A detailed analysis of deep learning-based techniques for automated radiology report generation

Prajakta Dhamanskar[1], Chintan Thacker[2]
[1]Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering, Mumbai, India
[2]Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujrat, India

Article Info

Article history:

Received Apr 22, 2024
Revised Jul 6, 2024
Accepted Jul 11, 2024

Keywords:

Automatic medical report generation
Chest X-ray images
Deep learning
Healthcare
Radiology reports

ABSTRACT

The automated creation of medical reports from images of chest X-rays has the potential to significantly reduce workloads for healthcare providers and accelerate patient care, especially in environments with limited resources. This study provides an extensive overview of deep learning-based techniques designed for radiology report generation from chest X-ray pictures automatically. By examining recent research, we delve into various deep learning architectures and techniques used for this task, including transformer-based approaches, attention mechanisms, sequence-to-sequence models, adversarial training methods, and hybrid models. We also discuss about the datasets used for evaluation and training, as well as future directions and research problems in this area. The significance of deep learning in revolutionizing radiology reporting is further emphasized by our review, which also highlights the need for additional research to address challenges such data accessibility, image quality variability, interpretation of complex findings, and contextual integration. The objective of this research is to present a comparative analysis of cutting-edge methods for developing automated medical report generation to enhance patient outcomes and healthcare delivery.

Corresponding Author:

Prajakta Dhamanskar
Department of Computer Engineering, Fr. Conceicao Rodrigues College of Engineering
Bandstand, Bandra west, Mumbai, India
Email: prajakta.dhamanskar@fragnel.edu.in

1. INTRODUCTION

For many years, doctors have used chest X-rays to identify and treat patients with disorders related to the chest. Chest X-ray is the most commonly used and least expensive diagnostic method for diagnosing a variety of conditions, such as lung cancer, heart failure, pneumonia, tuberculosis (TB), and lung interstitial disorders. Getting access to quality healthcare facilities is extremely difficult in rural areas of developing nations like India. Furthermore, radiologists may need to study thousands of X-ray images of various patients in order to prepare medical reports for the same due to the large population. Writing reports from images taken by X-rays is a laborious and boring task. This may lead to delays in medical treatment provided to the patient.

Getting access to quality healthcare facilities is extremely difficult in rural areas of developing nations like India. Furthermore, radiologists may need to study thousands of X-ray images of various patients in order to prepare medical reports for the same due to the large population. Writing reports from images taken by X-rays is a laborious and boring task. This may lead to delays in medical treatment provided to the patient. Due to the increased demand for X-rays, shortage in availability of medical professionals such as radiologists, lack of experience, lack of knowledge and faulty reasoning, this process of writing medical reports manually from chest X-rays has become error prone. To overcome this difficulty, different approaches that uses deep learning to automatically create medical reports from chest X-ray pictures is proposed by many researchers, one of which is shown in Figure 1.

A detailed review of different approaches used for automatically creating medical reports from X-ray images of the chest is presented here. This article also presents the research challenges or gaps found in the literature and further presents future directions in the research. The process of creating medical reports from input images of chest X-rays is comparable to the work of captioning images. Nevertheless, the image captioning method cannot be used to produce medical reports. i) Because abstract and complicated medical terminology, such as “lung atelectasis” and “hypertension”, are included in medical reports; ii) Natural image captioning has mostly one sentence, whereas findings in medical reports consist of four, five or even more than five sentences; and iii) Natural image captioning tasks use natural language evaluation metrics such as the BLEU score, but simply using such high text relevant scores would not be efficient. The diagnostic accuracy of the medical report should also be taken into consideration.

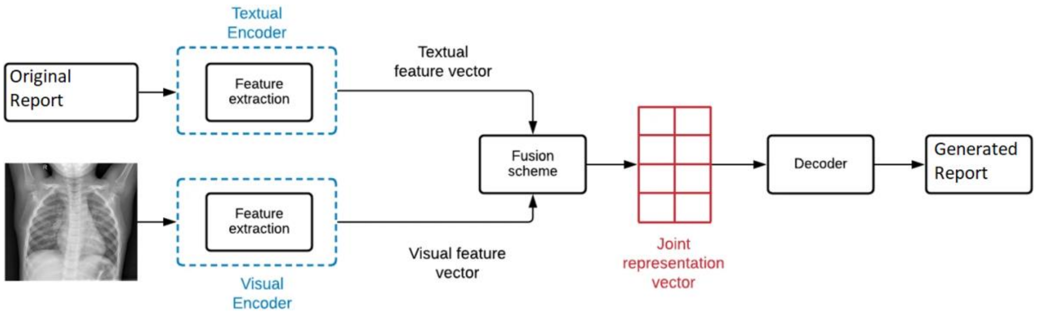


Figure 1. Automatic medical report generation process [1]

2. SEARCH CRITERIA AND STUDY SELECTION

The main objective of the study is to present a detailed and systematic review of deep learning approaches in autonomously creating medical reports from chest X-rays and provide answers to the following research questions (RQ). RQ1. What are the public datasets available for this study? RQ2. What are the different DL approaches in automatic medical report generation? RQ3. What are the challenges for incorporating automatic medical report creation systems into existing healthcare systems? RQ1 is answered in section 4, RQ2 in section 3 and RQ3 in section 6.

The search keywords used for searching through reputed databases like IEEE, Science Direct and PubMed are listed as follows. (“Automatic medical report generation” OR “Medical image captioning” OR “Deep learning for medical report generation”) AND (“Chest X-ray” OR “Radiology report generation”) AND (“Deep learning” OR “Convolutional neural networks”). A total of twenty-papers during 2019 to 2023 were selected for study, which include different approaches for report generation as shown in Figure 2 and all selected papers were from latest years as shown in Figure 3.

