

Computerized Chest X-ray Report Generation via Masked Multi-Head Attention & Pathology Tags

Shivanshi Gupta^{1,2*}, Vinay Mishra^{2,3†} and Avantika Singh^{1,2†}

^{1,2*} Dr. Shyama Prasad Mukherjee International Institute of Information Technology, , Naya Raipur, 493661, Chhatisgarh, India.

*Corresponding author(s). E-mail(s): shivanshigupta7@gmail.com;
Contributing authors: vinaymishra00012@gmail.com;
avantika@iiitnr.edu.in;

[†]These authors contributed equally to this work.

Abstract

Accurate and precise examination of X-ray images is important for facilitating good healthcare services. It requires years of clinical experience for radiologists to generate correct and detailed reports. Furthermore, it is a tiresome task. Henceforth, the generation of accurate and detailed X-ray reports is the need of the hour. In this paper, we focus on developing a framework that generates automatic chest X-ray reports. In our work, we emphasis on generating both linguistically clear and clinically precise reports. The proposed framework utilises hybrid natural language processing techniques and deep learning algorithms to analyse medical images. The framework mainly consists of four modules, namely: an image feature extractor, a text embedding extractor, a disease classification module, and a report generator, which work together to generate medical reports. We have analysed the proposed framework on a popularly used publicly available open-i dataset. The experimental results indicate that the proposed methodology has the ability to generate reports that are comparable in terms of linguistic performance as well as clinical accuracy performance over SOTA.

Keywords: Chest Radiology Report Generation, Visual Embedding, Text Embedding, Pathology Tags, Disease States, Transformer Encoder & Decoder, Masked Multi-head Attention

1 Introduction

Medical imaging is a technique that enables medical professionals to observe the inner details of the human body for clinical diagnosis and treatment([1]). Medical images provide a clear view of the functions of different tissues and organs. It is frequently used in hospitals and clinics to diagnose broken bones, wounds, and diseases([2]). In particular, chest X-rays (CXR) are frequently used in emergencies as they are fast, simple, and affordable. Apart from this, medical images can detect several abnormalities in various organs. Thus, it enables physicians to diagnose conditions such as pneumonia, fractured ribs, a swollen heart, and blocked blood vessels([3]). Furthermore, it also assists in detecting diseases like lung cancer, breast cancer, and numerous other ailments and diseases. Experts frequently write medical reports to explain their analysis of chest X-ray images and their findings. However, for expert radiologists and pathologists, this process can be time-consuming, especially in places with fewer clinicians per patient. For less experienced radiologists, especially those operating in rural locations with insufficient healthcare systems, the task might become much more difficult. A self-sustainable generator to generate chest X-ray reports, such as the one depicted in figure (1), can be utilized in these circumstances.

Report generation and image captioning both involve generating text, but they differ in their objectives and context. Image captioning is a crucial and challenging aspect of artificial intelligence (AI) that involves textual description, or caption, for an image ([4], [5]). Its significance lies in merging two other essential fields of artificial intelligence that involve the study of natural language processing and computer vision ([6], [7]). Creating links between visual and textual data has been a goal of computer vision for decades. The paper ([8]) treated this issue as a ranking task in which images are used to retrieve relevant captions from a database and vice versa. Due to the complexity and compositional nature of language, it is improbable that a database could contain every possible image caption. Therefore, an alternative method is to concentrate on explicitly generating captions. Initially, language generation([9]) was performed using handwritten templates. Afterward, it was generated by drawing


	Ground Truth	Generated Report
	Radiology Report: No acute cardiopulmonary process. The lungs are clear, and the cardio mediastinal silhouette is within normal limits for pneumothorax or pleural effusion. No acute osseous abnormality.	Radiology Report: No acute cardiopulmonary process. No focal lung consolidation. Heart size and pulmonary vascularity are within normal limits. No pneumothorax or pleural effusion. Osseous structures are grossly intact.
	MTI Tags: Cardiopulmonary Diseases, Abnormalities	MTI Tags: Cardiopulmonary Diseases, Intact Osseous Structures

Fig. 1 An illustration of a chest X-ray report generated automatically. The highlighted text shows a visual presentation of data that is consistent between generated and ground truth reports.

samples from conditioned recurrent neural network language models ([10], [11]) with image features. These techniques have also been adapted to generate video captions ([12]).

Previously, there had been an increase in research efforts to generate medical reports. It is fueled by advancements in technology. The progress in computer vision ([4]), specifically in image-based captioning technology, has been a driving force behind the recent surge in research efforts for generating medical reports. Several studies have indicated that these methods demonstrate satisfactory performance in terms of language fluency. However, empirical evaluations have revealed that these methods tend to exhibit lower accuracy in terms of clinical relevance. This discrepancy can be attributed to two main reasons. First, medical reports typically contain extensive sentences that describe disease-related symptoms using specialised and precise terminology specific to the medical field. This distinguishes the task of generating medical reports from image-based captioning. Secondly, there is a lack of full utilisation of rich contextual information. For instance, the patient’s clinical history, physician indications, and multiple scans with distinct 3D views. This type of data is typically abundant in real-world scenarios, as evidenced by X-ray benchmarks (datasets) such as open-i ([13]).

