Automatic Radiology Report Generator Using Transformer With Contrast-Based Image Enhancement

Publisher: IEEE

<< Results | < Previous | Next >
Hilya Tsaniya; Chastine Fatichah; Nanik Suciati
All Authors

- III.

  Method

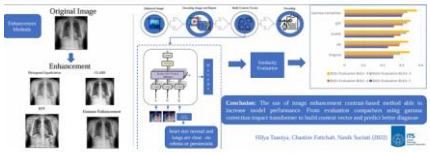
- IV.

  Result and Discussion

- V.

  Conclusion

Show Full Outline
Authors
Figures
References
Citations
Keywords
Metrics
More Like This



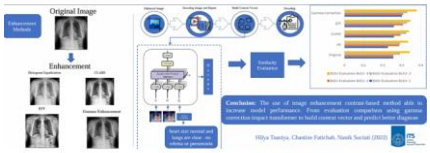
Different contrast enhancement methods like HE, CLAHE, EFF, and Gamma Correction are assessed for radiology image enhancement. The optimal method is integrated into a rad...Show More

Abstract:
Writing radiology reports based on radiographic images is a time-consuming task that demands the expertise of skilled radiologists. Consequently, the integration of techn...Show More

Metadata

Abstract:
Writing radiology reports based on radiographic images is a time-consuming task that demands the expertise of skilled radiologists. Consequently, the integration of technology capable of automated report generation would be advantageous. Developing a coherent predictive text is the main challenge in automatic report generation. It is necessary to develop methods that can increase the relevance of features in producing predictive text. This study constructed a medical report generator model using the transformer approach and image enhancement implementation. To leverage the visual and semantic features, an approach to enhance the noise-prone nature of the medical image is explored in this study along with the transformers method to generate a radiology report based on Chest X-ray images. Four contrast-based image enhancement methods were used to investigate the effect of image enhancement techniques on the radiology report generator. The encoder-decoder model is used with text feature embedding using Bidirectional Encoder Representation from Transformer (BERT) and visual feature extraction utilizing a pre-trained model ChexNet and Multi-Head Attention (MHA) mechanism. The performance of the MHA model with gamma correction is 5% in better with a 0.377 value using the Bilingual Assessment Understudy (BLEU) with 4 n-gram evaluation. MHA also produces 15% better results with a 0.412 value

than the baseline model. This method is able to outperform the baseline model and other previous works. It can be concluded that the use of transformer MHA encoder layer and BERT is effective in leveraging visual and text features. Additionally, the inclusion of an image enhancement approach has been found to have a positive impact on the model's performance.



Different contrast enhancement methods like HE, CLAHE, EFF, and Gamma Correction are assessed for radiology image enhancement. The optimal method is integrated into a rad...Show More

Published in: IEEE Access (Volume: 12)
Page(s): 25429 - 25442
Date of Publication: 09 February 2024
Electronic ISSN: 2169-3536
DOI: 10.1109/ACCESS.2024.3364373
Publisher: IEEE
Funding Agency:

Contents

SECTION I.

Introduction

Medical image diagnoses are getting more complicated as medical imaging technology develops, increasing the need for medical specialists. Data from the Medical Journal of Radiology publications in 2015 disclose a 26% increase in radiologists' workload compared to the previous decade. To get accurate and thorough diagnoses based on medical pictures, radiologists need to compare numerous information including a lot of factors than in the past due to medical developments [1].

This increased workload poses a significant risk of medical report errors, particularly when radiologists work under adverse conditions or experience fatigue. These errors, in turn, can compromise the accuracy of the patient's final diagnosis, especially in cases when radiologists lack experience, leading to an increase in diagnostic inaccuracies.

The need to develop automated radiology generator technology that can provide interpretation and information from the medical image input would beneficially save time and reduce the workload of radiologists efficiently. Radiologists can produce reports with the aid of the automated system, particularly in underdeveloped nations with limited professional resources. This system also can help radiologists by comparing the generated report to manual observation and use the information to make diagnostic judgement.

One of the challenge in the medical image processing is the noise during image acquisition that leading to a reduction in image quality [2]. In other computer vision

cases, such as classification and segmentation, it has been proven that employing enhanced images significantly impacts the performance of models compared to using the original images without enhancement [3]. Previous study also proven the effectiveness of enhanced images in radiology report generation using CNN-LSTM encoder-decoder models [4]. This observation motivated our investigation into the importance of image enhancement and its effect in the medical image captioning. Our exploration involved integrating the enhancement method with transformer attention mechanisms, aiming to discern how this combination could be utilized to elevate the performance of automated systems in the context of medical image interpretation.

This research introduces an innovative architecture for an X-ray report generator with the image enhancement in the pre-processing process and leveraging a Multi-Head Attention (MHA) transformer encoder layer inspired by Vaswani [5]. Combined by an additional pre-trained densenet-121 for initializing visual features' weights and BERT embedding for textual features, the proposed model is trained on the Indiana University X-ray dataset. To further enhance the quality of X-ray images, a gamma correction is carried in the pre-processing step, as the best enhancement based in comparison with other methods. The model incorporates a pre-trained ChexNet [6] encoder, initially trained on the ChestX-ray14 dataset [7], that fine-tuned with the removal of the top classification layer to extract and provide initial weight to the image features. For the semantic aspect, pre-trained BERT is fine-tuned to extract text embedding from the medical reports. The extracted features are then utilize using Multi-Head Attention (MHA) to effectively exploit context information from both modalities. Subsequently, these features are forwarded in the LSTM decoder to generate diagnostic reports. Emphasizing integrated CNN and transformer embedding layers from BERT that fine-tuned with the dataset and the MHA that allow the model to focus on different parts of the input sequence simultaneously. Comparative analysis with previous studies shown a better performance from our model, particularly evident in surpassing previous benchmarks in average BLEU score metrics.

To summarize, this paper has four contributions:

1. To address the noise that occurred in the X-ray image during the acquisition process, we conducted a comprehensive exploration into the effects of contrast-based image enhancement and its subsequent influence on the report generator system. The best result obtained from the comparison will be employed in the pre-processing phase of the report generator. Different from previous works in medical image captioning that mainly focus on model modifications to improve performance, we studied and compared the effect of several enhancement methods to improve image input quality and its effect on the report generator model performance, our study underscores the significance of improving raw input quality over high-level complexity. This approach aims to elevate feature quality by refining image input quality to obtain better results on the medical report generator model.

2. Encoding process using fine-tuned pre-trained model in both visual and semantic features to fit the small dataset. The lower layer of pre-trained

ChexNet is adjusted to effectively extract image features. Simultaneously, the last layers of the pre-trained BERT are used to extract the embedding from the report text. Usually, research in image captioning often concentrates on enhancing the quality of text representation, our approach involves fine-tuning pre-trained models to extract features from both modalities—images and text. This comprehensive approach contributes to an improved quality of features during the encoding process, facilitating more effective context learning. The use of the pre-trained model also optimizes the computation cost for the model training.

