JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

MEDICAL IMAGE CAPTIONING ON CHEST X-RAYS USING CHEXNET AND LSTM

Sruthi Hasini Neerukattu, Ashritha Bhavani Satyavarapu, Gnana Sai Bandi, Nagesh Kumar Korada, V.S.V.S. Murthy

Department of Computer Science Engineering, GITAM Deemed to be University, Visakhapatnam, India-530045

Abstract: As of today, doctors must spend lot of time for analyzing the scanned Chest x-rays images. To avoid this, generating a system which can detect the abnormalities from scanned chest x-ray images and converting them into understandable description can be an efficient way to examine the reports within short span of time. The generated report will be verified by the doctor. This can help doctors to easily diagnose the medical report. In this paper, we have proposed an Image Captioning system that is applied on Chest x-ray medical images using deep learning models such as Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM). Prominent features are extracted from chest x-ray images using ChexNet CNN model. These features are given as input to long short-term Memory framework to generate long descriptive captions for better understanding of internal medical problems faced by the patients.

Keywords: Chest x-rays, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), ChexNet, Image Captioning.

I.INTRODUCTION

Image captioning is a technology that combines computer vision and natural language processing to generate textual descriptions for images. The goal is to develop algorithms or models that can automatically describe the content of an image in a human-like manner. It involves extracting meaningful features from the image using computer vision techniques and then using these features to generate a coherent and relevant natural language description. Chest X-ray images of the chest help us check the presence of tumor cells, infection, chronic lung conditions or other abnormalities. In the medical field, image captioning could describe the content of medical images, including imaging modalities used for chest disease diagnosis. For instance, an image captioning model trained on medical images could generate descriptive text for a chest abnormality, highlighting essential features, abnormalities, or potential signs of Chest. It also assists radiologists and other medical professionals analyze and communicating diagnostic results. The relationship between image captioning and chest x-rays lies in applying image captioning techniques to describe and communicate information extracted from medical images in the context of chest and monitoring. For extracting the image features

The CNN Chexnet[1] is used and it detects lung opacities, pleural effusion, atelectasis, cardiomegaly, pneumothorax, consolidation, edema, and masses/nodules in chest X-ray images, aiding in the diagnosis of thoracic abnormalities. Its capabilities assist radiologists in identifying various pulmonary conditions, guiding treatment decisions, and improving patient outcomes. Additionally, Chexnet's performance can be further enhanced by leveraging advanced techniques such as transformers or BELU[2] and integrating with other medical imaging tools. Chexnet is a convolutional neural network (CNN) architecture explicitly designed to analyze chest X-ray images. Developed by researchers at Stanford University, Chexnet aims to assist radiologists in diagnosing various thoracic diseases by providing automated analysis of X-ray images. The architecture is tailored to handle the complexities of chest X-ray interpretation, which often involves subtle abnormalities and intricate patterns that can be challenging for traditional computer vision systems to detect accurately. The features that are extracted will be sent to the LSTM model for generating the captions.

II.LITERATURE SURVEY

○ In this paper titled "Report Generation on Chest X-Rays Using Deep Learning", the authors "M. A. Kumar, S. Ganta and G. R. Chinni" [1] proposed a system which demonstrates deep learning approaches that involves encoder and decoder and a pretrained chexnet model to generate reports automatically. The Chexnet model extracts visual features from the X-ray, which are then sent to an encoder for processing. The encoder is LSTM, while the decoder is GRU and Bi GRU, both recurrent neural networks. The generated text report is evaluated using the BLEU score.

○ In this paper titled "Chest X-Ray Caption Generation with CheXNet" [2], the author "R. V. Aswiga & A. P. Shanthi" proposed a model which represents an automatic chest X-ray captioning system with two primary components: an image feature extractor and a phrase generator. Here, CheXNet and a memory-driven transformer is used to extract features and generate sentences. The model was trained and evaluated using the IU chest X-ray dataset. The model was analyzed with the BLEU.