Based on previously discussed issues, we have been inspired to develop a framework that prioritizes clinical accuracy. This framework also ensures that the generated medical reports are understandable. It includes four main components, as shown in figure (2) and explained in brief as follows:

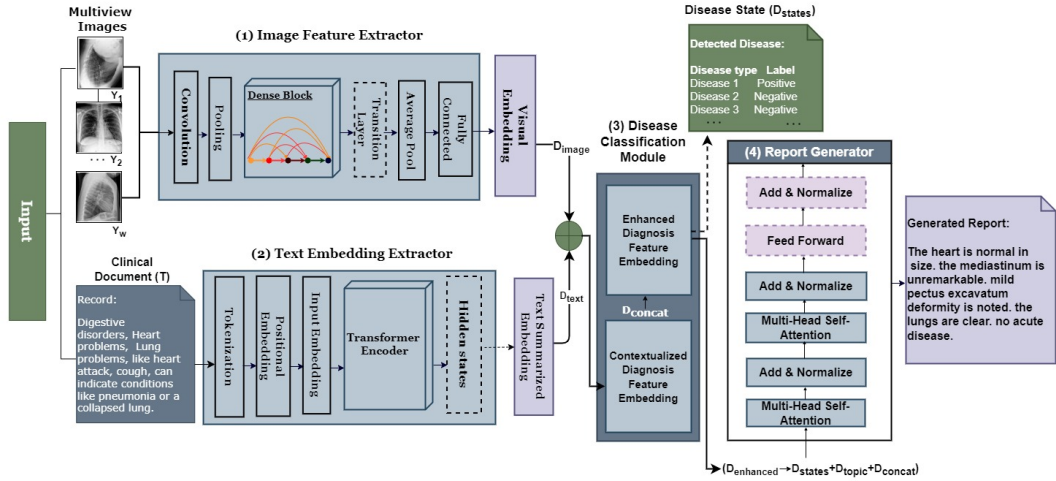


Fig. 2 An overview of the proposed hierarchical framework’s consistency of four components; (1) image features extractor: extracts the image feature by visual embedding. (2) text embedding extractor: reads clinical documents, such as expert’s notes, and creates a summary of the content in text-summarized embeddings. (3) disease classification module: combines visual and text embeddings to summarize disease-related topics. (4) Report Generator: generates text sequentially word by word.

- **Image Feature Extractor:** It accepts chest X-ray images as input and provides visual embeddings corresponding to those images.
- **Text Embedding Extractor:** It takes clinical documents (e.g., expert notes) as input and creates a summary of the content in the form of text-summarised embeddings.
- **Disease Classification Module:** This process combines visual and text features to create a contextualized embedding and classify enhanced diseases.
- **Report Generator:** This module uses enhanced feature embedding to generate report text word by word.

Specific contributions of this paper are as follows:

- For report generation, we propose a multi-stage architecture. This architecture is build upon a transformer with a masked multi-head attention mechanism unlike traditionally used naive transformer based architecture.
- Furthermore, our proposal includes a text summarization embedding system that condenses the information in a text document. This embedding specifically highlights disease-related content, focusing only on the most relevant words associated with the disease ("cold" or "soreness in the chest") within the text. This approach allows for a concise and targeted summary of the disease-related information within the text.

The remainder of the paper is structured as follows: section 2 provides an overview of related work. Section 3 is the main section that describes the proposed methodology in detail. In section 4, the experimental analysis and dataset details are described, followed by section 5 limitations and future work. The sixth and final section 6 concludes the paper.

2 Related Work

In this section, visual captioning, along with medical report generation, are explained.

2.1 Visual Captioning

In recent years, the discipline of visual captioning has made extensive use of techniques based on deep learning ([5],[14]). One of the most common approaches to generating image captions is the encoder-decoder model. Encoders are used to encode images into a latent representation, which is then decoded using a language generation model ([15]). For example, long-short-term memory (LSTM) ([16]), gated recurrent unit (GRU) ([17]), or a variation of these models.

Attention-based techniques ([18]) in image captioning have also been developed by researchers, with the goal of anchoring the words in the generated description to specific places within the image. Recent advances in natural language processing (NLP), particularly the use of the transformer architecture ([19]), have resulted in significant performance improvements in tasks such as translation ([20]), text generation ([21]), and language understanding ([22]). The transformer model was also used to produce image captions in a specific study ([23]). The success of image-based captioning ([5],

[14]) has had a considerable impact on recent advances in medical report generation ([24], [25]).

Inspired by the image captioning research ([26], [7]) and the dominance of the transformer-based framework, we adopt a transformer-based architecture with a masked multi-head attention mechanism for generating radiology reports from medical images.

2.2 Chest-X ray Report Generation

The revolution at CNN led to great success in the image-processing domain. In the healthcare sector, CNNs have been used in a variety of applications like disease classifiers, brain tumor segmentation, detection of skin lesions, detection of various types of cancer, and many more ([27]). However, in the field of medicine, when it comes to generating reports, there has been limited research. This is mainly because, unlike image captioning, the text required for medical reports tends to be lengthy, which results in increased sequential information dependency. [28] have employed convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to identify specific diseases in X-ray images and generate reports on them.

These studies describe related work in medical image analysis and natural language processing. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) were specifically employed by [29] to classify diseases from medical images and produce relevant descriptions. Moreover, it categorizes chest X-rays and produces associated reports, [30] suggested a text-image embedding network (TiNet) that makes use of a CNN-RNN architecture with multi-level attention. [31] developed a multi-task learning framework that predicts keywords related to critical information in medical findings and generates text descriptions using a hierarchical LSTM and co-attention network. ([32]) also worked on generating accurate chest X-ray reports. Their model first processed the image through an image encoder and then a sentence decoder. The main idea of this research work was to build a domain-aware model to detect relevant topics and areas to generate medical reports. They used a reward-based system to keep the clinical relevance and the natural language generation process intact. Table 1 represents some of the notable work in the chest X-ray report generation domain.