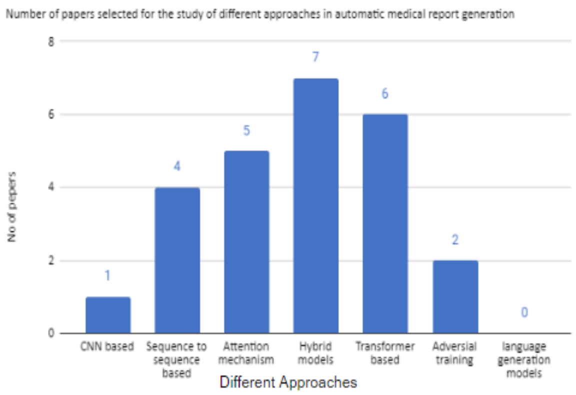


Figure 2. Different approached in automatic medical report generation

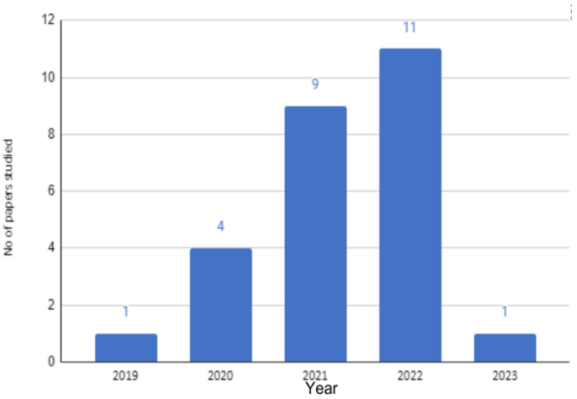


Figure 3. Year wise selection of research papers during 2019 To 2023

3. LITERATURE REVIEW

The objective of the study described in [2] is to design a model capable of learning versatile joint representations of vision and language in the medical field. The model is developed to carry out both visual language understanding and visual language generation tasks using a unified visual language pretraining model. To extract image features from medical images, a convolution neural network (CNN) model based on residual networks (ResNet) 50 is employed. For textual feature embedding, a bidirectional encoder representation from transformers (BERT)-based architecture comprising 12 transformer layers is utilized to encode textual information. The study employs two datasets, namely, MIMIC-CXR and Open-I.

Wang *et al.* [3] employ more than one criterion to train transformer models for report generation. They propose a method in which they assign weights to medical terms or words to adjust the importance of words during training. Instead of generating region proposals, the model employs a method that uses a transformer that extracts image features and considers the relationship among image regions to understand the image representation. The model utilizes an image-text matching loss to ensure that it learns correlation between image and text features for report generation very strongly. The study utilizes the MIMIC-CXR and Open-I datasets.

Amjoud and Amrouch [4] describes a study that creates an automated architecture for creating reports from X-rays of chest by combining “transfer learning” and “transformer” approaches. The authors use a “pretrained DenseNet-121” model trained on the “ImageNet” database for feature extraction, and the classification layer is removed. The IU Chest X-ray dataset is used in this study.

Hidden visual features are decoupled into meaningful disease embedding and disease states for clear and accurate disease representations in [5]. The transformer model receives these disease representations and uses them to produce high-caliber medical reports. The visual features are extracted using the CNN image encoder. By dividing the visual elements into distinct states such as “positive”, “negative”, and “uncertain”, the embedding block in disease is used to include informative properties while supporting with self-attention. Enriched disease embedding is created by connecting state-aware disease embedding. To produce medical reports, a transformer decoder is used. Chest Xray 14, Open-I and COVID-19 datasets are used in this study.

Two unresolved issues are tackled in [6]: i) Visual information becomes redundant or irrelevant during encoding due to low fraction of disease regions in an image. and ii) Because of multimodal representation, correlations that were modeled during the encoding step might not be successfully decoded during the decoding stage. CNN paired with semantic aware visual learning (SVL) as the encoder, and the transformer decoder's encoder layer with memory augmented semantic enhancement (MASE) incorporated as the decoder.

DenseNet-121 is used in [7] to extract spatial features. Terminology encoder is then run to obtain terminology-related features. For knowledge pretraining, a straightforward lookup table is used to generate textbook embedding.

The study presented in [8] introduces a method that that is based on template retrieval and relational-topic-driven for generating reports. They referred to this method as “Relation-paraNet”. The encoder used in this method represents relationships between medical terms by considering semantic consistency among them. The adaptive generator uses one of the approaches to generate the output: either retrieve an existing template from the database of templates or create a new sentence. To extract visual features, the study uses a deep CNN architecture and extracts text features from a long short-term memory (LSTM) layer with an attention mechanism. The study employs the IU X-ray and CX-CHR datasets.

The study presented in reference [9] utilizes a CNN as an encoder and an LSTM with an attention mechanism as a decoder. The system is trained and evaluated on two different datasets: the “IU CXR” dataset and the “MIMIC CXR” dataset. The use of attention mechanisms in the decoder can also help the system detect the most important parts of the input image and generate more contextually relevant reports.

To address the problem of data bias and provide reliable medical reports led by disease tags, the research effort in [10] proposes the AIMNet, which includes an importance-based merging method. The position, intensity, and shape of the lesion are only a few examples of the visual characteristics of the abnormalities that are included in the visual information in the images. Illness tags document specific details about the abnormalities and offer a comprehensive view of the abnormalities. Transformer serves as AIMNet's core module. In order to produce reliable and accurate reports, it introduces visual attention, Tag attention, and adaptive merging gate. This is because creating medical reports necessitates writing lengthy paragraphs.

