**‘Automatic Report Generation from Chest X-Rays’**

**By: Divyarani D**

**Under the guidance of:**

**Mr. Naga Pavan Kumar**

**Dr. Milan Joshi**

**TABLE OF CONTENTS**

1. [Introduction](#_bookmark0) 4
   1. [Background](#_bookmark1) 4
   2. [Problem Statement](#_bookmark2) 5
   3. [Image Captioning](#_bookmark3) 5
      1. [Medical Image Captioning](#_bookmark4) 5
   4. [Proposed Solution](#_bookmark7) 6
2. [Literature Survey](#_bookmark8) 7
   1. [Image Captioning](#_bookmark9) 7
   2. Medical [Image Captioning](#_bookmark10) 7
3. [System Requirements Specification](#_bookmark19) 9
   1. [System Environment](#_bookmark36) 9
      1. [Hardware Requirements](#_bookmark37) 9
      2. [Software Requirements](#_bookmark39) 9
4. [Proposed Methodology](#_bookmark41) 10
   1. [System Architecture](#_bookmark42) 10
   2. Add Token in text data 11
   3. Tokenization 11
   4. Image Features 11
   5. Encoder-Decoder Architecture 11
   6. Model Inference 11
5. [Implementation Details](#_bookmark62) 13
   1. [Software and Tools](#_bookmark63) 13
   2. [Modules](#_bookmark65) 13
      1. [Feature Extraction](#_bookmark67) 13
      2. [Text Generation](#_bookmark68) 14
6. [Intermediate Results and Discussion](#_bookmark74) 15
   1. [Datasets](#_bookmark75) 15
   2. [Exploratory Data Analysis](#_bookmark75) 17
      1. [Radiology Image Analysis](#_bookmark81) 17
      2. [Text Features Analysis](#_bookmark82) 19
   3. [Performance Metrics](#_bookmark83) 20
   4. [Results](#_bookmark88) 21
7. [Conclusions and Future Work](#_bookmark107) 24
   1. [Conclusions](#_bookmark108) 24
   2. [Future Work](#_bookmark109) 24

[References](#_bookmark110) 26

**TABLE OF FIGURES**

[FIGURE 1-3 A Medical Image Captioning Model](#_bookmark44) 6

[FIGURE 3-1 Futuristic View of the Proposed Solution](#_bookmark25) 9

[FIGURE 4-1 Overall System High Level Architecture 1](#_bookmark44)0

[FIGURE 4-2 CNN Architecture 1](#_bookmark44)1

[FIGURE 4-1 Encoder - Decoder Architecture 1](#_bookmark44)2

[FIGURE 6-1 Sample Data Point](#_bookmark45) 15

[FIGURE 6-2 Radiology Text Report Tree View in XML format](#_bookmark49) 16

[FIGURE 6-3 Structured Sample Data Point](#_bookmark52) 17

[FIGURE 6-4 Sample X-Ray Images](#_bookmark53) 17

[FIGURE 6-5 Total Images Present Per Report](#_bookmark54) 18

[FIGURE 6-6 Unique Word Count in Target Feature](#_bookmark60) 19

[FIGURE 6-7 PDF and CDF Word Count Distribution of Target Feature](#_bookmark71) 20

[FIGURE 6-8 Wordcloud](#_bookmark73) 20

[FIGURE 6-9 Sample Output - 1](#_bookmark73) 23

[FIGURE 6-10 Sample Output - 2](#_bookmark73) 23

**LIST OF TABLES**

[TABLE 3-1 Hardware Requirements for the Proposed System](#_bookmark38) 9

[TABLE 3-2 Software Requirements for the Proposed System](#_bookmark40) 9

[TABLE 5-1 Python Libraries and Their Version for the Various Tasks Used in the Project](#_bookmark64) 13

[TABLE 6-1 Quantitative Results and Comparison with Other Models](#_bookmark70) 22

# INTRODUCTION

X-ray report generation is the process of generating textual description from X-ray images. In the world of Deep learning, it is known as Image Captioning. Image captioning uses both Natural Language Processing (NLP) and Computer Vision (CV) to generate the text output. We as humans can look at an image and can describe whatever it is there in an image with an appropriate language. However, when it is X-ray, only an expertise radiologist can describe the X-ray accurately with years of experience. X-Rays are a form of Electromagnetic Radiation that is used for medical imaging. They can be used to spot the fractures, the presence of pneumonia, tuberculosis, blocked blood vessels, cancer such as lung or breast cancer, and many other conditions, bone injuries or any tumors especially in an emergency. Analysis of X-ray reports is an important task of radiologists to recommend the correct diagnosis to the patients. But with deep learning techniques we can predict medical reports by using just medical images.

