

Automated Generation of X-Ray Reports by Machine Learning Approach

by

Aigerim Bauyrzhanova

Submitted to the Department of Data Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

July 2025

© Nazarbayev University 2025. All rights reserved.

Author
Aigerim Bauyrzhanova
Department of Data Science
20 July 2025

Certified by
Min-Ho Lee
Assistant Professor of Computer Science
Thesis Supervisor

Certified by
Berdakh Abibullaev
Assistant Professor of Robotics
Thesis Co-Supervisor

Accepted by
Elizabeth Arkhangelsky
Dean, School of Engineering and Digital Sciences

Automated Generation of X-Ray Reports by Machine Learning Approach

by

Aigerim Bauyrzhanova

Submitted to the Department of Data Science
on 20 July 2025, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

Medical reports are essential for conveying conclusions and diagnoses from medical scans, such as X-ray scans, in order to proceed with the patient's treatment. Especially lung X-rays, which are used in diagnosing and treating diseases such as pneumonia, pulmonary diseases, COVID-19, lung cancer, etc. The traditional way of writing medical reports for such X-ray images is to write them manually, by specialists. This is a time-consuming, hard-to-keep-up with the increasing demand, and error-prone process since it is done by the specialists themselves after taking the scans. This prompts to make the process more automated so as to save time and lessen the probability of a mistake in diagnosis. Therefore, automating the process of writing the medical reports from chest X-ray images has become a critical need in the field of radiology as nowadays there are a lot of detrimental and even fatal diseases of the lungs that need to be treated as soon as possible. For this purpose, the machine learning approach using encoder-decoder transformers and vision transformers is a potential candidate for getting numeric results of an effective model to generate X-ray reports on two datasets: IU-Xray and MIMIC-CXR.

Keywords: automated X-ray reports, machine learning, medical scans, diagnosis, chest X-ray image

Thesis Supervisor: Min-Ho Lee

Title: Assistant Professor of Computer Science

Thesis Supervisor: Berdakh Abibullaev

Title: Assistant Professor of Robotics

Contents

1	Introduction	4
1.1	Motivation	4
1.2	Limitations of the previous approaches	5
1.3	Application of machine learning	5
2	Related works	7
2.0.1	Report Generation Frameworks	7
2.0.2	Visual Extraction	8
2.0.3	Knowledge-Enhanced Generation	10
2.0.4	Report Generation Through Natural Language Processing . .	11
2.0.5	Hybrid and Hierarchical Models	13
2.0.6	Memory-Augmented Architectures	14
3	Methodology	17
3.0.1	Datasets	17
3.0.2	Reference Approach	18
3.0.3	Proposed Approach	21
3.0.4	Experiments	24
4	Results	26
5	Conclusion	31

Chapter 1

Introduction

1.1 Motivation

Recently, with the increasing volume of medical imaging data, there is a growing need for the efficient and accurate generation of medical reports [12]. Especially, since the method used in accustomed report generation can take too much of the specialists' time and be prone to human-made errors. It is stated that on average, even the most experienced radiologists interpret the X-ray images for 5-10 minutes, and this time does not include the time consumed for writing the report [13]. Moreover, each healthcare professional may analyze the same X-ray image differently and write a different report, which in turn will lead to inconsistent or inaccurate diagnoses and delays in treatment [9], whereas the model trained via different datasets can get the experience and knowledge of several specialists of the field, making diagnosis of the rare or hard-detectable diseases much easier and, thus, increasing the likelihood of the successful treatment of the patient. Since X-ray images are used to diagnose and treat many potentially fatal diseases such as pneumonia, pulmonary diseases, COVID-19, lung cancer, and so on, it is crucial to identify the disease at its early stages and start the treatment plan as soon as possible. To address this challenge and improve efficiency, automated medical report generation has appeared as a promising research field to work on. This approach holds the potential to reduce radiologists' workload, minimize diagnostic errors as much as possible, accelerate clinical workflows, and

facilitate the early start of patient treatment since image captioning methods are often used in disease diagnosis and treatment [27]. It is especially crucial since artificial intelligence (AI) is able to process and generate massive amounts of data, which can be a solution to the increasing demand for writing reports for X-ray images and diagnosing them [18].

1.2 Limitations of the previous approaches

It should also be noted that for the problem of either classifying or generating reports from X-ray images, other image captioning models used for generic images cannot be used. While advancements in generic image captioning have been significant, directly applying these models to medical report generation demonstrates a notable decrease in performance. This difference in the performance can be stated to be because of the innate characteristics of medical images. Unlike generic images with high visual variability, radiographic images exhibit a higher degree of similarity or minor changes between healthy and unhealthy tissues in organs. Consequently, successful medical report generation hinges on the model's ability to capture and describe subtle, clinically significant and important visual details. To achieve this, a tighter coupling between the visual and textual representations learned by the model is necessary [29]. Another reason why the task of diagnostic report generation presents a significant challenge is that disease anomalies often occupy only a small portion of the whole X-ray image. They can also manifest in unpredictable locations, necessitating a high degree of focus on specific regions throughout the feature extraction and report generation process [28].

1.3 Application of machine learning

This research explores the utilization of machine learning methods and techniques to automate the generation of X-ray reports. I will delve into the limitations of existing image captioning methods in the context of generating reports for X-ray images in

related works part and offer a more efficient model that uses transformers and vision transformers, explaining the methodology in the next part. This is crucial since automated medical report generators will be trained to classify and diagnose the X-ray images, as well as automatically write the medical report for the patient based on prior knowledge. It is also noted that the execution time for some models is also crucial to be taken into account since the heavy load for the training unavoidably causes inefficiency and a high risk of errors during the prediction process [8]. The main objectives for this research are:

- Choose a reference model to test its results
- Modify the existing model to make the process more efficient
- Improve the feature extraction process
- Compare the results for the existing model and the offered model
- Visualize the results and feature extraction

The rest of the document is organized as follows: Section 2 discusses the literature review of existing studies; Section 3 presents the methodology of the proposed solution; Section 4 describes the results of the research project; and Section 5 concludes this document.

Chapter 2

Related works

2.0.1 Report Generation Frameworks

Building an effective model that is capable of generating automated medical reports that are accurate and clinically useful for making diagnoses, and integrates both images and text is a complex task with numerous challenges compared to the standard image captioning tasks. Unlike conventional image captioning, which focuses on brief scene descriptions, medical report generation requires a deeper understanding of radiological images and their correlation with clinical text. It needs a profound and multi-layered comprehension of the image and image features and their relation to the specific anomalies. The model must not only detect abnormalities in medical scans but also interpret how visual findings align with textual descriptions at varying levels of analysis. This multi-layered analysis is particularly crucial when identifying subtle tissue changes that may indicate early-stage diseases.

