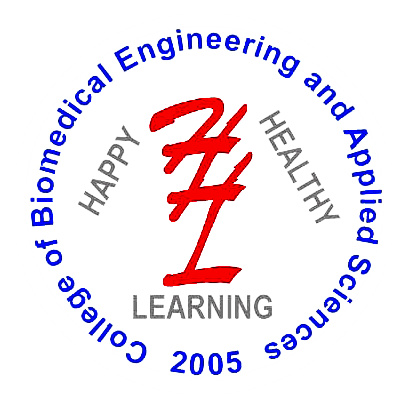
**Purbanchal University**

**Faculty of Engineering**

**College of Biomedical Engineering and Applied Sciences**

**Kathmandu, Nepal**



**Visual Language Pre-Trained Model on Medical Chest X-rays**

A Progress Report By:

Anuja Shrestha-BME/2019/A06

Esha Kharbuja-BME/2019/A11

Prabhakar Neupane-BME/2019/B21

Rishav Maharjan-BME/2019/A27

Shrijan Thapa-BME/2019/B32

Supervised by:

Dr. Kamal K.C

Er. Kushal Devkota

14th January, 2024

# Abstract

X-rays, CT scans, and MRIs serve as indispensable diagnostic tools, enabling healthcare professionals to meticulously examine the internal structures of the body for accurate diagnoses. These imaging modalities harness diverse forms of electromagnetic energy, including radio waves and X-rays, to generate detailed and insightful images of internal anatomical structures. Following the completion of these examinations composing radiology reports unfolds, demanding meticulous attention and expertise from radiologists. While these diagnostic tests have become common for patients today, the scarcity of skilled radiologists has resulted in delays in report generation, impacting the timely medical diagnosis of individuals. The prospect of automated radiology report generation holds promise in enhancing patient care and mitigating diagnostic delays, offering a potential solution to address these challenges in the healthcare landscape. This project focuses on creating an automatic radiology report generation model. This model harnesses the capabilities of LSTM networks and integrates features extracted by the VGG16 model. This approach involves analyzing X-ray images to generate descriptive captions, aiming to streamline and enhance the efficiency and accuracy of radiologists, tasks, ultimately contributing to improved patient care.

*Keywords****:*** *Deep learning, VGG-16, LSTM, Visual Language, Image Captioning, Medical Imaging Report, Generative AI, Impression, Wget*

# Abbreviations

AI: Artificial Intelligence

AP: Anterior-Posterior

BLEU: Bilingual Evaluation Understudy

CNN: Convolutional Neural Network

CT: Computed Tomography

DICOM: Digital Imaging and Communications in Medical Format

GPU: Graphic Processing Unit

JPG: Joint Photographic Experts Group

LAT: Lateral

LRN: Local Response Normalization

LSTM: Long Short-Term Memory

MIMIC-CXR: Medical Information Mart for Intensive Care Chest X-ray

MRI: Magnetic Resonance Imaging

NLP: Natural Language Processing

PA: Posterior-Anterior

PACS: Picture Archiving and Communication System

PNG: Portable Network Graphics

ReLU: Rectified Linear Unit

RNN: Recurrent Neural Network

VGG16: Visual Geometry Group

Wget: World Wide Web get

# 

**Table of Contents**

[**Abstract**…………………………….…………………………………………………...….i](#_Toc156131526)

[**Abbreviations** ii](#_Toc156131529)

[**List of Figures** v](#_Toc156131527)

[**List of Tables** vi](#_Toc156131528)

[**Chapter 1** **Introduction** 1](#_Toc156131530)

[1.1 Background 1](#_Toc156131531)

[1.2 Statement of Problem 2](#_Toc156131532)

[1.3 Objectives 2](#_Toc156131533)

[1.3.1 General Objective 2](#_Toc156131534)

[1.3.2 Specific Objectives 2](#_Toc156131535)

[**Chapter 2** **Overview of the Project** 3](#_Toc156131536)

[**Chapter 3** **Methodology** 4](#_Toc156131537)

[3.1 Dataset 4](#_Toc156131538)

[3.2 Data Processing 6](#_Toc156131539)

[3.2.1 Data Extraction 6](#_Toc156131540)

[3.2.2 Image Conversion 6](#_Toc156131541)

[3.2.3 Caption Focus 6](#_Toc156131542)

[3.2.4 Data Cleaning 6](#_Toc156131543)

[3.3 Model 7](#_Toc156131544)

[3.3.1 VGG-16 7](#_Toc156131545)

[3.3.2 LSTM 8](#_Toc156131546)

[3.4 Encoder-Decoder Networks 9](#_Toc156131547)

[3.5 Evaluation Metric 9](#_Toc156131548)

[3.5.1 BLEU 9](#_Toc156131549)

[**Chapter 4** **Results** 11](#_Toc156131550)

[**Chapter 5** **Discussion** 12](#_Toc156131551)

[**Chapter 6** **Conclusion** 17](#_Toc156131552)

[**Gantt Chart 18**](#_Toc156131553)

[**Estimated Budget 19**](#_Toc156131554)

[**References 20**](#_Toc156131555)

