



Computer Aided Deep Image Captioning for Medical Images

Hareem Ayesha^{1,*}, Mehreen Tariq^{2,*} and Sundas Israr³

¹Institute of Computer Science & Information Technology, The Women University, Multan, 60000, Pakistan

²Department of Computer Science, Bahauddin Zakariya University, Multan, 60000, Pakistan

³Department of Computer Science, NUML, Multan, 60000, Pakistan

*Corresponding Author: Hareem Ayesha. Email: hareem.ayesha@wum.edu.pk

Received: 18 November 2022; Revised: 22 December 2022; Accepted: 08 February 2023; Published: 6 March 2023

AID: 002-01-000018

Abstract: Drug development, illness diagnosis, prognosis, and prediction are just a few of the many different uses for medical imagery. Ultrasound, X-rays, CT scans, positron emission tomography (PET), and magnetic resonance imaging (MRI) are some of the most common forms of medical imaging. After expert medical professionals have examined these medical photos, they meticulously document their findings in detailed written reports, identifying whether the images are normal, abnormal, or potentially abnormal. This reporting procedure requires a lot of human labor, is prone to mistakes, and takes a lot of time. There have been several proposals for computer-aided report production systems to improve the efficiency and standardization of medical picture reporting. These programs mimic human doctors' work by automatically extracting information from medical images using image captioning techniques and then generating comprehensive written reports. Here, image captioning—which lies at the crossroads of AI's computer vision and natural language processing domains—is crucial. One potential way that medical practitioners may speed up their diagnostic processes is by automating the creation of medical reports. Creating medical reports for chest X-rays is the primary goal of this study's deep learning-based algorithm. A few examples of the many uses for the generated reports include verifying assumptions, keeping tabs on minor adjustments, getting a second opinion, helping with final decisions, getting quick and primary data on the issue under consideration, and much more besides. Therefore, the reports that are automatically created provide critical data for future medical treatments, instead of waiting for a report from an expert doctor.

Keywords: Medical imagery; Computer-aided systems; Image captioning; Deep learning; Diagnostic automation;

1. Introduction

There are numerous applications for medical images in the medical field. From helping chemists discover new drugs to providing surgeons with crucial information during pre-, post-, and intra-operative phases of procedures, these images are everywhere. Radiologists and other medical professionals depend on medical imaging for illness diagnosis and treatment. In the medical field, X-rays, CT-Scans, PET scans, ultrasonography, and MRIs are some of the most common imaging modalities used. After carefully analyzing these medical images, skilled doctors write extensive reports that summarize their findings. These reports are given in paragraph form and can be categorized as normal, abnormal, or potentially abnormal.

Writing medical reports from scratch requires an in-depth familiarity with the condition, medical imaging, and thorough review of the pictures; this can be especially difficult for novice examiners. It takes at least thirty minutes to review each picture and write up the results, which is a lot of time even for experienced doctors. This task is made worse in areas where medical practitioners are in short supply, like Pakistan, where the population is large, which increases the chances of wrong diagnosis [1]. In low-income nations like Pakistan, the situation is worsened since it costs more to follow up with physicians for report-related inquiries.

To make medical image reporting easier, many computer-aided report-generating systems have been suggested, all based on picture captioning. Like a seasoned doctor, these computers can automatically analyze medical photos for results and write out detailed reports. Medical professionals can save time and avoid employing extra staff to write reports manually thanks to this technology. Radiologists can use the reports for monitoring and cross-checking, getting a second opinion from other doctors, and giving techs quick information, among other uses. These reports are automatically created and provide crucial context for starting treatment quickly in situations of emergency when experienced doctors may not be easily accessible.

Although there are clear benefits, medical picture captioning does present a number of obstacles. The process of writing an extensive paragraph for a medical report is far from easy compared to making a caption consisting of only one phrase. In addition to various kind of textual descriptions, medical reports also include impressions (diagnoses), comparisons, and keyword tags generated from important discoveries. A number of steps are required to tackle these intricacies, including segmentation, feature extraction, classification, pre-processing (to reduce noise in medical pictures), and visual feature selection. Finding abnormality zones using segmentation is difficult in general, but especially when dealing with noisy modalities like ultrasound [2]. Furthermore, overfitting and possible discrepancies between produced and original captions or reports are consequences of the lack of high-quality datasets for medical picture captioning, which in turn hinders model generalization. Making sure the produced medical reports are accurate, legible, and free of grammar and spelling errors just adds to the difficulties already encountered in this field.

1.1. Objectives of Research

The primary goal of this thesis is to create a deep learning (DL) model for generating radiology reports from medical images. Other objective of this research is described below:

- Investigate the diverse applications of DL in the medical field, including image classification, segmentation, translation, retrieval, and disease detection. With particular attention to how important it is for DL to automate disease prognosis and diagnosis across all modalities so that medical professionals may make better decisions.
- Conduct a comprehensive review of existing literature to identify gaps and challenges in current methods. Examine the most important factors affecting how well automatic medical report generating works.
- Develop a DL-based model specifically designed for generating grammatically correct medical reports based on medical images.

2. LITERATURE REVIEW

The utilization of deep learning networks to accomplish natural language-like image captioning in recent years [25,26] has generated enthusiasm for the application of deep learning methodologies in the domain of medical image captioning. Existing literature frequently employs the encoder-decoder architecture, in which a CNN encodes and decodes image characteristics into fixed-length vector representations. Subsequently, the encoder Recurrent Neural Network (RNN) such as LSTM or Gated Recurrent Units (GRU) is supplied with these representations; it generates word sequences that possess precise syntax and elaborate semantics. The training process for these models commences and concludes, eliminating the

necessity to align individual modules. One drawback of this technique is that it combines multiple categories of retrieved image characteristics into a single fixed-length vector representation, which is then utilized by the decoder.