3 Proposed Methodology

As shown in figure (2), the proposed framework is a hierarchical framework comprising four main components, namely: (1) an image feature extractor; (2) a text embedding extractor, (3) a disease classification module; and (4) a report generator.

The image feature extraction module uses a multiview image encoder to extract global visual features divided into various low-dimensional visual embeddings from chest X-ray images. The text encoder simultaneously reads clinical records, such as doctor’s notes, and summarizes the content in text-summarized embeddings. The contextualized embeddings that pertain to disease-related topics are further formed by merging the visual and text-summarized embeddings. Lastly, the report generator module utilizes these disease embeddings to generate the report (one word at a time). The following subsections will discuss the proposed hierarchical framework in detail.

Chest X-Ray Report Generation		
Year	Approach	Remark
2015 ([10])	An approach based on visual-semantic alignment and visual-semantic attention mechanisms has been proposed. The approach outperforms existing methods, emphasizing the significance of integrating visual and semantic information for generating accurate and descriptive image captions.	Visual-semantic attention mechanisms often rely on accurate image segmentation to identify regions of interest. If the segmentation process is not robust or contains errors, it can affect the attention mechanism's performance.
2018 ([21])	A multimodal recurrent approach based on text-modality mechanisms has been proposed that leverages an attention mechanism to focus on import image regions.	The proposed approach has a complex architecture that requires joint training of both the visual and textual components. This complexity increases the computational resources and training time needed. It may also make the model more susceptible to overfitting or convergence issues, especially when dealing with insufficient multi-modal training data.
2019([32])	A CNN-A RNN-based approach has been proposed for generating text-based reports.	Usage of RNN for report generation struggles to capture long-range dependency and complex relationships between textual elements, potentially resulting in incomplete or inaccurate reports.
2020 ([33])	A hierarchical approach has been proposed that utilizes pathology tags along with multi-head attention mechanisms.	The hierarchical approach involves separate training and inference processes, which can significantly increase the computational resources and time needed.
2021 ([23])	A transformer-based approach has been proposed.	Transformers typically require extensive fine-tuning and hyperparameter tuning to achieve optimal performance.
2022 ([34])	A multi-attention-based sparse radiology report generation approach have been proposed.	Extensive pruning techniques were employed to reduce the computational complexity of the proposed framework

Table 1 Comparative study of latest existing Chest-X ray report generation work.

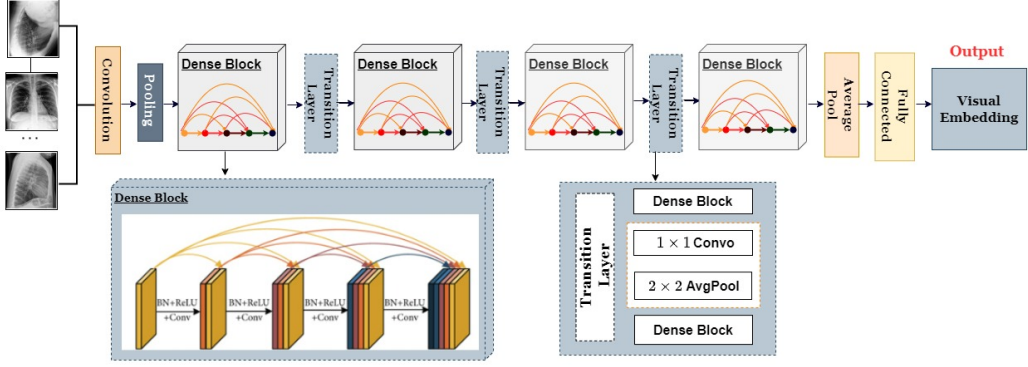


Fig. 3 Feature extraction by using Densenet-121

3.1 Image Feature Extractor

To extract general and regional visual features from the input chest x-ray images, a convolutional neural network (CNN)-based image encoder is utilized, as shown in figure (3). The proposed image encoder backbone is based on DenseNet-121 ([35]) pre-trained on an image-net dataset ([13]). It takes input in the form of 256×256 pixels and works as a multi-view encoder ([36]).

A multi-view chest X-ray encoder takes the chest X-ray images as input in the form of a set of $\{Y_i\}_{i=1}^w$ [where w is the number of chest X-ray images relating to the same subject] and gives output in the form of visual embedding $D_{\text{img}} \in R^c$. It can be summarized as follows:

$$\{Y_i\}_{i=1}^w \rightarrow \boxed{\text{Multiview Encoder}} \rightarrow \{D_{\text{img}} \in R^c\}$$

3.2 Text Embedding Extractor

It takes the clinical document T as input, which consists of k words. Henceforth, input is tokenized to generate individual tokens. Further positional embeddings and input embeddings are generated in such a manner that each token is mapped to generate word embeddings as follows: $\{\nu_1, \nu_2, \dots, \nu_k\}$, where each ν_i ($\forall i$ from 1 to k) is a vector of size e .