# List of Figures

[Figure 1 Overview of the Project 3](#_Toc156122529)

[Figure 2 Sample chest X-ray and report of p13 from MIMIC-CXR 4](#_Toc156122530)

[Figure 3 Sample chest X-ray and report of p19 from MIMIC-CXR 5](#_Toc156122531)

[Figure 4 Converting of image from DICOM to PNG format 6](#_Toc156122532)

[Figure 5 Extraction of Impression from Report 6](#_Toc156122533)

[Figure 6 VGG-16 Architecture 7](#_Toc156122534)

[Figure 7 LSTM Architecture 9](#_Toc156122535)

[Figure 8 Actual Caption and Model Predicted Caption of Chest X-ray 11](#_Toc156122536)

[Figure 9 Comparison of BLEU score of different Model 12](#_Toc156122537)

[Figure 10 Data loss using SSI 12](#_Toc156122538)

[Figure 11 Histogram plot of Pixel Intensity vs Normalized Frequency of DICOM 13](#_Toc156122539)

[Figure 12 Histogram plot of Pixel Intensity vs Normalized Frequency of PNG 13](#_Toc156122540)

[Figure 13 Loss vs Epoch graph 14](#_Toc156122541)

[Figure 14 BLEU score comparison between Models using MIMIC-CXR dataset 16](#_Toc156122542)

# List of Tables

[Table 1 Comparison of BLEU score of different models. 15](#_Toc156131272)

[Table 2 Cost Estimation 19](#_Toc156131273)

# 

# Introduction

## 1.1 Background

Chest radiography is the most common form of medical imaging in the world and can provide important diagnostic information on life-threatening cardiopulmonary conditions in a non-invasive manner [1]. However, interpreting X-rays and writing corresponding medical reports are time-consuming and difficult tasks that require experienced radiologists, which may not be available in various cases [2]. Providing accurate reports using deep learning will significantly reduce costs, ease burdens on radiologist workflows, and minimize diagnosis delays, especially in low-resource places like developing countries [3]. The computer-aided approach to diagnosis can improve the efficiency and accuracy of radiology diagnoses [4].

Deep learning has a wide range of applications in the medical field, as it can often capture complex relationships in all types of data with superior performance results [1]. Nevertheless, in medical practice, the accuracy of predictions is very important to determine the final diagnosis. Therefore, we should not consider the models as something that is unmistakably true but as an auxiliary tool that should help doctors examine X-rays.

Over recent decades, digitization of image and text data has increased rapidly. This development along with the improvement of computing resources, allows for building complex and high-performing models [5]. There have been several key advancements in deep learning, such as transformer models and unsupervised pre-training methods, which have pushed forward our ability to interpret data and automate tasks. Thus, the goal of our project is to work with vision-language deep learning methods, transferring them to the medical domain to automatically generate free-text radiology reports from patients’ radiology images.

There are two approaches for report generation, retrieval-based and generation-based. This report focuses on a Generation-based method for radiology report generation that can conceptualize the task as a medical image captioning task. This approach often takes inspiration from the state-of-the-art image-to-text generation approaches used for natural image captioning task benchmarks [6]. For example, a prior method for radiology report generation (R2Gen) trains a ResNet CNN image encoder and novel enhanced transformer-based decoder for report generation. The decoder is a Memory-Driven Transformer, which uses memory-driven conditional layer normalization to incorporate a relational memory that can enhance the transformer’s ability to learn patterns and generate text [6].

## 1.2 Statement of Problem

As per the American Journal of roentgenology and BMJ: British Medical Journal there are very few radiologists when compared to the population in a particular area, especially in the rural and smaller community settings, hence there are huge delays in the medical image interpretations and cataloging which delays the medical diagnosis and leaves patient care at risk. Our developed Model can speed up medical image interpretation and cataloging without any intervention of the radiologist and cataloguers addressing the issues effectively.

A successful deep learning application requires a very large amount of data i.e., images to train the model, as well as GPUs, or graphics processing units, to rapidly process data. In medical applications, there is a limited number of images. Therefore, one of the main challenges in applying deep learning to medical images arises from the limited number of available training samples to build deep models without suffering from overfitting.

## 1.3 Objectives

### **1.3.1 General Objective**

To develop and train a sophisticated deep learning model designed to autonomously generate textual medical reports based on chest x-ray images.

### **1.3.2 Specific Objectives**

* To conduct comprehensive image processing on the dataset to enhance the quality and relevance of chest x-ray images.
* To select appropriate neural network layers, consider transfer learning from pre-trained models, and optimizing the architecture for effective feature extraction.
* To conduct a comprehensive evaluation of the trained model to assess its efficiency and performance. Employ relevant evaluation metrics, such as BLEU, Meteor, Rogue, and CIDEr.
* To fine-tune the model to achieve optimal performance while addressing potential issues such as overfitting or underfitting.