○ In this paper titled " Multilevel Transfer Learning Technique and LSTM Framework For Generating Medical Captions For Limited Ct And DDT Images" [3], the authors "Pang, T., Li, P. & Zhao, L" proposed the MLTL framework aims to detect to classify rare diseases using limited medical datasets. It starts with a model trained on diverse non-medical images to learn generalized features. This knowledge is then transferred to an auxiliary domain relevant to target dataset, aiding in learning intermediate tasks. To enhance caption accuracy, an optimized multi-input CNN model is employed to extract image features and feed them into the LSTM for precise caption prediction, ensuring no repetition in generated captions.

○ In this paper titled [4] "A Survey on Automatic Generation of Medical Imaging Reports Based on Deep Learning" , the authors "Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo andía, Cristian Tejos, Claudia Prieto, Daniel Capurro" proposed a hierarchical LSTMs, combining correction LSTM and word LSTM to generate long chest X-ray reports. These LSTMs generate fine-grained words, forming the final medical report paragraph Reinforcement learning (RL) has also been used to optimize medical imaging report generation. Including a single reward module that captures the bias between normalcy and abnormality to produce more accurate chest X-ray results.

○ In this paper titled "Efficient Evolving Deep Ensemble Medical Image Captioning Network", the authors "Dilbag Singh; Manjit Kaur; Jazem Mutared Alanazi; Ahmad Ali AlZubi; Heung-No Lee " proposed an ensemble deep transfer network (DCNet) [5] using models like VGG16, ResNet152V2, and DenseNet201. Ensembling models improves results, enhances feature extraction, and enhances supervised model performance. The model has 128 neurons for input dense layers. A gated recurrent unit (GRU) is designed to obtain dependencies of time scales dynamically. The DCNet faces hyper-parameters tuning issues. Differential evolution (DE) is used for its robustness, performance, and simplicity in solving optimization problems.

○ In this paper titled "Image Caption Generation Using Deep Residual Learning", the authors "DAIBING HOU , ZIJIAN ZHAO ,YUYING LIU , FALIANG CHANG , AND SANYUAN HU" proposed a framework that uses ResNet to create a model for picture caption creation, utilizing skip connections to use layers efficiently. The system [6] uses residual blocks to calibrate the model before training and routes input through skip connections and weight connections. The dataset is stacked with ground truth captions before encoding, and the machine encodes the pictures into highlight vectors. The framework predicts the image and nearby activities, starting with the first word and continuing until the last word is not heard. The framework's accuracy and speed in developing competency with complex skills are superior to previous template augmentation methods.

○ In this paper titled "Dense Deep Transformer For Medical Image Segmentation: Ddtramis", the author "Ritika Jain , Arun Jhapate , Minal Saxena " proposed a DDTraMIS network [7] on a transformer architecture. This approach outlines low projection feature extraction, patching of the pictures, embedding tokens with positions, attention mechanisms, fully connected layers, and a fusion technique. High resolution local and global characteristics are required for the semantic interpretation of medical images. The feature extraction method, convolution layer and local scale-invariant process have created local representations. Following that, the transformer architecture advanced global representation. The proposed technique, called DDTraMIS, derives its name from a low projection deep layer depiction of the double attention mechanism.

○ In this paper titled "Deep Learning And Machine Learning With Grid Search To Predict Later Occurrence Of Breast Cancer Metasis Using Clinical Data" [8], the authors "Abhilasha Joshi , K. K. Sharma " proposed a fully connected feed forward network known as DFNN models. Deep learning has a huge number of hyperparameters that may be tuned, making it a potent machine learning technique. The hyperparameters are examined during the DFNN model training. In the process of preparing unstructured textual data for subsequent analysis, text preprocessing is a crucial NLP stage. Important data is extracted from radiology reports, especially reports on breast imaging. Data on the patient's demographics, the imaging modality, and the imaging findings can be extracted using a variety of pre-processing techniques. Performed pre-processing methods like stemming, tokenization, and stop word removal to investigate the use of deep learning for radiology report classification.

○ In this paper titled "Natural Language Processing For Breast Imaging: A Systematic Review", the authors "Xia Jiang ,Chuhan Xu" proposed an NLP [9] approach used to extract important information from radiology reports and statistical machine learning techniques are components of traditional NLP strategies that are used in this framework. Deep learning techniques, which have been shown to be extremely successful in a variety of applications, have, nevertheless, significantly increased in popularity in the NLP sector in recent years.

