VER
Vascular & Endovascular Review

Transforming Healthcare Reporting: Using Learning-Based Encoding-Decoding and Attention Mechanism For AI-Powered Radiology Report Generation

Deepika Gupta*[1], Suma Dawn[2]

[1]Jaypee institute of Information and Technology, Sector 62, Noida, Uttar Pradesh, 201306 India
[2]Jaypee institute of Information and Technology, Sector 62, Noida, Uttar Pradesh, 201306 India

ABSTRACT

Manual generation of medical imaging reports is usually tedious and prone to errors, particularly for junior physicians, and continues to be a clinical workflow bottleneck. Although previous approaches based on LSTM or BERT-based models have made advancements, they usually fail to integrate visual and textual modalities in an effective way, which hinders report accuracy. This paper introduces a new multi-modal deep learning architecture for end-to-end automated radiology report generation from X-ray images to fill the gaps. Our method integrates an encoder-decoder model with a co-attention mechanism to learn visual features and semantic context simultaneously. We investigate three settings: (1) ChexNet121-based visual encoding with LSTM decoding (BLEU: 0.268), (2) ChexNet121 with attention-augmented decoding (BLEU: 0.185), and (3) our proposed VGG16-based encoder with contextual word embedding (BLEU: 0.844). The significant BLEU improvement demonstrates the key role of stable visual feature extraction in report quality. The originality of this work is in combining VGG16 as a feature extractor, which has been underutilized in recent research, and a co-attention mechanism for efficient long-distance semantic generation. In contrast to previous research that mainly addresses language model enhancement, our approach improves image-text alignment, a critical necessity for clinical applicability. This system shows great promise for real-world application, providing improved accuracy, clinical usefulness, and scalability. It offers a useful tool to help radiologists produce accurate, consistent, and interpretable reports from medical images.

KEYWORDS: Medical Report Generation, LSTM, VGG16, Natural Language Processing, Pretrained Model.

How to Cite: Deepika Gupta, Suma Dawn, (2025) Transforming Healthcare Reporting: Using Learning-Based Encoding-Decoding and Attention Mechanism For AI-Powered Radiology Report Generation, Vascular and Endovascular Review, Vol.8, No.2s, 71-80.

INTRODUCTION

Radiology-based medical reports are critical documents that help patients, doctors and caregivers to understand and interpret medical imaging [1] studies. These aid in imaging findings, helping with a variety of medical diseases diagnosis and treatment. Such radiology reports [2] help in diagnosis, treatment planning, disease monitoring, communication, and documentation in patient care. Generically, radiology report writing involves detailed and accurate interpretations of medical imaging studies by radiologists. Generation of such reports requires a combination of clinical expertise, technical knowledge, attention to detail, and effective communication skills. By providing accurate and timely interpretations of medical imaging studies, radiologists are essential to the patient care continuum and the diagnostic procedure.

The radiology medical reporting process begins with the radiologist reviewing medical images from a range of imaging techniques, including nuclear medicine research, CT, MRI, ultrasound, and X-rays.

They carefully examine the images to identify anatomical structures, abnormalities, lesions, and other relevant findings. Based on their analysis of the medical images and clinical context, radiologists provide interpretations of the imaging findings. They describe any abnormalities observed, characterize their features (such as size, location, and morphology), and assess their significance in relation to the patient's condition. Radiology reports often include recommendations for additional imaging studies, follow-up examinations, or consultations with other specialists based on the findings. These recommendations help guide the referring physician in the appropriate management of the patient's condition.

Radiology reports typically follow a structured format to ensure consistency and clarity. This format may include sections such as patient demographics, imaging technique, clinical history, findings, interpretation, impression, recommendations, and technical details. Radiologists use specialized medical terminology and standardized language to accurately describe imaging findings and convey diagnostic information. Clear and precise communication is essential to ensure that referring physicians understand the content of the report and can make informed decisions regarding patient care. Radiology reports serve as legal documents that document the interpretation of imaging studies and support medical decision-making. They provide a permanent record of the radiologist's findings and recommendations for future reference.

Benefits of Automatic Radiology Report Generation include (i) Efficiency: Automatic report generation streamlines the reporting process, reducing the time and effort required for radiologists to create reports manually. (ii) Consistency: AI algorithms provide consistent and standardized interpretations of imaging studies, minimizing variations in reporting among different radiologists.

(iii) Speed: Automated report generation enables rapid turnaround times for reporting, allowing healthcare providers to receive timely diagnostic information for patient care. (iv) Quality: AI systems can help improve the quality and accuracy of radiology reports by leveraging advanced image analysis techniques and incorporating clinical context into the interpretation process. And (v) Scalability: Automatic report generation can scale to accommodate large volumes of imaging studies, ensuring efficient and timely reporting for a large number of patients.

Healthcare personnel are able to concentrate more on patient care and decision-making thanks to the efficiency that autonomous report generation brings, which eventually raises the standard of care provided. Automatic medical report generation requires careful consideration of regulatory and ethical considerations, including patient privacy and data security. Additionally, collaboration with healthcare professionals is essential to ensure that the generated reports are accurate, clinically relevant, and aligned with best practices in patient care. Automatic medical report generation typically involves the following steps: (i) Data Collection: Gather a large dataset of medical reports and associated patient data. This dataset should include a diverse range of medical conditions, imaging studies, and clinical contexts; (ii) Preprocessing: Clean and preprocess the data to ensure consistency and remove noise. This may involve tasks such as text normalization, tokenization, and data augmentation; (iii) Model Selection: Choose appropriate deep learning architectures for automatic report generation. Recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models such as GPT (Generative Pre-trained Transformer) are popular options. (iv) Training: Utilize the pre-processed dataset to train the chosen model.