# Overview of the Project

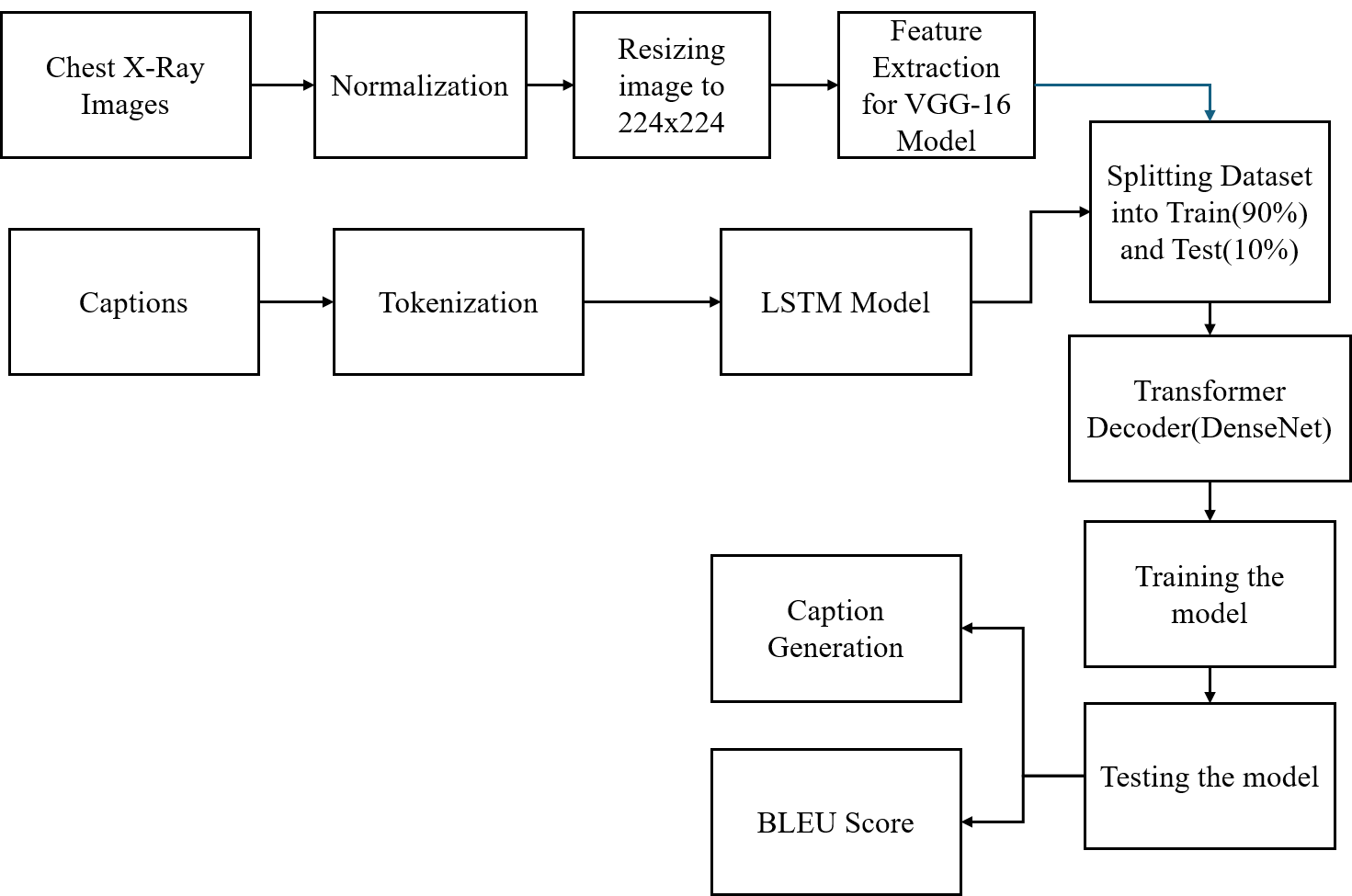


Figure 1 Overview of the Project

Feature extraction- Features are extracted from the images using a pre-trained model VGG-16. These features capture the visual content of the image.

Tokenization- The captions are tokenized into words or sub-words.

LSTM model- LSTM model is used to learn the relationships between the tokens in the captions.

Transformer decoder- A Transformer decoder is used to generate captions for new images based on the features extracted from the images and the knowledge of the language learned from the captions in the training dataset.

Training the model- The model is trained by minimizing the loss between the generated captions and the real captions in the training dataset.

Caption generation- After the model is trained, it can be used to generate captions for new images.

BLEU score- The quality of the generated captions is evaluated using a metric called BLEU score.

# **Methodology**

## 3.1 Dataset

We are using the MIMIC-CXR dataset for our project. The MIMIC-CXR database v2.0.0 is a large publicly available dataset of chest radiographs in DICOM format with free-text radiology reports [7]. The MIMIC-CXR dataset is suitable for our project as it is a collection of outpatient examinations with associated free-text radiology reports. The total size of the dataset on the website is 4.6TB. This dataset contains 377,110 images corresponding to 227,835 radiographic studies performed at the Beth Israel Deaconess Medical Center in Boston, MA [7]. Chest radiographs were sourced from the hospital picture archiving and communication system (PACS) in Digital Imaging and Communications in Medicine (DICOM) format.  DICOM is a common format for medical images which facilitates interoperability of many distinct medical devices [8]. Put simply, the DICOM format contains structured meta-data associated with one or more images, and the DICOM standard stipulates strict rules around the structure of this information. Reports are made up of Indication: symptoms, Findings: visual features noted by the radiologist in the X-ray scan, and Impression: pathology diagnosis [9]. Examples of a full exam consisting of one lateral and posteroanterior view chest X-ray image, together with the full radiological report are illustrated in Figure 2 and Figure 3. The ‘\_’ characters represent redacted information.