○ In this paper titled "Image Captioning Using Hybrid LSTM-RNN With Deep Features" [10], the authors "Kareem Mahmoud Diab ,Jamie Deng ,Yusen Wu ,Yelena Yesha ,Fernando Collado-Mesa ,Phuong Nguyen" proposed a hybrid deep learning architecture for automatic picture captioning has been developed. The proposed approach is divided into two parts: deep feature extraction and automatic picture captioning. A hybrid deep learning architecture for automatic picture captioning has been developed. The proposed approach is divided into two parts: deep feature extraction and automatic picture captioning. The soft max function is provided as a post-processing stage in order to obtain the output normalized probabilities. The proposed SI-RHSO approach also optimizes the RNN weights by setting the blue score as the maximum value. Back-propagation is used to calculate the gradients. The output of the optimized RNN is denoted by the notation OutRNN.

○ In this paper titled "A Review On Deep Learning In Medical Image Analysis", the authors "Kalpana Prasanna Deorukhkar, Satish Ket" proposed a model which involves a volume segmentation in CTs [11]. A patch generator and a volume renderer both get the image. The convolutional neural network (CNN) receives the sub-volumes that this patch generator divided the volume into. The volume is subsequently processed by the segmentation network, and a patch stitcher is used to reassemble it. The segmented tumor is highlighted

by rendering this processed volume. The most important phase is feature representation, which entails turning speech or text from the raw input into numerical representations that machine learning algorithms may use. Word embedding is a method for encoding words' meanings into real-valued vectors.

III.PROPOSED METHODOLOGY

3.1 IMPORTING THE DATASET:

The dataset used is Indiana University Chest X-rays dataset that represents the images which are related to the frontal and lateral part of the chest. The X-ray image contains some features related to some various diseases that are related to the chest part. The dataset contains an image folder which contains the number of images in png format. It also contains radiology reports for the chest x-ray images from the Indiana University hospital network which describes about various diseases related to the x-rays in the xml format. There are 563 reports present in this dataset. To identify images associated with the reports we need to use the xml tag. Various reports related to the chest x-rays are present which are like normal chest, No acute pulmonary findings, No acute cardiopulmonary abnormality identified.