3. Aligning both features from different inputs is achieved through the use of multi-head attention (MHA). Instead of directly using the output of the encoder as the input for the LSTM decoder, we leverage MHA to align and match features from both image and text. This facilitates the extraction of context information, as MHA allows the model to selectively focus on different parts of the input sequence simultaneously through multiple attention heads. This also helps to reduce the vanishing gradient during the training process.

This paper is organized in the following order: Section I introduces the background, motive and brief proposed solution, Section II explains the method and model used, Section III describes and compares the obtained results, and Section IV concludes this research.

SECTION II.
Related Works

Image captioning is an automated procedure aimed at interpreting an image through one or more sentences in natural language. Significant progress has been noted in image captioning over time, moving from early template-based models to more recent versions based on deep neural networks [8], [9]. Several approaches for image captioning have been made from deep learning encoder-decoder based models with CNN to extract the spatial and visual features and RNN to generate them in sequence [10], [11]. A spectrum of encoding models has been explored to enhance image captioning systems, encompassing diverse architectures such as Inception-v3, Visual Geometry Group Network (VGGNet), Inception-v3 augmented with LSTM as a decoder [12], Residual Network 152 layer (ResNet-152) [13], and VGG-16 [14]. Notably, employing transfer learning through pre-trained encoders, commonly derived from ImageNet, has demonstrated superior outcomes [15]. The strategic modification of weight embeddings has emerged as a pivotal factor in augmenting the model's performance, especially in the context of clinical image quality [16] some pre-trained such as ChexNet become the benchmark of feature embedding in chest x-ray data [6]. This wide range of approaches signifies the dynamic landscape within which image captioning research unfolds, continually exploring and improving methodologies to achieve optimal results. The trajectory of employing deep learning in the realm of medical report generation commenced with the adoption of Deep Neural Network (DNN) based models, primarily focusing on generative tasks associated with clinical interpretability of medical images [17], with one of the common baseline used are CNN-LSTM [18] that proven to be effective in generating clinical diagnostic. Over time, this fundamental method has been improved and refined. Notably, there has been a noteworthy evolution marked by the incorporation

of reinforcement learning mechanisms, addressing challenges related to gradient propagation within generative models dealing with discrete outputs [19]. Concurrently, the utilization of latent variables topics has been explored to further enhance the variety and specificity of generated reports [20]. Moreover, the integration of classifiers into the generative process has emerged as a strategic avenue, aiming to ensure the clinical accuracy of the generated medical reports [21]. This chronological progression underscores the improvement of research in medical report generation, continually incorporating advanced technology to refine and optimize the methodologies.

The advent of the Transformer architecture in 2017 [5] marked a transformative shift in computer vision and natural language processing (NLP) [22], setting a precedent for surpassing state-of-the-art benchmarks in domain-specific challenges. In the field of medical image captioning, this trend has taken off a bit later. One of the early research, such as the one undertaken in 2019 [23], focused on identifying the regions of interest within medical images. In particular, the prevailing approaches in medical image captioning leverage pre-trained models based on Convolutional Neural Networks (CNN), with improvements to address several problems such as data bias [24], feature alignment [25], and description coherence [26]. Specifically, in the chest X-ray, attention is also implemented to leverage semantic information [27], [28] and visual information [29]. These approaches, have obtained competitive results in generating comprehensive medical reports. Other researches, also put emphasizes in the importance of feature information in medical image data [30], with recent strides have been made through innovative techniques like meta-learning combined with attention mechanisms to enhance feature interactions for medical time-series classification [31]. Furthermore, a memory-driven transformer approach has been deployed to ensure the generation of extensive reports without omitting crucial information [32]. The implementation of attention mechanisms has shown remarkable improvements in retrieving global information from medical images, a pivotal aspect in facilitating clinical analysis [33], [34]. To further develop the system, multi-level alignment approaches have been used in conjunction with self-attention processes, this has improved the performance of information extraction by efficiently bridging the gap between picture and text data [35].

Medical images often face challenges related to noise introduced during the acquisition process, with additional quality degradation when images are converted from DICOM format to jpg or png. Enhancing medical images becomes crucial to not only mitigate noise but also improve overall image quality and spatial information. The objective of image enhancement is to achieve subjectively superior input quality compared to the original image. Several approaches have been employed to enhance medical image quality, particularly in chest X-ray data. A histogram equalization-based model has shown good result in enhancing chest X-ray data, providing better visual results for interpretability and noise reduction [36], [37], [38]. The implementation of enhancement method for specific cases was also done such as to boost deep learning model for classification tasks in Covid-19 data using histogram equalization-based method [39], [40], [41], Siracusano et al. conducted a detailed exploration, utilizing advanced contrast enhancement and CLAHE to enhance Covid-19 images [42], while Kanjanasurat et al. incorporated gamma correction alongside histogram equalization to improve Covid-19 classification model performance [43]. In the context of pneumonia classification, a comparison of histogram equalization-

based techniques was carried out by Abin et al to enhance chest X-ray image and improve model performance [44], the implementation of EFF was also carried out for the same cases by Setiawan [45]. Regarding the medical report generator, in a similar dataset a comparison of several enhancement methods also been conducted to obtain better model performance than using original image [4].

Inspired by these advancements, our study introduces a model that employs a pre-trained model and incorporates a transformer to synchronize features from both modalities, fostering a comprehensive contextual understanding. In contrast to previous researches that primarily concentrate on modifying models to enhance the quality of generated text, our approach diverges by integrating image enhancement in the preprocessing stage, influenced from previous research [4], to improve model performance.

SECTION III.
Method

A. Overview

In the Fig. 1 the proposed model schema is illustrated, the input data will be through five process which contribution in the Preprocessing process specifically the image enhancement process and the Context Building process. The detailed proposed method is illustrated in Fig. 2. Whereas, the data will be through pre-processing process in which the X-ray image will be enhanced using best contrast-based methods while text report are processed with basic preprocessing steps (cleaning character, deconstructed, lowering case). Then both data will be encoded, for the X-ray image using a pre-trained Densent121 ChexNet utilized as a feature extractor with an output of image feature vector, and as for the report text BERT is used to generate embedding before the decoding process. Both features then will be synchronized to get the context information. To amplify the model's ability to recognize subtle features in images, we use Multi-Head Attention (MHA) as an addition to building the context that aligns between image features and text features combined with LSTM as a decoder.