Furthermore, the model must incorporate external medical knowledge to ensure that generated reports are clinically accurate and resemble those written by radiologists [15]. This integration of knowledge is crucial to ensure that the generated medical report are not merely plausible and semantically probable descriptions but also accurate clinically and could be used in patient treatment. Traditional image captioning methods have major limitations for this task, as they are optimized for general-purpose descriptions rather than detailed, medically precise reports. They

quite often lack the capacity of deeply analyzing each feature and comparing usual features and abnormal ones. Given the high stakes of medical diagnostics, where even minor inaccuracies, omissions or ambiguities in report generation could potentially lead to misdiagnoses or delayed healthcare, underscoring the need for highly specialized and advanced techniques tailored to the demands and standards of clinical contexts. Therefore, it is crucial to implement both the advanced visual extraction technique and the report generation architecture that would be able to capture medically relevant details and features and suit the task appropriately by its complexity, structure and efficiency.

2.0.2 Visual Extraction

Convolutional neural networks (CNNs) have significantly advanced computer vision by effectively extracting discriminative and crucial features directly from images. These architectures achieve remarkable performances by enabling efficient pattern recognition across diverse visual domains and thus capturing the relationships between the image and the text.

The ResNet [10] architecture addressed the vanishing gradient problem, which is a critical limitation and challenge encountered by many models during the training of images and optimization of deep neural networks, especially when using backpropagation methods with certain activation functions such as sigmoid or tanh activation functions, through residual connections, where the gradients diminish and become negligible exponentially during the propagation process. ResNet’s solution enabled some deeper architectures for improved feature extraction by implementing residual connections through identity mapping and thus decreasing the possibility of the gradient becoming too small and stabilizing it.

Building upon this foundation, DenseNet [14] introduced and employed a new paradigm that uses dense skip connections with iterative feature map concatenation in order to deal with the same issues by making each layer of the model receive direct feature inputs from all preceding layers, not just the immediately preceding layer to the current one, through channel-wise concatenation operations. This architecture

leverages three advantages, such as enhancing gradient propagation throughout the network depth, implicit deep supervision via multi-level feature reuse and significant parameter number decrease by condensing the features [14].

Transformers and Vision Transformers (ViTs), which were initially successful in NLP, then have been adapted for being used in vision tasks such as image classification, segmentation, and so on. Vision Transformer (ViT) [7] in itself processes images by dividing them into several patches and treating them as patch-based tokens, using self-attention for any possible global relationships. Recent research focuses on efficient designs, including pyramidal architectures, optimized training, and streamlined self-attention mechanisms for effective usage of ViTs in feature extraction.

Some successful hybrid architectures combine CNNs and Transformers, such as CoAtNet [5], D-TrAttUnet [2] and ConvNeXt [30], to leverage their complementary strengths. They can integrate self-attention into ResNet’s final blocks, merge depth-wise convolutions with lightweight Transformer modules, enhance self-attention with convolutional projections, and replace patch embedding with a convolutional stem for faster inference. By leveraging the strengths of both base architectures, this method can optimize the efficiency of the classification while maintaining global-local feature integration.

CNNs, reliant on low-level features, are often vulnerable to adversarial attacks. Robustness improvements include specialized architectures [22] and activation/pooling optimizations [33]. While these methods bolster CNNs, their efficacy for Vision Transformers (ViTs) remains underexplored.

Recent studies suggest ViTs exhibit superior adversarial robustness compared to CNNs [25], especially against certain types of adversarial attacks, which makes the model more error-free and lessens the possibility of incorrect predictions. This happens due to the model prioritizing global dependencies and connections between image patches over local features. This holistic perspective reduces the model’s vulnerability to localized adversarial attacks. This, in its turn, highlights ViTs’ resilience to natural distribution shifts and image corruptions. However, rather than merely comparing CNNs and ViTs, this research advances robust hybrid architectures (MedViTs)

and introduces a novel augmentation technique to enhance report generation through efficient feature extraction.

2.0.3 Knowledge-Enhanced Generation

To overcome these challenges, some works propose a framework consisting of three modules: a classification module, a generation module, and an interpretation module, which generates medical texts from chest X-ray images and other clinical documents with relevant information. The classification module extracts feature from X-ray images of the human’s chest and clinical documents and combines them to form a contextualized disease embedding, basically taking disease types as distinct classes. The clinical documents are analyzed via an encoder part of the transformer to summarize the disease-related information. The generation module takes the contextualized disease embedding as input and writes a medical report one word at a time using a transformer decoder. It uses a transformer decoder to produce a medical report word-by-word, leveraging the contextualized disease embeddings as input. The interpretation module takes the generated medical report as input and classifies the reported disease-related topics. It then compares the classification results with the ground truth labels and uses the discrepancy to fine-tune the generated report [23]. There is another work that emphasized three modules in its framework. It works on disease classification by integrating a knowledge graph to encode relationships between different disease labels [32]. This structured representation helps the model better understand correlations between medical conditions, improving diagnostic accuracy. This research, however, was more focused on their classification module to classify the images to different diseases. Their classification module mostly utilizes a knowledge graph to encode relationships between disease labels [32]. A method of generation of a ready radiology report that leverages a knowledge base and multi-modal alignment could also be used. The proposed method utilizes a deep learning architecture with three main components: a generator with a knowledge-enhancing part, a knowledge-updating module, and a multi-modal alignment module. Knowledge-enhancing generator generates radiology reports by considering both visual features

extracted from the image and relevant knowledge retrieved from the knowledge base. It combines visual features (extracted via CNN) with medical knowledge retrieved from a structured knowledge base (stored as a matrix). Textual embeddings are derived from reference reports using a BERT-like transformer. The knowledge base is a matrix that stores medical domain knowledge relevant to radiology report generation. Here, a convolutional neural network (CNN) is employed to extract significant visual features of the input radiology image, and a Transformer-based model similar to BERT is used to extract textual embeddings from reference reports. Knowledge-updating module dynamically updates the knowledge base during training. It learns from the relationships between reference reports written by radiologists and the corresponding radiology images. Multi-modal alignment module ensures consistency between the visual features of the image, the textual content of the report, and the disease labels associated with the image. Multi-head attention is used to model the relationships between the knowledge base, textual features from reports, and visual features from images. A standard transformer model is then used to write the final radiology report by considering the visual features and the retrieved knowledge from the knowledge base [31].

