INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

Generation of Medical Reports from Chest X-Ray Images using Multi Modal Learning Approach

Anupama Navuduri[1]*, **Siddhivinayak Kulkarni**[1]

1 School of Computer Engineering and Technology, Dr. Vishwanath Karad MIT World Peace University, Pune, India

OPEN ACCESS

Received: 06-06-2023

Accepted: 04-08-2023

Published: 12-09-2023

Citation: Navuduri A, Kulkarni S (2023) Generation of Medical Reports from Chest X-Ray Images using Multi Modal Learning Approach. Indian Journal of Science and Technology 16(34): 2703-2708. https://doi.org/10.17485/IJST/v16i34.1397

* **Corresponding author**.

anu.navuduri@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: The purpose of this research is to use a multimodal learning approach to perform suggestive diagnosis and generate reports based on chest X-rays and associated data. This research falls under the Vision Language Generation, or VLG, which in this case produces reports given a chest X-ray. **Methods**: We use a Transformer model with CNN and RNN as part of a multi-modal architecture in addition to greedy beam search to generate report impressions in order to construct a proper transformer model capable of producing precise report impressions. We will also collect reports and chest X-rays from the dataset in order to evaluate the adaptability of the model: Indiana University's Open-I CXR[(1)]. This will be done so that the results can be evaluated and the model's ability to produce accurate and grammatically correct impressions of reports can be improved. **Findings**: We achieved better BLEU-1 and BLEU-2 scores compared to the research selected for this research. We have been able to achieve following BLEU scores through our proposed model: BLEU-1 = 0.592, BLEU-2 = 0.422, BLEU-3 = 0.298, BLEU-4 = 0.205. **Novelty:** We propose a Transformer model to generate report impressions. This transformer model has CNN as an encoder and RNN as a decoder with attention mechanism on top of it. Additionally, greedy beam search has been used to get grammatically correct sentences.

Keywords: Chest XRay; Transformers; OpenI CXR; CNN; RNN

1 Introduction

Clinical procedures use medical imaging modalities like X-rays, ultrasounds, MRIs, and CTs to diagnose the human body in detail. X-rays, which are easily available, are commonly used by medical experts to inspect and assess a patient's health in order to create thorough medical reports. Analysing and interpreting the chest X-ray is a very significant and difficult task. A hospital specialist creates the radiology/pathology report with the intention of reviewing the patient's medical history and conducting a thorough examination. To accurately evaluate the patient and take into account their medical history, a detailed medical report is required.

Through this literature survey, we attempt to report on recent attempts to generate radiologist reports for chest X-ray images using several machine learning and deep neural frameworks. Recent benchmark performance has been noted by Jong Hak Moon et al.[2] by introducing a CNN-BERT based model, MedVILL, for performing four downstream tasks, such as diagnosis classification, report generation, image-report pairing and image question answering. They have achieved 86% accuracy overall for all these tasks although they have used only single view Chest X-Rays for simplification in model training. In research paper[3] by Jun Li et al., Self-Guided Framework was proposed for this purpose which uses CNN-RNN as a base architecture and reported 0.467 BLEU score against the reports generated. They have noted that the trained model is able to tell differences between fine grained images and meaningless text sentences. Weakly supervised contrastive model was proposed by An Yan et al.[4] and they have reported that the reports generated were semantically close to correct grammatical sentences. They have reported around 0.467 BLEU score and have mentioned future scope for it as a way to extend this diagnosis generation using other X-Rays. CNN-LSTM model was introduced in a paper by Sirshar et al.[5] along with an attention mechanism to generate reports. They have reported 0.580 BLEU score, indicating fairly good results. They have mentioned using a small dataset as a drawback and included using bigger datasets and better training parameters as future scope. In another paper[6] by Baoyu Jing et al, they proposed a two-step strategy where first, a relationship is established between Findings and Impressions and second, a novel cooperative multi-agent system that implicitly captures the imbalanced distribution between abnormality and normality was described. Yikuan Li et al in their paper[7] have incorporated 4 pretrained models: LXMERT, VisualBERT, UNIER and PixelBERT over the MIMIC-CXR dataset. They have observed VisualBERT, LXMERT and UNITER have achieved the highest values under AUC (0.958). In a research paper by Nelson Filipe Rodrigues Nunes[8], a combination of ResNet and Bi-LSTM to report 95% accuracy in generating reports however seeking more attention to the working of these models is required as mentioned. Guanxiong Liu et al in their paper[9], they have proposed a reinforcement-based system and was assessed by a custom-made evaluation system called Clinically Coherent Reward. Open-I and MIMIC-CXR were used and CheXnet were used to compare the results with the baselines. In Baoyu Jing et al.[10], the authors have proposed a state-of-the-art framework named Co-operative Multi-Agent System or CMAS with Reinforced Learning. They have described downstream tasks to detect normalities and abnormalities. The downstream tasks defined had different outputs which would affect their final output.