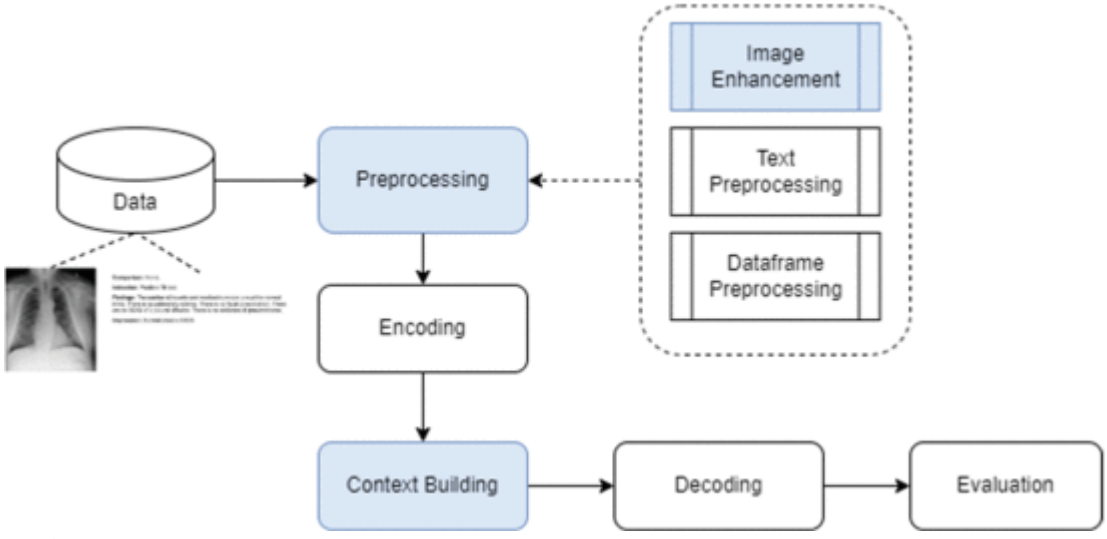


FIGURE 1.
Our Medical Image Captioning method scheme.

Show All

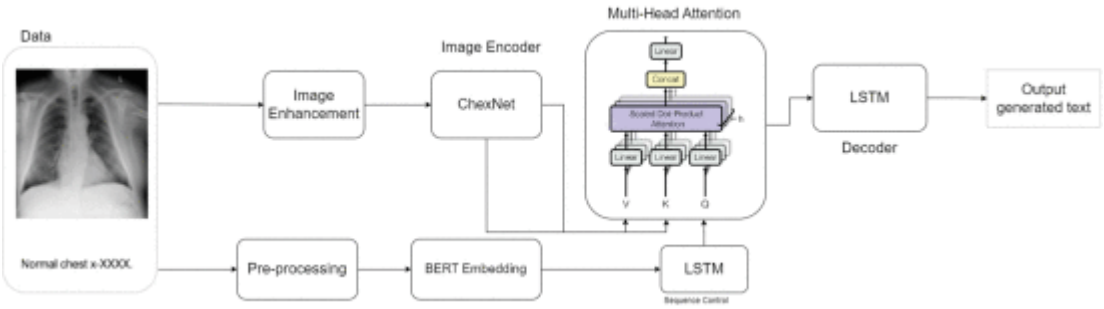


FIGURE 2.
Proposed radiology report generator architecture.

Show All

Fig. 1 depicts the research methodology's schema, and the model architecture used in this research can be seen in Fig. 2.

B. Dataset

The dataset used in this study is open data from the Indiana University Hospital Link collection. The collected data consists of 9199 chest X-ray images and 3973 XML format of medical reports written in english by radiologists [46]. The medical reports in the dataset include comparison and indication data, which are indication and comparative data based on the patient's symptoms and medical background. The diagnosis of the patient is represented by the impression, whereas the finding is a descriptive examination of the radiological image.

Several images may be linked to one medical report or no report at all. Basic pre-processing will be applied to the image and text, then sampling will be done to the data frame before being split into train, validation, and test data.

C. Image Enhancement Method

Image enhancement is an important image-processing technique which highlights key information in an image and reduces or removes certain secondary information to improve the identification quality in the process. It aims to make the objective images more suitable for a specific application than the original images. Based on previous research, implementing an enhancement to the radiographic image has a positive impact on feature extraction to clinical diagnoses [47], in another study on similar data radiograph x-ray, comparison and analysis on the enhancement implementation also showed good impact in segmentation and classification [3]. In the medical report generator, our initial research with a convolutional encoder and recurrent decoder has proven to have a good impact on the generated report model performance [4]. In this study, we continued the enhancement processing with an additional transformer as an improvement to build a context vector.

Four different enhancement techniques are employed in this research, namely HE, CLAHE, EFF, and Gamma Correction. First method is Histogram Equalization which is a technique designed to enhance image quality by adjusting the contrast and brightness of low-contrast and dark images [48]. This method achieves improvement by redistributing the image's grey levels, leading to enhanced image clarity and a more uniformly stretched histogram. The discrete function formulation of the image histogram is represented as $h(r_k)=n_k$, where h is the histogram function, r is the intensity value in image pixel k , and n is the number of pixels with intensity value r in the image. Normalization of the histogram value is accomplished using the total number of pixels in the image. The resulting even distribution of pixel values in the final image enhances the clarity of the grayscale image. The second method is Contrast Limited Adaptive Histogram Equalization (CLAHE) which is an enhanced version of HE that improves contrast in specific image sections. It differs from traditional histogram equalization by enhancing local contrast and edges based on the local distribution of pixel intensities. This method operates in the HSV color space, focusing solely on the value component, and utilizes a threshold parameter to control contrast enhancement within selected zones. The image undergoes HSV processing to create a color rendering more closely aligned with human perception, with the final result converted back to the RGB color space [4]. The third method is the Exposure Fusion Framework (EFF) which is an advanced method for enhancing image contrast through exposure ratio adjustments, offering superior contrast improvement compared to other techniques. The algorithm, intelligently fuses pixel regions with varying exposure levels, considering pixel values, weight maps, and color channels to produce a well-exposed result. Brightness Transform Function (BTF) is used to manage exposure differences, ensuring a harmonious blend of exposures. The final enhancement process combines weighted pixel values, BTF application, and exposure ratios to achieve a visually appealing and computationally efficient contrast enhancement [4]. The last method is Gamma correction, a non-linear operation crucial for adjusting exposure or tristimulus in images and videos [49], which involves altering pixel values based on the gamma constant (γ). The gamma function expressed as $g(x)=x_\gamma$, transforms the pixel value x , yielding a new value in the image. Applying gamma correction within a pixel range of 0–255 to adjust the image gamma value. Gamma values > 1 lighten the image, while $\gamma = 1$ shifts the image towards the darker spectrum.

D. Image Encoder

A 121-layer dense convolutional network (DenseNet) called ChexNet was trained using data from 14 chest X-ray datasets. Existing networks can be optimized with DenseNet, which improves gradients and information throughout the network [48]. On the pre-trained ChexNet, the DenseNet model has been trained on the data that consists of more than 100,000 X-ray images [6].

In this study, image features are extracted from the data as convolutional features using ChexNet as an encoder. The pre-trained ChexNet will transmit the weight to the image through the unfroze last layer of the model. To customize the output into 1024 as a dimensional image initial features value, we remove the top layer on ChexNet. In this study, the parameters used were a batch size of 100, a dropout of 0.2, and learning rate of 10^{-2} and sigmoid activation. Then, the pooling of initial

weight initiation will be forward with ReLU activation and output vector size 512. This dimension will be adjusted with BERT embedding to get the context information along with the text features in the multi-head attention.