2.0.4 Report Generation Through Natural Language Processing

Mostly, many previous works use Natural Language Processing (NLP), since with NLP radiologists would be able to generate reports faster, leading to improved quality and consistency of care [17]. Given the textual nature of medical reports, NLP plays a pivotal role in automating their generation. NLP-based systems enable radiologists to produce reports more efficiently, enhancing both the speed and consistency of patient care. For example, some works investigated methods for auto-writing medical reports from chest X-ray images. The authors explored various models, varying from simple baselines to more complex neural networks. They commonly evaluate the created reports using standard language generation metrics (BLEU and CIDEr) as

well as clinical relevance metrics based on agreement with human-labeled findings [1]. The authors of one paper tried using a hierarchical generation strategy with a CNN-RNN-RNN architecture for generating medical image reports and further improving the reports’ clinical accuracy with reinforcement learning. The authors mentioned the limitations of existing approaches in medical report generation, which are the facts that optimizing solely for automatic NLG metrics like CIDEr might not guarantee clinically accurate reports and Chest X-ray reports often follow a template structure, and focusing on automatic metrics can lead the model to prioritize keywords over accurate disease state descriptions. To address these limitations, they introduced two key contributions. Firstly, Clinically Coherent Reward (CCR) which is a function that leverages a rule-based disease mention annotator, CheXpert, to directly optimize for clinical accuracy. CCR aims to maximize the correlation between disease states mentioned in the generated report and the ground truth report. Secondly, Weighted Loss Function which is the model that combines the traditional NLG metric with the newly introduced CCR in a weighted fashion during reinforcement learning. This allows the model to prioritize both semantic similarity between the created text and the text that is chosen as a reference and clinical coherence with the ground truth reports [20]. There is also a work that proposes to leverage Reinforced Learning (RL) for cross-modal alignment. The model treats report generation as a sequence-to-sequence task, similar to prior works. It incorporates a Cross-modal Memory (CMM) module to minimize the difference between visual features extracted from the image and textual features of the report. An encoder-decoder structure based on Transformer architecture is employed for sequence processing. RL is introduced to guide the learning process by providing rewards based on Natural Language Generation (NLG) evaluation metrics. The model acts as an agent interacting with the visual and textual features of the environment. RL provides supervision by assigning rewards based on the improvement in NLG metrics (BLEU, METEOR, ROUGE) as the report is generated. The agent aims to maximize the expected reward by adjusting its policy (model parameters) to choose better word predictions. The reinforcement algorithm with a reference reward is operated so that the model could learn and

reduce variance in the learning process [24].

2.0.5 Hybrid and Hierarchical Models

Tiwari et al. used a framework for automated medical report generation from X-ray images of human chest which addresses the limitations of standard RNNs for long sequences by employing Hierarchical LSTMs (sentence and word-level LSTMs). The paper acknowledges the well-known vanishing gradient problem hindering standard RNNs from learning long-term dependencies in sequences, which makes them quite unsuitable for tasks like the generation of whole paragraphs and full-length medical reports. To address this limitation, the authors leverage Hierarchical LSTMs, a two-level architecture comprising sentence and word-level LSTMs. This approach allows the model to comprehend both long-term dependencies between sentences and finer-grained dependencies within sentences, ultimately leading to improved performance in the generation of detailed medical reports [26]. There is a paper that proposes a novel method for generating medical image reports using Adversarial Reinforcement Learning (ARL). The model consists of three main parts, namely: an encoder, a decoder, and a reward module. The encoder extracts features from the input medical images using two separate branches. A CNN branch extracts visual features and a Multi-Label Classification (MLC) branch indicates the standard observations that were made and medical information for the X-ray images. After that, these predicted labels are embedded and integrated and then sent into the decoder. This design addresses the issue of the MLC learning being more uncomplicated than the text generation step, potentially causing the shared network to prioritize the learning part and perform poorly on the text generation step. The decoder employs a hierarchical Long Short-Term Memory (LSTM) architecture with multilevel attention. The sentence-making part of LSTM generates topic vectors step-by-step, while the word-making LSTM generates words separately according to a specific topic vector. The attention mechanism leverages the information on the logic and definition that was pulled out (observations and MeSH labels) to guide the generation of reports that are aware of the abnormal findings in the images. Also, this method uses the Re-

ward module and ARL training to improve the accuracy of the reports [11]. There is also CMCL, a framework for medical report generation based on curriculum learning. Curriculum learning addresses the challenge of training models on complex datasets by gradually introducing them to more difficult examples. Medical report generation involves multimodal data (images and text) learning and capturing abnormalities [19].

2.0.6 Memory-Augmented Architectures

Another paper proposes a new approach for the same task of writing radiology reports from medical images using a deep-learning model. The authors argue that this task can be framed as an image-to-text generation problem, similar to image captioning. The proposed model leverages a Transformer architecture with several key modifications that are done by adding relational memory and Memory-driven Conditional Layer Normalization (MCLN). The model incorporates a relational memory module to capture recurring patterns and trends observed in radiology reports for similar medical images. This memory module helps the model learn from these trends and enhance the accuracy and consistency of generated texts and medical reports. The MCLN layer allows the model to dynamically adjust the normalization parameters based on the memory content, potentially leading to more targeted and informative report generation [4]. Another use of deep learning method in order to generate reports as an image-to-text translation task was used in different research. The framework uses Cross-modal Memory Networks (CMN) and Multi-thread querying and responding. The main idea for this is to introduce a memory component to bridge the gap between visual features extracted from radiology images and textual representations of reports. This memory helps the model learn the complex relationships between image content and corresponding descriptions. The memory component then utilizes a multi-threaded approach for querying and responding. This allows the model to get significant information from different memory places, potentially leading to richer representations [3].

As can be seen in Table 1, most works use CNN in order to extract the most important features of the images provided. The offered approach, however, proposes

to use a vision transformer for the process of feature extraction to improve the capturing of the details in disease detection. Since the tissues where the disease is located might be hard to distinguish from other tissues, therefore the feature extraction model should be more thorough. The offered approach tries to improve the feature extraction process.

Table 2.1: Comparing the approaches

Literature	Visual Extractor	Encoder+Decoder	ViT	Approach
[29]	CNN	Y	Y	Transformer with relational memory
[9]	CNN	Y	Y	Cross-Modal Memory Networks
[15]	N/A	Y	Y	Vision Transformer with multi-modal fusion module
[18]	N/A	Y	Y	Transformer
[12]	N/A	Y	Y	Tree reasoning
[17]	CNN	Y	Y	Transformer with spatial and temporal records keeping
[27]	CNN	Y	Y	Cross-Modal Transformer
[13]	CNN	Y	Y	Knowledge Graph Tree Reasoning
[24]	CNN	Y	Y	CMM
[3]	CNN	Y	Y	CMN
[19]	CNN	Y	Y	Competence-based multi-modal learner
[4]	CNN	Y	Y	Transformer with different memory blocks
[31]	CNN	Y	Y	BERT
[20]	N/A	Y	Y	NLG reward
[32]	N/A	Y	Y	Graph Convolution
				Network
[1]	CNN	N/A	N/A	RNN
[11]	N/A	Y	Y	Reward Module
[26]	CNN	N/A	N/A	LSTM
[23]	N/A	Y	Y	Vision Transformer
Proposed approach	Vision Transformer	Y	Y	Transformer with different memory blocks

Chapter 3

Methodology

In this section, the methodology of the proposed approach is described. The section consists of the following parts: Datasets, Reference Approach, and Proposed Approach. The datasets subsection describes the datasets that were used in the research. Reference Approach mentions the reference model that is used as a base for the new architecture, and the Proposed Approach subsection presents the new model and how it is structured.