Furthermore, the generated word embeddings are transmitted to the transformer encoder ([20]). The transformer encoder architecture described above comprises six identical layers, with two sub-layers per layer. The first sub-layer employs an attention mechanism, whereas the second sub-layer is a straightforward, position-wise, fully connected feed-forward network. It gives us a set of hidden states $X = \{x_1, x_2, \dots, x_k\}$, each x_i ($\forall i$ from 1 to k) is a vector of size e representing the features of the i^{th} word, based on other textual elements.

$$x_i = \text{Transformer Encoder}(\nu_i \mid \nu_1, \nu_2, \dots, \nu_k). \quad (1)$$

After obtaining the hidden states X , the document T is further summarized to generate text-summarized embeddings as follows:

$$D_{\text{txt}} = \text{Softmax}(QX^\top)X. \quad (2)$$

Where $D_{\text{txt}} \in \mathbb{R}^{n \times e}$, where e is the embedding dimension, and n is the number of diseases.

The matrix Q is generated by arranging the set of random-initialized vectors $\{q_1, q_2, \dots, q_n\}$, where $q_i \in \mathbb{R}^e$. The term $\text{Softmax}(QX^\top)$ represents the report's word attention heatmap for the n diseases queried. The idea is to query the diseases (e.g., pneumonia) from the text document T , emphasizing primarily on the most pertinent words (e.g., cold or lack of breathing) pertaining to that disease by utilizing a similarity dot product for vectors.

3.3 Disease Classification Module

This module consists of two sub-parts, namely: (1) the contextualised diagnosis feature embedding module and (2) the enhanced diagnosis feature embedding module. The former sub-module combines the image and text features, while the later one enhances the disease information. The following subsection discusses them in detail.

3.3.1 Contextualized Diagnosis Feature Embedding

It takes visual embedding ($D_{\text{img}} \in \mathbb{R}^{n \times e}$, where e is the embedding dimension & n is the number of diseases) generated by image feature extractor and text summarized embedding ($D_{\text{txt}} \in \mathbb{R}^{n \times e}$) as input to generate contextualized disease representation $D_{\text{concat}} \in \mathbb{R}^{n \times e}$ as:

$$D_{\text{concat}} = \text{LayerNorm}(D_{\text{img}} + D_{\text{txt}}). \quad (3)$$

This combination of visual & textual features enhances the report's performance by integrating various information sources ([33]).

3.3.2 Enhanced Diagnosis Feature Embedding

The main idea behind enhanced diagnosis feature embedding is to encode additional information about disease states (examples of categories include positive, negative, uncertain, or unmentioned). Formally, contemplate state embedding $S_e \in \mathbb{R}^{r \times e}$, where r denotes the total number of states and e refers to the embedding dimension. The matrix S_e is initialized at random and then learned through the classification of D_{concat} , which represents the concatenated features for multi-label classification.

For each disease, the likelihood that it is one of r disease states is determined using a softmax function as follows:

$$p = \text{Softmax}(D_{\text{concat}} S_e^\top)$$

Here, the classification loss is computed as:

$$\mathcal{L}_C = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^r g_{ij} \log(p_{ij})$$

where $p_{ij} \in (0, 1)$ represents the predicted values and g_{ij} represents the ground truth values for the i disease. In addition, the state conscious embedding $D_{\text{states}} \in \mathbb{R}^{r \times e}$ is computed by element-wise multiplication, as follows:

$$D_{\text{states}} = \begin{cases} gS_e, & \text{if training phase} \\ pS_e, & \text{otherwise} \end{cases}$$

where p is the predicted value and g is the ground truth.

The ground-truth labels for disease-related topics are represented using one-hot vectors. These vectors have a size of $n \times r$, where n represents the number of disease instances and r represents the number of possible disease states. Each disease instance is linked to a disease state through a one-hot vector, where the true state is indicated by a value of 1 and other states are represented by 0.

Ultimately, the enhanced diagnosis embedding D_{enhanced} combines the state-conscious diagnosis embedding (D_{states}) with disease names (D_{topics}), and the disease features (D_{concat}) as:

$$D_{\text{enhanced}} = D_{\text{states}} + D_{\text{topics}} + D_{\text{concat}}$$

Here D_{topics} matrices, initialized randomly with dimensions $n \times e$, serve to generate diseases or topics. They are learned during training in the medical report generation pipeline.

3.4 Report Generator

This module takes enhanced diagnosis feature embedding (D_{enhanced}) as an input and generates a report as shown in figure (4). This module is based on masked multi-head attention transformer-based encoder-decoder approach and is discussed in detail in the following sections.

3.4.1 Encoder

Its architecture is inspired by ([20]). It takes the input in the form of enhanced feature disease embeddings as $D_{\text{enhanced}} = \{d_i\}_{i=1}^n$. These feature embeddings are augmented with positional encoding, which provides information regarding the relative positions of elements within a sequence. In addition, it is transmitted through twelve layers of multi-head self-attention, add & norm, and feed-forward neural networks. By layering multiple layers of these operations, the encoder gradually produces concealed states as its output. Hidden states for each word at position $x_i \in \mathbb{R}^e$ are computed in the medical report derived from preceding words and enhanced feature embeddings