E. Multi-Head Attention

The initial weight from the encoder will forwarded as an input to multi-head attention as a key and value to align the information with the semantic features as a query in a parallel process. Attention technically is mapping a query to the keys from data with a value as an output. It consists of three components: query, key, and value. Multi-head attention is multiple single attention that works repeatedly and simultaneously.

The multi-head attention function can be represented as (1).

$$multihead(k,q,v)=concat(head_{1..n})W(1)$$

View Source where k,q , and v are the three components, query from the report embedding, key, and value from the image encoder. Each head is a single attention process that can be defined as (2), (3), (4).

$$attention(k,q,v)\alpha_{q,k_i}e_{q,k_i}=\sum_i \alpha_{q,k_i}v_{k_i}=softmax(e_{q,k_i})=q\cdot k_i(2)(3)(4)$$

View Source where α in function (2) is the output of the softmax activation from function (3). Meanwhile, e is the attention weight that has been aligned from both features as a new value (4).
Attention is defined as mapping a query with key-value pairs to get a new value. Where query, keys, and values are all vectors. The weights assigned to each value are determined by the compatibility function between the query and keys. The final output is computed as the sum of the weighted values. This mechanism is notably efficient for aligning features from both images and text to establish context. The resulting attention weight serves as the context vector, which is subsequently fed into the decoder as input.

F. BERT Embedding

The semantic features are extracted from pre-trained BERT embedding. BERT is a neural network-based training technique released by Google known as Bidirectional Encoder Representations from Transformers (BERT) [49]. This method is an open source after Google released it publicly in 2018. BERT can train a language model based on an entire set of words in a sentence or query in a two-way manner. BERT allows language models to study the context of words based on the words around them, not just the words that precede or follow them.

BERT uses an attention mechanism in the data encoding process to study the contextual relation between words in the medical report. The encoder will read the whole words, so it can learn the context of the word. An example of input representation in BERT can be seen in Fig. 3.

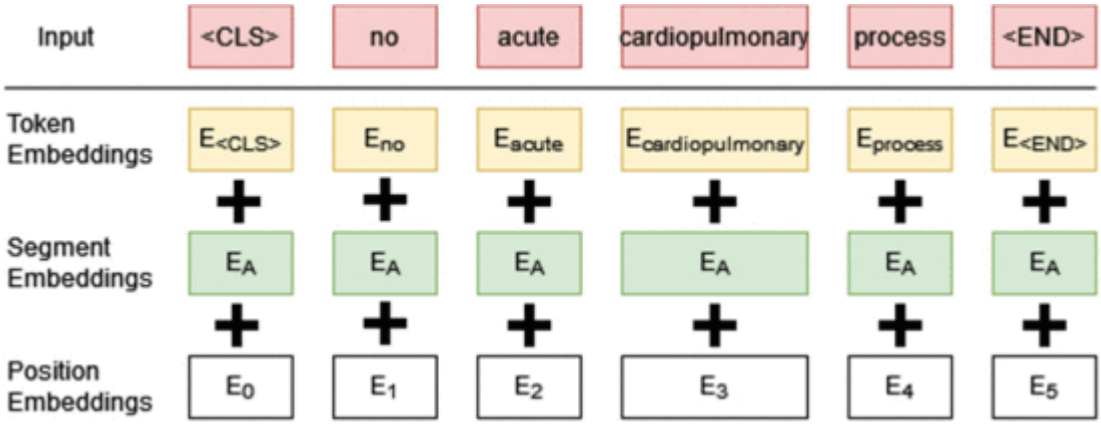


FIGURE 3.
BERT input representation.

Show All

For the token in sequences, [CLS] is used as the first token in every input, and the [SEP] token is used to separate between sentences in the input. From the input, the token will be embedded based on the vocabulary id. Then, to distinguish between sentences, it will be embedded by the segment, and the position embedding to indicate the precise position of the token. The input embedding is the sum of the three embedding processes. This embedding will be used as an input for context building in multi-head attention with fixed dimensions.

In this research, we employed the “bert-based-uncased” model from Huggingface for BERT. Subsequently, we utilized BERT’s tokenization to tokenize medical report data, and the BERT embeddings were loaded as initial weights for all words in the medical report data. The dimensions of BERT were fine-tuned specifically for word embedding.

G. Decoder

Long Short-Term Memory (LSTM) is utilized as a decoder. Three gates make up the LSTM: an input gate, a forget gate, and an output gate. The gates perform several tasks related to data collection, classification, and processing. An input modulator is used by an LSTM cell to modify the input to the memory as it continuously learns the weights from the input gate. This memory cell is also used to learn the weights for the output gate. In each process step, the LSTM stores and deletes some data, which is subsequently utilized in the process step.

In this study, the visual data from multi-head attention is utilized as input and preserved in the input modulator. Additionally, the memory cell stores text embeddings through an embedding matrix, employing BERT embedding in text prediction to improve the selection of the next word. The hidden state’s output is then employed in multi-head attention as a query to align with the visual features and get the context information from both modalities.

H. Evaluation

Bilingual Evaluation Understudy (BLEU) will be used to assess how well the model described the image. Using actual data as a reference, the BLEU method approach counts word occurrences in the model-generated text [50].

The method of determining the match between the words produced by the model and the actual data is known as the n-gram calculation. To compare the predicted text quality value to the reference data, the value of the n-gram words in the sentence will be calculated. The equation to calculate the BLEU score can be seen in (5).

$$BLEU=BP.\exp\sum_{n=1}^{N} w_n \log p_n \quad (5)$$

View Source where:

p_n: the predictive text precision per n word,

w_n: the predictive text weight on the n word,

N: gram value used,

BP: penalty value of error prediction

Whereas BP is a brevity penalty for predicted text errors and stands as a crucial factor, meticulously designed to evaluate and penalize errors in predicted text length. The precision value p of the predicted text, the weight assigned to words w , and the chosen gram value N collectively contribute to the nuanced calculation of BP. In the context of evaluating candidate sentences against reference sentences, a standard practice involves employing a gram value of 4 for candidate sentences. This means that each of the prediction's four words is examined to see if it is similar to the reference sentence. The gram value, denoted by the letter N , plays a pivotal role in determining the extent of the comparison, thus adding a layer of complexity to the brevity penalty computation. In essence, the n-gram value becomes a pivotal parameter in this evaluation, guiding the matching process and shaping the precision assessment of predicted text with reference text. This detailed understanding of the interplay between gram values, brevity penalty, and precision offers a comprehensive view of the intricacies involved in assessing the quality and conciseness of predicted text in natural language processing tasks.

SECTION IV.

Result and Discussion

A. Dataset

From the Indiana University dataset exploration in Open-I, there was a disparity between the normal and abnormal diagnoses of the X-ray images. Because this can lead to model bias to normal diagnoses, several processes to reduce imbalance are done. In the normal diagnoses, the duplication data for a similar diagnosis is reduced, and then the minor diagnoses in normal and abnormal sampled to reduce

the imbalances. The diagnoses as medical reports are subjected to pre-processing before being embedded.