This involves feeding input data (such as medical imaging studies or textual descriptions) into the model and optimizing its parameters using backpropagation and gradient descent; (v) Assessment: Examine the performance of the trained model using an independent validation dataset.

Metrics such as BLEU score, ROUGH, METEOR and CIDEr (for text generation tasks) or accuracy (for classification tasks) can be utilized to evaluate the correctness and effectiveness of the model; (vi) Fine-tuning: To enhance the model's performance, make necessary adjustments. This could entail changing the architecture, adding more training data, or altering the hyperparameters; (vii) Deployment: Install the trained model in a working setting so it can automatically provide medical reports. This could entail creating a stand-alone report-generation application or incorporating the model into already-in-use healthcare systems; and (viii) Monitoring and maintaining it: Keep an eye on the deployed model's performance and update it as necessary. This could entail updating the model's algorithms in response to user comments or retraining it on fresh data.

LITERATURE SURVEY

In the last few years, deep learning for image processing has advanced remarkably [1], with encouraging findings seen in a range of disciplines, including radiology [2]. Improvements in patient care [2] and radiological diagnostics have been made possible by automatic radiology report generation. As a result, Consequently, automated systems capable of processing the growing quantity of medical images in a correct and effective manner. Image captioning presents a major artificial intelligence difficulty in the realm of medicine science, involving the generation of textual descriptions from image content. This task requires the AI system to effectively interpret and convey information from images into coherent and meaningful captions.

In regions with limited healthcare resources, less-experienced radiologists and pathologists may find crafting medical imaging reports challenging. Meanwhile, for seasoned professionals, the task can be both time-consuming and tedious. To address these challenges comprehensively, envision a scenario where a computer could analyze an X-ray image of chest, mirroring the expertise of a radiologist, and effortlessly produce detailed findings in textual format. Presented approach involves leveraging the two models which are Encoder-Decoder Model and Attention model to address our specific problem.

There are many researches which have been done with handling many challenges like report generation which contains many heterogeneous forms of information, abnormal regions in x-ray imaging etc. Jing et al. [3] introduce a multi-task learning approach to predict tags and generate descriptions concurrently. Using a co-attention mechanism, this approach investigates both visual and semantic data in order to accurately locate anomalies and describe them. Hierarchical LSTM network used for improved capture of long-range semantics and generation of high-quality text. VGG19 was used for feature extraction and got bleu score 0.247, Meteor 0.217, Rough 0.447 and CIDER 0.327 respectively. A CNN-RNN structure is used by Shin et al. [4] to forecast labels in images of the chest, such as positions and seriousness. In [5] Zhang et al., the objective is to produce semi-structured pathology reports, limited to 5 predefined topics. Crafting descriptions within imaging reports often entails lengthy passages spanning multiple sentences, posing significant challenges. Rather than relying on a single-layer LSTM (Hochreiter and Schmidhuber, 1997) [6], which may struggle with modeling extended word sequences, capitalize on the structured nature of the report. Employing a hierarchical LSTM enables it to generate lengthy texts more effectively. Enhanced by the co-attention mechanism, this hierarchical LSTM initially generates overarching topics and subsequently refines them to produce detailed descriptions aligned with the identified topics Recently, image captioning has shown the efficiency of attention mechanisms [7]. A spatial-visual attention method is presented by Xu et al. (2015) and is used to extract picture features from CNN intermediate layers. In the meanwhile, You et al. (2016) propose a semantic attention method that centers on tags related to given images.Add a co-attention technique to report creation to improve the use of both visual characteristics and semantic tags. Transformers, attention-only models that may benefit from GPU simultaneous execution and show quicker to train than recurrent models due to their sequential character, are currently replacing recurrent models in the Natural Language Processing (NLP) domain [9–11]. Pre-trained transformers with generative capabilities that are already available, like GPT2 [11], can be trained more quickly, require no vocabulary to be specified, and make use of previously learned word structures and punctuation. For these reasons,

conditioning a pre-trained transformer has been shown to have benefits. Natural reporting on the IU-Xray dataset was generated in [13] by merging the attention

processes on both the visual and predicted label embeddings, or co-attention, and extracting visual features from a VGG network that had been pre-trained on Imagenet. To create the reports, the co-attention output is subsequently fed into two hierarchical LSTMs, one for phrases and one for words. used the multi-view data from the IU-Xray dataset to train a Resnet152 on the Chexpert dataset [13] to predict visual attributes and tags from the patient's front and side images.

The ports were then generated by hierarchical LSTMs that resembled those in [13]. [14] makes use of knowledge graphs that incorporate prior information about chest observations; [15] uses Chexnet models to extract graph node attributes from IU-Xray pictures; Layered LSTMs that paid attention to the graph were used in the construction throughout the study.