3.0.1 Datasets

The current work uses two widely used datasets in this field of research, the IU X-Ray dataset [6] and the MIMIC-CXR dataset [16]. The IU X-ray dataset is a collection of chest X-ray images with a significant number paired with diagnostic reports accordingly. The dataset contains 7,470 chest X-ray images with 3,955 corresponding medical reports that were published by Indiana University. It is an open-source dataset that can be accessed on the official website of the publisher. The MIMIC-CXR Database is another big dataset with chest X-ray images and their written medical reports. This dataset, sourced from Beth Israel Deaconess Medical Center in Boston, MA, comprises 377,110 images paired with 227,835 radiological studies. To ensure patient privacy, the dataset has been anonymized in compliance with HIPAA Safe Harbor regulations.

3.0.2 Reference Approach

R2GEN

R2Gen[4] is a model that uses a memory-driven transformer in order to create texts and medical reports on X-ray images. The model was trained and taught on two major datasets of the field, IU X-Ray, and MIMIC-CXR, using a self-supervised loss function that operates on the similarity between chosen image representations and the contextual meanings of the created text report. The model uses an image encoder to extract visual features from the input medical image using a pre-trained convolutional neural network (CNN). The extracted features are then transformed into a sequence of image tokens, representing the spatial and semantic information contained within the image. The text decoder of the model generates the radiology report one word at a time, prepared on the current input image and the previously generated text. A transformer-based decoder is employed to model long-range dependencies within the generated text and to align the created text with the visual part and meaning of the image. The model also has different memory blocks that ensure that the relevant information of the images and texts is stored and retrieved from the input image and previously generated text, comparing the latter to the ground truth, the true reports. This allows the model to maintain context and generate coherent and informative reports while keeping the semantic relations between the words in sentences. The main advantages of the R2GEN model are that the memory-driven transformer architecture allows the model to grasp long-range dependencies between features and text and generate more accurate and informative reports, while its self-supervised learning objective helps to mitigate the problem of hallucination, where the model generates text that is not true or relevant to the input image.

Given a medical image, the model first extracts patch features $\mathbf{X} = x_1, x_2, \dots, x_S$ using a visual feature extractor, where S denotes the number of patches and each $x_i \in \mathbb{R}^d$ represents a d -dimensional feature vector. The target report constitutes a token sequence $\mathbf{Y} = y_1, y_2, \dots, y_T$, where $y_t \in \mathcal{V}$ represents a token from vocabulary $\mathcal{V}$ at position t , and T is the total report length.

The framework employs an encoder-decoder Transformer architecture enhanced with two novel components: a relational memory module that captures clinical patterns across reports and Memory-driven Conditional Layer Normalization (MCLN) that dynamically modulates decoder behavior.

The encoder module $\text{fe}(\cdot)$ processes patch features to generate contextualized hidden states:

$$\mathbf{H} = \{h_1, h_2, \dots, h_S\} = \text{fe}(x_1, x_2, \dots, x_S)$$

where $\mathbf{H} \in \mathbb{R}^{S \times d_h}$ contains encoded representations of visual features, implemented as a standard Transformer encoder with multi-head self-attention and position-wise feedforward layers.

The decoder $\text{fd}(\cdot)$ generates tokens autoregressively, enhanced by our relational memory system:

$$\hat{y}_t = \text{fd}(\mathbf{H}, \mathbf{M}_{t-1}, y_{1:t-1})$$

where $\mathbf{M}_{t-1} \in \mathbb{R}^{m \times k}$ represents the relational memory matrix at step $t-1$ with m memory slots of dimension k . Unlike standard Transformers, we replace conventional layer normalization with MCLN to integrate memory states.

The relational memory $\mathbf{M}$ captures cross-report clinical patterns through a dynamically updated matrix where each row represents a clinical pattern prototype. The memory updates at each decoding step:

$$\mathbf{M}_t = \text{Update}(\mathbf{M}_{t-1}, \mathbf{h}_{t-1}^{\text{dec}})$$

where $\mathbf{h}_{t-1}^{\text{dec}}$ is the decoder’s previous hidden state. The update mechanism uses multi-head attention with $\mathbf{M}_{t-1}$ as query and its concatenation with decoder outputs as key/value.

MCLN modulates decoder layer normalization parameters using memory states:

$$\text{MCLN}(\mathbf{z}) = \frac{\mathbf{z} - \mu}{\sigma} \odot \gamma(\mathbf{m}_t) + \beta(\mathbf{m}_t)$$

where $\mathbf{z}$ is the pre-normalization activation, μ/σ are mean/std, and $\gamma(\cdot)/\beta(\cdot)$ are affine parameters predicted from the flattened memory vector $\mathbf{m}_t = \text{vec}(\mathbf{M}_t)$ via MLP. Three MCLN blocks are deployed: the first transforms decoder output into the query for cross-attention with encoder states, while others modulate residual connections.

The model maximizes the conditional log-likelihood of the target report.

However, as a feature extractor, the proposed approach does not use CNNs as a reference model, but a different transformer, the MedViT visual transformer.

MedViT

MedViT [21] is a hybrid convolutional-transformer architecture designed for robust medical image classification. The architecture comprises of Patch Embedding Layer, which converts input images into patch tokens, Efficient Convolution Blocks (ECB) that capture local features and short-range dependencies, Local Transformer Blocks (LTB), which model long-range dependencies and fuse multi-frequency signals, Transformer Augmentation Block (TAB) that implements Patch Momentum Changer (PMC) for feature-space augmentation.

Spatial resolution is progressively reduced by $32\times$ through $[4\times, 2\times, 2\times, 2\times]$ down-sampling, while channel dimensions double after each ECB stage. This design enables efficient multi-scale feature extraction, critical for medical imaging.

ECB integrates convolutional efficiency with attention mechanisms to overcome limitations of traditional residual blocks and ConvNext. The ECB formulation is:

$$\begin{aligned}\tilde{z}_l &= \text{MHCA}(z_{l-1}) + z_{l-1} \\ z_l &= \text{LFFN}(\tilde{z}_l) + \tilde{z}_l\end{aligned}\tag{3.1}$$

where: MHCA is Multi-Head Convolutional Attention (mixes tokens), LFFN is Locally Feed-Forward Network (introduces locality), z_{l-1} is Input from the preceding block.