The pre-processing of the medical report includes word deconstruction, character and number deletion, and letter conversion to lowercase form. Then, it is trained with BERT to get the embedding of the text and feature extraction from text data. The medical report is also filtered based on the occurrence in the data. The data split based on occurrence percentage into training, validation, and test data.

The data are partitioned based on the number of reported occurrences, with the highest frequency group being under-sampled and the lowest frequency group being oversampled. The data is then separated into train, validation, and test data based on the percentage of each minor and major data. Table 1 shows the quantity of data used for the training, validation, and test.

TABLE 1 Data Split for Modelling

Data	Data Quantity
Train	3605
Validation	901
Test	566

B. Image Enhancement

Radiographic images tend to have noise during the acquisition process. The common uneven contrast values can cause damage to the extracted information from the image. This could affect model learning in generating captions. In addressing this problem, approaches to enhancing image quality and analyzing the effect of improving the quality of learning models in generating captions are carried out. Several enhancement methods were compared based on the effect of the enhancement on the model improvement with the baseline model as an individual enhancement. The best result obtained will be used as a preprocessing method before visual feature extraction.

In investigating this, X-ray images were enhanced using several contrast methods, namely HE, CLAHE, EFF, and Gamma Correction. Each enhanced image will be used with baseline model CNN as a feature extractor and LSTM as a decoder. The

effect in the generating learning model is evaluated by the predicted diagnoses similarity with the ground truth using BLEU.

HE method is used to distribute pixel value to enhance image contrast. Calculating the histogram value in the range 0-255, then creating the cumulative distribution and re-assigning the pixel value to become a linear function. For CLAHE, a default function from OpenCV with clip limit value 2 to avoid amplification of extreme pixel value is used. The Ying method, Exposure Fusion Framework (EFF), works by combining multiple images of the same scene taken with different exposures to create a single image that retains the best details and luminance from each input image. These frameworks analyze the pixel values in each image and choose the most appropriate values for the final output. They prioritize well-exposed regions and preserve details in both dark and bright areas, resulting in an image with a wider range of tones and details than a single exposure could capture. In this study, we set the threshold for minimum exposure illumination with the same value as the lambda which is 0.5, alpha -0.3, which is also the value used to calculate the max entropy of the image features, and beta value 1 to get the weight matrix of the image. The beta and gamma values are also used to calculate the exposure ratio between under-exposure images and over-exposure images based on the minimum illumination threshold. For the gamma method, adaptive gamma with a threshold value of 0.3 was utilized to determine whether the image was too bright or too dimmed. The gamma value of 0.7 was used to brighten the dimmed image and 1.5 was used for the opposite.

The difference between the original image and the enhanced image for each method can be seen in Fig. 4. The baseline model used in the comparison is ChexNet as a feature extractor and LSTM as a report generator, with the best result used as an enhancement method in the proposed model. The result of the comparison can be seen in Table 2.

TABLE 2 Result Comparison of Different Enhancement Methods With the Baseline Model

Model	BLEU Evaluation			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4
ChexNet-LSTM	0.25567	0.25576	0.29077	0.34157
ChexNet-LSTM + HE	0.22783	0.28682	0.29190	0.33237
ChexNet-LSTM + CLAHE	0.25893	0.29327	0.32329	0.33782
ChexNet-LSTM + EFF	0.25892	0.30681	0.32879	0.34482
ChexNet-LSTM + Gamma Correction	**0.30147**	**0.30698**	**0.32969**	**0.37792**

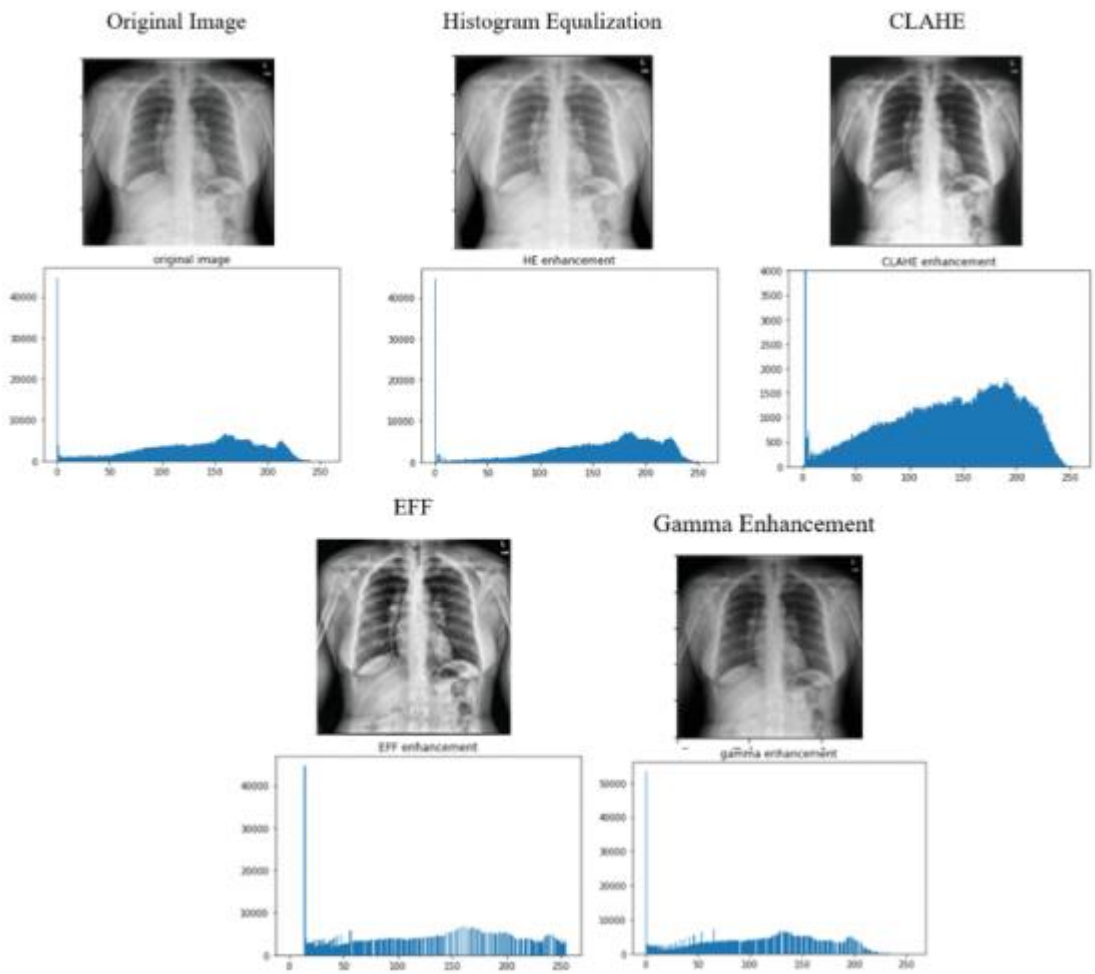


FIGURE 4.
Histogram comparison for original X-ray image and the image with different enhancement.