Complete deep neural network that generates radiological reports from CXR pictures that are therapeutically helpful by using contextual word representations. The suggested network [16], called `RadioBERT`, makes use of transfer learning and `DistilBERT` for contextual word representation. The following performance scores have been attained by them: CIDEr = 0.5563, ROUGE = 0.897, BLEU-1 = 0.772, BLEU-2 = 0.770, BLEU-3 = 0.768, and BLEU-4 = 0.767 which is really remarkable. Despite these advancements, existing methods often prioritize either linguistic quality or visual interpretation, but rarely optimize both in a unified manner. Many rely heavily on pretrained transformers or graph-based models without carefully tailoring the visual feature extraction process, leading to suboptimal alignment between the image content and generated text. Additionally, some models lack interpretability, struggle with long-range semantic coherence, or are not computationally efficient for real-time clinical use.

Our work addresses these gaps by integrating a carefully chosen visual encoder (VGG16) with a hierarchical LSTM and co-attention mechanism that jointly captures spatial and semantic dependencies. Unlike prior approaches that treat report generation as a purely text generation task or rely heavily on large pretrained models, our method emphasizes precise image-text grounding, computational simplicity, and clinically relevant report composition. This makes it more suitable for deployment in real-world, resource-constrained clinical environments where both accuracy and efficiency are critical.

PROPOSED METHOD

A thorough description of our suggested model is provided in this section. Figure.1 shows work flow of proposed model which follows a structured approach consisting of five key steps is presented. The model is trained using the OpenI IU-dataset, which includes images and corresponding reports. The proposed model for automated radiology report generation follows a systematic approach consisting of several essential steps. To begin, preprocessing was performed that prepared the OpenI IU-dataset[20, 21], which included a collection of medical images and their corresponding reports. This dataset served as the foundation for training our model, ensuring it learns from a diverse set of clinical scenarios and imaging findings.

Special type of Convolutional Neural Network (CNN) model which is VGG16 a pre-trained used to extract features from the medical images. VGG16 provides a strong representation of the visual content seen in x-ray images by effectively extracting structured and informative features from images. These extracted features serve as the initial input to our model. The extracted features from VGG16 were used as input to generate textual reports. This involved converting the encoded visual information into natural language descriptions.

The next critical step involved converting these extracted features into coherent textual descriptions. Here in, LSTM (Long Short-Term Memory) networks, known for their effectiveness in handling sequential data such as text was employed. The LSTM models are integrated into a sequence-to-sequence architecture, where they generate the radiology report one word at a time. LSTM networks were employed for their effectiveness in handling sequential data, particularly text. These models generated the report one word at a time, ensuring coherence and relevance to the image features. This process was employed to ensure that the generated reports maintain coherence and accurately reflect the visual features extracted by VGG16.

To enhance the produced output's linguistic relevance and accuracy in the form of text, pre-trained GLOVE word embeddings was implemented. These embeddings map each word into a dense vector space, capturing semantic relationships and contextual meanings. By integrating GLOVE embeddings into our model, we ensure that the generated reports are linguistically informed and clinically relevant. Every word had a 300-dimensional vector representation thanks to the pre-trained GLOVE word embeddings. This embedding is crucial for translating the encoded features into meaningful textual descriptions.

The model was trained with the Adam optimizer at a learning rate of 0.0001, with the parameters of the optimizer being $\beta 1 = 0.9$, $\beta 2 = 0.999$, and $\varepsilon = 1e\text{-}8$. The batch size was 32 to satisfy computational speed and gradient stability. The major loss function utilized was the cross-entropy loss, optimized at each decoding step to predict the next word in the sequence correctly. In addition, an additional attention regularization loss was introduced to promote a more uniform distribution of attention weights over the image regions to further improve the model's capability to process important areas in the medical images. The model was trained for a maximum of 50 epochs with early stopping implemented based on validation BLEU score improvement with a patience of 5 epochs to avoid overfitting. In addition, a learning rate scheduler was applied in order to decrease the learning rate by a factor of 0.1 whenever the validation BLEU score failed to improve for three consecutive epochs in order to ensure more stable and efficient training convergence.

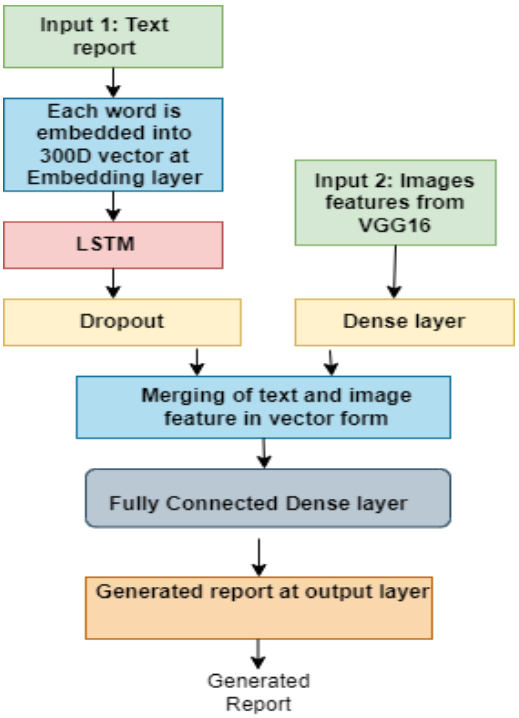

Figure 1: Proposed working model

Ultimately, the BLEU (Bilingual Evaluation Understudy) ,ROUGH, METEOR and CIDEr measure was utilized to evaluate the quality of the reports that were produced. By measuring the degree of resemblance between the reports that are generated and references that are written by humans, BLEU offers an evaluation of how well the model performs in generating precise and therapeutically relevant radiology reports.