This endeavor centers around the creation of captions and textual labelling. The subsequent section presents an exhaustive analysis of caption generating methods based on deep learning. Figure 1 illustrates the taxonomy of medical image captioning systems that utilize DL.

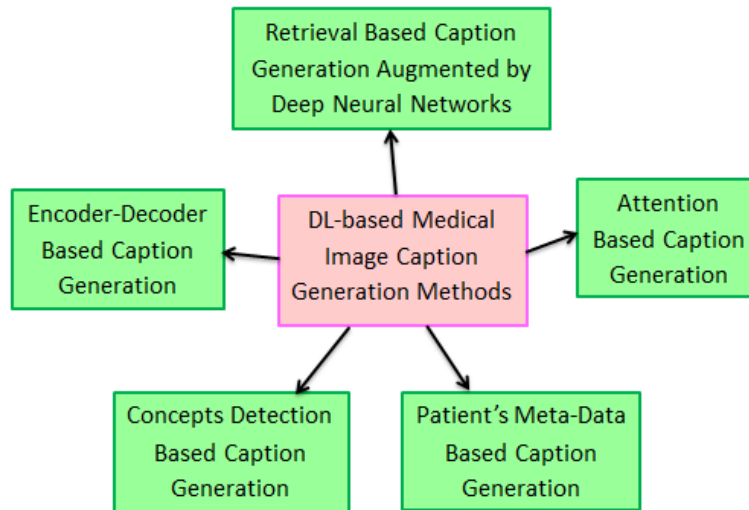


Figure 1: DL based medical image caption generation methods

2.1. Encoder-decoder based caption generation

In order to create their NN-based algorithms, the examined study makes use of Transfer Learning (TL). An encoder-decoder architecture works by first using a model that has been pre-trained on real-world pictures and then refining it using data from a specific domain. With the exception of Lyndon et al. [19], all of the other studies employ single-layer LSTM decoders. To prevent overfitting, regularization was utilised by Shin et al. [5], Wu et al. [10], Su et al. [16], and Lyndon et al. [19]. No regularisation approach was mentioned in the studies of Zeng et al. [2], Pelka et al. [18], and Spinks et al. [24]. All of these strategies provide brief captions. Despite its success, the captions produced by Shin et al. [5] are not a cohesive report but rather a 5-word "bag of words" that just explain the environment of a defined condition. The captions produced in the study by Wu et al. [10] solely included picture anomalies. Since the only focus of Wu et al. [10] was on the produced caption's relationship to the picture, specificity, sensitivity, and accuracy are the evaluation metrics utilised. Even though they included regularisation in their model, the dataset was too tiny to rule out the possibility of overfitting. In their study, Liang et al. [12] demonstrated that training on multiple models using different sub-datasets with different properties and finally classifying through SVM improved performance. However, their model is not well-suited for producing complex and fully descriptive captions, such as natural language descriptions. In the absence of a complete sentence caption, Pelka et al. [18] likewise produced just keywords. Faster RCNN was employed by Zeng et al. [2] to accomplish both picture area identification and encoding inside a single model. It outperformed captioned methods used for ultrasound pictures, which take full-size images into account, and two independent models for detection and encoding. Their model was more efficient and required fewer parameters, saving time. Nevertheless, this approach does have a few drawbacks, such as the fact that it generates brief explanations and makes mistakes in language and word class prediction when creating captions.

Table 1: Encoder Decoder based caption generation.

Authors	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[5]	GoogleNet	Image classification and detection	LSTM, GRU	Yes (fine-tuning)	Dropout, Batch normalization
[2]	VGG-16	Region detection, Classification Regression	LSTM	Yes (fine-tuning)	N/A
[10]	Batch Normalized CNN [17]	No	LSTM	Yes	Dataset Expansion, Dropout
[19]	Inception-V3	Concepts detection	3- layer LSTM	Yes	Dropout
[12]	VGGNet	SVM classification	LSTM	Yes	N/A
[18]	Inception-V3	No	LSTM	Yes (fine-tuning)	N/A
[16]	ResNet-152	Experiment on VGGNet	LSTM	Yes (fine-tuning)	Dataset Expansion
[24]	N/A	Coding Images into continuous representation	N/A	N/A	Early stopping

2.2 Attention based caption generation

Our evaluation revealed that the majority of research models for captioning are constructed utilising transfer learning. As a rule, these models make use of pre-trained encoders that have been either fine-tuned on domain datasets or utilised without any fine-tuning at all; however, in the case of Zhang et al. [4], the authors train their own encoder from the ground up. A comprehensive radiology diagnostic report is more difficult to develop than the five sorts of bladder characteristics described in the reports produced by Zhang et al. [4] based on cellular appearance (Fig. 6). Jing et al. [3] is the only paper in the attention-based caption creation category that calculates semantic attention over extracted information; all other papers in this category compute attention over solely visual spatial features. The majority of the research presented here found that employing sentence-LSTM and word-LSTM together as a decoder enhanced performance. Jing et al. [3] used hierarchical LSTM to get decent results, although the reports they made use of repeated terms. It is possible that these repeats are caused by their hierarchical model ignoring contextual coherence. Similar to Jing et al. [3], Wang et al. [9] conducted experiments on the OpenI dataset; however, they did not provide the final assessment findings. By examining the produced reports and concentrating on their experiment with the Chest X-Rays dataset, it is evident that the findings are inferior to the OpenI results reported by Jing et al. [3]. The usage of flat LSTM to decode the textual reports could be the cause of this. The reports produced by Xue et al. [15] are well-organized, although they may benefit from fewer repeats. Additionally, some of the produced reports do not catch all of the irregularities. Training the model on a limited dataset with a small number of aberrant samples might be the reason of this erroneous behaviour. It is not easy to develop well-formed sentences when they lack underlying facts. The difficulty of acquiring grammatical accuracy from tiny samples is another possible explanation. The authors Gale et al. [6] were unable to pinpoint the exact site of the fracture, but their model was able to convey the fracture's characteristics in straightforward language. According to Yuan et al. [17], combining fusion mechanism and concept information with decoding process enhances performance, with late fusion outperforming early

fusion. Training an encoder using a dataset that is particular to a domain improves its performance. Unfortunately, their model isn't superior at producing unseen words because of the tiny scale dataset.