Show All

From the results of the comparison using the baseline Table 2 shows that the gamma correction gives the best result from the BLEU evaluation number. In the radiographic image of the chest x-ray, the data was enhanced using the gamma correction method according to (2). The gamma value between 0.7 and 1.5 used the enhancement process to balance the contrast in the image.

C. Implementation Details

From the comparison result, gamma correction for the enhancement technique employs a threshold value of 0.3, strategically determining whether an image leans towards excessive brightness or undue dimness. The gamma correction process is guided by a nuanced choice of gamma values. In instances where an image exhibits excessive dimness, a gamma value of 0.7 is applied to brighten and restore visual clarity. Conversely, for images veering towards excessive brightness, a gamma value of 1.5 is adeptly employed to achieve a harmonious balance and enhance overall quality. This methodical use of gamma correction, coupled with a discerning

thresholding strategy, contributes to a refined and adaptive enhancement process, ensuring that each image receives a tailored treatment based on its unique luminosity characteristics.

For the visual extraction process, the images that have been enhanced with Gamma Correction were resized to 224×224 following the size of the pre-trained ChexNet model for initial weight initiation. The parameters were batch size of 100, dropout of 0.2, learning rate of 10^{-2} and sigmoid activation. Then, the pooling of initial weight initiation with ReLU activation was done, and the output vector size was 512.

The vector generated by the weight initiation using ChexNet becomes the input to the MHA as a key and value in parallel with the head attention value of 8, with an output tensor array size of 512×512 and dropout of 0.2. In the transformer layer, the input image is positioned encoding with the output in the form of a vector, which will be used as input in the form of a key and value for the query from the decoder on the attention module. Attention weight is calculated using softmax activation on the module.

Extracted diagnoses from radiology reports will be embedded using BERT. There are 3 combinations of embedding used as a final representation starting from token embedding for each specific word and with [CLS] token as a tag to indicate the beginning of the diagnosis, and [END] to indicate the end of the diagnosis, then segment embedding to provide word position information in a sentence, and position embedding for detailed information of word position in the text.

The model performance between the baseline and transformer models was compared using BLEU as a natural language generation evaluation. These comparison results can be seen in Table 3.

TABLE 3 Result Comparison With Different Model

Model	BLEU Evaluation			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4
CNN-LSTM	0.25567	0.25576	0.29077	0.34157
Multi-head Attention	0.30681	0.30214	0.33182	0.39264
CNN-LSTM + gamma enhancement	0.30147	0.30698	0.32969	0.37792
Multi-head Attention + gamma enhancement	**0.36343**	**0.37199**	**0.38846**	**0.41259**

Based on the BLEU evaluation with n-gram values 1-4, the Transformer model is able to outperform the baseline model with an increase of 15% from using the transformer

MHA mechanism of the BLEU-4 score compared to the ChexNet-LSTM model. Furthermore, adding gamma enhancement boosts the BLEU-4 score by 11% for ChexNet-LSTM and 5% for MHA models. The better result was obtained by the MHA model with gamma correction with an increase of 9% of the BLEU-4 score compared to ChexNet-LSTM with gamma correction.

Based on the test data, the predicted text with a high BLEU score effectively predicted that it would be nearly identical to the average description on the ground truth. Meanwhile, descriptions of images can be found in data with low BLEU scores that do not match the reference word from ground truth. Fig. 5 displays a few instances of predicted outcomes based on the BLEU score. From the prediction results, in the prediction text with a high BLEU value, the model successfully predicts according to the ground truth per word with the same definition, especially data with a ground truth of more than one sentence.

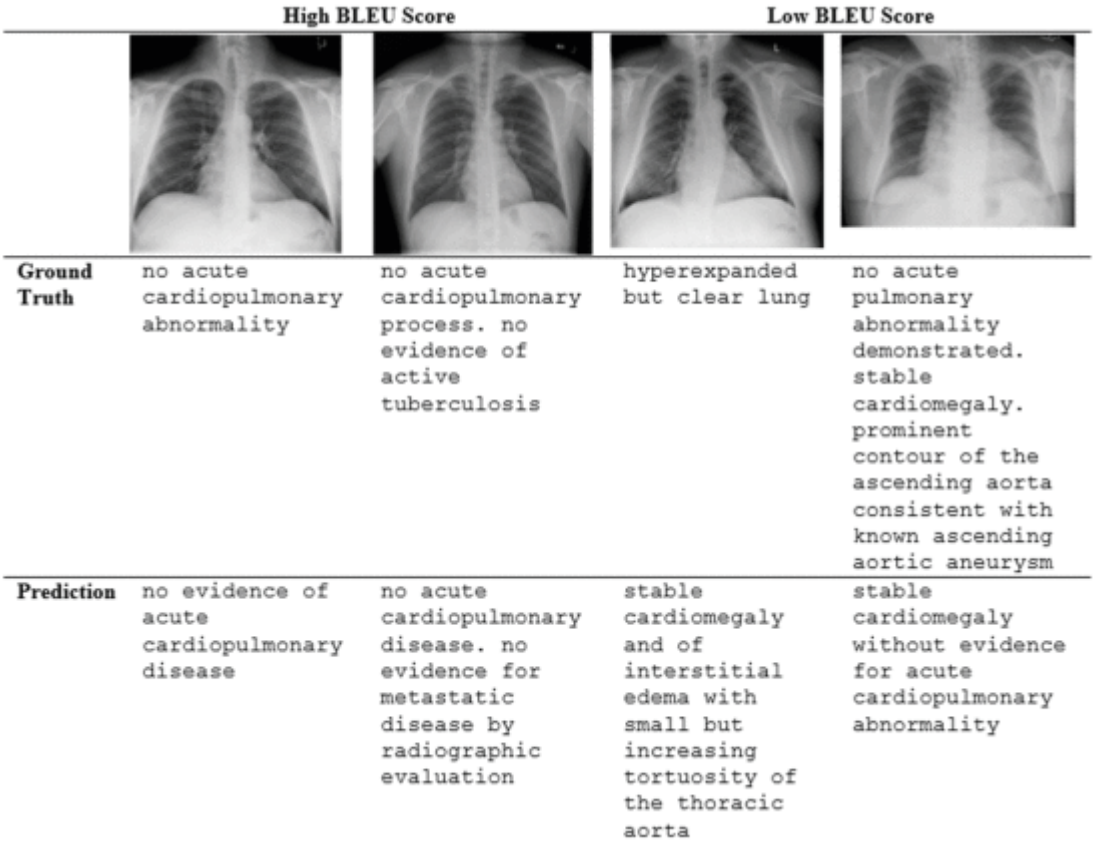


FIGURE 5.
Histogram comparison for original X-ray image and the image with different enhancement.