A x-ray of a person's chest

Description automatically generated

**FINAL REPORT EXAMINATION: CHEST (LAT)**

**INDICATION:** \_\_\_ year old man with pleural effusion // eval

**COMPARISON**: Prior chest radiographs since \_\_\_ most recently \_\_\_.

**IMPRESSION:** Moderate to large right pleural effusion has increased since \_\_\_. No pneumothorax. Atelectasis at the left base in the left upper lobe have not improved since \_\_\_. Heart size indeterminate. Right subclavian infusion catheter ends in the region of the superior Cavo atrial junction. No pneumothorax. Severe thoracolumbar scoliosis alters the thoracic anatomy.

Figure 2 Sample chest X-ray and report of p13 from MIMIC-CXR

X-ray of a chest with a lung

Description automatically generatedX-ray of a person's chest

Description automatically generated

**FINAL REPORT**

**INDICATION:** Breast cancer, on chemotherapy, evaluate for pneumonia

**COMPARISON:** Chest radiograph on \_\_\_.

**FINDINGS:** AP and lateral views of the chest. In the mid-right lung, there is a new round opacity that is concerning for a mass however may represent focal infection. There is no pleural effusion or pneumothorax. No focal consolidation. Cardio mediastinal and hilar contours are normal.

**IMPRESSION:** Right mid-lung rounded opacity may represent a new mass or infection. Recommend CT for further evaluation.

These findings were emailed to the \_\_\_ nurses by Dr. \_\_\_ at 747am on \_\_\_.

Figure 3 Sample chest X-ray and report of p19 from MIMIC-CXR

The dataset has different sub-folders named from p11 to p19. These sub-folders consist of medical chest X-rays with their respective radiographic report from 2011 to 2019. We faced storage constraints, so we downloaded only a portion of the MIMIC-CXR dataset. Specifically, we retrieved the 2013 data, which was 274GB in size. We used Wget tool to download it from *https://physionet.org/files/mimic-cxr/2.0.0/files/p13/.*

## 3.2 Data Processing

**3.2.1 Data Extraction** -Initially, only image-caption pairs with a single image and caption were extracted from the dataset using Python script.

**3.2.2 Image Conversion** -The DICOM images were converted to JPEG and PNG formats for easier handling, significantly reducing the file size from 67.3GB to 1.73GB and 9.42GB respectively.

A close-up of x-ray images

Description automatically generated

Figure 4 Converting of image from DICOM to PNG format

**3.2.3 Caption Focus** -Radiology reports typically contain a “Findings” section with detailed observations and an “Impression” section with a concise summary of the most salient findings. For the caption, we focus on the Impression section and explore using LSTM to clean the text before generating instruction-following data.

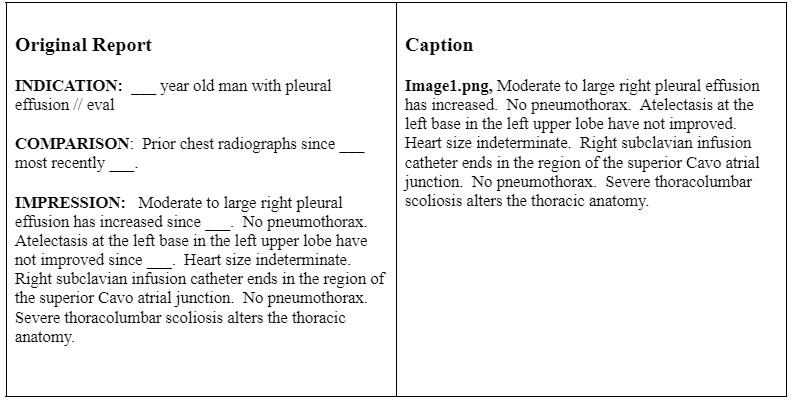


Figure 5 Extraction of Impression from Report

**3.2.4 Data Cleaning** -Images without impressions were removed, resulting in 3377 images with corresponding captions.

## 3.3 Model

The architectures we used in our project are described below:

### **3.3.1 VGG-16**

VGG-16 is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper “Very Deep Convolutional Networks”. It is for large-scale Image Recognition. This model achieves 92.7% top-5 test accuracy in ImageNet (which is a dataset of over 14 million images belonging to 1000 classes) [10]. It improves Alex-Net by replacing large kernel-sized filters (11 and 5 in the first and second convolutional layer respectively) with multiple 3×3 kernel-sized filters one after another. VGG-16 was trained for weeks by using NVIDIA Titan Black GPU’s. The architecture of VGG-16 is shown in Figure 6.