Table 2: Attention based caption generation.

Authors	Encoder	Additional Processing	Type of Attention	Decoder	Transfer Learning	Regularization
[3]	VGG-19	Multi-label Classification to predict tags	Visual spatial and semantic	Hierarchical LSTM (Sentence , Word)	Yes (fine-tuning)	Yes (Early Stopping)
[4]	Own designed new ResNet	Symptom description-based image retrieval	Visual spatial	LSTM	No	Yes (Gradient Optimization)
[15]	ResNet-152	--	Visual attention over image and semantic attention over sentence	Hierarchical BiLSTM	Yes (Pre-trained)	--
[6]	DenseNet	Manual labeling of images	Visual spatial	BiLSTM	Yes (Pre-trained)	Yes (Dropout , augmentation)
[9]	ResNet-50	Auto-annotation and classification of images	Visual spatial	LSTM	Yes (Pre-trained, fine-tuning)	Yes (Dropout, L2 regularization)
[17]	ResNet-152	Classification of chest radiographic observations	Visual spatial	Hierarchical LSTM (Sentence , Word)	Yes (pre-trained on CheXpert dataset)	--
[14]	--	Disease classification	Visual spatial	Hierarchical LSTM (Sentence , Word)	--	--
[13]	VGGnet-19	Concepts generation	Visual spatial	LSTM	Yes (fine-tuning)	Yes (Dropout)
[20]	ResNet-101	Concepts detection	Visual spatial	LSTM	Yes (Pre-trained)	Yes (Dropout, early stopping)
[21]	RseNet-151	Disease classification, localization	Visual spatial	LSTM	Yes (Pre-trained)	--

2.3. Patient's meta-data Based Caption Generation:

A medical report's background or indication part will often go over patient information, symptoms of the ailment, and past therapies. You may include this data by merging a training model with the results section, but you'll need to train your model well to tell them apart. In order to circumvent this overhead and

direct the decoder to provide more precise results, the patient's background information is encoded independently.

Incorporating background information for report generation requires little effort and yields good results. While Zhang et al. [23] did employ regularisation and transfer learning in their model, they only produced an impression sentence and did not take medical images as input. Instead, they used a results paragraph. Rather than focusing just on visual attention, Huang et al. [11] calculated spatial and channel attention. The resulting report differed greatly from the original report, though, since the dataset was too tiny. Details of research that used patient meta-data in their reviews The results may be seen in Table 5.

Table 3: Patient's meta-data Based Caption Generation.

Authors	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[11]	ResNet-152	Spatial and Channel Attention	Hierarchical BiLSTM (Word ,Background, Sentence LSTM)	Yes	Yes (Dropout)
[23]	BiLSTM	Attention mechanism	LSTM	No	--

2.4. Retrieval based caption generation augmented by deep neural networks

Generate new descriptions for input photos using pre-existing captions in a database using retrieval-based image caption creation. Captions that aren't related to the scene or item are the result of this method's inability to adapt. Scientists are working to enhance its capabilities by merging retrieval-based caption creation with deep neural networks.

While a combination of retrieval-based captioning and deep neural networks can improve performance, there is still room for improvement when it comes to medical report generation and medical picture captioned. Liang et al. [12] found that while results may vary depending on the dataset used to train the model, utilising a combination of support vector machines (SVMs) and a convolutional neural network (CNN) improves performance, and making use of the extracted caption can yield even better results. However, complicated captions that are entirely descriptive will not be generated using the suggested paradigm. Li et al. [22] suggested a strategy where RNN-derived captions linked to sentence topics outperform new captions created using encoder-decoder architecture. However, they also noted that obtaining captions from template corpora does not adequately describe some unusual discoveries. However, with great accuracy, the caption creation module produces captions that include aberrant results. In their papers, none of them mentioned regularisation.

Table 4: Summarises the details of the research that used this approach for caption generating.

Authors	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[7]	GoogleNet for multi-label classification	Concepts prediction	--	Yes (fine-tuning)	--
[12]	VGGNet	SVM classification	LSTM	Yes	--
[22]	DenseNet	Reinforcement learning, attention mechanism	Hierarchical RNN	Yes (fine-tuning, pre-trained)	--
[8]	Inception-V3	LIRE retrieval System	--	Yes (pre-trained)	--

2.5. Concepts detection Based Caption Generation

Using the visual look of medical photographs as a starting point, this technique aims to develop medical concepts that may then be used as captions. These ideas are seen as separate components that may be used to create captions.

Little work has been discovered on producing medical captions using this approach, similar to other categories. While several studies made advantage of transfer learning, only Hasan et al. [13] developed a model that included an attention mechanism. Their better score compared to Ben abacha et al. [7] might be explained by this. The ICLEFcaption2017 competition includes both of these works. Despite using the same dataset, they omitted information about the dataset's pre-processing from their publications. In their regularisation approach, Hasan et al. [13] utilised dropout, however Ben abacha et al. [7] made no mention of regularisation at all. Unfortunately, neither author has produced a comprehensive medical report, and the captions they have produced are just one or two sentences long. Table 7 provides a summary of the research that were examined and used concepts detection-based caption creation.

Table 5: Using deep neural networks to enhance caption production based on concepts detection

Authors	Encoder	Additional Processing	Decoder	Transfer Learning	Regularization
[7]	GoogleNet	Concepts detection	--	Yes (fine-tuning)	--
[13]	VGGnet-19	Visual soft attention	LSTM	Yes (fine-tuning)	Yes (Dropout)

3. Research Methodology

The subsequent approach was implemented in order to optimize the process of medical image captioning.