CNNs (ResNet 50) is utilized in [11] to extract visual features from multi view medical images. Visual embedding with a visual attention mechanism is employed for subject encoding. The process of predicting corresponding term frequency-inverse document frequency (TF-IDF) features of the corpus yields visual to semantic embedding. Encoder-decoder used for generating corresponding sentences or captions.

The pre trained show attend and tell model is utilized in [12] to extract the attention matrix. To acquire fine-grained features, the dot product of the attention matrix and the features retrieved by ResNet101 is done. Noise in features is reduced by object drop out strategy using attention matrix. To increase the model's efficiency, multilayer perceptron (MLP) is utilized to replace the transformer's encode module.

Biswal *et al.* [13] used a multilevel multi attention (MLMA) approach to generate highly contextual and coherent medical reports for X-ray images of chest. They used the "encoder decoder" framework and combined different deep learning models, such as CNN, LSTM, and bidirectional LSTM, to learn diverse and semantic patterns of the report. The proposed "MLMA" approach outperformed the current methods for report generation from medical images.

The study mentioned in reference [14] focuses on creating medical reports by making use of doctors' partially completed phrases and disease-related words. To achieve this, dense CNN and LSTM models are used. The dataset used in the study is the IU X-ray. Another dataset used in the study is the Temple University Hospital electroencephalogram (EEG) data, which includes EEG recordings of variable length and their corresponding EEG reports. The Massachusetts General Hospital (MGH) EEG data are were also utilized as an evaluation dataset in the study. It is worth noting that the study only considered chest X-ray and EEG recordings for report generation, and other radiology modalities, such as computerized tomography (CT) scan and magnetic resonance imaging (MRI), were not included.

The study described in reference [1] aims to generate impression sections of textual medical reports from chest X-rays by first generating findings and then summarizing them. To accomplish this, an encoder-decoder framework that utilizes CNN and LSTM is employed. One potential issue with the study is that errors in the image classification module can be transmitted to the generation module, which may disturb the quality of the report generated as output. There is no study available that reflects to what extent higher values of evaluation metrics, such as BLEU or ROUGE, indicate better radiology reports. Therefore, to judge the quality of the report generated, it is important using a combination of evaluation metrics and radiologist evaluation.

Babar *et al.* [15] presents a new method for assessing the quality of the diagnostic content of radiology reports generated by artificial intelligence (AI). The authors use two different types of dictionaries in their work: one contains words and another contains sentences. The sentence-based dictionary used in this study provides greater flexibility in generating more natural language reports than traditional word-based dictionaries. The authors train and test their model on the IU Chest X-rays dataset. However, the use of a dictionary containing sentences has a drawback in that it cannot generate new sentences that have not been seen in the training set. Therefore, it assumes that the training set is large enough and covers all important information such that it can cover all possible sentences.

The study described in reference [16] focuses on detecting rare diseases and generating reports for them. To accomplish this, the study utilizes a generative adversarial network that is designed to perform well in few-shot learning scenarios. Furthermore, it is used for disease classification and is used to address the challenge of detecting rare diseases for which there may be limited training data available. A hierarchical LSTM is used for report creation, which consists of a "topic LSTM", which is responsible for generating the main topics of the report, and a "sentence LSTM", which generates the corresponding sentences.

The study presented by Harzig *et al.* [17] focuses on detecting normal and abnormal findings. Similarly, it also detects biological points in gastrointestinal (GI) tract images using a deep learning approach. Two different CNNs, namely, MobileNetV2 and DenseNet-121, are used for extracting image features from the input image. The authors used two datasets for training and evaluation: Medico 2018 and Kvasir-v2. Medico 2018 contains 16 classes with 5,293 images in the development set and 8,740 images in a test set. The Kvasir-v2 dataset contains an additional 8,000 images of an additional 8 classes. The study also includes an additional dataset consisting of six videos for the report generation subtask. Finally, found that their proposed method achieved high accuracy. Furthermore, suggested that improvements could be made by generating paragraphs of text using the English language to improve automatic report generation and by using multiclass labels to improve disease detection accuracy.

Different frameworks in recent research work [18], [19] include "encoder-decoder framework" also known as Sequence to sequence, "encoder-decoder with self-attention", "transformer-based framework", "Hierarchical RNN/LSTM-based framework", and "adversarial reinforcement learning-based framework". By studying the latest research papers, it can be concluded that automatically generating medical reports from X-ray images of the chest by applying deep learning can present several challenges. Some of the main challenges are listed below.

- Limited availability of data: Limited medical datasets are available that contain medical images and their corresponding reports. Examples include IU X-RAY and MIMIC-CXR [1].
- Variability in image quality: The quality of chest X-ray images varies significantly. X-ray images vary in terms of resolution, exposure, and noise levels. Anatomical features can also appear differently due to positional variations and patient characteristics like age and posture. Deep learning models are challenged by these variances since they have to learn how to handle a large variety of image qualities and variations.

- Complex findings: Chest X-ray pictures can show abnormalities and complex anatomical features that are difficult to identify and properly interpret without specialist training. For deep learning models to produce accurate and insightful medical reports, these anomalies must be captured. Producing precise results for the abnormalities found in chest X-ray pictures is a challenging task.
- Long tail distribution problem [3]: In the training dataset, there is a notable discrepancy in the frequency of occurrences of common or typical cases compared to unusual or less commonly encountered cases.
- Interpreting context and clinical information: In order to generate medical reports, it is necessary to comprehend a wider range of information, including the patient's symptoms, medical history, and other diagnostic test results. The process of incorporating this data into the deep learning model to produce reports that are suitable for the given context is complex and necessitates careful modelling and data representation.