Show All

The assumption on the influence of the low BLEU score is due to the fact that ground truth descriptions on the dataset are written. Even though they have the same meaning, the words used differ, as in the descriptions of ‘no acute’, ‘no evidence, and ’no abnormality’ have the same meaning which is a healthy normal chest condition

based on x-ray results. Because the BLEU method calculates based on word similarity, the model's BLEU score is low as a result of the different word selections.

In the encoder-decoder model with MHA and gamma enhancement, the model performance is also seen based on the loss and accuracy values in the training process. Fig. 6 shows that the loss and accuracy values in the validation data are not too far from the training data during the 10-epoch training iteration. This indicates that during the training period, the model has good performance, and there is no overfitting of the data.

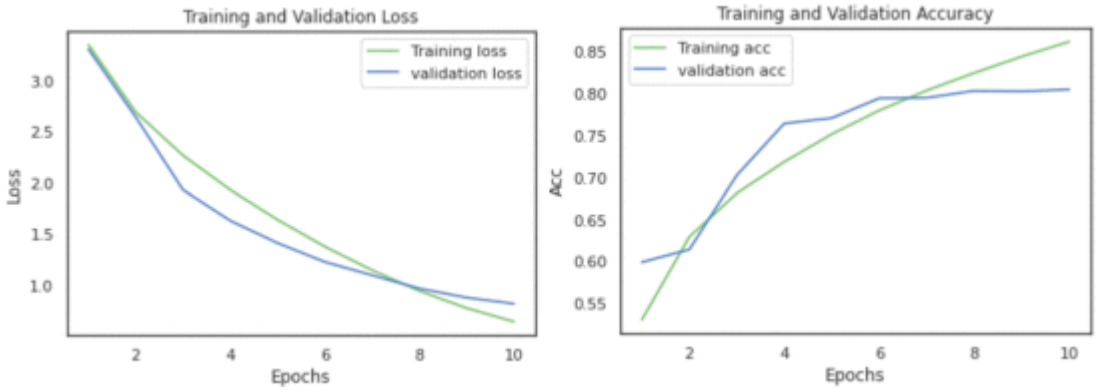


FIGURE 6.
Proposed model performance in loss and accuracy.

Show All

A comprehensive comparison of our approaches using enhanced image data was conducted in terms of computational cost. The first approach employs a Convolutional Neural Network (CNN) to extract image features, followed by a Long Short-Term Memory (LSTM) network to generate captions. With the same batch size, the CNN-LSTM model took training time for 10 epochs 4159s, and the inference time for testing data is 436ms. The computational cost of this approach primarily depends on the model's size, batch size, sequence length, and the depth of the CNN-LSTM architecture. The second approach utilizes an encoder-decoder architecture with an attention mechanism applied to the image features. While this attention mechanism enhances caption quality, it introduces additional computational complexity. This model approach took training time in the same epoch 4284s and inference time for testing data 2340ms. The computational cost can be influenced by factors such as the choice of attention mechanism and the granularity of attention as attention architecture has a more complex computational and longer sequence in inference time.

It was observed that the CNN-LSTM encoder approach tends to have a lower computational cost, making it more suitable for scenarios with limited computational resources. However, the encoder with attention, despite its higher computational demands, exhibits superior captioning quality and is preferred when high-quality captions are paramount, and ample computational resources are available.

The model undergoes a detailed comparative analysis with diverse architectures from previous studies on the same dataset. This assessment employs the BLEU evaluation metric across various n-gram dimensions (1, 2, 3, and 4), as outlined in Table 4. The results offer insights into the model's proficiency in generating coherent medical image reports, serving as a benchmark against prior approaches.

TABLE 4 Comparison With Previous Research Using Bleu

Model	BLEU Evaluation			
	BLEU-1	BLEU-2	BLEU-3	BLEU-4
LRCN [18]	0.369	0.229	0.149	0.099
Att-RK [27]	0.369	0.226	0.151	0.108
CDGPT2 [28]	0.387	0.245	0.166	0.111
VisualGPT [29]	**0.388**	0.333	0.231	0.226
Proposed model without gamma	0.307	0.302	0.332	0.393
Proposed model	0.363	**0.371**	**0.388**	**0.412**

As per the comparison results, the proposed model using pre-trained ChexNet MHA and gamma correction outperformed other methods in the BLEU evaluation. The increase in image quality that is also conducted could be the reason for the result.

There are different mechanisms used in the previous model, for the LRCN [18], the authors highlighted using the recurrent model as a way to decode the image interpretation for long-term dependencies learning. Att-RK [27] combined top-down and bottom-up computation using the attention mechanism in focus to leverage the semantic information by selectively attending to the features and using it in the recurrent hidden state as the decoder. CDGPT2 [28] also tries to leverage the semantic information using a self-attention mechanism and combine the weighted visual embedding with the report embedding. VisualGPT [29] uses a masked attention mechanism with the pre-trained language model weight combined before calculating the attention weight to balance the visual information processed in the model. From the previous works compared in this article, only CDGPT2 was originally implemented on similar data from the author. The model proposed in this paper was also inspired by the LRCN to use the recurrent model as a decoder for long-term dependencies and use multi-head attention instead of self-attention or masked attention to align the visual information and semantic information in parallel.

In comparison with LRCN [18], Att-RK [27], CDGPT2 [28], and VisualGPT [29] our model both with and without the addition of image enhancement got averagely better results in predicting x-ray image interpretation. We investigate the test result in BLEU-1 evaluation and BLEU-4, our model is able to outperform others from BLEU evaluation n-gram 2, 3, 4, while still behind VisualGPT [29] on BLEU n-gram 1 with a difference of 0.02. We further investigate it by comparing the predicted diagnosis of VisualGPT in BLEU evaluation n-grams 1 and 4. In the context of BLEU-1 evaluation, VisualGPT demonstrates superior proficiency in predicting individual

words from the Ground Truth compared to our model, particularly in succinct diagnoses. However, it is susceptible to inaccuracies in predicting words for longer diagnoses, sometimes leading to predictions longer than the ground truth. This misalignment can result in penalties during BLEU evaluation. On the other side, our model can predict according to ground truth better in longer diagnosis with BLEU-4. A sample of comparison can be seen in Fig. 7.