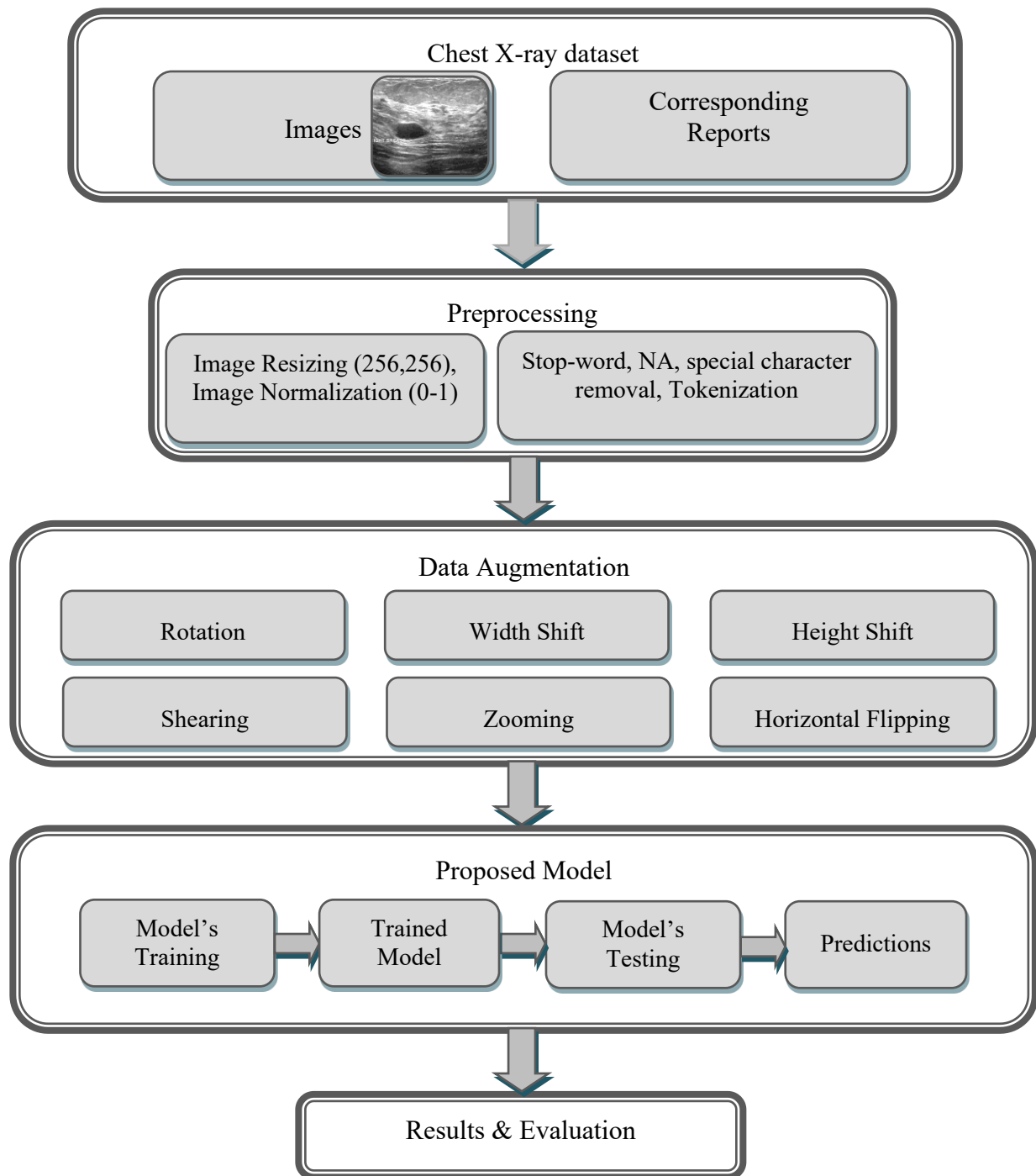


Figure 2: Research Methodology

3.1. Dataset Description

Through OpenI, we were able to obtain the IU Chest X-ray dataset, which contains 7,470 pictures of the chest and its sides, together with 3,955 radiology reports. Each image is linked to a radiology report comprising five sections: Impression (final diagnosis), Comparison (previous treatment details), Indication

(patient symptoms and meta-data), and Findings (radiologists' observations). The Tags section contains keywords derived from critical information in Impression and Findings, manually encoded with MeSH terms and auto-encoded using MTI. These tags are vital for generating terms in the final caption process. Despite researchers often using only Impression and Findings, our generated descriptions are more detailed [3].

3.2. Dataset Preparation and Preprocessing

Data preparation is a key step in developing an efficient, high-quality, and competitive model. Each algorithm has specific prerequisites that must be met in order to produce the intended results. As a result, data is preprocessed before being assigned to an algorithm. This study's data preparation and preprocessing is detailed below.

3.2.1. Image Preprocessing

In the preprocessing phase, addressing variations in image sizes within the selected dataset was essential for compatibility with deep learning networks. Two strategies were considered: padding, involving the addition of extra columns and rows to achieve uniform size but potentially incurring additional computational costs, and image resizing, where images were rescaled to a standardized (256*256) pixel size, reducing computational overhead. Image normalization was used to achieve optimal performance throughout model training. Deep model weights are often initialized with tiny random values (0-1), but the intensity range of input pictures is greater (0-255). It was essential to resize input images to a normalized range of 0–1 in order to minimize problems like bursting gradients that can impair learning.