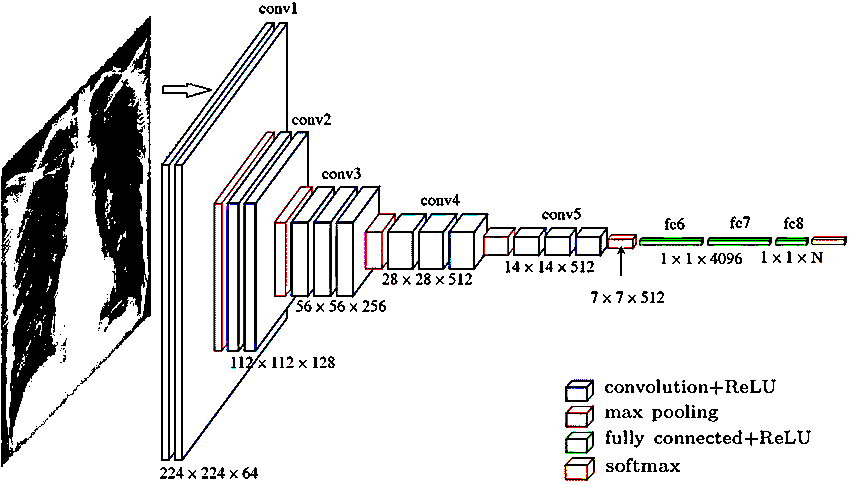


Figure 6 VGG-16 Architecture

The input to the convolution layer is a fixed size 224×224 RGB image. The image is passed through a stack of convolutional layers. Here the filters are used with a very small receptive field of 3×3 (which is the smallest size to capture the notion of up/down, left/right, center). In one of the configurations, utilizes 1×1 convolution filters, which can be seen as a linear transformation of the input channels (followed by non-linearity). The convolution stride is fixed to one pixel; the spatial padding of convolution layer input is such that the spatial resolution is preserved after convolution i.e. the padding of convolution layer input is such that the spatial resolution is carried out by five max-pooling layers, which follow some of the convolution layers (not all the convolution layers are followed by max-pooling). Max-pooling is performed over a 2×2 pixel window with stride 2 [11].

Three fully connected layers follow a stack of convolutional layers (which has a different depth in different architectures): the first two have 4096 channels each, the third performs 1000-way ILSVRC classification and thus contains 1000 channels (one for each class). The final layer is the SoftMax layer for multiple classification and sigmoid for binary. The configuration of the fully connected layer is the same in all networks. All hidden layers are equipped with ReLU. It is noted that none of the networks (except for one) contain LRN, such normalization does not improve the performance on the ILSVRC dataset but leads to increased memory consumption and computation time.

### **3.3.2 LSTM**

Long short-term memory (LSTM) models are widely used in machine translation and natural image and video captioning due to their ability to capture long-term dependencies and reduce the problem of vanishing gradients in vanilla RNNs [12]. LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data, apart from single data points such as images. This finds application in speech recognition, machine translation, etc. LSTM is a special kind of RNN, which shows outstanding performance on a large variety of problems. Each LSTM unit has three sigmoid gates to control the internal state: input, output, and forget. At each time step, the gate controls how much of the previous time steps are propagated through to determine the output. The forget gate takes the current input and previous hidden state and decides whether to keep the information from the previous time stamp or forget it [13]. The input gate does the same but quantifies the importance of the current input (new information). The cell hidden state *h(t)*) is updated based on the input and forget gates, and the output gate determines the value of the next hidden state *m(t*).

For an input word sequence *{x1, . . ., xn}*, the internal hidden state *h*,and memory state *mt* are updated as follows:

(3.1)

(3.2)

) (3.3)

(3.4)

(3.5)

where *xt* is the input at time step *t, W(hx)* and *W(hm)* are the trainable weight parameters, and *it , ot* and *ft* are the input, output and forget gates respectively.

A diagram of a computer process

Description automatically generated

Figure 7 LSTM Architecture

## 3.4 Encoder-Decoder Networks

Automated caption generation draws on both computer vision and natural language processing techniques of image and text representation [14]. LSTMs have been shown to generate human-like text by training on large textual dataset [15]. Given a starting token (a character or a word), these models predict which tokens are likely to follow. LSTM language generation model has been applied to tasks such as machine translation whereby the language generation is conditioned on a representation of a word or a sentence. The words and sentences in one language are encoded into a single representation using LSTM and decoded using another.

This idea of encoder-decoder networks has been extended to include encoding other data representations to be decoded into text: summarizing electronic medical records, text summarization and image caption generation [15]. The various methods and architectures of encoding the image information, passing the image features to the VGG-16, and generating the outputs.