Some of the research gap identified were usage of small dataset or using only a subset of the dataset selected, incoherence or grammatical errors of the output achieved. When we took a proactive approach to the process of creating reports from chest X-rays in this endeavour, two significant tasks were completed. We will initially conduct a series of tests utilizing various Deep Learning techniques and determine the better model in order to create a perfect transformer model that can generate accurate report impressions. Second, we will utilize dataset, Open-I CXR[1], to collect chest X-rays and reports to test the model's adaptability. This will be done to assess the results and strengthen the model's capacity to generate precise and grammatically correct report impressions.

2 Methodology

The proposed system architecture and workflow is given in Figure 1. Architecture of the systemp roposed for this research; we take the sample X-Ray and reports from Open I CXR Dataset, preprocess them fortraining and using VisualBERT, we are creating a model to generate properimpressions for any X-Ray image given.

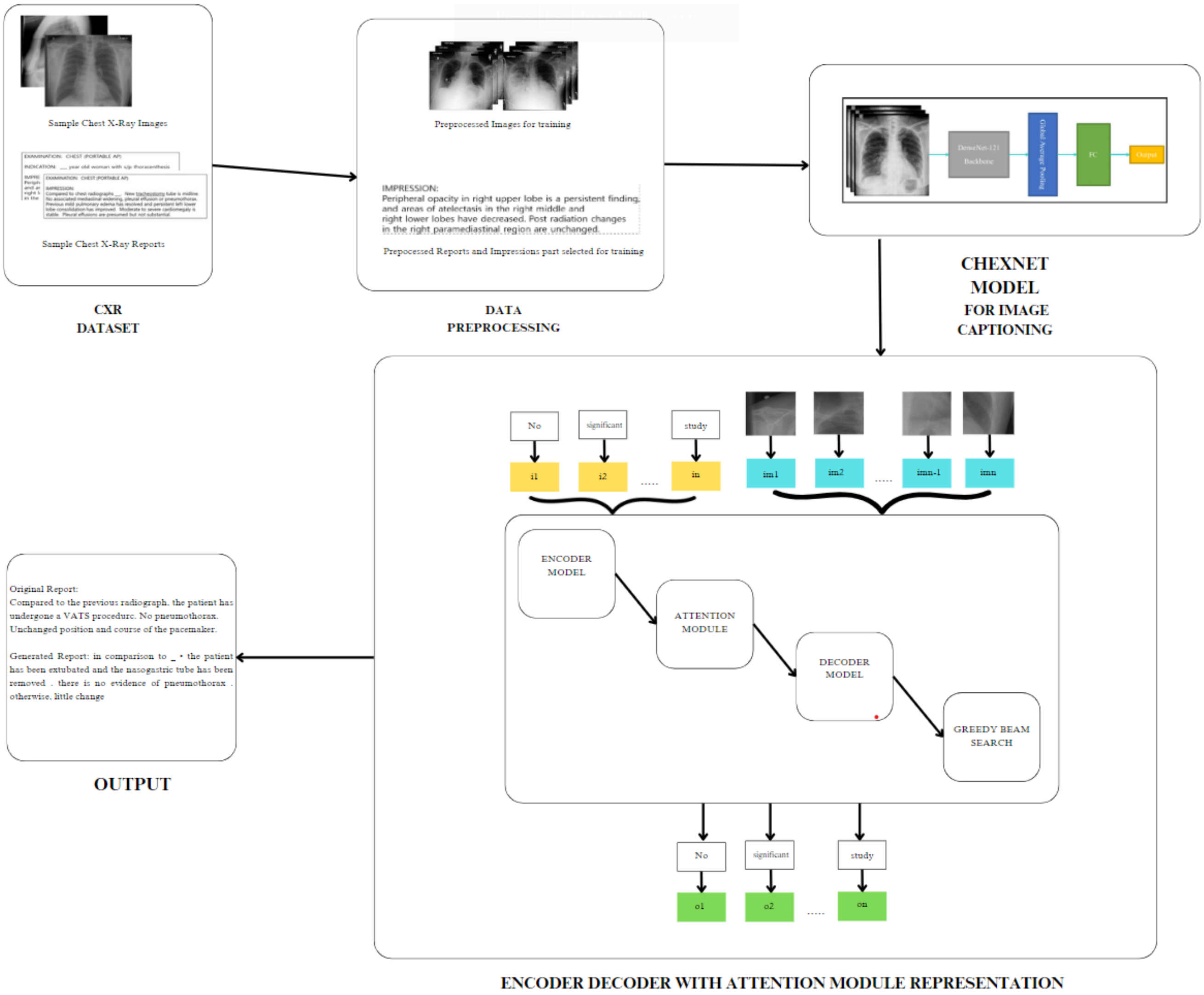
It has the following stages:

A. Data Acquisition: The dataset identified for this project is an open-source dataset from Indian University[1]. It is called the Open-I CXR dataset. The dataset contains 3,851 reports and 7,466 Chest X-Ray images. It is identified that there are more than one x-ray images associated with one report. The reports are in XML format. Reports have following sections to it, which are identified to be universally similar: Comparison, Indications, Findings and Impressions. Figure 1 shows a sample of the chest X-Ray and shows a sample of medical report impressions corresponding that is used in this research.

B. Pre-processing: The next stage is to preprocess the images and reports. The XML files of reports are converted into dataset with different columns created for each section. The file path for corresponding images for each report is also added as a column to this dataset. Maximum number of images associated with a report is 5 and minimum number is 0. Hence, baseline for the number of images permitted with single report is decided, which is 2.

C. Feature extraction: The Impressions column and images column are considered as required columns as the other columns have too many null values. The impressions are then tokenized to further feed it to the model. The Impressions column is considered the target column, i.e., given an X-Ray image, the model should be able to generate an impression from it. CheXnet model is DenseNet model with 121 convolutions, which is used to classify the images into 14 different classes for 14 different chest conditions[11]. The output of this model would be a feature vector classifying the images into 14 classes.

D. Transformer model: Once after we get this vector, we are feeding this input into a transformer model. The transformer has three components: Encoder, Decoder and Attention Layer. The encoder and the decoder layers are a combination of neural



EXAMINATION: CHEST (PORTABLE AP)

INDICATION: ___ year old woman with s/p thoracenthesis

EXAMINATION: CHEST (PORTABLE AP)

IMPRESSION:
Compared to chest radiographs ___. New tracheostomy tube is midline. No associated mediastinal widening, pleural effusion or pneumothorax. Previous mild pulmonary edema has resolved and persistent left lower lobe consolidation has improved. Moderate to severe cardiomegaly is stable. Pleural effusions are presumed but not substantial.

IMPRESSION:
Peripheral opacity in right upper lobe is a persistent finding, and areas of atelectasis in the right middle and right lower lobes have decreased. Post radiation changes in the right paramediastinal region are unchanged.

Original Report:
Compared to the previous radiograph, the patient has undergone a VATS procedurc. No pneumothorax. Unchanged position and course of the pacemaker.

Generated Report: in comparison to _ • the patient has been extubated and the nasogastric tube has been removed . there is no evidence of pneumothorax . otherwise, little change

Fig 1. Architecture of the system proposed for this research

networks that will process the information. The attention layer here makes key value pairs to map the feature vectors of the images to the word vector of the reports to help the decoder understand the next words required to be predicted for output. This trained model is then used to generate report impressions. The encoder goes over each item in the input sequence and aggregates the information it gathers into a vector called the feature context. This feature context of images is created with the help of pretrained CheXnet model. The encoder provides the context to the decoder after processing the complete input sequence. The decoder receives the feature context and tokenized report text as inputs, which begins creating the output sequence item by item. In our scenario, the encoder is a Convolutional Neural Network or CNN that generates a context vector from our picture features. A Recurrent Neural Network or RNN is used as the decoder. The dot product attention mechanism used in the cross-modal attention module is defined as:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors[(12)].

The model is developed using Masked Region Prediction (MRP), a self-supervised learning approach in which a portion of the visual input is masked off and the model is trained to predict the masked region based on the textual input. MRP's binary cross-entropy loss function is defined as follows:

$$L_{mrp} = -\frac{1}{N}\sum\nolimits_{i=1}^{N}\left(y_i log(\widehat{y}_i) + (1-y_i)log(1-\widehat{y}_i)\right)$$

where N is the number of masked regions, y$_i$ i is the ground truth label for the i-th masked region (1 if the region should be predicted correctly, 0 otherwise), and hat{y$_i$} is the predicted probability for the i-th masked region.