Clinical imaging captures enormous amounts of information but most radio-logic data are reported in qualitative and subjective terms. In this project, we are tackling the image captioning problem for a data set containing Chest X-ray images with the help of the state of the art deep learning architecture and Natural Language Processing.

A background on the project is provided in Section [1.1](#_bookmark1). A concise problem statement is described in Section [1.2](#_bookmark2), Section [1.3](#_bookmark3) briefly describes the steps involved during image captioning, and Section [1.4](#_bookmark7) introduces the scope and methods used in this study.

# Background

Medical imaging techniques are widely used in hospitals worldwide. The detailed information generated from medical images is necessary for diagnosing illnesses or tracking patients’ progress. However, every image requires a radiologist to carefully examine and write a full-text report to describe the findings. Diagnosing medical images requires an appropriate amount of experience from the radiologists to develop more confident and accurate reports. Nevertheless, many reports conclude with indecisive findings that require the patient to take further tests, including pathology or other advanced imaging methods, as the spectrum of possible cases is too broad. The cost incurred to hire proficient radiologist would be expensive.

Furthermore, a more glaring issue is the amount of time it takes the radiologist to write a full-text report. It would take on an average of 15 min or more based on the radiologist’s degree of experience and also the case of patient, so this would prove very time-consuming when considering the number of cases a radiologist should investigate per day, and in crowded hospitals, regions, and cities, this would be problematic. Radiologists would get over burdened due to manual analysis. It would impact the quality of analysis and also patients will have to wait for longer time and causes delay in treatment. And also human error would occur in case of urgency. There are a chances of loss of patient due to absence of Radiologists in an emergency cases.

These reasons combined provided good motives to research on deep learning models capable of automating report writing.

# Problem Statement

Clinical imaging captures enormous amounts of information but most radio-logic data are reported in qualitative and subjective terms. X-Rays are a form of Electromagnetic Radiation that is used for medical imaging. Analysis of X-ray reports is a very important task of radiologists and pathologists to recommend the correct diagnosis to the patients. The interpretation of these chest x-ray images and their results are communicated in medical reports written by expert physicians. Writing medical reports is usually a time-consuming task for experienced radiologists and pathologists, especially those who practice in regions where doctors’ ratio to patients is relatively low. This task becomes even more tedious for less experienced radiologists, especially those working in rural areas with a poor medical care system. Image captioning is one of the most challenging and important aspects of artificial intelligence. In this project, we are tackling the image captioning problem for a data set containing Chest X-ray images with the help of the state of the art deep learning architecture and Natural Language Processing.

The problem statement here is to predict the ‘Findings’ from the given chest X-Ray images. These images are of two types: Frontal and Lateral view of the chest. With these two types of images as input we need to find the findings for given X-Ray. To resolve this problem statement, we will be building a predictive model which involves both image and text processing to build a deep learning model.

# 1.3 Image Captioning

Image captioning is a research field that focuses on methods for automatically generating text to explain the contents of an image. This area involves the convergence of computer vision to understand images and natural language processing to generate word sequences. Image captioning offers various applications, such as text-based image retrieval, related keyword assignment, human–robot interactions, and support for visually impaired people. Several methods have been developed for image captioning, including retrieval-based, template-based, and deep learning-based methods.

Most deep learning studies have used an encoder–decoder architecture with an attention mechanism. The encoder transforms the image into a feature vector, whereas the decoder converts the feature vector into word sequences. We propose a new model that uses a convolutional neural network (CNN) as the encoder and hierarchical long short-term memory (LSTM) or a transformer as the decoder. In recent years, deep learning-based methods have gained popularity in the image captioning field.