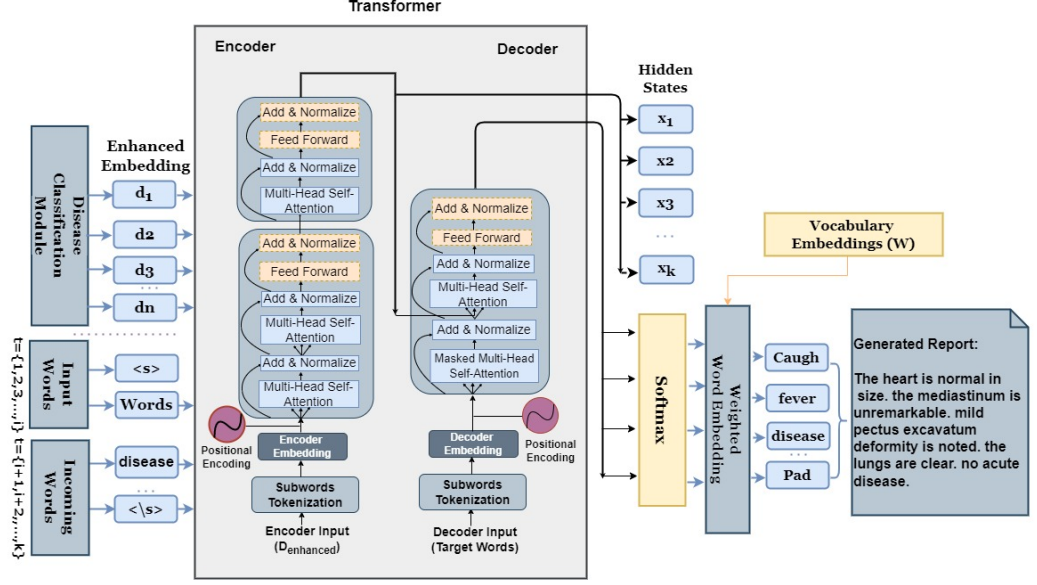


Fig. 4 Shows the overall pipeline of the report generator module consisting of an encoder and decoder architecture with multi-head self-attention mechanism

$D_{\text{enhanced}} = \{d_i\}_{i=1}^n$, as:

$$x_i = \text{Encoder}(\nu_i \mid \nu_1, \nu_2, \dots, \nu_{i-1}, d_1, d_2, \dots, d_n).$$

3.4.2 Decoder

Transformer decoders, like encoders, are constructed with multiple layers. These layers are constructed by sandwiching a masked multi-head self-attention component between a normalizer and a feed-forward layer, as shown in figure (4). The decoder takes the output hidden states $X = \{x_i\}_{i=1}^k \in \mathbb{R}^{k \times e}$ from the encoder as its initial input, where k is the document length and e is the number of states. It also incorporates a separate input, referred to as the "target sequence," which represents the desired output sequence. The decoder generates the output sequence step by step, attending to the encoder's hidden representations and using its own attention mechanisms. At each layer of the decoder, it predicts tokens for each word position that contains information about the generated output sequence.

Afterwards, the decoder computes the probability distributions over the vocabulary W for the next word using a softmax function:

$$p_{\text{word}} = \text{Softmax}(XW^T).$$

Here, $W \in \mathbb{R}^{v \times e}$ is the vocabulary embedding, v is the vocabulary capacity, and k is the document length. The likelihood of choosing the j^{th} word from the vocabulary

W for the i^{th} position in the generated medical report is represented by $p_{\text{word},ij}$. The decoder loss is defined as the cross-entropy of the ground truth words g_{word} and the predicted words p_{word} :

$$\mathcal{L}_{\mathcal{G}} = -\frac{1}{k} \sum_{i=1}^k \sum_{j=1}^v g_{\text{word},ij} \log(p_{\text{word},ij})$$

The generated report is represented by the weighted word embeddings $\hat{W} \in \mathbb{R}^{k \times e}$ obtained by multiplying the probability distribution p_{word} by the embedding matrix W :

$$\hat{W} = p_{\text{word}} W$$

Our model is exhaustively trained by minimizing the total loss, which consists of two components: the classification loss $\mathcal{L}_{\mathcal{C}}$ and the decoder loss $\mathcal{L}_{\mathcal{G}}$. The aggregate loss $\mathcal{L}_{\text{total}}$ is computed as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\mathcal{C}} + \mathcal{L}_{\mathcal{G}}$$

This approach ensures that the classification and report generation tasks are optimized together during the training process. By minimizing the total loss, our model learns to effectively balance the performance of both tasks and achieve overall better results.

4 Experimental Analysis

The following section provides details about experimental settings such as databases, experimental setup, and evaluation metrics. Furthermore, this section also presents a comparative analysis with state-of-the-art contemporary techniques.

4.1 Dataset Description

We used a publicly accessible open-i dataset ([13]) to determine the validity of the proposed method. This dataset contains 3,955 studies in radiology. As depicted in figure (5), these investigations are associated with 7,470 frontal and lateral chest X-ray images. 70% of our experimental data was used for training, 10% for validation, and 20% for assessment. Combining the impression and findings sections to produce the desired output was the method used in previous studies ([31]) and ([33]) for the generation of medical reports. We have accumulated 2069 reports for training, 590 for testing, and 296 for validation.

4.2 Implementation Setup

For implementing the proposed framework, we have utilized PyTorch2 ([37]). We trained the network using an Nvidia Quadro RTX 5000. The learning rate was set to $3e-5$, and a time delay was included to improve the training. We have used the Adam optimizer with a batch size of 8.



Fig. 5 Dataset frontal and lateral chest X-ray images

4.3 Evaluation Metrics

The proposed framework has been evaluated based on the following metrics:

- **Bleu-score:** This metric assesses the similarity between a hypothesis sentence and multiple reference sentences by analyzing individual N-grams ([38]). It assigns a value between 0 and 1, with a value closer to 1 indicating a greater level of similarity. The formula to calculate the Bleu-score is as follows:

$$\text{Bleu} = \text{BP} \cdot \exp \left(\sum_{n=1}^N W_n \log p_n \right)$$

Where,

BP : Brevity penalty,

N : No. of n-grams,

W_n : Weight for each modified precision,

p_n : Modified precision.