4. DIFFERENT TECHNIQUES FOR AUTOMATIC MEDICAL REPORT GENARTION

This paper presents state of the art techniques based on deep learning for automatically generating medical reports from chest-X ray images. The different techniques are as follows: i) transformer-based approach, ii) attention mechanism-based approach, iii) sequence-to-sequence based approach iv) adversarial training-based approach and v) hybrid model-based approach.

4.1. Transformer based approach

Transformer based approach makes use of transformer model. This method is best suitable for image captioning task which requires understanding and generating captions from visual inputs.

- Step 1: Fine-tune CheXNet: The IU-Xray dataset's chest X-ray pictures are used to fine-tune the pre-trained CheXNet model to predict particular tags. A convolutional neural network called CheXNet was created specifically for classifying chest X-ray images.
- Step 2: Compute weighted semantic features: Following CheXNet's fine-tuning, weighted semantic features are computed using the pre-trained embeddings of the predicted tags. The confidence scores of the tags are multiplied by the associated pre-trained embeddings to calculate the semantic features.
- Step 3: Condition pre-trained GPT2 model: To provide comprehensive medical reports, a pre-trained GPT2 model is conditioned on both visual and semantic characteristics as shown in Figure 4. The term “CDGPT2” refers to the conditioned GPT2 model.

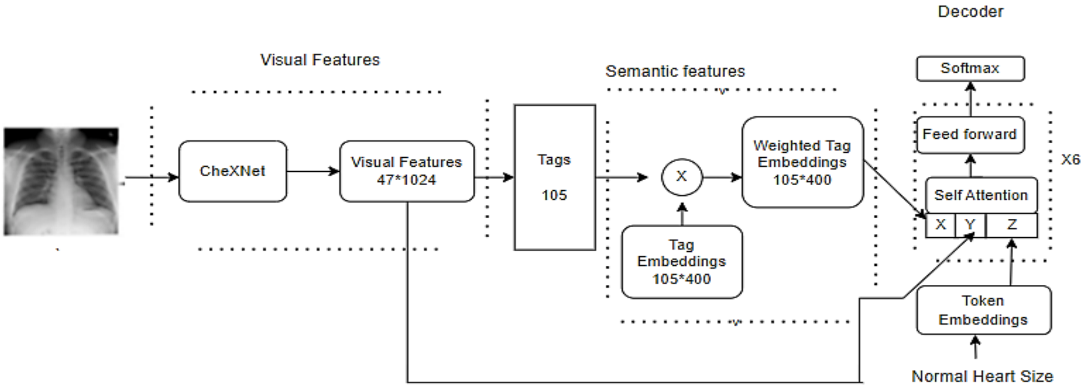


Figure 4. CDGPT2 model architecture proposed in [20] for automatically medical report creation

4.2. Attention mechanism-based approach

Attention mechanism approach is used for focusing on most relevant parts of the input data. This method improves the performance of in tasks like image captioning.

- Step 1: Encoder (CNN): Convolutional neural networks are employed as the encoder to characterize the information contained in the CXR images. CNN's VGG16 architecture is chosen because it performs well in visual classification tasks.
- Step 2: Decoder (LSTM): Using the learnt image features, a LSTM network is employed as the decoder to produce sentences. Because of its performance in tasks involving sequences, like translation, LSTM is selected.

- Step 3: Attention mechanism: The encoder's input is first sent through an attention mechanism before being fed to the LSTM. Only the areas of the pictures that are interesting and have the most information are highlighted by the attention mechanism as shown in Figure 5.

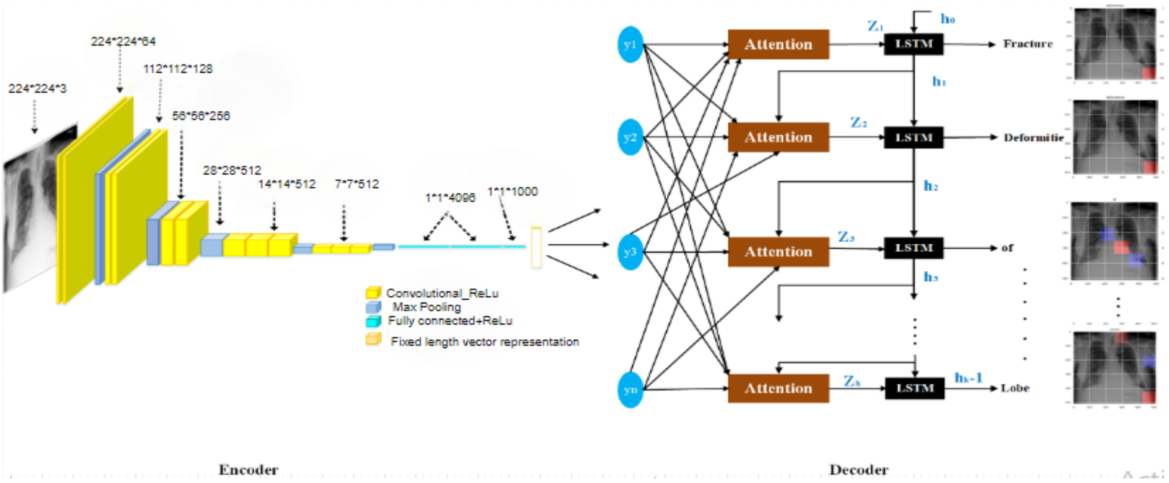


Figure 5. Attention based approach proposed in [21]

4.3. Sequence-to-sequence based approach

The sequence-to-sequence model maps an input sequence to an output sequence. This method is used for extracting image features from input image.