3.2.2. Report Preprocessing

In the process of preparing the dataset for caption generation, XML reports downloaded from OpenI were initially converted into an Excel format, with only essential details extracted for input to the model. Subsequently, certain preprocessing steps were applied to refine the textual data. 'xxx' or 'xxxx', used to replace patient information in the original X-ray reports before public release, were deemed irrelevant for our purposes and thus removed. Additionally, all punctuation marks, special characters, and digits (0-9) were eliminated. Reports lacking a findings section were excluded from the dataset. To ensure consistency, all reports were converted to lowercase. Special tokens, 'startseq' and 'endseq', were introduced at the beginning and end of captions, aiding the model in recognizing sentence boundaries. Tokenization was employed, breaking down the entire report into words, with each word assigned a unique numerical token. Every unique word was kept in an array called vocabulary, and each word had a token number assigned to it. Since deep learning algorithms need numerical data, word embedding was done outside the model by giving each word a distinct token number and generating a word embedding matrix that could be integrated into the model.

3.3. Data Augmentation

During the training of deep learning architectures, a prevalent challenge is overfitting, where the model memorizes input data, yielding better results during training compared to testing. This problem is frequently caused by a lack of training data, which is a prevalent scenario in the automated diagnosis of numerous diseases due to the scarcity of large-scale medical imaging datasets. In order to reduce overfitting, we present a methodology that uses data augmentation. Specifically, we perform six different modifications to enlarge the dataset:

- Rotation: Images are rotated at random within a 0.2-degree range.
- Width Shift: Input photos are randomly shifted horizontally to the left or right by 0.05% of their entire width.
- Height Shift: Random vertical changes uphill or downward within a range of 0.05% of the overall height occur in input photos.
- Shearing: Input pictures are sheared anticlockwise within a 0.05-degree range.

- Zoom: The input photos are zoomed at random within the range $[1-0.05, 1+0.05]$.
- Horizontal Flipping: Input pictures are flipped horizontally at random.

3.4. Proposed Architecture

The proposed architecture is divided into two distinct categories. A CNN performs multi-label classification of chest x-rays as the initial model. LSTM is utilized as a captioning algorithm in the second section to generate textual paragraphs. The diagram below illustrates the strategic architecture.

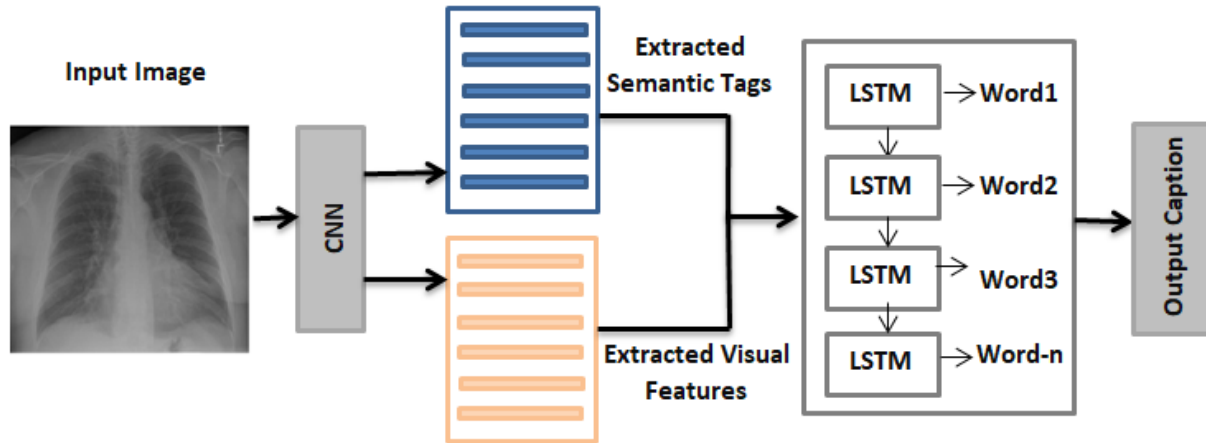


Figure 3: Proposed architecture

3.4.1. Multi Label Classification (MLC)

Convolutional neural networks (CNNs) clearly shine in binary or multi-label classification situations, according to the current research. The Keras programme allows several CNN variants to save their weights for future use. These variations are trained using large picture datasets. Researchers love these pre-trained models because they are easy to use, improve performance, and cut down on computing time. We used ChexNet [6], a pre-trained network that is indicative of its kind, in our investigation.

Specifically trained on the 14-class ChestXray14 dataset, ChexNet is a convolutional neural network. To detect 14 anomalies from 112,120 ChestX-rays, ChexNet was built using a 121-layer architecture. To make tag predictions, we used this model for both feature extraction and multi-label classification. In this case, given a picture, the task was to extract characteristics from the second-to-last layer. To make the model work for tag prediction, we added 210 nodes to the last dense layer and treated it as an MLC object. We used the top ten tags, which were determined to be the most semantically meaningful by the MLC model, to train our caption generating model. The model was trained from the ground up using a dataset of 6512 lateral and frontal chest X-rays for feature extraction and tag prediction.

3.4.2. Caption Generation

With the use of ChexNet features, an LSTM language model can create captions. Following feature extraction, a 256-unit LSTM is supplied with a word embedding matrix. In order to generate new output, the LSTM looks at previous ones. The caption model takes three things into consideration while creating a word-by-word caption: picture characteristics, extracted tags, and LSTM output.

3.5. Training Parameters

Image types in the original dataset were 70% training, 20% validation, and 10% testing. ChexNet trained with batch size 10, 100 epochs, Adam optimizer ($lr=1e-4$, binary_crossentropy). Optimal model weights preserved according to validation loss. CNN weights that have been trained are used for tag prediction and

feature extraction. Language model trained on 70% dataset with batch size 10, 100 epochs, Adam optimizer ($lr=1e-4$, categorical_crossentropy). Both models in the proposed architecture underwent image augmentation.