By following this structured approach, the proposed workflow allowed for attempting automation and improvement in the efficiency of radiology report generation, providing valuable support to healthcare professionals in diagnostic processes.

Architecture of Proposed Model

The sequence-to-sequence (Seq2Seq) deep learning framework is the foundation of the architecture of the suggested medical report generation model. This model is designed to transform a sequence of input features (derived from medical images) into a sequence of outputs (comprehensive medical reports). The process begins with the encoder, which systematically processes every word in the series of inputs. The encoder gathers and condenses the information it has acquired during the sequence into a context vector, which is a fixed-size vector. The essential elements of the input sequence are contained in this context vector, which also contains the data required for the subsequent processing step. Once the encoder has processed the entire input sequence, The Long Short-Term Memory (LSTM) network, which serves as the decoder, receives the context vector. The LSTM network starts to generate the output sequence, producing each item one by one. This step-by-step generation allows the model to create detailed and coherent medical reports according to the input image attributes.

Figure 2 presents a visual representation of the model's architecture, illustrating the flow from input image features through to the final generated medical report. The proposed architecture is depicted in Figure 3 detailing the use of Seq2Seq model, emphasizing the roles of the encoder, context vector, and LSTM in the process.

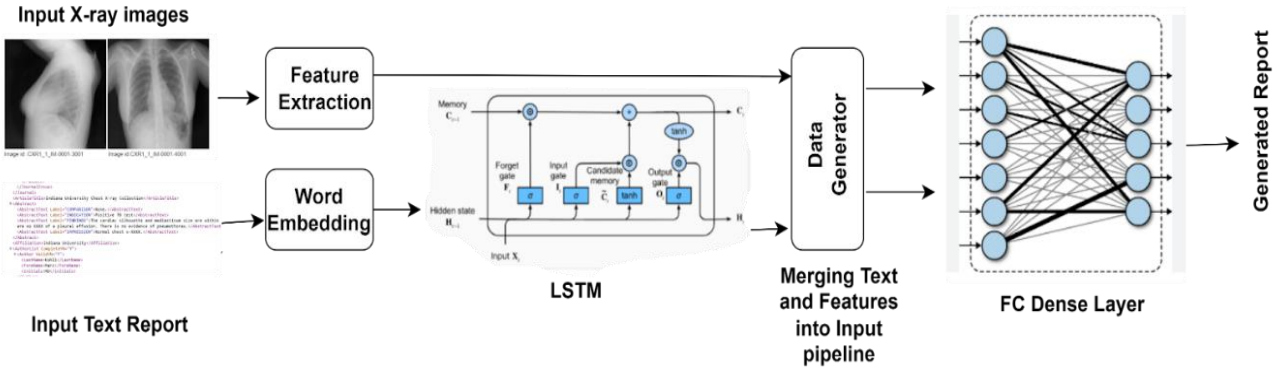

Figure 2. Architecture of proposed model.

There are two main building blocks of the proposed model. First is language generation and another one is converting input image into report form using Encoder-Decoder model.

Language Generation

After feature extraction from the images (x-rays) the next step is feature translation into reports. To extract the information from the image, a pretrained convolutional neural network is required. To generate reports, these attributes ought to be fed into a series of sequence model as illustrated in below figure 3. One word at a time will be output by these models. These terms will be encoded in one hot form. To obtain the real words, argmax of these one-hot encoded vectors are taken. The model will produce the final report after some epochs.

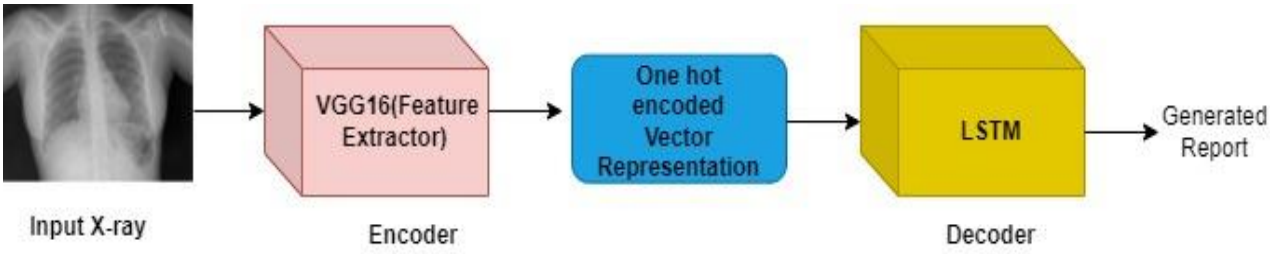

Figure 3: Encoder-decoder(seq2seq) architecture for image captioning

Modeling

Encoder-Decoder will be used for this purpose. A sequence-to-sequence model is a type of deep learning model that generates reports from one sequence of items (that is, in this example, the features of animage).

The encoder's main job is to process every item in the input sequence and gather data so that it can be transformed into a vector known as the context. The decoder receives the context from the encoder and begins creating the output sequence item by item after analysing the entire input sequence. Proposed working model is shown in Figure 4. This model is specifically designed to transform a sequence of input features (extracted from medical images) into a corresponding sequence of outputs (detailed medical reports).