## 3.5 Evaluation Metric

### **3.5.1 BLEU**

BLEU (Bilingual Evaluation Understudy) is a type of modified precision metric that evaluates how closely a model-generated text matches that of the (human-generated) ground truth [16]. It was developed for evaluating the quality of machine translation and has been reported to have a high correlation with human judgment. It has since been used to evaluate image caption generation, text summarization, and speech recognition. The BLEU score is always a number between 0 and 1, with numbers closer to 1 representing a closer match to the ground truth, or reference, text [16]. A unigram BLEU score, or BLEU-1, is the modified precision based on single word matches, penalized by candidate sentence length in comparison to reference length. For instance, although a candidate sentence [‘the cat’] matches [‘the cat sat on the mat’] when calculating word-level precision, this would favor shorter sentences, therefore a ‘brevity’ penalty is applied to lower the score of a candidate sentence if it is shorter than all the reference sentences. Similarly, bigram, 3-gram and 4-gram individual BLEU scores can be calculated [17]. Typically, these scores are taken cumulatively by taking a geometric mean of scores up to BLEU n-gram.

The formulae used to calculate the BLEU score are:

(3.6)

(3.7)

Where *c* is predicted length and *r* is target length.

(3.8)

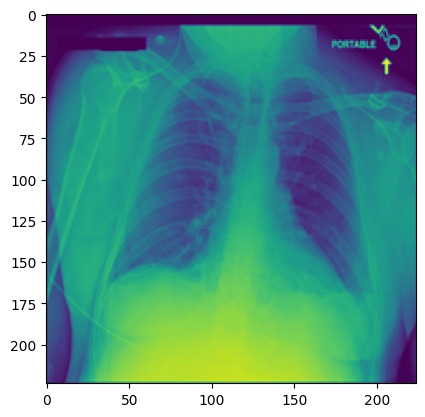
BLEU score can be computes for different values of N. typically, we use N=4. BLEU-1 uses the unigram Precision score, BLEU-2 uses the geometric average of unigram and bigram precision and so on.

# Results

The DICOM images were meticulously preprocessed by conversion to PNG format with normalization. The Structural Similarity Index (SSI) between the original DICOM images and their PNG counterparts was calculated, resulting in a value of 0.0891.

Subsequently, a model was developed utilizing the VGG16 and LSTM components, and rigorous evaluation ensued with the optimization of various hyperparameters, including batch size and epoch adjustments. Through systematic experimentation, the model achieved notable improvements in BLEU scores. The optimal configuration yielded a BLEU-1 score of 0.2013535 and a BLEU-2 score of 0.081486, reflecting enhanced performance in the image captioning task. This empirical investigation underscores the significance of fine-tuning hyperparameters in optimizing model performance for complex tasks such as image captioning.

It is noteworthy that while our model demonstrated substantial progress, achieving notable BLEU scores, it is acknowledged that the predicted captions were not fully accurate.



---------------------Actual---------------------

startseq bibasilar atelectasis. no focal consolidation or pneumothorax. Endseq

--------------------Predicted--------------------

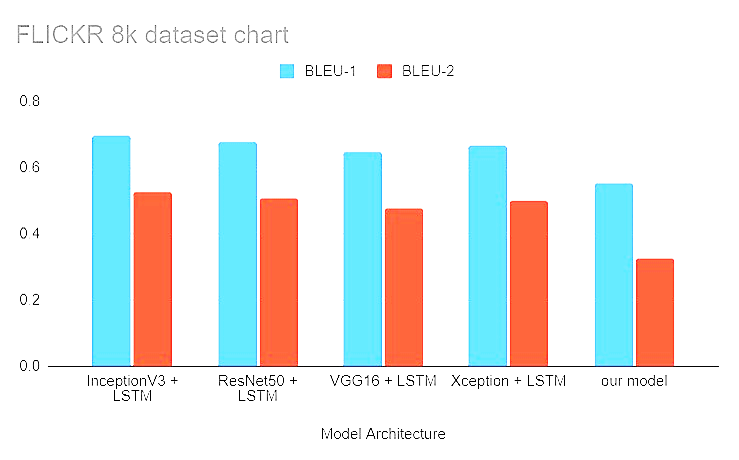
startseq no acute cardiopulmonary process endseq

Figure 8 Actual Caption and Model Predicted Caption of Chest X-ray

# Discussion

In the initial phase of our study, the Flicker8k dataset served as the foundation for image caption generation. A model was constructed, incorporating the VGG-16 and LSTM architectures. To gauge the efficacy of our approach, a comprehensive comparison was undertaken between our model and those documented in relevant papers.

The evaluation criteria centered around BLEU-1 and BLEU-2 scores, which were computed for both the previously established models in the literatures and our model.



BLEU Score

Figure 9 Comparison of BLEU score of different Models

Following the conversion of DICOM images to PNG format, a meticulous evaluation was conducted to assess potential data loss using the Structural Similarity Index (SSI). A custom script was employed for this purpose, facilitating a quantitative comparison between DICOM and PNG formats.