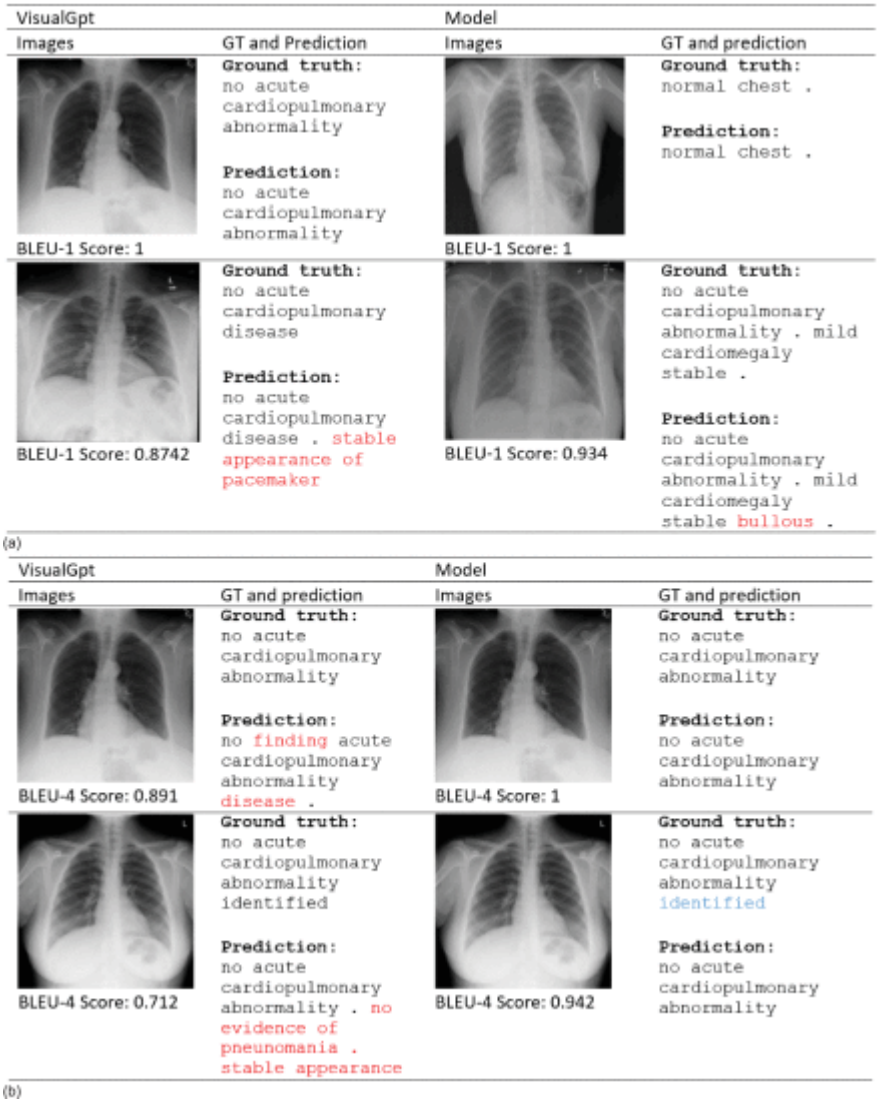


FIGURE 7.
Comparison VisualGPT and our proposed in BLEU-1 (a) and BLEU-4 (b) evaluation. Red color texts are the word that misspredicted by the model, and the blue texts are the words that unable to be predicted by the model.

Show All

D. Limitation and Future Works

In this study, while implementing our proposed model for a medical image report generator, several challenges emerged. The dataset used has a diverse report based on similar conditions, which caused difficulty in obtaining convergence in the model

training process. These challenges underscore the complexity of predicting medical reports in order to get high similarity in BLEU score.

There are also several limitations that should be acknowledged. First, building a good report generator model is really dependent on the diversity of data. Even though we tried to curate and filter data to be as effective as possible, the diversity in the dataset itself for a few abnormal conditions remains limited. Secondly, as our report generation model is based on visual features, it may not capture subtle details present in the images that an experienced radiologist can interpret, although enhancing the image has proven to help increase the similarity of predicted reports with ground truth.

Furthermore, there are several promising avenues for future work, such as in the dataset enrichment for rare abnormalities to increase diversification. In clinical relevance, such as image enhancement and abnormality labelling enhancement also have the potential to improve model report generators. There is also a potential to develop an evaluation method that is comparable to radiologist clinical analysis.

SECTION V.
Conclusion

In this research, we introduce an automatic radiology diagnosis generator leveraging the transformative capabilities of a transformer architecture. Our primary focus is to unravel the intricate dynamics between image enhancement processes, model architecture, and evaluation metrics within the realm of medical image captioning. The proposed model incorporates a Multi-Head Attention (MHA) mechanism and employs BERT embedding for extracting intricate text features.

To enhance the model's overall efficacy, we conducted an exploration of various enhancement processes and their impact on radiograph images. Among the four implemented enhancement methods, all showcased performance improvements compared to the original images, with the Gamma Correction method emerging as the most effective. This nuanced understanding of image enhancement techniques and their implications on improving the model performance.

Radiographic images often suffer from contrast noise, a challenge that significantly influences the evaluation outcomes of generated captions. Our findings highlight the pivotal role of enhancing image contrast, enabling the model to extract richer and more contextually relevant features. This nuanced insight aligns with contemporary discussions on the importance of preprocessing steps in optimizing model performance for medical image analysis.

Following the encouraging results from the contrast method tests, we integrated Gamma Correction as a pre-processing stage for our proposed transformer model. This architecture, augmented with the MHA mechanism and BERT embedding, represents a cutting-edge approach to medical image captioning. Comparative experiments were conducted, both with and without gamma correction, revealing a superior performance when gamma correction was applied as a pre-processing step. The best result was with the addition of pre-processing gamma correction, which

produced text prediction with 9% better coherence in the BLEU evaluation result than the conventional CNN-LSTM method. Similarly, the BLEU evaluation of the multi-head attention approach was also 15% better than the conventional CNN-LSTM method in comparing models without gamma correction. This underscores the significance of tailoring pre-processing techniques to the specific characteristics of medical images, a practice that is gaining prominence in recent research endeavours.

The proposed model not only outperforms previous works on the same dataset but also demonstrates superior BLEU evaluation results for n-grams 2, 3, and 4. Furthermore, it excels in predicting more accurate diagnoses, providing longer ground truth information. This comprehensive investigation advances our understanding of the intricate interplay between image enhancement, model architecture, and evaluation metrics in the context of medical image captioning. Our proposed model is able to outperform other previous works with the same dataset and BLEU evaluation n-gram 2, 3, 4. It also predicts better diagnosis with longer ground truth.

Author Contribution

All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Conflict of Interest

The authors declare no conflict of interest.

Additional Information

No additional information is available for this paper.

Show More

References

References is not available for this document.

IEEE Personal Account

- Change username/password

Purchase Details

- Payment Options
- View Purchased Documents

Profile Information

- Communications Preferences
- Profession and Education
- Technical interests

Need Help?

- US & Canada: +1 800 678 4333
- Worldwide: +1 732 981 0060
- Contact & Support

Follow

IEEE Account

- Change Username/Password
- Update Address

Purchase Details

- Payment Options
- Order History
- View Purchased Documents

Profile Information

- Communications Preferences
- Profession and Education
- Technical Interests

Need Help?

- **US & Canada:** +1 800 678 4333
- **Worldwide:** +1 732 981 0060
- Contact & Support

- About IEEE *Xplore*
- Contact Us
- Help
- Accessibility
- Terms of Use
- Nondiscrimination Policy
- Sitemap
- Privacy & Opting Out of Cookies