The process starts with the encoder, a crucial component of the Seq2Seq model. The important information is extracted and condensed into these vectors by the encoder as it analyzes every single item in the given input sequence. In this instance, the encoder is a VGG16, widely-used convolutional neural network (CNN) architecture renowned for its efficiency at extracting features from images. The VGG16 model processes the input images and extracts high-level features, which are then compiled into a context vector. Since each patient has two images, the features from both images need to be combined to form a comprehensive representation. As there are two images per patient, so it is required to combine both features of images. Here concat function is used to take images. To combine the features from the two images, a concatenation function is employed. This function merges the feature vectors from both images into a single, unified vector. The concatenated vector thus contains the combined information from both images, providing a richer and more detailed context for the subsequent processing stages. Next, a dense layer is traversed by the concatenated vector, which serves to reduce its dimensionality. This step is crucial for managing the complexity and size of the feature vector, making it more computationally efficient for the model to handle. The dense layer effectively compresses the combined feature vector into a more compact form while preserving the essential information. The final context vector that will be fed into the decoder is this reduced-dimension vector. Finally, the decoder's role is to transform the context vector into a target sequence with a variable length, which in this case is the detailed medical report. The decoder, typically implemented using a Long Short-Term Memory (LSTM) network, begins by receiving the context vector from the encoder. It then generates the output sequence item by item, allowing for the creation of a coherent and contextually accurate medical report. The variable length of the target sequence accommodates the need for detailed descriptions and findings in medical reports, ensuring that the output is both comprehensive and relevant to the input features.

Once the encoder output is obtained, the next step is to convert this encoded information into text. For this purpose, LSTM networks are used, which are highly effective for handling sequential text data. In this case, a sequence-to-sequence (Seq2Seq) model using LSTM networks is utilized. The Seq2Seq model processes the encoded vector to generate the text sequence, which in this scenario is a detailed medical report. The LSTM network within this model generates single word at a time instance, with every word being produced as an output at a specific time step. The generation process is a four-step undertaking. This initiates with input at time steps, implying that at each time step t−1, the LSTM network receives inputs that include the encoder outputs (context vector) and the word embedding vector from the previous time step t−1. These inputs help the LSTM predict the next word in the sequence. This is followed by prediction of word vectors using the provided inputs, wherein the LSTM layer predicts a vector representation for the word at the current time step t. This predicted vector represents the likelihood of various words in the vocabulary being the next word in the sequence. In the next step, the predicted vector is then passed through a softmax layer. The softmax layer processes this vector to produce a probability distribution over the entire vocabulary. This transforms the predicted vector into a one-hot encoded form, which indicates the most probable word based on the highest probability. The final step in the generation is the word selection action. To choose the word with the highest probability, the softmax output is passed on to the argmax function. After then, this word is utilized as both an input and an output for the current and subsequent time steps.

An embedding layer is used to represent words in the sequence as dense vectors. This layer maps each word to a continuous vector space, enabling the model to capture the semantic relationships between words. A GloVe (Global Vectors for Word Representation) model that has already been trained is utilized for this assignment. The GloVe model has been trained on extensive text corpora and provides a 300-dimensional vector representation for each word. These dense vectors encapsulate the semantic meanings of words, allowing the model to produce language that is accurate and coherent within its context.

This encoder output has to be converted to text now. LSTM networks are employed, which is excellent for handling text data, for that purpose. In this case, the sequence-to-sequence model is used i.e. LSTM. When inputs are received by LSTM networks in time steps, one word is obtained as an output at a time. At time t. By using the encoder outputs and the word embedding vector (t-1), the LSTM layer predicts a vector representation for the word as inputs at each time step. After that, a softmax layer processes the vector to turn it into a one-hot encoded form. Relevant terms from vocabulary can be retrieved by utilizing the argmax function. Embedding Layer: Words are represented as dense vectors thanks to the embedding layer. Here, a GLOVE model that has already been trained has been used to map each word into a 300-dimensional representation.

Algorithm 1: ReportGen: Using Attention Model

Input: Ep ⊲ No. of epochs
I_{mg} ⊲ Image from the dataset
Sent=w1,w2,w3,...,wn ⊲ Sentence consisting of words
VGG16 ⊲ Pre-trained CNN over ImageNet
DatagenData_{gen}Datagen ⊲ Data generator
GLOVE ⊲ GloVe Vector Embedding
M ⊲ Model
Procedure:
Start

VGG16 ←$Initialize$
$N \leftarrow 0$
while $N \leq Ep$ do
$V_{features} \leftarrow VGG16(I_{mg})$
$E_{output} \leftarrow Encoder(V_{features})$
$A_{weights} \leftarrow Attention(E_{output}, D_{hidden})$
$C_{vector} \leftarrow ContextVector\ (A_{weights}, E_{output})$
$Concat_{vector} \leftarrow Concatenate(C_{vector}, Embedding(D_{inputprev}))$
$GRU_{output} \leftarrow GRU(Concat_{vector})$
$Report \leftarrow Dense(GRU_{output})$

end while
End
Output:
Report ⊲ The report generated from the image

The image is fed into a caption generator using the "merge" idea. When the RNN has completed thoroughly encoding the prefix, the image is incorporated into the language model. Because it is a late binding architecture, an image representation is not changed with each time step.