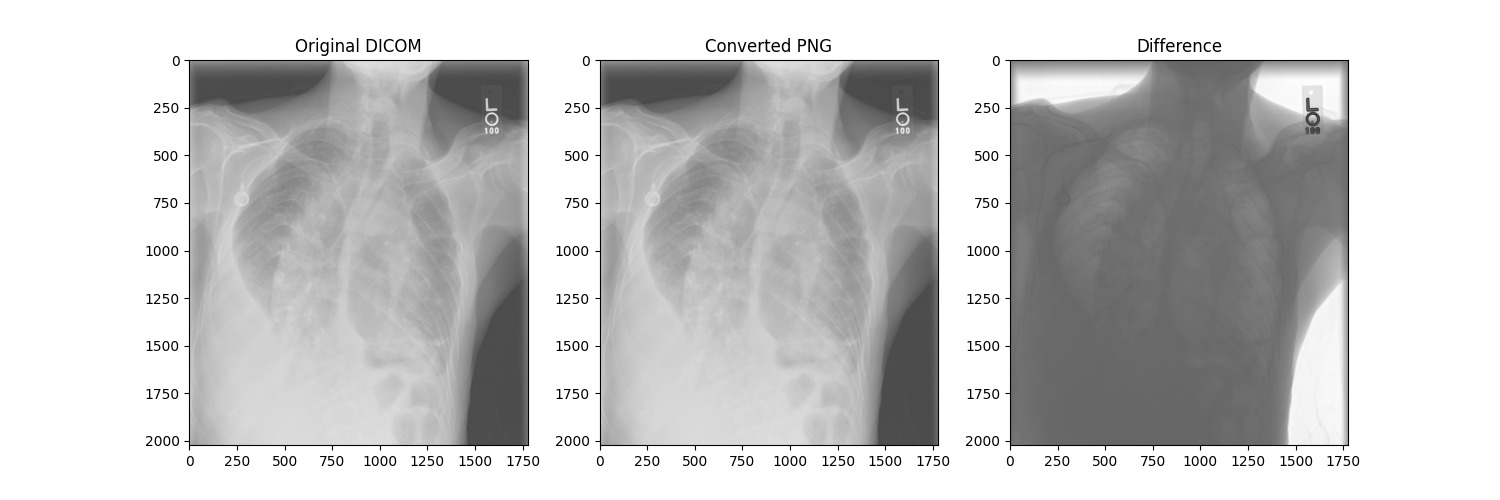


Figure 10 Data loss using SSI

The conversion of medical images from DICOM to PNG format is a crucial step in various medical imaging applications. The assessment of the fidelity of this conversion is vital to ensure that essential information is preserved during the process. In our study, we employed the Structural Similarity Index (SSI) to quantitatively compare DICOM and PNG images, with the obtained SSI value of 0.0891.

To further scrutinize the impact of the conversion, a detailed analysis of pixel intensity distribution was performed. Histogram plots were generated for both DICOM and PNG images, offering insights into the preservation of pixel information during the conversion process.

A graph showing a number of dicom

Description automatically generated

Figure 11 Histogram plot of Pixel Intensity vs Normalized Frequency of DICOM

A graph of a graph showing the size of a number of objects

Description automatically generated

Figure 12 Histogram plot of Pixel Intensity vs Normalized Frequency of PNG

Both graphs show the pixel intensity distribution of an image in DICOM and PNG format.

The x-axis of the graph represents the pixel intensity, which is a measure of the brightness of each pixel in the image. The y-axis represents the normalized frequency, which is the number of pixels that have a particular intensity value, divided by the total number of pixels in the image.

Additionally, throughout the model fitting phase, the loss metric was diligently monitored over the course of 100 epochs. The resulting loss values were plotted against the respective epochs to create a visualization that effectively captured the dynamics of the model training process. This graphical representation serves as a valuable tool for assessing convergence, identifying optimal epochs, and gaining a comprehensive understanding of the training trajectory.

A graph of training loss

Description automatically generated

Figure 13 Loss vs Epoch graph

The x-axis represents the epoch, which is one iteration over the entire training dataset. The y-axis represents the training loss, with lower values indicating better performance. The blue curve shows the training loss decreasing over epochs. This suggests that the model is learning from the training data and improving its performance.

The table 1 provides a comprehensive overview of various models and their performance metrics, in terms of BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores.

Table 1 Comparison of BLEU score of different models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model | Paper | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
| Baseline Model | Radiology Report Generation Using Transformers Conditioned with Non-imaging Data | 0.314 | 0.193 | 0.140 | 0.089 |
| TieNet | ICARC2023 Automated Radiology Report Generation Using Transformers | 0.190 |  |  |  |
| RATCHET | ICARC2023 Automated Radiology Report Generation Using Transformers | 0.232 |  |  |  |
| Baseline  (Densenet-121) | ICARC2023 Automated Radiology Report Generation Using Transformers | 0.224 |  |  |  |
| ResNet + without segmentation | ICARC2023 Automated Radiology Report Generation Using Transformers | 0.251 |  |  |  |
| ResNet + Segmentation | ICARC2023 Automated Radiology report Generation Using Transformers | 0.296 |  |  |  |
| RGRG | Interactive and Explainable Region-guided Radiology Report Generation | 0.373 | 0.249 | 0.175 | 0.126 |
| Our model |  | 0.20207 | 0.10295 | 0.04660 | 0.02648 |

A graph showing different types of models

Description automatically generated

BLEU Score

Figure 14 BLEU score comparison between Models using MIMIC-CXR dataset

The Bar graph shows comparison of different models for chest x ray caption generation with our model. The lines on the graph represent the performance of each model for BLEU-1, BLEU-2, BLEU-3 and BLEU-4. Our model appears to be performing well on most metrics, although it's not necessarily the best on all of them

# Conclusion

In this report, we present a simple yet efficient approach to radiology caption generation utilizing MIMIC-CXR dataset resulting in a BLEU score of 0.202070 which has a notable effort in the context of medical image captioning. The obtained results suggest that the model has demonstrated a moderate proficiency in generating captions, yet it is essential to consider the specific challenges inherent in the medical domain. The significance of our study lies in its contribution to the ongoing discourse on medical report generation. However, it is crucial to acknowledge the limitations of our approach, including potential biases and the need for more diverse and qualitative datasets.