- Methods: A range of deep learning models, including CNNs, LSTMs, and attention processes, to produce medical reports that are both coherent and relevant to their context. Other methods include creating impression parts of written medical reports from chest X-rays, using doctors' half completed phrases and disease-related words, and the MLMA method.
- Model architectures: CNNs are utilized for feature extraction and LSTMs are used for sequence creation in most models, which are based on the encoder-decoder framework. In order to discover a variety of semantic patterns inside the report, several researches have integrated CNNs, LSTM, and bidirectional LSTM models [1], [13]–[15].

4.4. Adversarial training based approach

Adversarial training facilitates the generator to generate high quality data since it receives feedback straight from the discriminator.

- Step 1: Encoder, consists of two separate components (MLC and CNN) for independent extraction of semantic and visual data. Common observations and clinical concepts are predicted by the multi-label classification (MLC) branch and then embedded and fed into the decoder.
- Step 2: Decoder, LSTM with multiple levels of hierarchy and attention. Topic vectors are generated via sentence LSTM. Word LSTM uses topic vectors as a basis for word generation.
- Step 3: Module of reward, it has two discriminators in it. An accuracy discriminator (AD) assesses the extent to which a report includes important chest observations. The report's resemblance to expert reports is evaluated using the fluency discriminator (FD). It provides a reward based on the quality of the report; this incentive is utilized in reinforcement learning to train generators. During training cycles, the reward module and decoder are updated in order as shown in Figure 6.

4.5. Hybrid model based approach

Hybrid model based approach is utilized in study presented in [4], [5]. The study presented in [22] concentrates on enhancing contextual and visual features. The authors employ a pretrained ResNet101 as the encoder and extract features from medical images. The hierarchical decoder, called the “H-decoder”, is composed of two-level LSTMs. The LSTM at the first level is used to obtain tag features, while the LSTM at the second level generates the output, which is the target paragraph.

Messina *et al.* [23] provides a detailed review of different approaches used for generating medical reports from medical images using deep learning. It highlights the use of different algorithms, such as, including encoder-decoder, CNN, LSTM, and gated recurrent unit (GRU), as well as different performance measures, such as, including ROUGE, BLEU, and CIDER. The review also points out that different researchers are using different datasets, such as IU X-ray, ChestX-ray 14, and CheXpert.

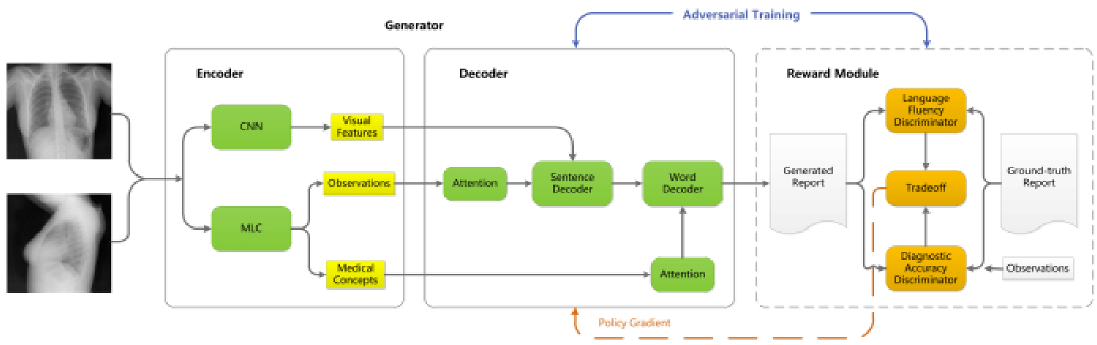


Figure 6. Adversarial reinforcement learning based approach proposed in [24]

5. DATASETS

The datasets available for producing descriptive medical reports from the chest X-ray images are as shown the Table 1. Comparison of existing datasets is provided in this table. These datasets are especially used for medical image captioning purpose. These datasets can be evaluated for appropriateness for medical imaging applications by using comparison provided in this table.

Indiana University Chest X-ray dataset is used for medical report generation task. There are 7,470 images and related reports in the Indiana University X-ray Data (IU X-ray) dataset [12], [25]. Two images, a frontal view and a lateral view are available for each patient. Each report consists of different sections, such as impressions, findings, tags, and comparisons and indications. The representative image of a frontal view from the dataset is shown in Figure 7 and the corresponding 'findings' section from the report is shown in Figure 8. Meta data for each image are available in the dataset and contain information such as patient demographics, clinical information and radiology reports.