The LFFN operates via spatial token rearrangement:

$$\mathbf{Y}^r = f(\mathbf{Z}^r \circledast \mathbf{W}_1^r) \circledast \mathbf{W}_2^r, \quad \mathbf{Y} = \text{Img2Seq}(\mathbf{Y}^r) \quad (3.2)$$

where $h = H/p$, $w = W/p$ for input resolution $H \times W$ and patch size p . This enables 1×1 convolutions:

$$\mathbf{Y}^r = f(\mathbf{Z}^r \circledast \mathbf{W}_1^r) \circledast \mathbf{W}_2^r, \quad \mathbf{Y} = \text{Img2Seq}(\mathbf{Y}^r) \quad (3.3)$$

preserving spatial relationships while maintaining computational efficiency.

LTB addresses the high-frequency information degradation in conventional transformers through multi-frequency signal mixing. And LTb incorporates Efficient Spatial Attention (ESA) with spatial reduction:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(Q\mathcal{P}_s(K)^T/\sqrt{d}\right)\mathcal{P}_s(V) \quad (3.4)$$

where $\mathcal{P}_s$ performs average pooling with stride s to reduce computational complexity. BatchNorm and ReLU replace LayerNorm and GELU for deployment efficiency.

PMC enhances adversarial robustness through feature-space augmentation at the token level and asymmetrically combines features from two random samples ($\mathbf{Z}_A^l, \mathbf{Z}_B^l$):

$$\mathbf{Z}_{\text{mix}}^l = \frac{\mathbf{Z}_A^l - \mu_A}{\sigma_A} \odot \gamma_B + \beta_B \quad (3.5)$$

where γ_B, β_B are affine parameters derived from $\mathbf{Z}_B^l$. This token-level fusion creates smoother decision boundaries while preserving global context relationships.

3.0.3 Proposed Approach

The approach that is used in the current work is inspired by the R2GEN model; however, instead of using CNN to extract the features from the CNN model, the offered approach will use a different Vision Transformer to do so. It integrates a hierarchical Vision Transformer (MedViT) for visual feature extraction with an R2GEN-based report generation model. The framework consists of two core components that will be

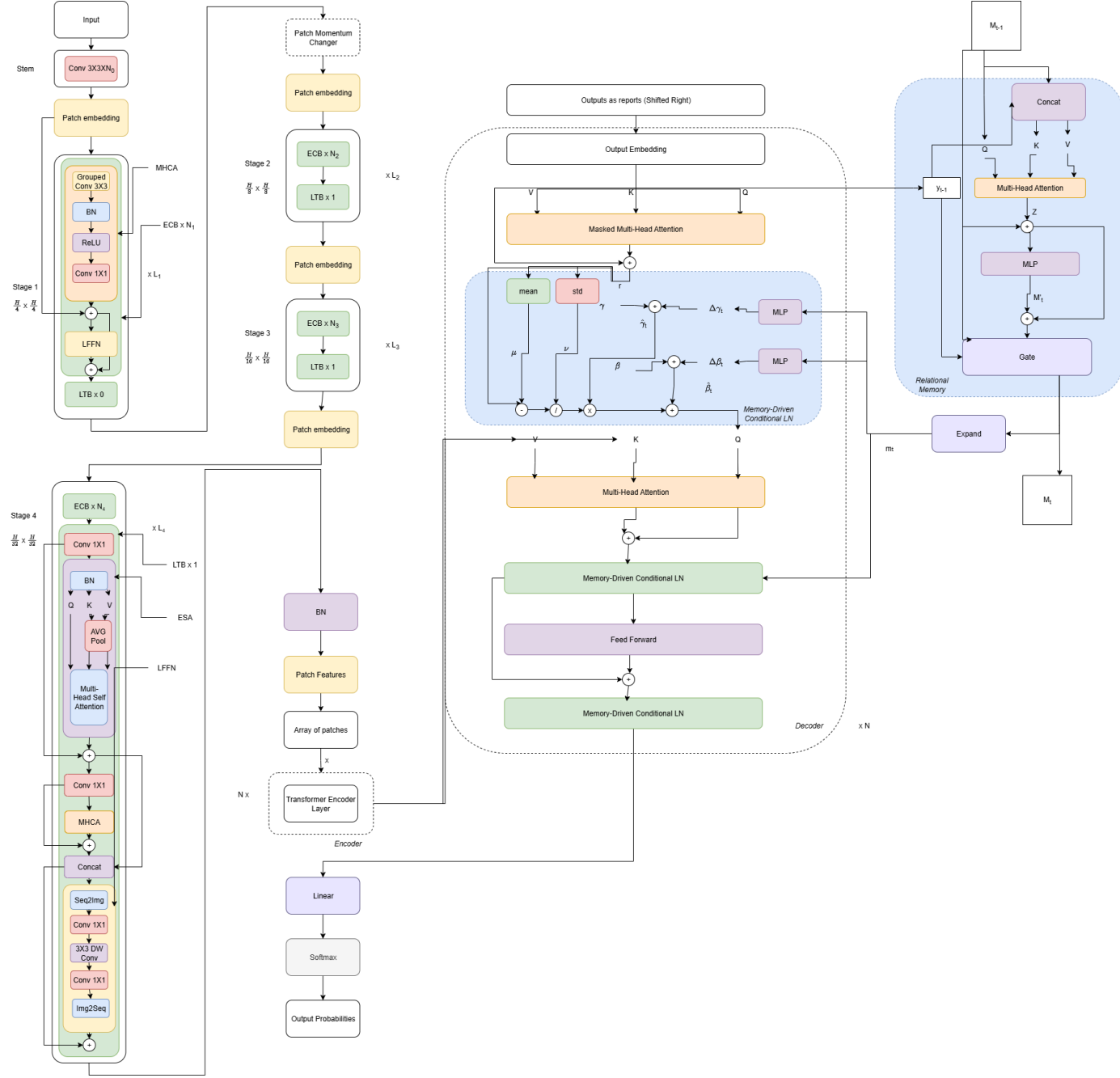


Figure 3-1: The structure of the proposed approach

explained further. The core components are the MedViT Feature Extractor that is used to process input medical images through hybrid convolutional-transformer blocks and R2GEN Report Generator, which generates diagnostic reports using memory-augmented transformers. Here, the framework employs MedViT’s hierarchical pyramid architecture for robust feature extraction, and then the augmented features feed

into the R2GEN model for report generation.