**Future Work**

Our current model, leveraging VGG16 and LSTM on the MIMIC-CXR dataset, has shown promising results in medical image captioning. However, to further enhance its capabilities and applicability, several avenues for future investigation and improvement have been identified.

* The exploration of advanced transformer models, such as Vision Transformer (VIT) and Bidirectional Encoder Representations from Transformers (BERT), holds significant potential. Integrating these models could elevate feature extraction and contextual awareness which might improve BLEU SCORE.
* Exploring evaluation metrics beyond BLEU, such as METEOR, ROUGE, and CIDEr, could provide a more comprehensive result.

# Gantt Chart

# Estimated Budget

Table 2 Cost Estimation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SN** | **Materials** | **Monthly Cost/person (NRs)** | **Months** | **Quantity** | **Amount (NRs)** |
| 1 | Google Collaboratory pro | 1335.78 | 10 | 5 | 66,789/- |
| 2 | Solid-State Drive (SSD) |  |  | 1 | 6,500/- |
| 3 | Miscellaneous | 200 | - | - | 10,000 /- |
| Total | 83,289/- |

# References

[1] P. Rajpurkar *et al.*, “Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists.,” *PLoS Med*, vol. 15, no. 11, p. e1002686, Nov. 2018, doi: 10.1371/journal.pmed.1002686.

[2] A. E. W. Johnson *et al.*, “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” *Sci Data*, vol. 6, no. 1, p. 317, Dec. 2019, doi: 10.1038/s41597-019-0322-0.

[3] D. R. Beddiar, M. Oussalah, and T. Seppanen, “Retrieved Generative Captioning for Medical Images,” in *20th International Conference on Content-based Multimedia Indexing*, New York, NY, USA: ACM, Sep. 2023, pp. 48–54. doi: 10.1145/3617233.3617246.

[4] Y. Sim *et al.*, “Deep Convolutional Neural Network-based Software Improves Radiologist Detection of Malignant Lung Nodules on Chest Radiographs.,” *Radiology*, vol. 294, no. 1, pp. 199–209, Jan. 2020, doi: 10.1148/radiol.2019182465.

[5] I. Allaouzi, M. Ben Ahmed, B. Benamrou, and M. Ouardouz, “Automatic Caption Generation for Medical Images,” in *Proceedings of the 3rd International Conference on Smart City Applications*, New York, NY, USA: ACM, Oct. 2018, pp. 1–6. doi: 10.1145/3286606.3286863.

[6] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, “Generating Radiology Reports via Memory-driven Transformer,” pp. 1439–1449, Oct. 2020.

[7] A. E. W. Johnson *et al.*, “MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,” *Sci Data*, vol. 6, no. 1, p. 317, Dec. 2019, doi: 10.1038/s41597-019-0322-0.

[8] W. D. Bidgood, S. C. Horii, F. W. Prior, and D. E. Van Syckle, “Understanding and using DICOM, the data interchange standard for biomedical imaging.,” *J Am Med Inform Assoc*, vol. 4, no. 3, pp. 199–212, 1997, doi: 10.1136/jamia.1997.0040199.

[9] N. Aksoy, N. Ravikumar, and A. F. Frangi, “Radiology Report Generation Using Transformers Conditioned with Non-imaging Data,” Nov. 2023, doi: 10.1117/12.2653672.

[10] Q. Guan *et al.*, “Deep convolutional neural network VGG-16 model for differential diagnosing of papillary thyroid carcinomas in cytological images: a pilot study.,” *J Cancer*, vol. 10, no. 20, pp. 4876–4882, 2019, doi: 10.7150/jca.28769.

[11] S. Tammina, “Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images,” *International Journal of Scientific and Research Publications (IJSRP)*, vol. 9, no. 10, p. p9420, Oct. 2019, doi: 10.29322/IJSRP.9.10.2019.p9420.

[12] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[13] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” Sep. 2014.

[14] A. Gasimova, “Automated Enriched Medical Concept Generation for Chest X-ray Images,” Oct. 2019, doi: 10.1007/978-3-030-33850-3\_10.

[15] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, “Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN),” Dec. 2014.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics  - ACL ’02*, Morristown, NJ, USA: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

[17] K. Joshi and M. B. Rajarshi, “Modified BLEU for Measuring Performance of a Machine-Translation Software,” 2011, pp. 294–298. doi: 10.1007/978-3-642-19403-0\_51.