- **ROUGE-L:**It determines the length of the longest common subsequence between the generated and reference summaries. It considers the length of the common subsequence and the total length of the reference summary.

The formula for ROUGE-L is as follows:

$$\text{ROUGE-L} = \frac{(\text{Length of longest common subsequence})}{(\text{Length of reference summary})}$$

- **Meteor score:**It combines precision, recall, and alignment-based measures to evaluate translation quality. It considers not only exact word matches but also stem matches, synonym matches, and paraphrase matches ([39]). The precision and recall values for unigrams, which correspond to individual words, are computed, and a harmonic mean F_1 is then calculated in the following manner:

$$F_1 = \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})}$$

To calculate the final METEOR score, a penalty factor is applied to address excessive or insufficient uni-grams and the process is carried out as follows:

$$\text{METEOR} = (1 - \text{penalty}) * F_1$$

The penalty factor depends on the percentage difference between the uni-gram precision and recall.

- **Area Under the Curve (AUC):** It is a quantitative measure of the performance of a binary classification model ([40]). It provides a concise summary of the model's ability to differentiate between positive and negative instances.

Calculating the area under the Receiver Operating Characteristics (ROC) curve yields the AUC. It ranges from 0 to 1, with greater values denoting improved classification performance.

Apart from the above-mentioned evaluation metrics, we have also evaluated our proposed framework on commonly used metrics like precision, recall, F1-score, and accuracy.

4.4 Experimental Results

For validating our proposed framework, we have conducted two types of performance evaluation : (1) Linguistic performance and (2) Clinical Accuracy performance.

Linguistic Performance: For measuring this, we have used standard language evaluation metrics, namely Bleu-1 to Bleu-4 ([38]), ROUGE-L([41]), and METEOR scores([39]). The results of this analysis are presented in (2). It can be inferred from the table (2) that our proposed approach yields competitive results as compared to state-of-the-art approaches working on the same dataset. Very recently, generating chest X-ray reports, a method based on a multi-attention framework ([34]) has been proposed. This method adopts several pruning techniques to eliminate the value of the weight tensor in order to reduce training time. One of the major limitations of pruning techniques is that they are highly dependent on the nature of the data set. Thus, in our implementation, we have avoided it to provide a generic framework irrespective of the dataset under consideration.

Dataset	Models	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE-L
Open-i	S & T([4])	0.273	0.144	0.116	0.082	0.125	0.226
	SA&T([18])	0.328	0.195	0.123	0.080	0.167	0.323
	Liuet. al.([32])	0.359	0.237	0.164	0.113	0.175	0.344
	Hierarchical Method ([42])	0.437	0.323	0.221	0.172	0.224	0.325
	Y. Xue et al.([21])	0.464	0.358	0.270	0.195	0.175	0.344
	am.2021 al. ([23])	0.479	0.359	0.219	0.160	0.205	0.380
	Ours	0.476	0.350	0.271	0.205	0.220	0.385

Table 2 A comprehensive evaluation of the clinical accuracy on various state-of-the-art methods

Dataset	Models	Accuracy	AUC	F1 Score	Precision	Recall
Open-i	S&T ([4])	0.915	N/A	N/A	N/A	N/A
	SA&T([18])	0.908	N/A	N/A	N/A	N/A
	Liu et. al.([32])	0.918	N/A	N/A	N/A	N/A
	Ours	0.975	0.946	0.454	0.668	0.385

Table 3 A comprehensive evaluation of the linguistic performance on various state-of-the-art methods

Clinical Accuracy Performance: For measuring this performance, we have used standard measures such as accuracy, AUC, F_1 score, precision, and recall. The corresponding results are depicted in the table (3). It can be inferred from the table (3) that our model achieved at-par results as compared to state-of-the-art techniques working on the same dataset and with the same training and testing protocol.

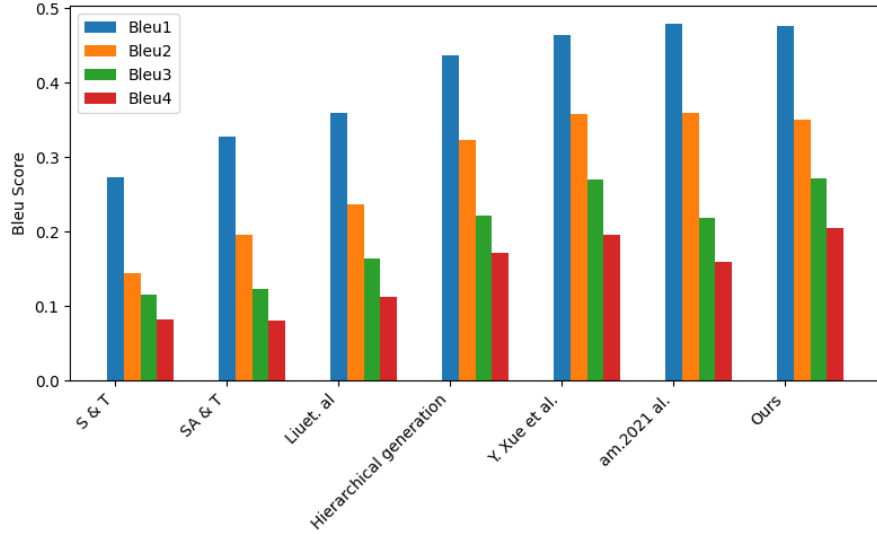


Fig. 6 Bar graph depicting comparative analysis of Bleu score on various state-of-the-art techniques

Figures (6) and (7) depicts the comparative analysis of our proposed approach with various state-of-the-art techniques. It is evident from figures (6) and (7) that our proposed approach has attained relatively high performance as compared to SOTA.