Algorithm 1: Algorithm relating the attention mechanism used as a decoder to generate medical report from given X-ray images Dropout is used to regularize the output of the RNN. Neither a non-linear activation function nor dropout regularization is present in the image vector. Prior to feeding the image input vector into the neural network, it needs to be normalized. This was accomplished while extracting features from the VGG16.

In recent advancements, attention mechanisms have been pivotal in helping sequence-to-sequence models perform better since they allow the model to flexibly concentrate on pertinent portions of the input sequence when needed. Building upon the conventional encoder-decoder architecture, introduce a model that integrates attention within the decoder, specifically employing a one-step attention mechanism. Proposed architecture has two primary components. First one is encoder and second one is

decoder. However, unlike traditional models, our decoder incorporates an additional attention layer. At each stage of decoding. The attention mechanism aids in the model's ability to concentrate on particular parts of the input sequence.

Encoder-Decoder with One-Step Attention

In another approach for the image feature extraction, we leverage the pre-trained CheXNet network. The final convolutional layer of CheXNet yields the features, which give the input images a rich representation. The goal of the decoder is to generate the output sequence at each time step by consecutively decoding the input sequence. The encoder output is then produced by feeding these extracted features by ChexNet into the encoder. The encoder output and the decoder's previous hidden state are then fed into the attention model, which uses them to calculate the attention weights. These attention weights are combined with the encoder output to compute the context vector. After that, a Gated Recurrent Unit (GRU) processes the context vector by concatenating it with the embedding vector of the previous decoder input. Ultimately, a dense layer processes the GRU output to provide the model's predictions. Predicted outputs from all these methods are compared in Table 1.

Model Metrics

BLEU Score — Greedy Search

The goal of the proposed systems is to producing radiological reports that look and feel like they were written by human specialists. A number of metrics, including the BLEU (Bilingual Evaluation Understudy) score, are utilized to evaluate the caliber of these reports that are generated automatically

The BLEU score compares generated reports to real reports, with a score of 1 indicating identical reports and 0 indicating complete mismatch. It uses one-hot encoded vectors as output from the decoder model, making categorical cross-entropy a suitable loss function. A precision-based statistic called BLEU evaluates how closely text produced by machines and text translated by humans' line up. The following formula is used to compute it:

$$BLEU = BP \cdot \sum_{n=1}^{N} w_n \log p_n \qquad (1)$$

where N is the number of n-grams, wn is the weight for each modified precision, and pn is the modified precision. BP stands for the brevity penalty. Due to its quickness and ease of use, BLEU is frequently employed; nonetheless, it favours shorter phrases and may provide higher scores to reports that fail to identify anomalies. Therefore, better quality reports are not necessarily indicated by a higher BLEU score. Although results are quite promising as shown in Table 2, so here we can consider achieved BLEU score to be good enough.

RESULTS AND DISCUSSION

Table 1: Results of Bleu_Score of generated reports using different methods.

Method	Feature Extraction Method	Average Bleu Score	ROUGH	METEOR	CIDEr
Dense Encoder+LSTM	ChexNet121	0.268	0.345	0.290	0.550
Attention Model	ChexNet 121	0.185	0.260	0.220	0.400
Dense Encoder+LSTM	VGG16	0.844	0.870	0.820	1.850

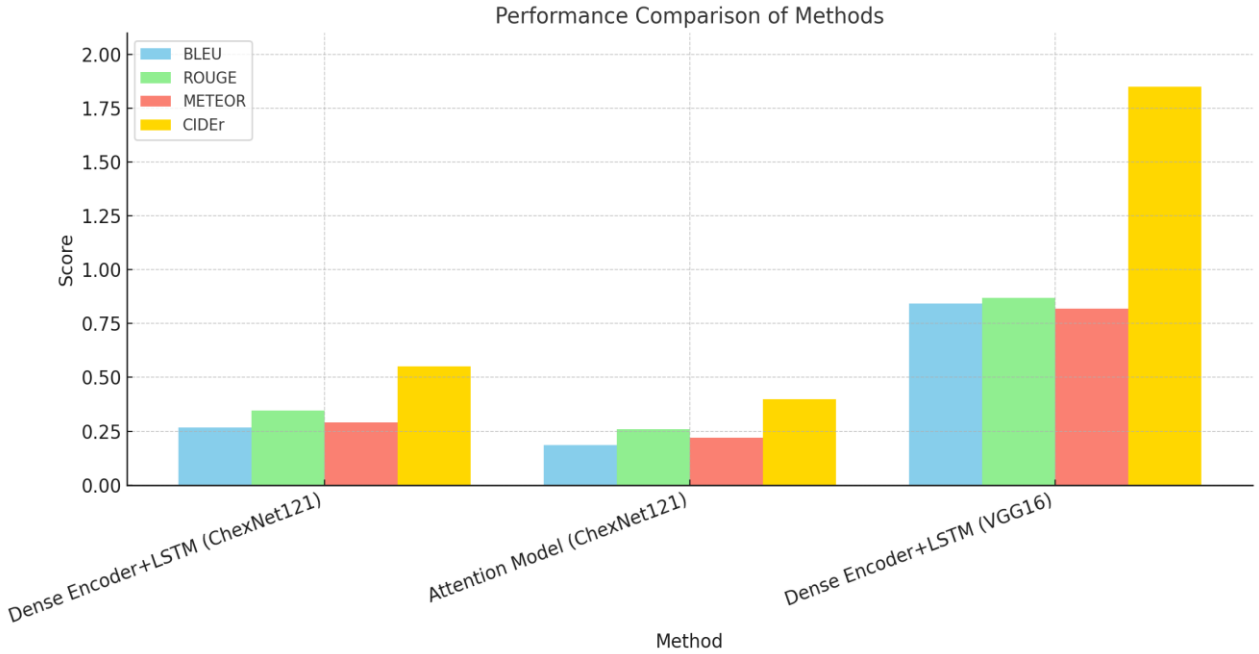

Figure 4: Bar chart comparing the BLEU, ROUGE, METEOR, and CIDEr scores across three methods.