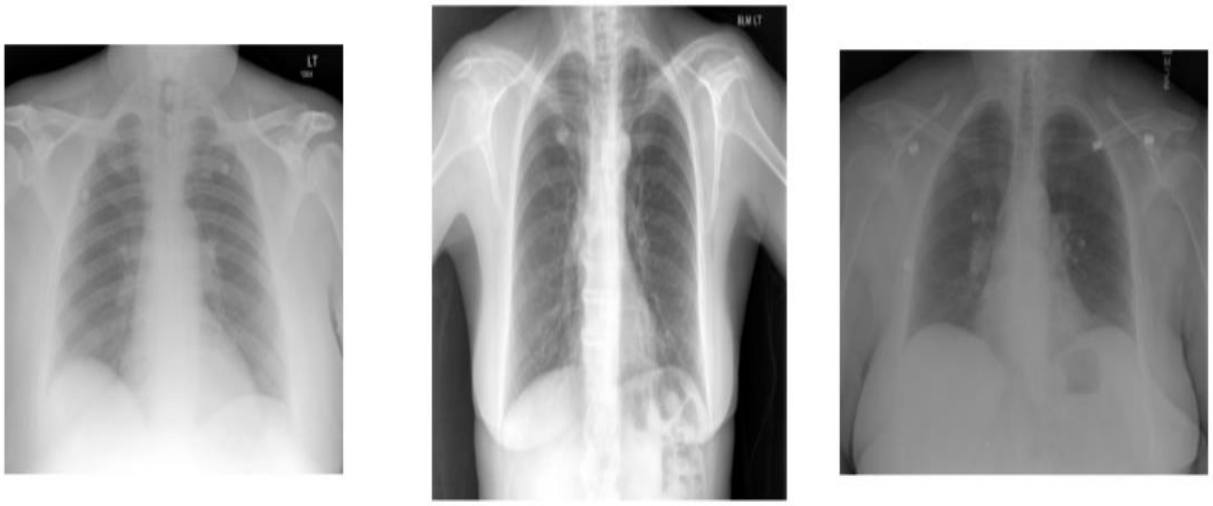

Fig 3.1: Chest X-ray Image

3.2 Pre-Processing the Dataset: Tokenization, filtering, OOV token handling, and tokenizer fitting are all part of pre-processing, which is critical for fast NLP model training. Conversion to sequences improves numerical representation, whereas vocabulary quantity influences model complexity. The calculation of caption and padding lengths ensures that sequence data is uniform. Image loading, resizing, and caption preprocessing ensure data consistency. Data augmentation broadens training instances, whilst batching and shuffling improve training efficiency and generalizability. Setting maximum padding values dependent on caption length improves neural network stability and performance with sequence data.

3.3 Build the Convolutional Neural Networks: Building a convolutional neural network (CNN) with a sequential model has advantages in terms of readability, simplicity, and effectiveness on picture data. The CNN design typically consists of convolutional, pooling, fully connected, and dropout layers. Convolution layers extract various information from input photos using mathematical operations and filters. Pooling layers reduce feature maps to capture important information. Fully connected layers handle high-level feature extraction and categorization. Dropout layers reduce overfitting by randomly deactivating neurons. CNNs like CheXNet specialize in tasks like recognizing thoracic disorders from chest X-ray pictures by using pre-trained weights from designs like DenseNet121. Transfer learning and data augmentation improve model efficiency, while careful architectural design provides optimal feature extraction for correct diagnosis.

3.4 Build the Long Short-Term Memory (LSTM): The encoder function, which uses the Image_encoder class, gathers features from input images and merges them using dimension reduction via a Dense layer to achieve unified representation. Regularization, batch normalization, and dropout help to refine features for downstream tasks like classification and similarity evaluation. For sequence-to-sequence models, the global_attention class implements attention by modulating encoder and decoder states via trainable Dense layers. This method allows for the concentration of key segments during sequence creation, which is critical for activities such as machine translation. The One Step Decoder class facilitates single-token decoding by combining encoder output, decoder hidden state, and an attention mechanism to quickly synthesize sequential tokens, which aids tasks such as language translation. The decoder class of LSTM models iteratively produces successive outputs based on encoded input and supplied caption data, employing LSTM principles. The many-to-many LSTM architecture, chosen for its appropriateness in applications such as picture captioning, converts sequential inputs into outputs while capturing complicated temporal correlations and allowing for variable caption durations. The addition of an attention mechanism improves the decoder's precision by aligning generated words with key image elements, resulting in more informative captions that accurately reflect the visual content. This comprehensive framework provides effective image captioning by capturing the essence of input photos with meaningful descriptions.

3.5 Compiling and Training: In the training pipeline, EarlyStopping, ModelCheckpoint, TensorBoard, and ReduceLROnPlateau components collectively optimize training efficiency and monitor performance. These tools ensure that the model trains effectively by stopping early if performance metrics stagnate, saving the best-performing model, providing visualization capabilities, and adjusting learning rates dynamically. With 10 epochs set for training, the model iterates over the training dataset, updating parameters through forward and backward passes to prevent overfitting or underfitting. This balanced approach in epoch selection ensures that the model converges optimally, striking the right balance between generalization and fitting the training data.

3.6 Evaluating and Testing: To evaluate the caption generating model's performance on an independent test dataset, generated captions are compared to ground truth captions. The get_bleu function is used to calculate BLEU scores for 1-gram to 4-gram sequences, allowing us to analyze the quality of machine-generated predictions. The mean_bleu function computes average BLEU scores across the dataset by iterating through data points and assigning BLEU ratings to each forecast. Greedy search predicts tokens repeatedly using output probabilities in the greedy_search_predict function, whereas final_caption_pred uses either beam search or greedy search for caption prediction during testing. These routines help to measure model effectiveness by printing photos with both true and predicted captions, allowing for a comparison of prediction quality across search algorithms. Overall, these evaluation methods simplify the assessment process by quantifying alignment.