Table 1. Comparison of different datasets for producing medical reports from Chest X Rays

Sr. No.	Data Set	No of Chest X ray images	No of patients	No of reports	Language of reports	Public/Private
1.	MIMIC CXR	377,110	65,379	227,835	English	Public
2.	IU Chest X ray	7,470	-	3,955	English	Public
3.	CX-CHR	45,598	35,609	-	Chinese	Private
4.	Chest X ray 14	112,120	30805	14,000	English	Public
5.	PadChest	160,000	67,000	Not available	-	Public
6.	CheXpert	224,316	65,240	Not available	-	Public

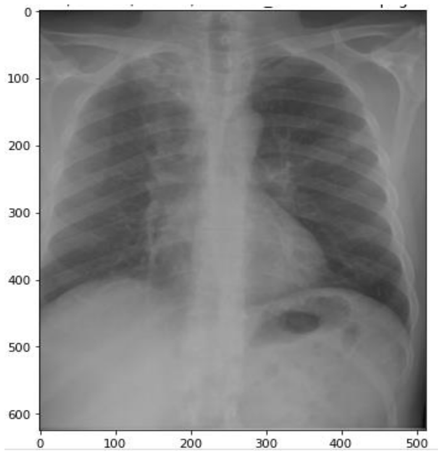
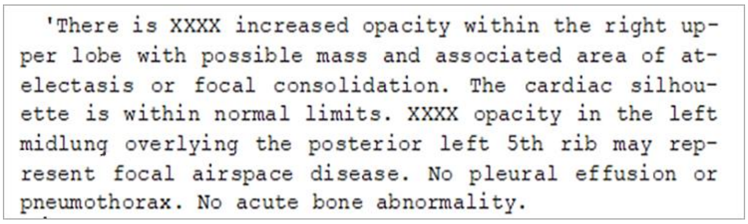


Figure 7. An example of frontal view of a chest X-Ray in the IU-CXR dataset



'There is XXXX increased opacity within the right upper lobe with possible mass and associated area of atelectasis or focal consolidation. The cardiac silhouette is within normal limits. XXXX opacity in the left midlung overlying the posterior left 5th rib may represent focal airspace disease. No pleural effusion or pneumothorax. No acute bone abnormality.

Figure 8. An example of a 'findings' section from the radiology report in the IU-CXR dataset

6. COMPARATIVE STUDY

The existing research performed in recent years is mostly based on the encoder-decoder architecture. This encoder-decoder architecture uses different models at the encoder and decoder sides. Table 2 presents a comparison of existing research performed in recent years using different comparison parameters, such as: i) models used at the encoder and decoder sides, ii) dataset used in the study, and iii) results achieved. The results are presented using different performance evaluation metrics, such as: BLEU 1, BLEU 2, BLEU 3, BLEU 4, METEOR, ROUGE and CIDEr scores.

Table 2. Comparative analysis of results obtained in the latest research work

Encoder	Decoder	Dataset	B1*	B2*	B3*	B4*	R*	M*	C*	Used by paper
ResNet 101, DenseNet 121	BERT	MIMIC-CXR and Open-I datasets	-	-	-	0.126	-	-	-	[2]
Transformer	Transformer	IU X-ray MIMIC CXR	0.496 0.351	0.319 0.223	0.241 0.157	0.175 0.118	0.377 0.287	- -	0.449 0.281	[3]
Transformer (DenseNet 121)	Transformer	IU X-ray	0.479	0.359	0.219	0.160	0.380	0.205	-	[4]
CNN	Transformer	Open-I (IU X ray) and COVID-19	0.466	0.307	0.218	0.158	0.358	-	-	[5]
CNN	Transformer	IU X-ray and MIMIC-CXR	0.461	0.285	0.196	0.145	0.367	-	-	[6]
Densnet 121 (terminology encoder)	BERT (language decoder)	CX-CHR COVID-19 CT	0.700 0.618	0.627 0.551	0.570 0.511	0.534 0.484	0.655 0.591	- -	3.220 1.158	[7]
Deep CNN (VGG 19, DensNet 121)	LSTM with attention Layer	IU X-ray, CX-CHR	0.503 0.711	0.333 0.637	0.236 0.586	0.175 0.548	0.360 0.675	- -	0.331 3.249	[8]
CNN	LSTM followed by attention	Indiana University CXR and MIMIC-CXR dataset	0.580	0.342	0.263	0.155	-	-	-	[21]
ResNet 50	Transformer	MIMIC-CXR, IU-Xray	0.492	0.320	0.236	0.179	0.395	0.218	-	[9]
ResNet-50	LSTM	IU X-ray MIMIC-CXR	0.478 0.362	0.344 0.251	0.248 0.188	0.180 0.143	0.398 0.326	- -	0.439 0.273	[10]
ResNet-101	Pre-trained SAT	IU X-ray MIMIC-CXR	0.498 0.383	0.336 0.246	0.241 0.174	0.192 0.121	0.391 0.299	0.204 0.169	0.414 0.302	[11]
CNN, LSTM	Bidirectional LSTM, Attention mechanism, word level soft max prediction	IU chest X-ray dataset	0.500	0.380	0.317	0.278	0.440	0.281	1.067	[12]
Dense CNN	LSTM	IU X-ray	0.489	0.386	0.225	0.234	-	-	0.374	[13]
CNN	LSTM	Indiana University (IU-CXR) MIMIC-CXR	0.3680 0.3538	0.2285 0.2960	0.1682 0.2549	0.1250 0.2234	0.3339 0.3532	.0.1768 0.1765	0.7589 1.3207	[14]
ResNet 50	Self-attention GRU	MIMIC-CXR	0.3538	0.2960	0.2549	0.2234	0.3532	0.1765	1.3207	[1]
This study evaluates the diagnostic contents of AI using word-based approach		Indiana University Chest X-rays	0.35 ±0.01	0.22 ±0.01	0.15 ±0.00	0.10 ±0.00	0.29 ±0.01	0.16 ±0.00	-	[15]
This study evaluates the diagnostic contents of AI using sentence-based approach			0.35 ±0.01	0.23 ±0.02	0.16 ±0.02	0.12 ±0.02	0.27 ±0.01	0.16 ±0.01	-	
Generative adversarial network	Hierarchical LSTM	IU X-ray	0.448	0.343	0.231	0.178	0.371	-	0.378	[16]
ResNet 152	Hierarchical LSTM and multilevel attention	IU X-ray, MIMIC-CXR, CheXpert labeler	-	-	-	0.125	0.262	0.171	0.366	[24]
ResNet 101	Hierarchical two-level LSTM	IU X-ray PEIR Gross dataset.	0.508 0.466	0.356 0.323	0.259 0.233	0.191 0.169	0.408 0.374	0.225 0.199	0.415 0.269	[22]

* B1, B2, B3, B4, R, M and C indicates BLEU 1, BLEU 2, BLEU 3, BLEU 4, ROUGE, METEOR and CIDEr, respectively.