The findings of this study are presented in Table 1, which illustrates the average BLEU scores of generated reports using three different methods.

In the first method, features were extracted using the ChexNet121 pretrained model. These features, along with their respective reports, were fed into an encoder-decoder model to generate medical reports using natural language processing. In this setup, an LSTM network was utilized as the decoder. Used approach yielded an average BLEU score of 0.268. The use of ChexNet121 for feature extraction likely contributed to the model's ability to capture relevant medical image details, while the LSTM decoder helped in generating coherent and contextually appropriate reports.

The second approach employed an attention mechanism to enhance the report generation process. Similar to the first method, features were extracted using the ChexNet121 model. The attention model however, added an additional layer of complexity by enabling the decoder to generate each word in the report by concentrating on different areas of the input image. Despite the advanced mechanism, this process got a lower BLEU score of 0.185. This might be due to the model's inability to effectively leverage the attention mechanism or potential issues with the training process.

In the third approach, the encoder-decoder framework was retained, but the feature extraction was performed using the VGG16 pretrained model instead of ChexNet121. he model's BLEU score significantly improved as a result of this modification, reaching an astounding 0.844. The VGG16 model, known for its effectiveness in image classification tasks, likely provided more robust and discriminative features, which significantly improved the predicted report's quality.

The high BLEU score indicates that this method was particularly successful in producing accurate and fluent medical reports. Table 2 showcases a sample of the generated medical report using the third approach, highlighting the quality and detail of the reports produced with features extracted from VGG16. This example underscores the potential of the encoder-decoder framework, combined with effective feature extraction techniques, in automating the generation of medical reports from imaging data. Ground truth report and generated reports are compared and wrong word predicted was written in red color. Figure 4 visually shows how VGG16-based model significantly outperforms the others.

Table 2: Examples of Reports generated by proposed model (Words that are not correctly predicted are indicated in red.)

Ground Truth	**Generated Report**
the heart again enlarged. aorta tortuous. the lungs are hypoinflated but clear. no pleural effusion pneumothora seen.	the heart size enlarged. aorta tortuous. the lungs are normal and clear. no pleural effusion. no pneumothora.
cardio mediastinal silhouette pulmonary vascular pattern are within normal limits. mildly low lung volumes. no focal infiltrate pleural effusion pulmonary edema. no pneumothora.	the heart size normal limits. the lungs are clear and low volume. no pneumothora pleural effusion. no pneumothora
the heart and lungs have the interval. both lungs are clear and expanded. heart and mediastinum normal.	A heart and lungs have the interval. the lungs are clear and expanded. heart and countors normal.
heart size and pulmonary vascularity appear within normal limits. the lungs are free focal airspace disease. no pleural effusion pneumothora seen. degenerative changes are present the spine	heart size and lung size appear within normal limits. the lungs are free from airspace disease. no pleural effusion pneumothora . many changes are are appear in spine.
the lungs are clear. heart size normal. no pneumothora.	the lungs are clear. heart size normal. no pneumothora.
lungs are clear. no pleural effusions pneumothoraces. heart and mediastinum normal size and contour. degenerative changes the spine.	the lungs are clear. no pleural effusion pneumothorace. heart and counter normal size and clear. many changes in spine.

CONCLUSION

Our framework's main objective is to assist medical practitioners by increasing the accuracy and efficiency of medical image report generating. This framework is designed to address several key challenges such as Information Synthesis by integrating diverse types of information from medical images into a unified framework, and Coherent Text Composition by Generating comprehensive and coherent texts that span multiple sentences or paragraphs, effectively mimicking the way radiologists describe and diagnose conditions.

In this study, we provide a strong framework for simultaneous learning aimed at automating the textual report generation for medical images. The main objective is to increase the correctness and effectiveness of generating medical report in order to empower medical practitioners. Our approach addresses several critical challenges in this domain, including the synthesis of

diverse types of information within a unified framework and the coherent composition of comprehensive texts spanning multiple sentences or paragraphs.

A key innovation in this research framework is the incorporation of a co-attention mechanism, which enables simultaneous exploration of both visual and semantic information, and incorporates precision in description. By jointly attending to visual and semantic information, the model can more accurately describe abnormalities observed in medical images. This is crucial for generating precise and reliable diagnostic reports, enhancing the quality of automated interpretations. Further, the proposed mechanism also plays a crucial role in effectively capturing intricate relation in the long-range semantic description, as well as describing abnormalities observed in medical images, thereby improving the precision of diagnostic reports. This is important because medical reports often contain detailed and interrelated information that spans multiple sentences, reflecting the complexity and nuances of medical diagnoses.