IV.BLOCK DIAGRAM AND SEQUENCE DIAGRAM FOR THE PROPOSED METHODOLOGY:

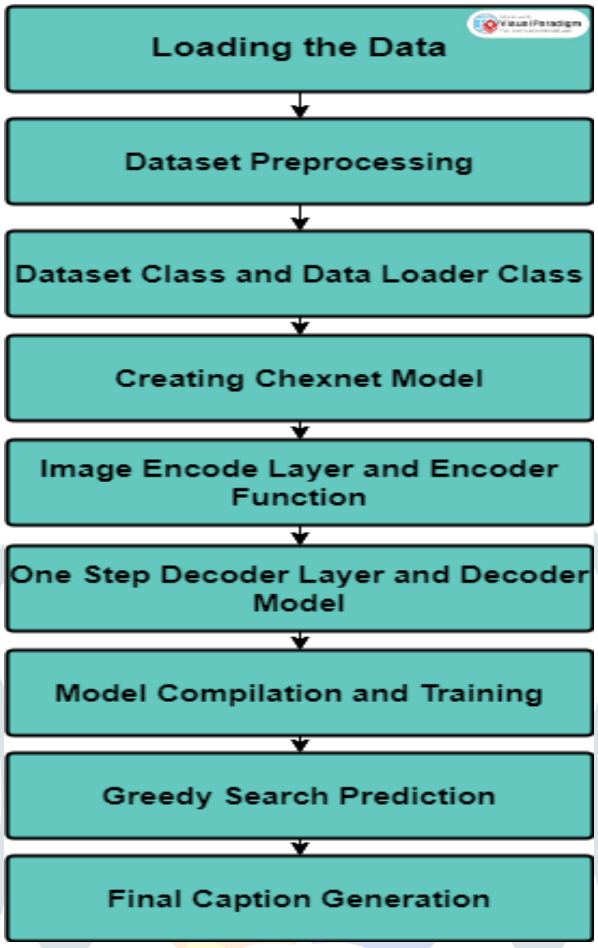

Figure 4.1: Block diagram for the proposed methodology

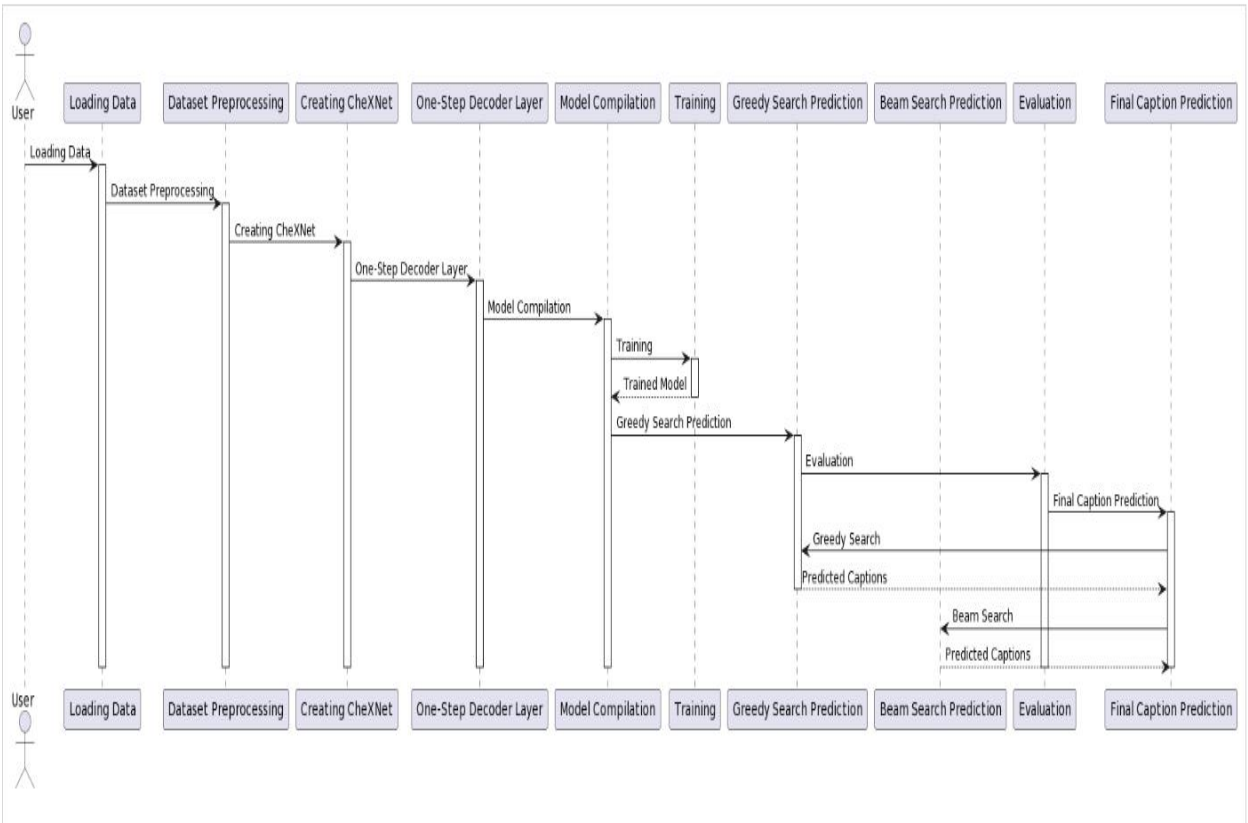

Figure 4.2: Sequence diagram for the proposed methodology

V.RESULTS AND DISCUSSIONS:

We have tested our model by implementing a testing procedure that evaluates the model's performance in generating captions for chest X-ray images by comparing the predicted captions with the ground truth captions. The model is tested in the following ways:

Test case-1:

Test case Objective: Display of an x-ray image that doesn't have any disease must give the caption 'no acute cardiopulmonary abnormality'

True caption: 'no acute cardiopulmonary findings .'

Predicted caption: 'no acute cardiopulmonary abnormality .'

Result: Pass

Figure 5.1: Test Case 1

Test case-2:

Test case Objective: Display of chest x-ray image which shows '' heart size normal. mediastinal silhouette and pulmonary vascularity are within normal limits. there is no focal airspace consolidation pleural effusion or pneumothorax.'

True caption: 'heart size normal. mediastinal silhouette and pulmonary vascularity are within normal limits. there is no focal airspace consolidation pleural effusion or pneumothorax.'

Predicted caption: 'no acute cardiopulmonary disease. no evidence of metastatic disease by radiographic evaluation.'

Result: Fail

Figure 5.2: Test Case 2

Test case-3:

Objective of the test case: Display of chest x-ray image which shows ‘stable appearance of the chest . no acute process’.