The vision transformer will divide the input images into patches, and each of them is flattened and then projected into a high-dimensional embedding space. Then the model will learn the long-range dependencies between every pixel and capture the global context of the images by running each patch through convolutional blocks and multi-head convolutional attention. Then, a sequence of tokens will be rearranged into a different the sequence of tokens that consists of a 2D lattice. This will help to introduce locality into the network by retrieving the close distance between tokens and pixels. After repeating such a process several times, the global average pool is constructed, and the features will go into the R2GEN model as it is shown in Figure 3-1.

MedViT introduces an innovative hybrid architecture that integrates convolutional and transformer blocks to enhance medical image classification [21]. The model employs a hierarchical pyramid structure comprising a patch embedding layer, transformer blocks, and multiple stacked convolutional layers at each stage. It proposes the Efficient Convolutional Block (ECB), which combines the deployment efficiency of residual blocks with the representational strength of transformers. ECB adopts a hybrid architecture to leverage multi-scale information while incorporating an efficient attention-based token-mixing mechanism, which is then used to be able to efficiently extract the features from the given images. Since the used approaches have issues with relying on efficient components of the model, neglecting adversarial robustness while prioritizing only model complexity, and flexibility of the long-ranged dependency learning, ECB was proposed to overcome these challenges by combining the efficiency of several other components, such as residual blocks, and adopting hybrid architecture, which enhances the components' strength and mitigates their weaknesses.

While the Efficient Convolution Block (ECB) effectively captures local representations, integrating global context remains crucial for robust feature extraction. MedViT uses Local Transformer Block (LTB), a lightweight yet efficient module designed to jointly model local and global features while preserving computational efficiency.

Local Transformer Block (LTB), a lightweight yet efficient module designed to jointly model local and global features while preserving computational efficiency. LTB first captures low-frequency signals via Efficient Self-Attention (ESA), which incorporates channel reduction and batch normalization [21].

At the same time, R2Gen model, consisting of three main components, gets the features extracted by the transformer, which is done by feature augmentation [4]. The process is operated on intermediate token features within ViT layers, leveraging the observation that self-attention mechanisms rely on both global and local token relationships. Unlike pixel-level methods (e.g., Mixup, CutMix), the feature extraction of MedViT asymmetrically combines normalized features from one image and moments from other figures [21]. Then, the encoder of the R2Gen maps visual features to hidden states, and the decoder integrates relational memory and MCLN to refine the generation of the text using the features and ground truths [4].

3.0.4 Experiments

As a scope of the research, a systematic experiment was conducted to evaluate the impact of changing the visual feature extraction on medical report generation. The first thing that was done was that the dataset was tested on the original pretrained R2Gen model, and the baseline values for the evaluation metrics were determined by evaluating the original R2Gen model with ResNet-101 as a visual extractor on the dataset. Having finished the establishment of baseline metrics (BLEU, CIDEr, ROUGE, METEOR), the dataset was then tested again on the model by self-training it in order to make sure that the results were repeatable.

After that, the visual extractor of the R2Gen was changed from the Resnet101 to several transformers in order to find if the change would benefit the text generation process and enhance the results of the metrics. The modified models were fine-tuned by setting up the hyperparameters of the model and then training all the layers by the hyperparameters that were chosen. After each hyperparameter learning, the validation results of the metrics were compared, and the set of parameters with the best metric scores was chosen. The model with the chosen parameter set was then

deployed and tested. The results of the experiments can be seen in the results section of the research.

Chapter 4

Results

Config	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Pretrained	0.470	0.304	0.219	0.165	0.187	0.371
Fine-tuned BLIP	0.459	0.303	0.221	0.169	0.185	0.386
Fine-tuned MedViT	0.472	0.313	0.228	0.176	0.192	0.371

Table 4.1: Performance metrics for different training configurations on IU-XRAY.

Config	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Pretrained	0.353	0.218	0.145	0.103	0.142	0.277
Fine-tuned BLIP	0.320	0.193	0.125	0.097	0.139	0.264
Fine-tuned MedViT	0.350	0.220	0.144	0.107	0.144	0.278

Table 4.2: Performance metrics for different training configurations on MIMIC-CXR.

As it can be seen in the Table 4.1, the performance of different training configurations that were used in the experiments was evaluated using several standard metrics, including BLEU (1-4), METEOR, and ROUGE-L since these are the most common methods used to evaluate the quality and relevance of the text which was generated by the model. BLEU measures the precision of n-grams which are the contiguous sequences of n words in the generated text compared to the reference text, where n usually ranges from 1 to 4. METEOR evaluates text generation by considering precision, recall, and a harmonic mean of the two while also taking into account the synonyms used, the stemming of the words, and the provision of the penalty for the poor sentence structure. ROUGE-L metric evaluates the longest common sub-

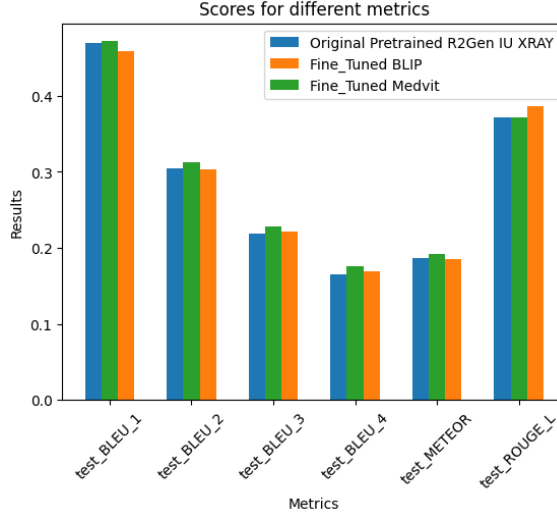


Figure 4-1: Visualization of the Results for IU-XRAY

quence (LCS) between the generated text and the ground truth. Subsequence here in itself is a sequence of some text and string derived from another text.

Thus, the configuration with the pretrained model achieved a BLEU-1 score of 0.470, BLEU-2 score of 0.304, BLEU-3 score of 0.219, and BLEU-4 score of 0.165. The METEOR score was 0.187, and the ROUGE-L score was 0.371. This was the baseline value for the research. The self-trained original model showed slightly lower performance in BLEU-1 (0.463) and BLEU-2 (0.300) compared to the pretrained configuration. However, it matched the pretrained model in BLEU-3 (0.219) and slightly improved in BLEU-4 (0.170). The METEOR score remained the same at 0.187, while the ROUGE-L score was marginally lower at 0.370. This might be due to the different parameters used during the training phase. The modified configuration of the model that had the transformer as a visual tractor instead of the common Resnet101 neural network model demonstrated improvements in several metrics. It matched the pretrained configuration in BLEU-1 (0.470) and showed higher scores in BLEU-2 (0.310) and BLEU-3 (0.227). The BLEU-4 score remained at 0.170, and the METEOR score was consistent at 0.187. Notably, the ROUGE-L score increased to 0.400, indicating better performance in capturing longer sequences.