Moreover, the implementation of a hierarchical LSTM network, which excels in capturing intricate long-range semantic relationships, is useful for extracting the complex information inherent in medical reports. By leveraging this architecture, the proposed framework ensures that the generated texts maintain coherence and accurately reflect the characteristics that are taken out of the X-ray images.

To check the accuracy of the proposed framework, we conducted comprehensive evaluations using the IU (Indiana University) medical imaging datasets. These datasets are widely recognized and used in the medical imaging community, by which to evaluate the effectiveness of automated report production systems.

Using the IU datasets, extensive quantitative and qualitative experiments were conducted to verify the efficacy of the recommended approach. The evaluations included both quantitative and qualitative assessments. These evaluations demonstrated the robust performance of the proposal, achieving a significant BLEU score of 0.844 for generated reports, showing a strong degree of consistency between the reference reports and the generated reports written by medical professionals. In addition to the numerical metrics, qualitative evaluations by experts showed that the generated reports were not only accurate but also meaningful and useful in a clinical setting. This means that the reports were evaluated for their practical utility and relevance in real-world medical diagnostics. This metric underscores the reliability and relevance of the approach in the context of medical diagnostics and clinical decision-making.

The automated radiology report generation field has made significant progress with the introduction of the proposed multi-task learning architecture. With key benefits such as enhanced Precision and Efficiency, Support for Clinical Decision-Making and Scalability and Adaptability, the proposed framework can be adapted and scaled to other types of medical imaging data and different medical specializations, making it a versatile tool in the healthcare domain.

AI models can inherit biases from training data, which can influence accuracy, especially in underrepresented groups or conditions. Having diverse and representative datasets is important for reducing such biases and improving fairness in real-world use. Though our framework works well in controlled environments, it should be validated further in various clinical settings. Testing its generalizability across multiple hospitals, specialties, and patient groups is critical to ensure it functions well in actual practice. Reports generated by AI, although precise, risk misdiagnosis, particularly in complicated situations. The system can act as a complementary tool with human monitoring to minimize the possibility of mistakes and enhance patient safety in everyday clinical processes.

Overall, this multi-task learning framework offers a valuable solution for the healthcare industry, providing robust support for interpreting and communicating complex medical imaging data with greater precision and efficiency. It represents a substantial advancement in automated radiology report generation, offering valuable assistance to healthcare professionals in interpreting and communicating complex medical imaging data with greater precision and efficiency.

Statements and Declarations

- **Competing Interests**: The authors declare that they have no competing interests related to the content of this research.

-Funding Information: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author Contributions

Deepika Gupta Conceived and designed the analysis, contributed data or analysis tools, Performed the analysis, Wrote the paper, Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data. Suma Dawn Conceived and designed the analysis, Collected the data, Performed the analysis, Wrote the paper Preparation, Oversight and leadership responsibility for the research activity planning and execution.

Data Availability Statement

The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Research Involving Human and/or Animals

This study did not involve any experiments on humans or animals.

Informed Consent

Not applicable. No human participants were involved directly in this study requiring informed consent.

REFERENCES

1. Y. LeCun, Y. Bengio, and G. Hinton (2015) "Deep learning," Nature, vol. 521, pp. 436-444.
2. J. H. Thrall, X. Li, Q. Li, C. Cruz, S. Do, K. Dreyer, and J. Brink (2018) "Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success," J. Am. Coll. Radiol., vol. 15, no. 3, pp. 504-508.
3. B. Jing, P. Xie, and E. Xing (2017) "On the automatic generation of medical imaging reports," arXiv preprint arXiv:1711.08195.
4. H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers (2016) "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.
5. Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang (2017) "MDNet: A semantically and visually interpretable medical image diagnosis network," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6428-6436.
6. S. Hochreiter and J. Schmidhuber (1997) "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780.
7. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio (2015) "Show, attend and tell: Neural image caption generation with visual attention," in International Conference on Machine Learning, pp. 2048-2057.
8. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo (2016) "Image captioning with semantic attention," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4651-4659.
9. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2017) "Attention is all you need," in Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, pp. 5998-6008.
10. J. Devlin, M. Chang, K. Lee, and K. Toutanova (2019) "BERT: pre-training of deep bidirectional transformers for language understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, vol. 1, pp. 4171-4186.
11. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019) "Language models are unsupervised multitask learners," arXiv preprint arXiv:1901.05408.
12. O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy (2021) "Automated radiology report generation using conditioned transformers," Informatics in Medicine Unlocked, vol. 24, p. 100557.
13. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng (2019) "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 590-597.
14. Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu (2020) "When radiology report generation meets knowledge graph," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, pp. 12910-12917.
15. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng (2017) "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," arXiv preprint arXiv:1711.05225.
16. N. Kaur and A. Mittal (2022) "RadioBERT: A deep learning-based system for medical report generation from chest X-ray images using contextual embeddings," Journal of Biomedical Informatics, vol. 135, p. 104220.
17. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu (2002) "BLEU: a method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 311-318.
18. C.-Y. Lin (2004) "ROUGE: A package for automatic evaluation of summaries," in Text Summarization Branches Out, pp. 74-81.
19. R. Vedantam, C. L. Zitnick, and D. Parikh (2015) "CIDEr: Consensus based image description evaluation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566-4575.
20. http://academictorrents.com/details/5a3a439df24931f410fac269b87b050203d9467d
21. https://academictorrents.com/details/66450ba52ba3f83fbf82ef9c91f2bde0e845aba9