4. Results and Evaluation

In this chapter, we assess the outcomes of the deep learning-based architecture that we have proposed for the automated generation of captioning pertaining to chest X-rays. The experimental procedures and the system utilised in this study are detailed below:

4.1. Experimental Analysis

Two models for Multi-Label Classification (MLC) and tag prediction were trained, using several training and testing rounds to optimise parameters for our unique scenario. The model was trained in the first experiment by replacing the final layer of the pre-trained ChexNet with 210 nodes, which corresponded to the 210 tags in our research. ChexNet was trained from scratch on our chosen dataset for 80 epochs in the second experiment, with continuous monitoring of training and validation loss. Weights were saved for both models after effective training. These weights were loaded during testing, and tags were expected. The model created from scratch beat the pre-trained model, which served as the encoder in our design, according to the results. The decoder gathered features and tags from this model in order to create the final caption. Using the training datasets, the whole network was trained across 80 epochs.

4.2. Training Loss and Validation Loss

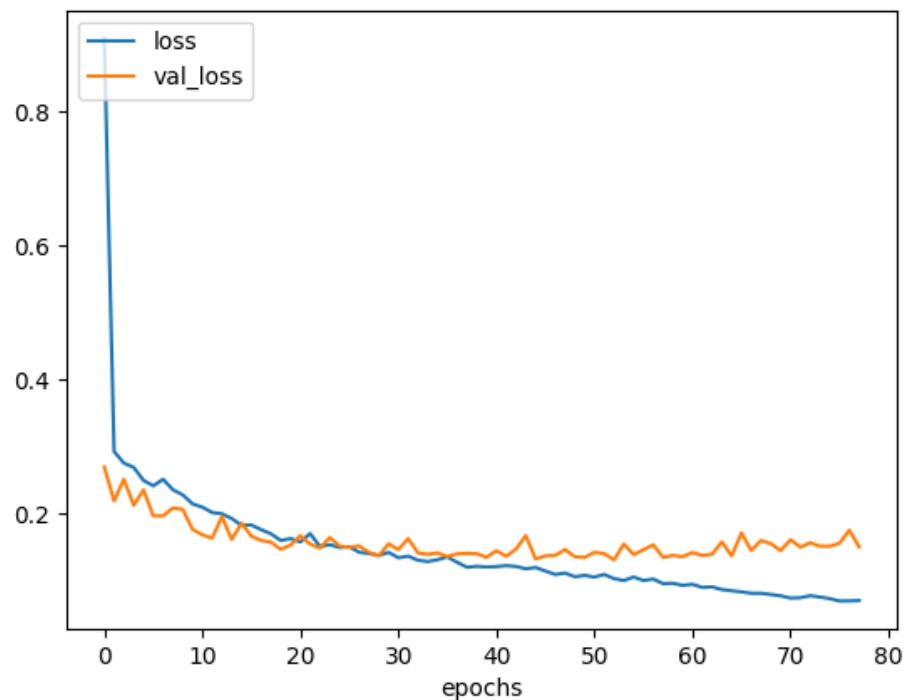


Figure 4: Training Loss and Validation Loss of Pre-trained ChexNet

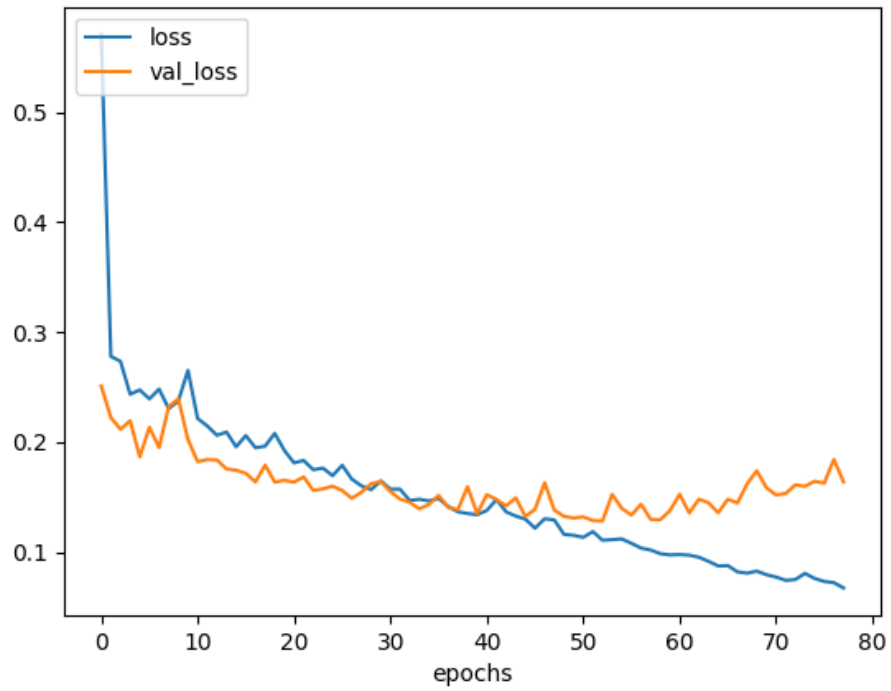


Figure 5: Training Loss and Validation Loss of ChexNet Trained from scratch

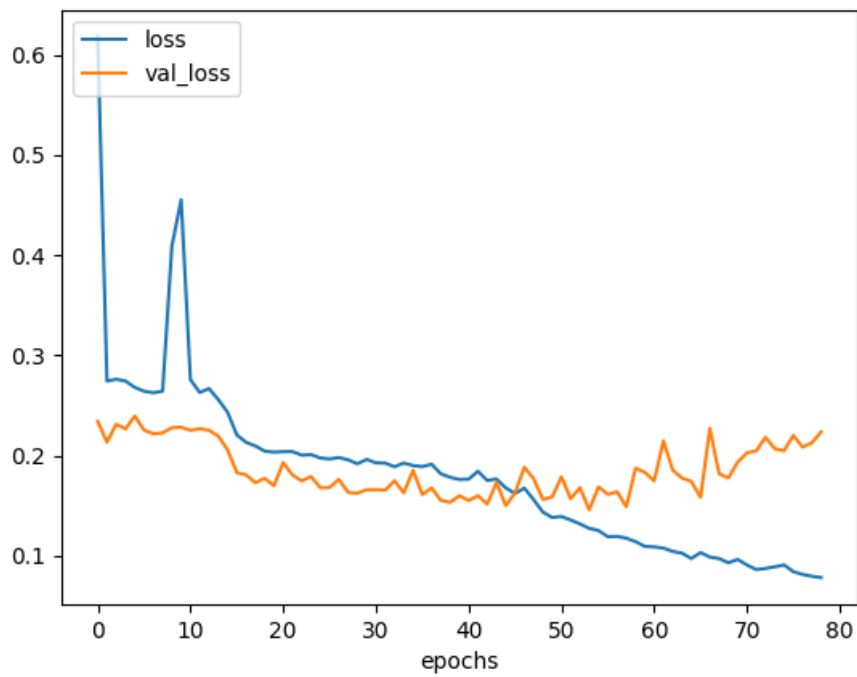


Figure 6: Training Loss and Validation Loss of Whole Network without Tags

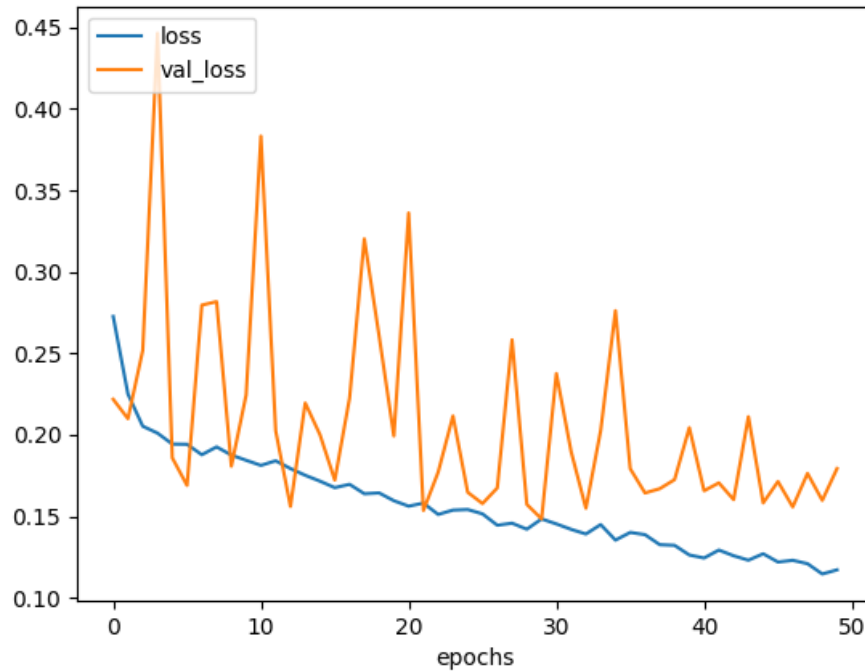


Figure 7: The Impact on Training and Validation Deletion of Entire Network Using Tags

4.2.1. Predictions of Proposed Model

Below are the figures displaying the projected reports and their matching original reports that were produced from our suggested model using 2 test images:

	Original Report	Predicted Report
	the lungs appear clear the heart and pulmonary are normal the pleural spaces are clear mediastinal contours are normal	lungs appear clear the heart is normal the pleural spaces are clear mediastinal contours are normal