True caption: ‘stable appearance of the chest . no acute process .’

Predicted caption: ‘no acute cardiopulmonary disease . calcified right paraesophageal versus intrapulmonary lymph node . modera te hiatus hernia .’

Result: Pass

Figure 5.3: Test Case 3

Test case-4:

Objective of the testcase: Display of chest x-ray image which shows multiple diseases as “heart size normal . mediastinal silhouett e and pulmonary vascularity are within normal limits”

True caption: ‘heart size normal . mediastinal silhouette and pulmonary vascularity are within normal limits . there is no focal airs pace consolidation pleural effusion or pneumothorax .’

Predicted caption: ‘no acute cardiopulmonary disease . no evidence of metastatic disease by radiographic evaluation .’

Result: Pass

Figure 6.1.4: Test Case 4

Based upon the above test procedures we have observed the following:
The accuracy of the trained model is 0.9312 with a loss of 0.4010, it specifies that the model has learned the training data well, achieving high accuracy with relatively low loss. The validation accuracy of the model is 0.7818 with a loss of 0.8300 indicate a noticeable drop in performance on unseen data, suggesting potential overfitting or generalization issues. The BLEU scores of 0.30,

0.30, 0.33, and 0.38 for 1-gram, 2-gram, 3-gram, and 4-gram respectively imply moderate performance in generating captions compared to the reference captions. Further analysis, including hyperparameter tuning and model architecture adjustments, may be necessary to improve validation performance and caption generation accuracy.