Overall, the modified configuration showed slightly higher performance than the reference model, particularly in the ROUGE-L metric, suggesting that the modifi-

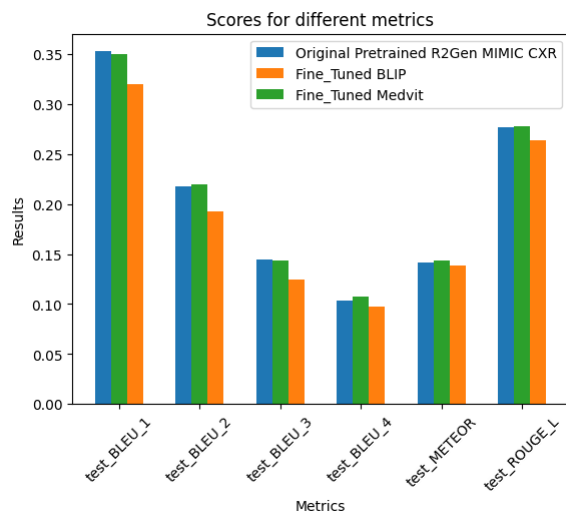


Figure 4-2: Visualization of the Results for MIMIC-CXR

cations enhanced the model's capability to write more relevant and comprehensive sequences of texts. The Figure 4-1 shows the visualization of the results.

In addition to the quantitative results, the reports from the original model and the proposed approach were taken and compared. For the dataset image (from IU-Xray) "CXR3030_IM-1405", the results are as such:

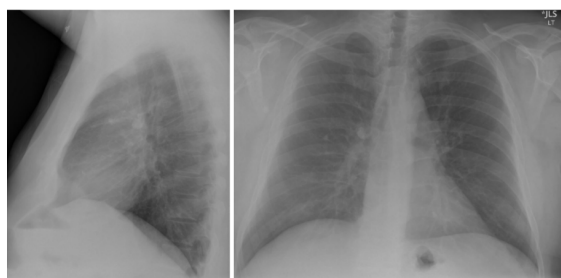


Figure 4-3: Image 1

Original report	New report
"the lungs are clear bilaterally. specifically no evidence of focal consolidation pneumothorax or pleural effusion. cardio mediastinal silhouette is unremarkable. visualized osseous structures of the thorax are without acute abnormality."	"normal cardiomediastinal silhouette. there is no focal consolidation. there are no xxxx of a large pleural effusion. there is no pneumothorax. there is no acute bony abnormality seen."

Table 4.3: Result Table 1

As for the dataset image (from IU-Xray) "CXR38_IM-1911", the results are described in Table 4.4.

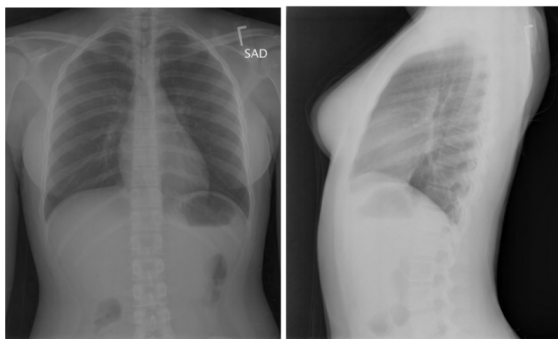


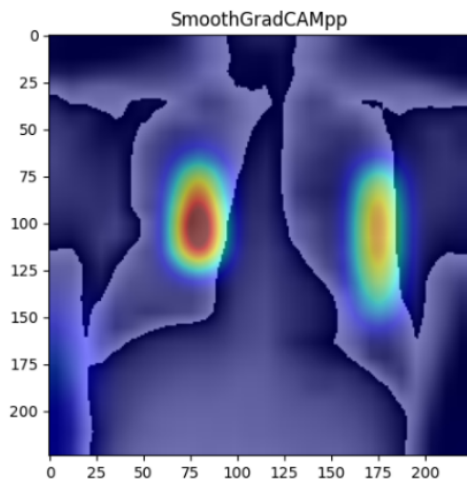
Figure 4-4: Image 2

Original report	New report
"the cardiomedastinal silhouette is within normal limits for size and contour. the lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax. osseous structures are within normal limits for patient age."	"normal cardiomedastinal silhouette. there is no focal consolidation. there are no xxxx of a large pleural effusion. there is no pneumothorax. there is no acute bony abnormality seen."

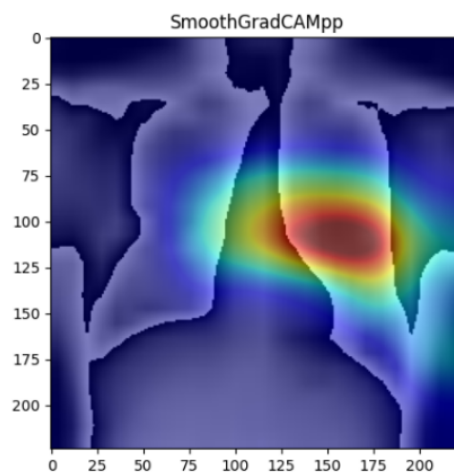
Table 4.4: Result Table 2

As can be seen from the tables, the reports from the proposed model still may have unknown words from the unlearned or ambiguous tokens from the learning process. Even with these additions, the reports from the proposed approach are clear and adequately precise.

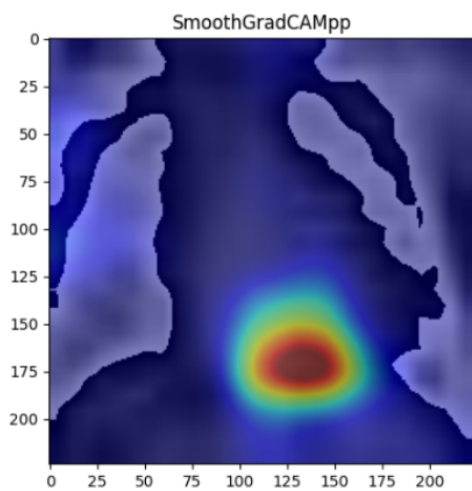
As for the feature extraction, it was shown from the Grad-CAM plotting that the new feature extractor captured more specific bits of the input images than the ResNet101, which was used in the original model.