	images heart size and pulmonary vascular engorgement appear within limits of normal mediastinal contour is unremarkable no focal consolidation pleural effusion or pneumothorax identified no convincing acute bony findings	Heart size and pulmonary vascular engorgement within limits of normal mediastinal contour is unremarkable no focal consolidation pleural effusion or pneumothorax identified no convincing acute bony findings findings or pneumothorax

Figure 8: Predictions from the Proposed Model on two test Images

4.2.2. BLEU Scores

Table 6: Experimental BLEU score and the final model proposal

Dataset	Method	BLEU1	BLEU2	BLEU3	BLEU4
IU Chest X-ray	Pre-trained Encoder	0.212	0.110	0.034	0.033
	From scratch trained Encoder	0.256	0.130	0.056	0.047
	With tags	0.307	0.296	0.327	0.282

The table above displays the BLEU Scores of the four models in relation to the predictions of the Chest X-ray dataset. The "startseq" and "endseq" elements are eliminated from the final prediction produced by each of the four architectures before this score is computed. We outperformed the other two studies and attained a higher BLEU score with the implementation of our proposed design.

CONCLUSION

The integration of artificial intelligence has brought about a transformative impact on the medical field, offering various applications to support medical practitioners in tasks such as automated disease detection and diagnosis, real-time patient monitoring, automated report generation, automated surgeries, and administrative responsibilities. Among these applications, automated report generation or image captioning for different diseases stands out as a particularly challenging endeavor. Unlike simple object recognition, segmentation, or classification, this task necessitates a comprehension of the relationships between different items in a picture and the behaviors that these objects represent. Traditionally, medical picture analysis and the detection of numerous types of anomalies, as well as manual report preparation, take a significant amount of time and effort. Medical image captioning emerges as a solution to address these challenges, aiming to save experts' time, mitigate subjectivity errors, and ensure accurate interpretation of both major and minor abnormalities. The resulting reports can serve as valuable resources for a range of subsequent tasks.