VI.CONCLUSION AND FUTURE WORK:

Conclusion:

This paper presents a transformative approach to healthcare which addresses the several critical challenges and drawbacks in the aspect of medical image captioning. We created a machine learning model that performs image captioning of x-ray images where features are extracted from chest x-ray images and relevant captions are generated. ChexNet is used for extracting the features from the images which are given as input to LSTM for mapping the features to corresponding captions. We achieved an accuracy of 0.7818. We are successful in extracting diverse properties from x-ray images for captioning, such as cysts, high masses, high density of tissues, fluid build-up in lungs, enlarged heart, white spots, smoky lungs, visceral pleural edge, clearly visible lungs, broken ribs, cardio related problems, airspace shadowing etc.

Future Work:

In the future, there are numerous avenues for enhancing the capabilities of the model.

1.Augmenting the training dataset with a more extensive collection of Chest x-ray images could significantly improve the model's proficiency. This expanded dataset would provide the model with exposure to a wider range of image features, thereby enhancing its ability to generate precise and insightful reports.

2.Conducting thorough hyper-parameter tuning could refine the model's architecture and boost its overall performance. Exploring advanced techniques such as transformers or BELU may further enhance the model's accuracy and efficiency.

3.This Model can be enhanced in such a way that generated captions can be grammatically analysed.

4.We can also improve the model by integrating LSTM with beam-search algorithm for more accurate captions.

VII.REFERENCE:

[1] M. A. Kumar, S. Ganta and G. R. Chinni, "Report Generation on Chest X-rays using Deep Learning," 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2023, pp. 376-381, doi: 10.1109/ICICCS56967.2023.10142637

[2] R. V. Aswiga & A. P. Shanthi, “A Multilevel Transfer Learning Technique and LSTM Framework for Generating Medical Captions for Limited CT and DBT Images” Journal of digital imaging, pp. 564–580, 2022

[3] Pang, T., Li, P. & Zhao, L. A survey on automatic generation of medical imaging reports based on deep learning. BioMed Eng Online 22, 2023.

[4] Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo andía, Cristian Tejos, Claudia Prieto, Daniel Capurro, “A Survey on Deep Learning and Explainability for Automatic Report Generation from Medical Images” arXiv:2010.10563, 2022

[5] Dilbag Singh; Manjit Kaur; Jazem Mutared Alanazi; Ahmad Ali AlZubi; Heung-No Lee, “Efficient Evolving Deep Ensemble Medical Image Captioning Network” IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 2, pp. 1016- 1025, 2022

[6] DAIBING HOU, ZIJIAN ZHAO, YUYING LIU , FALIANG CHANG , AND SANYUAN HU , “Automatic Report Generation for Chest X-Ray Images via Adversarial Reinforcement Learning” IEEE access, vol. 9, pp. 21236- 21250 , 2021

[7] Ritika Jain, Arun Jhapate , Minal Saxena , “Image Caption Generation using Deep Residual Learning” 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET)

[8] Abhilasha Joshi, K. K. Sharma, “Dense deep transformer for medical image segmentation: DDTraMIS” Multimedia Tools and Applications, 2023

[9] Xia Jiang, Chuhan Xu, “Deep Learning and Machine Learning with Grid Search to Predict Later Occurrence of Breast Cancer Metastasis Using Clinical Data”, Journal of clinical medicine, vol. 11, Issue 19 ,2022

[10] Kareem Mahmoud Diab, Jamie Deng, Yusen Wu, Yelena Yesha, Fernando Collado-Mesa, Phuong Nguyen, “ Natural Language Processing for Breast Imaging: A Systematic Review”, Diagnostics ,vol 13, issue 8, 2023

[11] Kalpana Prasanna Deorukhkar, Satish Ket "Image Captioning using Hybrid LSTM-RNN with Deep Features", Sensing and Imaging ,2022

[12] Hao Cheng, Kaijie Wu, Jie Tian, Kai Ma, Chaocheng Gu & Xinping Guan, "Colon tissue image segmentation with MWSI-NET" Medical & Biological Engineering & Computing, pp.727–737 ,2022