6.1. Analysis and discussion

After analyzing the results obtained by different researchers, it is clear that the generated medical reports still have very low BLEU 4 and other scores that need to be improved. These metrics demonstrate how inaccurately the generated reports captured and described the anomalies seen in the chest X-ray pictures. A potential answer to these problems is to integrate a Clinical BERT [26] model as a text encoder. Clinical BERT is able to collect complex clinical terminology and information because it has been pre-trained on extensive medical corpuses. The capability of Clinical BERT to comprehend the relationships and context found in clinical writing can improve the model's ability to produce coherent and acceptable descriptions of medical imagery.

7. CONCLUSION

This paper presents a thorough analysis of deep learning methods for automating the creation of radiology reports from chest X-ray pictures. Different deep learning strategies have demonstrated notable advancements in the field, such as hybrid models, adversarial training approaches, attention mechanisms, transformer-based approaches, and sequence-to-sequence models. Despite these developments, current methods are still unable to provide diagnostic language that satisfies the strict requirements for accuracy and efficacy in medical reporting. Challenges such as limited data availability, inconsistent image quality, result interpretation difficulties, and context integration remain significant obstacles. Looking ahead, to improve text feature and image feature extraction, future research paths should concentrate on utilizing pre-trained models on biomedical datasets and adding extra metadata. Furthermore, research should be done into integrating pre-trained language models that have been trained on biomedical data and enhancing deep learning models with the help of large datasets like the ChestX-14 dataset.

8. FUTURE RESEARCH DIRECTION

There are several interesting directions for further research in the area of automated medical report generation from chest X-ray scans utilizing deep learning algorithms. Future research should primarily focus on creating and curating larger, diverse datasets for the purpose of training and evaluating deep learning models. This comprises datasets from annotated medical reports covering a broad range of illnesses, clinical features, and imaging variants. Furthermore, methods for enhancing the robustness and generalization of deep learning models to untested data and clinical contexts should be explored in future research. This may involve techniques such as domain adaptation, transfer learning, and meta-learning, alongside rigorous evaluation procedures assessing the model's performance across different institutions, imaging modalities, and patient groups.

ACKNOWLEDGEMENTS

The authors would like to acknowledge that no external funding was obtained for this study.

REFERENCES

[1] R. Beddiar and M. Oussalah, “Explainability in medical image captioning,” in *Explainable Deep Learning AI*, Elsevier, 2023, pp. 239–261.

[2] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, “Multi-modal understanding and generation for medical images and text via vision-language pretraining,” *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6070–6080, Dec. 2022, doi: 10.1109/JBHI.2022.3207502.

[3] Z. Wang, H. Han, L. Wang, X. Li, and L. Zhou, “Automated radiographic report generation purely on transformer: a multicriteria supervised approach,” *IEEE Transactions on Medical Imaging*, vol. 41, no. 10, pp. 2803–2813, Oct. 2022, doi: 10.1109/TMI.2022.3171661.

[4] A. B. Amjoud and M. Amrouch, “Automatic generation of Chest X-ray reports using a transformer-based deep learning model,” in *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)*, Oct. 2021, pp. 1–5, doi: 10.1109/ICDS53782.2021.9626725.

[5] H. T. N. Nguyen *et al.*, “EDDIE-transformer: enriched disease embedding transformer for X-ray report generation,” in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, Mar. 2022, pp. 1–5, doi: 10.1109/ISBI52829.2022.9761459.

[6] X. Jia *et al.*, “Radiology report generation for rare diseases via few-shot transformer,” in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2021, pp. 1347–1352, doi: 10.1109/BIBM52615.2021.9669825.

[7] G. Liu *et al.*, “Medical-VLBERT: medical visual language BERT for COVID-19 CT report generation with alternate learning,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 9, pp. 3786–3797, Sep. 2021, doi: 10.1109/TNNLS.2021.3099165.

[8] F. Wang, X. Liang, L. Xu, and L. Lin, “Unifying relational sentence generation and retrieval for medical image report composition,” *IEEE Transactions on Cybernetics*, vol. 52, no. 6, pp. 5015–5025, Jun. 2022, doi: 10.1109/TCYB.2020.3026098.

[9] J. Shi, S. Wang, R. Wang, and S. Ma, “AIMNet: adaptive image-tag merging network for automatic medical report generation,”

in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 7737–7741, doi: 10.1109/ICASSP43922.2022.9747702.

[10] Y. Yang, J. Yu, J. Zhang, W. Han, H. Jiang, and Q. Huang, "Joint embedding of deep visual and semantic features for medical image report generation," *IEEE Transactions on Multimedia*, vol. 25, pp. 167–178, 2023, doi: 10.1109/TMM.2021.3122542.