## Medical Image Captioning

Medical artificial intelligence applications are undergoing rapid expansion, from reading images to diagnosing diseases. Image captioning has also been applied in the medical field (Fig. 1). As chest x-rays are the most common types of medical images, and are important

for screening and diagnosis, we conduct experiments using chest x-ray images. The number of medical images increases continually, which imposes a tremendous burden on doctors in terms of reading and writing reports. Medical image captioning can assist doctors by accelerating the reporting process and reducing their workload. However, little progress has been made in medical captioning compared to image captioning in other fields, and multiple factors have limited the performance of image captioning for chest x-rays.

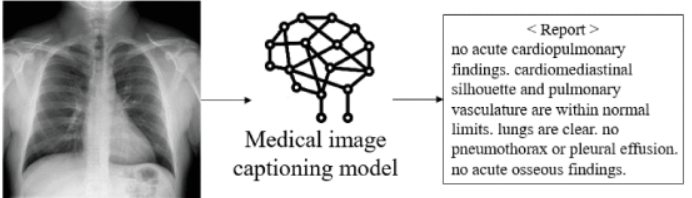


Fig 1. A medical image captioning model generates a draft report of the corresponding medical image.

# Proposed Solution

The methodology for medical image captioning proposed in this thesis works in several stages, and the problem is divided into two parts.

* First is the Encoder part for feature extraction from image data, using transfer learning method we extract information from x-ray images.
* Second part is Decoder, we will give those image information to the decoder model and get a medical report

In simple words, we have to extract bottleneck features from images using CNN from scratch or using transfer learning. The latter approach is preferable, as we have less data. Then use these extracted features to predict the captions using LSTM or GRU. The output would be a sequence of words.

# LITERATURE SURVEY

# 2.1 Image Captioning:

Image captioning is the problem of generating text to describe the input image. The problem gained popularity with the rise of deep learning techniques, as many works adopted the CNN-RNN architecture (Vinyals, Toshev, Bengio, & Erhan, 2015). Following the success of the attention concept introduced in Ref. (Dzmitry Bahdanau, 2015), more works started adding visual attention to their CNN-RNN architecture like in Refs.[(Xu, 2015), (Jeffrey Donahue, 2015)]. Moreover, works like (Quanzeng You, 2016) added semantic attention with the visual attention. Furthermore, hierarchical recurrent models were introduced to solve the problems that occur when generating long captions like in Ref. (Jonathan Krause, 2017). Few papers attempted to use transformer-based models (Jacob Devlin, 2019) as decoders in the image captioning domain, although they are faster to train and produce state-of-the-art results in most NLP problems. One of the papers that attempted to condition a GPT2 model was (Ziegler ZM, 2019) by adding new key and value weights in the self-attention module to be projected into the decoder’s self-attention space.

# 2.2 Medical Image Captioning:

There have been several works attempting to generate medical reports from their corresponding images. The first of which to use a CNN-RNN approach was (Hoo-Chang Shin, 2019) to predict tags used to form a structured report for chest X-ray images from the IU-X-ray dataset. The first work to use attention on the medical image was (Zizhao Zhang, 2017), long-short-term-memory cells (LSTM) (S. Hochreiter and J. Schmidhuber, 1997) were used to produce a report of five sentences on a private dataset of pathology images.

A framework to generate natural reports for the Chest-Xray14 dataset (Xiaosong Wang, 2017), using private reports, was introduced in Ref. (Xiaosong Wang Y. P., 2018), using a non-hierarchical CNN-LSTM architecture and attention on semantic and visual features. Natural reports on the IU-X-ray dataset were generated in Ref. (Jing B, 2017) by getting visual features from a VGG network pre-trained on ImageNet and introducing the concept of co-attention which is a combined attention mechanism on both the visual and predicted tags embeddings. The co-attention output is then passed to a hierarchical LSTM, one for sentences and one for words, to generate the reports. The multi-view information of the IU-X-ray dataset was leveraged in Ref. (Yuan, 2019) by getting the visual features and tags’ prediction from the front and side images of the patient using a Resnet152 trained on the CheXpert dataset then using hierarchical LSTMs like that of (Jing B, 2017) to generate the reports.