There are two main categories of medical image captioning models. One type uses deep learning to generate captions using different types of neural networks. The other type uses retrieval to find the most similar image-caption pairs and then associates the best one with the input image. Here, we provide a state-of-the-art deep encoder-decoder architecture and use a deep learning-based approach to build captions for medical images. The encoder in the proposed model is composed of two parts: a pre-trained deep features extractor based on the ChexNet network and the extraction of the top ten tags from the Convolutional Neural Network (CNN)'s fully connected layer. These tags will almost certainly act as the encoder's second input. The second section includes a Recurrent Neural Network (RNN), which uses the sequence created up to the (n-1) step as input at the nth step. Following each of the two input layers, a dense layer is added, and the outputs from all three portions are concatenated and given to the decoder. A completely linked hidden layer and an output layer compose the decoder. The proposed model was trained and evaluated using the publicly available chest X-ray dataset from Indiana University. Our methodology received a BLEU1 score of 0.307, confirming its efficacy in the area of medical picture captioning.

Future Work

In the future, we want to use LSTM and an attention mechanism in the proposed network's decoder.

References

- [1] Brady, Adrian, Risteárd Ó. Laoide, Peter McCarthy, and Ronan McDermott. "Discrepancy and error in radiology: concepts, causes and consequences." *The Ulster medical journal* 81, no. 1 (2012): 3.
- [2] Zeng, Xianhua, Li Wen, Banggui Liu, and Xiaojun Qi. "Deep learning for ultrasound image caption generation based on object detection." *Neurocomputing* 392 (2020): 132-141.

- [3] Jing, Baoyu, Pengtao Xie, and Eric Xing. "On the automatic generation of medical imaging reports." *arXiv preprint arXiv:1711.08195* (2017).
- [4] Zhang, Zizhao, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. "Mdnet: A semantically and visually interpretable medical image diagnosis network." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6428-6436. 2017.
- [5] Liu, Xinglong, Fei Hou, Hong Qin, and Aimin Hao. "Multi-view multi-scale CNNs for lung nodule type classification from CT images." *Pattern Recognition* 77 (2018): 262-275.
- [6] Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding et al. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." *arXiv preprint arXiv:1711.05225* (2017).
- [7] Cai, Yiheng, Yuanyuan Li, Changyan Qiu, Jie Ma, and Xurong Gao. "Medical image retrieval based on convolutional neural network and supervised hashing." *IEEE access* 7 (2019): 51877-51885.
- [8] Qayyum, Adnan, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. "Medical image retrieval using deep convolutional neural network." *Neurocomputing* 266 (2017): 8-20.
- [9] Azam, Sheikh Shams, Manoj Raju, Venkatesh Pagidimarri, and Vamsi Kasivajjala. "Q-Map: clinical concept mining from clinical documents." *arXiv preprint arXiv:1804.11149* (2018).
- [10] Soldaini, Luca, and Nazli Goharian. "Quickumls: a fast, unsupervised approach for medical concept extraction." In *MedIR workshop, sigir*, pp. 1-4. 2016.
- [11] Liu, Guanxiong, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. "Clinically accurate chest x-ray report generation." In *Machine Learning for Healthcare Conference*, pp. 249-269. PMLR, 2019.
- [12] Bustos, Aurelia, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. "Padchest: A large chest x-ray image dataset with multi-label annotated reports." *Medical image analysis* 66 (2020): 101797.
- [13] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I 13*, pp. 818-833. Springer International Publishing, 2014.
- [14] Zhang, Yu, Xuwen Wang, Zhen Guo, and Jiao Li. "ImageSem at ImageCLEF 2018 caption task: Image retrieval and transfer learning." In *CLEF CEUR Workshop, Avignon, France*. 2018.
- [15] Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. "Bleu: a method for automatic evaluation of machine translation." In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311-318. 2002.
- [16] Banerjee, S., and A. Lavie. "Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments." *Proceedings of ACL-WMT*: 65-72.
- [17] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." In *Text summarization branches out*, pp. 74-81. 2004.
- [18] Vedantam, Ramakrishna, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566-4575. 2015.
- [19] Anderson, Peter, Basura Fernando, Mark Johnson, and Stephen Gould. "Spice: Semantic propositional image caption evaluation." In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part V 14*, pp. 382-398. Springer International Publishing, 2016.
- [20] Wang, Xiaosong, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049-9058. 2018.
- [21] Wu, Luhui, Cheng Wan, Yiquan Wu, and Jiang Liu. "Generative caption for diabetic retinopathy images." In *2017 International conference on security, pattern analysis, and cybernetics (SPAC)*, pp. 515-519. IEEE, 2017.
- [22] Huang, Xin, Fengqi Yan, Wei Xu, and Maozhen Li. "Multi-attention and incorporating background information model for chest x-ray image report generation." *IEEE Access* 7 (2019): 154808-154817.
- [23] Hasan, Sadid A., Yuan Ling, Joey Liu, Rithesh Sreenivasan, Shreya Anand, Tilak Raj Arora, Vivek V. Datla et

- al. "PRNA at ImageCLEF 2017 Caption Prediction and Concept Detection Tasks." In *CLEF (working notes)*. 2017.
- [24] Frangi, Alejandro F., Julia A. Schnabel, Christos Davatzikos, Carlos Alberola-Lopez, and Gabor Fichtinger. "Lecture Notes in Computer Science: proceedings of the Medical Image Computing and Computer Assisted Intervention-MICCAI 2018." (2018).
- [25] Kilickaya, Mert, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. "Re-evaluating automatic metrics for image captioning." *arXiv preprint arXiv:1612.07600* (2016).
- [26] Elliott, Desmond, and Frank Keller. "Comparing automatic evaluation measures for image description." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 452-457. 2014.
- [27] Zeng, Xianhua, Li Wen, Banggui Liu, and Xiaojun Qi. "Deep learning for ultrasound image caption generation based on object detection." *Neurocomputing* 392 (2020): 132-141.