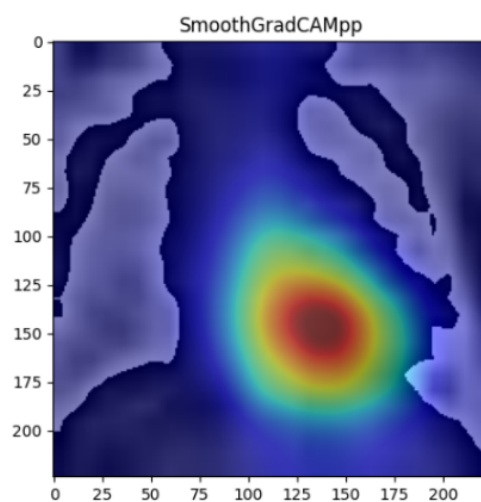
(a) a. Grad-CAM from MedViT chest image



(b) b. Grad-CAM from ResNet101 chest image



(c) c. Grad-CAM from MedViT pneumonia image



(d) d. Grad-CAM from ResNet101 pneumonia image

Figure 4-5: Grad-CAM comparisons: (a) and (c) New extractor focuses on finer details, (b) and (d) ResNet101 shows coarser attention.

Chapter 5

Conclusion

This research investigates the use of machine learning to automate X-ray report generation. It addresses the limitations of existing approaches and current image captioning methods in the medical domain and proposes a more efficient model leveraging transformers and vision transformers.

Therefore, this research aimed to enhance the performance of radiology report generation by proposing a modified architecture based on the R2GEN model. The study began by utilizing two widely recognized datasets, the IU X-Ray and MIMIC-CXR datasets, which provided a robust foundation for training and evaluating the models. The reference approach, R2GEN, served as the baseline, leveraging a memory-driven transformer architecture to generate coherent and informative radiology reports. Building upon this, the proposed approach introduced a Vision Transformer as the extractor for visual features of the images, replacing the traditional CNN-based Resnet101. This modification aimed to grasp and understand long-range dependencies of the word sequences more effectively by processing image patches with multi-head convolutional attention.

The experiments conducted in this study demonstrated the significance of the suggested modifications. The results, evaluated via commonly used standard metrics such as BLEU, METEOR, and ROUGE-L, demonstrated that the modified model achieved equivalent or slightly better performance than the original R2GEN model. The most notably, the ROUGE-L score improved from 0.371 to 0.400, indicating

that the modified model was better at capturing longer sequences and generating more coherent reports. The BLEU scores also showed improvements, particularly in BLEU-2 and BLEU-3, suggesting that the Vision Transformer enhanced the model’s proficiency in generating and writing accurate and applicable text in medicine.

In conclusion, the proposed approach demonstrated the potential of integrating vision transformers into radiology report generation models as a feature extraction tool. The improvements in some evaluation metrics highlight the benefits of capturing dependencies between the image and text in medical image analysis. Future work could explore further refinements to the architecture, such as leveraging larger datasets or imaging from different fields, to continue progressing the field of automated report generation. In addition, generated reports occasionally contain unknown words (from unlearned tokens) or ambiguous terms (arising during training). Nevertheless, the reports maintain clarity and adequate precision for clinical use. As well as, Grad-CAM visualizations confirm that the new feature extractor captures more specific, relevant regions of input images compared to the original ResNet101 backbone. This enhancement improves focus on diagnostically significant features.

Bibliography

- [1] William Boag, Tzu-Ming Harry Hsu, Matthew Mcdermott, Gabriela Berner, Emily Alsentzer, Peter Szolovits, Adrian V Dalca, Sam Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones. Baselines for chest x-ray report generation, 2020.
- [2] Fares Bougourzi, Cosimo Distanto, Fadi Dornaika, and Abdelmalik Taleb-Ahmed. D-trattunet: Dual-decoder transformer-based attention unet architecture for binary and multi-classes covid-19 infection segmentation, 2023.
- [3] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation.
- [4] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. 10 2020.
- [5] Zihang Dai, Hanxiao Liu, Quoc V. Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes, 2021.
- [6] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.
- [8] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Injecting semantic concepts into end-to-end image captioning.
- [9] Shashank Gupta, Yuhang Jiang, and Abdullah-Al-Zubaer Imran. Transforming radiology workflows: Pretraining for automated chest x-ray report generation, 2023.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

- [11] Daibing Hou, Zijian Zhao, Yuying Liu, Faliang Chang, and Sanyuan Hu. Automatic report generation for chest x-ray images via adversarial reinforcement learning. *IEEE Access*, 9:21236–21250, 2021.
- [12] Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. Recap: Towards precise radiology report generation via dynamic disease progression reasoning, 2023.
- [13] Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. Organ: Observation-guided radiology report generation via tree reasoning.
- [14] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks, 2018.
- [15] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation.
- [16] Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019.
- [17] Kaveri Kale, Pushpak Bhattacharyya, and Kshitij Jadhav. Replace and report: Nlp assisted radiology report generation.
- [18] Chang Liu, Yuanhe Tian, and Yan Song. A systematic review of deep learning-based research on radiology report generation. 11 2023.
- [19] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation.
- [20] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew Mcdermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation, 2019.
- [21] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B Shokouhi, and Ahmad Ayatollahi. Medvit: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, 2023.
- [22] Omid Nejati Manzari, Hossein Kashiani, Hojat Asgarian Dehkordi, and Shahriar Baradaran Shokouhi. Robust transformer with locality inductive bias and feature normalization, 2023.
- [23] Hoang T. N. Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. Automated generation of accurate & fluent medical x-ray reports. 8 2021.

- [24] Han Qin and Yan Song. Findings of the association for computational linguistics reinforced cross-modal alignment for radiology report generation.
- [25] Rulin Shao, Zhouxing Shi, Jinfeng Yi, Pin-Yu Chen, and Cho-Jui Hsieh. On the adversarial robustness of vision transformers, 2022.
- [26] Vivek Tiwari, Krutika Bapat, Kushashwa R. Shrimali, Saurabh K. Singh, Basant Tiwari, Swati Jain, and Hemant Kumar Sharma. Automatic generation of chest x-ray medical imaging reports using lstm-cnn. pages 80–85. Association for Computing Machinery, 8 2021.
- [27] Siyuan Wang, Bo Peng, Yichao Liu, and Qi Peng. Phenotypeclip: Phenotype-based contrastive learning for medical imaging report generation.
- [28] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens, 2022.
- [29] Zhanyu Wang, Luping Zhou, Lei Wang, and Xiu Li. A self-boosting framework for automated radiographic report generation, 2021.
- [30] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders, 2023.
- [31] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S. Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86, 5 2023.
- [32] Dehai Zhang, Anquan Ren, Jiashu Liang, Qing Liu, Haoxing Wang, and Yu Ma. Improving medical x-ray report generation by using knowledge graph. *Applied Sciences (Switzerland)*, 12, 11 2022.
- [33] Richard Zhang. Making convolutional networks shift-invariant again, 2019.