5 Limitations & Future Work

While conducting this research, we observed that generating new sentences not in the training data was a challenging task. In the future, researchers can enhance disease identification techniques and localization by incorporating visual-semantic embeddings associated with direction, orientation, and location. Additionally, our current research does not consider the time-series relationships between medical examinations

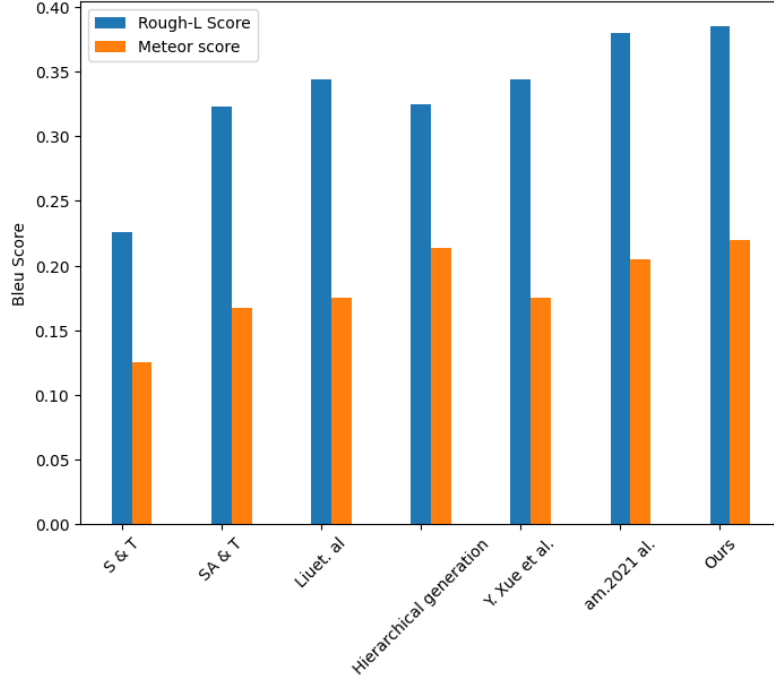


Fig. 7 Bar graph depicting comparative analysis of rough-L & meteor score on various state-of-the-art techniques

of a patient, which is essential for analyzing disease progression. Incorporating such information can enable the system to determine changes in the size or structure of a disease, indicating its improvement or worsening. Addressing these limitations can make the automatic medical report generation system more reliable for practical use in real-world applications.

6 Conclusion

Our approach involves four distinct modules to generate medical reports based on X-ray scans. We have demonstrated that our method surpasses existing techniques in widely used benchmarks through extensive testing and evaluation using multiple metrics. Our approach is flexible and can be adapted to incorporate other types of input data, leading to consistent improvements in performance. Moreover, our model can generate detailed and comprehensive medical reports and seamlessly integrate them into healthcare systems to support medical decision-making. Our work can serve as a valuable resource for various medical applications. Moving forward, we plan to extend our approach beyond X-ray-based medical report generation to other related tasks.

Research Data Availability and Data Policy Statement

The Open-i dataset ([13]) is publicly available for academic and research purposes, providing researchers with access to a vast collection of chest X-ray images and associated metadata. The dataset can be accessed through the Open-i website (<https://openi.nlm.nih.gov/faq>) or designated repositories.

To utilize the Open-i dataset, researchers must adhere to the dataset creator’s terms and conditions, including guidelines for data usage and patient privacy protection. The dataset undergoes regular updates and improvements to maintain its relevance and effectiveness. The accessibility of the Open-i dataset empowers researchers to make progress in automated chest X-ray analysis and report generation, thereby making valuable contributions to the overall understanding of the field.

Competing Interests

The authors of this article are declaring that they have no conflicts of interest to disclose that are relevant to the content presented. The authors have conducted this study with integrity and impartiality, ensuring that their findings and conclusions are based solely on the merits of the research and the data collected.

Compliance with Ethical Standards

It involves upholding principles of patient privacy, data security, transparency, and fairness. To protect patient privacy and confidentiality, strict controls have been implemented to limit access to patient data and comply with relevant privacy regulations. Strong security measures, including encryption and access controls, protect patient information, enabling researchers to develop technology that prioritizes patient well-being and maintains integrity.