[11] Z. Chen and Y. Tang, "Improving radiology report generation via object dropout strategy and MLP-based captioner," in *2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC)*, Dec. 2022, pp. 316–322, doi: 10.1109/IMCEC55388.2022.10019809.

[12] G. O. Gajbhiye, A. V. Nandedkar, and I. Faye, "Automatic report generation for chest X-ray images: a multilevel multiattention approach," in *Computer Vision and Image Processing: 4th International Conference, CVIP 2019, Jaipur, India, September 27–29, 2019, Revised Selected Papers, Part I 4*, 2020, pp. 174–182.

[13] S. Biswal, C. Xiao, L. M. Glass, B. Westover, and J. Sun, "CLARA: clinical report auto-completion," in *Proceedings of The Web Conference 2020*, Apr. 2020, pp. 541–550, doi: 10.1145/3366423.3380137.

[14] S. Singh, S. Karimi, K. Ho-Shon, and L. Hamey, "Show, tell and summarise: learning to generate and summarise radiology findings from medical images," *Neural Computing and Applications*, vol. 33, no. 13, pp. 7441–7465, Jul. 2021, doi: 10.1007/s00521-021-05943-6.

[15] Z. Babar, T. van Laarhoven, F. M. Zanzotto, and E. Marchiori, "Evaluating diagnostic content of AI-generated radiology reports of chest X-rays," *Artificial Intelligence in Medicine*, vol. 116, Jun. 2021, doi: 10.1016/j.artmed.2021.102075.

[16] X. Jia, Y. Xiong, J. Zhang, Y. Zhang, and Y. Zhu, "Few-shot radiology report generation for rare diseases," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2020, pp. 601–608, doi: 10.1109/BIBM49941.2020.9313563.

[17] P. Harzig, M. Einfalt, and R. Lienhart, "Automatic disease detection and report generation for gastrointestinal tract examination," in *Proceedings of the 27th ACM International Conference on Multimedia*, Oct. 2019, pp. 2573–2577, doi: 10.1145/3343031.3356066.

[18] T. Pang, P. Li, and L. Zhao, "A survey on automatic generation of medical imaging reports based on deep learning," *BioMedical Engineering OnLine*, vol. 22, no. 1, May 2023, doi: 10.1186/s12938-023-01113-y.

[19] Y. Liao, H. Liu, and I. Spasić, "Deep learning approaches to automatic radiology report generation: a systematic review," *Informatics in Medicine Unlocked*, vol. 39, 2023, doi: 10.1016/j.imu.2023.101273.

[20] O. Alfarghaly, R. Khaled, A. Elkorany, M. Helal, and A. Fahmy, "Automated radiology report generation using conditioned transformers," *Informatics in Medicine Unlocked*, vol. 24, 2021, doi: 10.1016/j.imu.2021.100557.

[21] M. Sirshar, M. F. K. Paracha, M. U. Akram, N. S. Alghamdi, S. Z. Y. Zaidi, and T. Fatima, "Attention based automated radiology report generation using CNN and LSTM," *PLOS ONE*, vol. 17, no. 1, Jan. 2022, doi: 10.1371/journal.pone.0262209.

[22] Q. Tang, Y. Yu, X. Feng, and C. Peng, "Semantic and visual enrichment hierarchical network for medical image report generation," in *2022 Asia Conference on Algorithms, Computing and Machine Learning (CACML)*, Mar. 2022, pp. 738–743, doi: 10.1109/CACML55074.2022.00128.

[23] P. Messina *et al.*, "A survey on deep learning and explainability for automatic report generation from medical images," *ACM Computing Surveys*, vol. 54, no. 10s, pp. 1–40, Jan. 2022, doi: 10.1145/3522747.

[24] D. Hou, Z. Zhao, Y. Liu, F. Chang, and S. Hu, "Automatic report generation for Chest X-ray images via adversarial reinforcement learning," *IEEE Access*, vol. 9, pp. 21236–21250, 2021, doi: 10.1109/ACCESS.2021.3056175.

[25] R. Li, Z. Wang, and L. Zhang, "Image caption and medical report generation based on deep learning: a review and algorithm analysis," in *2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)*, Sep. 2021, pp. 373–379, doi: 10.1109/CISAI54367.2021.00078.

[26] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: chest radiology report generation with general and specific knowledge," *Medical Image Analysis*, vol. 80, Aug. 2022, doi: 10.1016/j.media.2022.102510.

BIOGRAPHIES OF AUTHORS

Prajakta Dhamanskar has received M.E. degree in information technology from V.E.S.I.T. Chembur, Mumbai, Maharashtra, India and pursuing Ph.D. in computer engineering from Parul University, Vadodara, Gujrat. She has 12+ years of experience in Academia. Currently she serves as an assistant professor, in the Department of Computer Engineering at Fr. Conceicao Rodrigues College of Engineering, Bandra, Mumbai, Maharashtra, India. Her research interests are in the areas of machine learning, deep learning, computer vision and design and algorithms. She can be contacted at email: prajakta.dhamanskar@fragnel.edu.in.

Chintan Thacker received the Ph. D. degree in the domain of artificial intelligence and computer vision from Gujrat Technical University in the year 2021. He had served as Head of the Department of Computer Science and Engineering Department at HJD Institute of Technical Education and Research, Kera, India. He has 1+ years of experience in industry and 12+ years of experience in academia. Currently, he serves as an Associate Professor in Computer Science and Engineering Department in Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujrat. In addition, he has also guided several doctorate students and has been active in conducting several workshops in the domain of computer vision. He can be contacted at email: chintan.thacker19435@paruluniversity.ac.in.