Using the max-pooling technique, the pooling layer in the model integrates information from both modalities into a joint embedding space:

$$maxpooling\,(x) = \underset{i=1}{max}(x_i)$$

where x$_i$ is the i-th element in the input vector x.

The result of this transformer model is refined using greedy beam search. Greedy beam search is a decoding approach often employed in natural language processing jobs to create coherent and accurate phrases after using transformers. Transformers are sophisticated sequence-to-sequence models that excel at comprehending text context and semantics, making them excellent for language translation and text production. Greedy beam search combines the ease of use of greedy decoding with the power of beam search. Instead of merely examining the most likely token, it keeps a beam of K candidates and updates it at each step, allowing the model to explore various alternatives and select more contextually suitable tokens.

E. Output: Evaluation metrics for this output involves accuracy of the model and BLEU score. BLEU score stands for Bilingual Evaluation Understudy Score[13], which calculates the score of similarity between the reference sentence and the generated sentence and measures grammatic accuracy. The score lies between 0 and 1.

The formula for calculating the BLEU score is:

$$BLEU = BPexp\left(\sum_{n=1}^{N} w_n logprecision_n\right)$$

where:

N is the maximum n-gram length considered (usually 4)

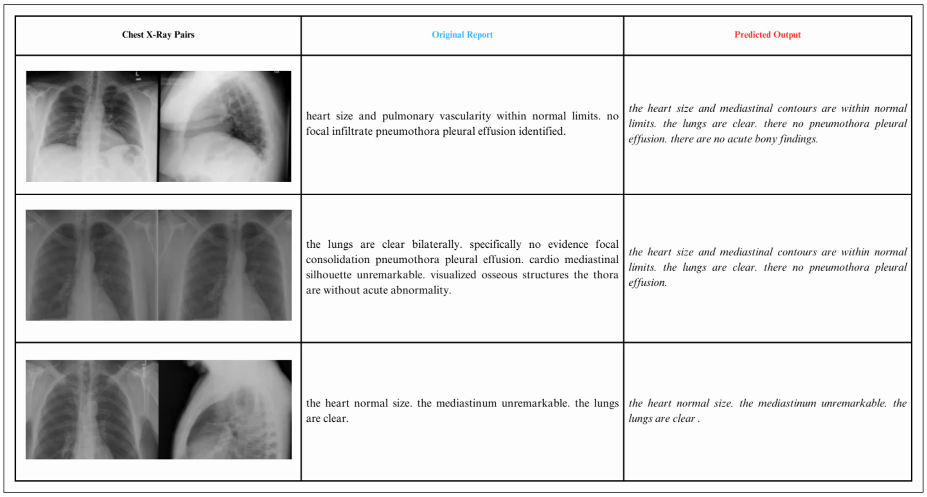
precision_n is the precision score for n-gram length. BP stands for Brevity Penalty. Then, to accommodate for machine-generated translations that are shorter than the reference translations, a brevity penalty is applied. If the length of the machine-generated translation is higher than or equal to the length of the shortest reference translation, the shortness penalty is 1, otherwise it is calculated as the ratio of the length of the machine- generated translation to the length of the shortest reference translation.

3 Results and Discussion

We compared our model with the models proposed by Jun Li et al[3], Sirshar M et al[5], Baoyu Jing et al[6], Guanxiong Liu et al[9] and Baoyu Jing et al[10]. We have trained the model for 20 epochs, depending on the hardware requirement we have and the time it takes to process. Table 1 shows the results we have achieved from our model. On the selected dataset, our model performs well on almost all metrics as compared to other models used for analysis. We have achieved better BLEU-1 and BLEU-2 scores as compared to other baseline models. However, LSTM with Co-attention[6] shows better BLEU-3 and BLEU-4 scores as compared to our model. Although the architecture of how the models used is similar to all the other baseline models, Greedy Beam search used in our model has given us better output in terms of getting impressions. As shown in Figure 2, we get decent results based on the inputs given. In general, our models tend to generate descriptions that correspond to the logical flow of radiology reports, which begin with basic information (such as views, prior comparison), positive, then negative discoveries, in the sequence of lung, heart, pleura, and others. Similar architecture can be probed further to accommodate a greater number of modalities and can be expanded to x-rays of other body parts.