References

- [1] Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems* **31** (2018)
- [2] Carter, J.D., Patel, S., Sultan, F.L., Thompson, Z.J., Margaux, H., Sterrett, A., Carney, G., Murphy, N., Huang, Y., Valeriano, J., *et al.*: The recognition and treatment of vertebral fractures in males with chronic obstructive pulmonary disease. *Respiratory medicine* **102**(8), 1165–1172 (2008)
- [3] Dey, N., Zhang, Y.-D., Rajinikanth, V., Pugalenth, R., Raja, N.S.M.: Customized vgg19 architecture for pneumonia detection in chest x-rays. *Pattern Recognition Letters* **143**, 67–74 (2021)
- [4] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164 (2015)
- [5] Xiong, Y., Du, B., Yan, P.: Reinforced transformer for medical image captioning. In: *Machine Learning in Medical Imaging*, pp. 673–680 (2019). Springer
- [6] Kisilev, P., Walach, E., Barkan, E., Ophir, B., Alpert, S., Hashoul, S.Y.: From medical image to automatic medical report generation. *IBM Journal of Research and Development* **59**(2/3), 2–127 (2015) <https://doi.org/10.1147/JRD.2015.2393193>
- [7] Tran, A., Mathews, A., Xie, L.: Transform and tell: Entity-aware news image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13035–13045 (2020)
- [8] Hodosh, M., Young, P., Hockenmaier, J.: Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* **47**, 853–899 (2013)
- [9] Berg, A.C., Berg, T.L., Daume, H., Dodge, J., Goyal, A., Han, X., Mensch, A., Mitchell, M., Sood, A., Stratos, K., *et al.*: Understanding and predicting importance in images. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3562–3569 (2012). IEEE
- [10] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–3137 (2015)
- [11] You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4651–4659 (2016)

- [12] Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4584–4593 (2016)
- [13] Demner-Fushman, McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* **23**(2), 304–310 (2016)
- [14] Al-Malla, M.A., Jafar, A., Ghneim, N.: Image captioning model using attention and object features to mimic human image understanding. *Journal of Big Data* **9**(1), 1–16 (2022)
- [15] Singh, S., Karimi, S., Ho-Shon, K., Hamey, L.: From chest x-rays to radiology reports: a multimodal machine learning approach. In: *Digital Image Computing: Techniques and Applications*, pp. 1–8 (2019). IEEE
- [16] Schmidhuber, J., Hochreiter, S., *et al.*: Long short-term memory. *Neural Comput* **9**(8), 1735–1780 (1997)
- [17] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014)
- [18] al., .: Show, attend and tell: Neural image caption generation with visual attention. In: *International Conference on Machine Learning*, pp. 2048–2057 (2015). PMLR
- [19] Author, A., Author, B.: Automated chest x-ray report generation using deep learning. In: *Proceedings of the International Conference on Artificial Intelligence in Medicine*, pp. 123–135 (2020)
- [20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
- [21] Xue, Y., Xu, T., Rodney Long, L., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: *Medical Image Computing and Computer Assisted Intervention*, pp. 457–466 (2018). Springer
- [22] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.*: Language models are unsupervised multitask learners. *OpenAI blog* **1**(8), 9 (2019)
- [23] Amjoud, A.B., Amrouch, M.: Automatic generation of chest x-ray reports using a transformer-based deep learning model. In: *2021 Fifth International Conference*

- On Intelligent Computing in Data Sciences (ICDS), pp. 1–5 (2021). IEEE
- [24] Yuan, J., Liao, H., Luo, R., Luo, J.: Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: Medical Image Computing and Computer Assisted Intervention, pp. 721–729 (2019). Springer
 - [25] Gasimova, A., Seegoolam, G., Chen, L., Bentley, P., Rueckert, D.: Spatial semantic-preserving latent space learning for accelerated dwi diagnostic report generation. In: Medical Image Computing and Computer Assisted Intervention, pp. 333–342 (2020). Springer
 - [26] Wijerathna, V., Raveen, H., Abeygunawardhana, S., Ambegoda, T.D.: Chest x-ray caption generation with chexnet. In: Moratuwa Engineering Research Conference (MERCon), pp. 1–6 (2022). <https://doi.org/10.1109/MERCon55799.2022.9906263>
 - [27] Sarvamangala, D., Kulkarni, R.V.: Convolutional neural networks in medical image understanding: a survey. *Evolutionary intelligence* **15**(1), 1–22 (2022)
 - [28] Dong, Y., Pan, Y., Zhang, J., Xu, W.: Learning to read chest x-ray images from 16000+ examples using cnn. In: IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), pp. 51–57 (2017). IEEE
 - [29] Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M.: Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2497–2506 (2016)
 - [30] Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9049–9058 (2018)
 - [31] Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
 - [32] Liu, G., Hsu, T.-M.H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. In: Machine Learning for Healthcare Conference, pp. 249–269 (2019). PMLR
 - [33] Srinivasan, P., Thapar, D., Bhavsar, A., Nigam, A.: Hierarchical x-ray report generation via pathology tags and multi-head attention. In: Proceedings of the Asian Conference on Computer Vision (2020)
 - [34] Kaur, N., Mittal, A.: Chexprune: sparse chest x-ray report generation model using multi-attention and one-shot global pruning. *Journal of Ambient Intelligence and*

- [35] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
- [36] Liang, Q., Li, Q., Zhang, L., Mi, H., Nie, W., Li, X.: Mhfp: Multi-view based hierarchical fusion pooling method for 3d shape recognition. Pattern Recognition Letters **150**, 214–220 (2021)
- [37] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
- [38] Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
- [39] Banerjee, S., Lavie, A.: 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, pp. 65–72 (2005)
- [40] Ling, C.X., Huang, J., Zhang, H., *et al.*: Auc: a statistically consistent and more discriminating measure than accuracy. In: Ijcai, vol. 3, pp. 519–524 (2003)
- [41] Lin, C.-Y.: Rouge: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
- [42] Krause, J., Johnson, J., Krishna: A hierarchical approach for generating descriptive image paragraphs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 317–325 (2017)