 2024). ISSN: 1546-170X. DOI: 10.1038/s41591-024-03185-2. URL: http://dx.doi.org/10.1038/s41591-024-03185-2.

[38] Chaoyi Wu et al. *Towards Generalist Foundation Model for Radiology by Leveraging Web-scale 2D3D Medical Data*. 2023. arXiv: 2308.02463 [cs.CV]. URL: https://arxiv.org/abs/2308.02463.

[39] Qianqian Xie et al. *Me LLaMA: Foundation Large Language Models for Medical Applications*. 2024. arXiv: 2402.12749 [cs.CL]. URL: https://arxiv.org/abs/2402.12749.

[40] Shruthi Bannur et al. *MAIRA-2: Grounded Radiology Report Generation*. 2024. arXiv: 2406.04449 [cs.CL]. URL: https://arxiv.org/abs/2406.04449.

[41] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[42] Shi-Qi Yan et al. *Corrective Retrieval Augmented Generation*. 2024. arXiv: 2401.15884 [cs.CL]. URL: https://arxiv.org/abs/2401.15884.

[43] Chang Liu et al. "Bootstrapping Large Language Models for Radiology Report Generation". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 38.17 (2024), pp. 18635–18643. DOI: 10.1609/aaai.v38i17.29826. URL: https://ojs.aaai.org/index.php/AAAI/article/view/29826.

[44] Liwen Sun et al. *Fact-Aware Multimodal Retrieval Augmentation for Accurate Medical Radiology Report Generation*. 2024. arXiv: 2407.15268 [cs.CL]. URL: https://arxiv.org/abs/2407.15268.

[45] Xiao Wang et al. *R2GenCSR: Retrieving Context Samples for Large Language Model based X-ray Medical Report Generation*. 2024. arXiv: 2408.09743 [cs.CV]. URL: https://arxiv.org/abs/2408.09743.

[46] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2016. arXiv: 1506.01497 [cs.CV]. URL: https://arxiv.org/abs/1506.01497.

[47] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[48] Kihyun You et al. "CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Springer Nature Switzerland, 2023, 101–111. ISBN: 9783031438950. DOI: 10.1007/978-3-031-43895-0_10. URL: http://dx.doi.org/10.1007/978-3-031-43895-0_10.

[49] Ming Chen et al. "Simple and Deep Graph Convolutional Networks". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 1725–1735. URL: https://proceedings.mlr.press/v119/chen20v.html.

[50] Zonghan Wu et al. "A Comprehensive Survey on Graph Neural Networks". In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (Jan. 2021), 4–24. ISSN: 2162-2388. DOI: 10.1109/tnnls.2020.2978386. URL: http://dx.doi.org/10.1109/TNNLS.2020.2978386.

[51] Qiang Fu et al. "TLPGNN: A Lightweight Two-level Parallelism Paradigm for Graph Neural Network Computation on Single and Multiple GPUs". In: *ACM Trans. Parallel Comput.* 11.2 (2024). ISSN: 2329-4949. DOI: 10.1145/3644712. URL: https://doi.org/10.1145/3644712.

[52] Saptarshi Sinha, Hiroki Ohashi, and Katsuyuki Nakamura. *Class-Wise Difficulty-Balanced Loss for Solving Class-Imbalance*. 2020. arXiv: 2010.01824 [cs.CV]. URL: https://arxiv.org/abs/2010.01824.

[53] Tianyi Zhang et al. *BERTScore: Evaluating Text Generation with BERT*. 2020. arXiv: 1904.09675 [cs.CL]. URL: https://arxiv.org/abs/1904.09675.

[54] Joy T. Wu et al. *Chest ImaGenome Dataset for Clinical Reasoning*. 2021. arXiv: 2108.00316 [cs.CV]. URL: https://arxiv.org/abs/2108.00316.

[55] Kishore Papineni et al. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[56] Satanjeev Banerjee and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005, pp. 65–72.

[57] Lin Chin-Yew. "Rouge: A package for automatic evaluation of summaries". In: *Proceedings of the Workshop on Text Summarization Branches Out, 2004*. 2004.

[58] Mercy Ranjit et al. *Retrieval Augmented Chest X-Ray Report Generation using OpenAI GPT models*. 2023. arXiv: 2305.03660 [cs.CL]. URL: https://arxiv.org/abs/2305.03660.