4 Conclusion

The concept of using Multi Modal Learning in the field of Radiology is comparatively new. Recent developments and interest have focused on the automatic generation of radiography reports through the interpretation of chest X-ray pictures. By utilizing such technology, radiologists' workloads would be reduced, clinical procedures would move more swiftly, and the necessary medical intervention would take place more quickly, preventing serious illnesses or negative outcomes. In this paper, we present



Chest X-Ray Pairs	Original Report	Predicted Output
	heart size and pulmonary vascularity within normal limits. no focal infiltrate pneumothora pleural effusion identified.	*the heart size and mediastinal contours are within normal limits. the lungs are clear. there no pneumothora pleural effusion. there are no acute bony findings.*
	the lungs are clear bilaterally. specifically no evidence focal consolidation pneumothora pleural effusion. cardio mediastinal silhouette unremarkable. visualized osseous structures the thora are without acute abnormality.	*the heart size and mediastinal contours are within normal limits. the lungs are clear. there no pneumothora pleural effusion.*
	the heart normal size. the mediastinum unremarkable. the lungs are clear.	*the heart normal size. the mediastinum unremarkable. the lungs are clear .*

Fig 2. Results from the proposed research; we have discussed the outputs generated by the model

Table 1. Results from the proposed research; we have discussed the outputs by evaluating them using BLEU score

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Open-I Chest X-Ray Dataset[1]	SGF[3]	0.467	0.334	0.261	0.215
	CNN/LSTM with Attention[5]	0.580	0.342	0.263	0.155
	LSTM with Co-attention[6]	0.517	0.386	**0.306**	**0.247**
	Reinforced RNN[9]	0.359	0.237	0.164	0.113
	Co-operative Multi-Agent System with Reinforced Learning[10]	0.464	0.301	0.210	0.154
	Our Model (CNN/RNN with Attention and Greedy Beam Search)	**0.592**	**0.422**	0.298	0.205

a transformer architecture along with attention mechanism to generate medical report impressions of chest x-rays and greedy beam search is used to fine tune the results. We have achieved better BLEU-1 and BLEU-2 scores than the baseline models selected for comparison. The model can be trained more, depending on the resources available, to get more accurate results. This project can be further extended by using time series format data such as EEG and ECG to create more accurate reports.

References

1) Demner-Fushman D, Antani S, Simpson M, Thoma GR. Design and Development of a Multimodal Biomedical Information Retrieval System. *Journal of Computing Science and Engineering*. 2012. Available from: https://lhncbc.nlm.nih.gov/LHC-publications/PDF/pub2012019.pdf.
2) Moon HJH, Lee W, Shin YH, Kim E, Choi. Multi-modal Understanding and Generation for Medical Images and Text via Vision-Language Pre-Training. 2022. Available from: https://arxiv.org/abs/2105.11333.
3) Li J, Li S, Hu Y, Tao H. A Self-guided Framework for Radiology Report Generation. *Lecture Notes in Computer Science*. 2022;p. 588–598. Available from: https://arxiv.org/abs/2206.09378.
4) Yan A, He Z, Lu X, Du J, Chang E, Gentili A, et al. Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation. *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021. Available from: https://arxiv.org/abs/2109.12242.

5) Sirshar M, Paracha MFK, Akram MU, Alghamdi NS, Zaidi SZY, Fatima T. Attention based automated radiology report generation using CNN and LSTM. *PLOS ONE*. 2021;17(1):e0262209. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262209.

6) Jing B, Xie P, Xing E. On the Automatic Generation of Medical Imaging Reports. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2021. Available from: https://doi.org/10.18653/v1/P18-1240.

7) Li Y, Wang H, Luo Y. A Comparison of Pre-trained Vision-and-Language Models for Multimodal Representation Learning across Medical Images and Reports. *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2020. Available from: https://arxiv.org/abs/2009.01523.

8) Nunes NFR. Deep Learning for Automatic Classification of Multi-Modal Information Corresponding to. 2019. Available from: https://www.semanticscholar.org/paper/Deep-Learning-for-Automatic-Classification-of-to-Nunes/7b67cc99ef5fca071b18be0f6ab8aabae298cf1a.

9) Liu G, Hsu TMH, Mcdermott M, Boag W, Weng WH, Szolovits P, et al. Clinically Accurate Chest X-Ray Report Generation. 2019. Available from: https://doi.org/10.48550/arXiv.1904.02633.

10) Jing B, Wang Z, Xing E. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. 2019. Available from: https://aclanthology.org/P19-1657.pdf.

11) Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *Computer Vision and Pattern Recognition*. 2017. Available from: https://doi.org/10.48550/arXiv.1711.05225.

12) Vaswani A, Shazeer N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, et al. Attention Is All You Need. 2017. Available from: https://doi.org/10.48550/arXiv.1706.03762.

13) Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a Method for Automatic Evaluation of Machine Translation. 2002. Available from: https://aclanthology.org/P02-1040.pdf.