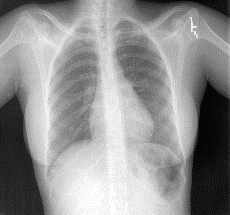
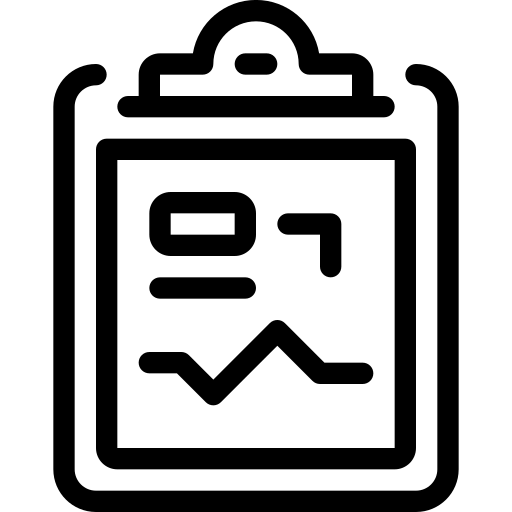
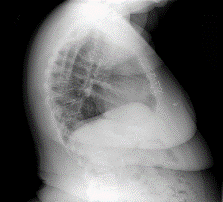
Knowledge graphs with prior knowledge on chest findings were utilized in Ref. (Zhang, 2020), the graph node features are extracted from the IU-X-ray images using CheXNET models, and hierarchical LSTMs, with attention over the graph, were used to generate the full report. A cross-modal retrieval method was used in Ref. (Ni J, 2020) to retrieve abnormal findings given images from the IU-X-ray dataset by learning visual-semantic embeddings on the images and the reports, which are then used to measure the similarity between a new image and the existing findings to retrieve the closest report.

A feature pyramid network was used in Ref. (Syeda-Mahmood T, 2020) to concatenate features from different models, pre-trained on ImageNet, given chest X-ray images to detect labels that depict the image; these labels are used to create a vector that describes the major findings in the image. The semantic distance between the image’s vector and the encoded reports in the IU-X-ray dataset was used to retrieve the most relevant report and remove parts from the report with no evidence. Few works used transformers as decoders in radiology report generation like (Xiong Y, 2019), which used a custom transformer as the decoder with a bottom-up region detector and a top-down visual encoder to generate reports for the IU-X-ray dataset. A custom transformer was also used in Ref. (Chen Z, 2020) with an extra relational memory unit to generate reports for the IU-X-ray dataset. The front and side chest images are passed through a visual extractor to get a sequence of visual features, from pre-trained models like VGG and ResNet that will be passed to the encoder and decoder to generate the reports. We will also be conditioning on visual and semantic features.

# SYSTEM REQUIREMENTS SPECIFICATION

This chapter describes the system to be designed for automatic report generation from chest X-Rays.

The main intention behind developing the system is an attempt to solve the problem of time consumption by radiologist, with specialized training to manually evaluate x-rays and note their findings in a radiology report. This manual evaluation is time-consuming and providing an automated solution for this task would help streamline the clinical workflow and improve the quality of care. This study proposes to generate text report from X-ray images. The futuristic visualization of the system is illustrated in [Figure 3-1](#_bookmark25).



**Designed Model**

**Figure 3-1 Futuristic view of the proposed solution;**

# 3.1 System Environment

For the initial phase, the entire system will be run on a laptop computer.

## Hardware Requirements

The minimum requirements to run the software developed as part of this study is summarized in [Table 3-](#_bookmark38)2.

|  |  |
| --- | --- |
| **Processor** | Intel Core-i3, 2GHz |
| **No. of CPUs** | 4 (2 physical CPUs, with hyper-threading enabled) |
| **RAM** | 8GB |
| **Cache** | L1 Cache: 32KB, L2 Cache: 256KB, L3 Cache: 3MB |

**Table 3-1 Hardware requirements for the proposed system**

## Software Requirements

The platform and software frameworks used are tabulated in [Table 3-](#_bookmark40)3.

|  |  |
| --- | --- |
| **Operating System** | Windows 10, 64bit |
| **Development Language** | Python 3.6 |
| **Deep Learning Framework** | Tensorflow (Keras Backend) |
| **Image Processing utilities** | OpenCV |
| **Third-party software and libraries** | Numpy SciKit Matplotlib, |

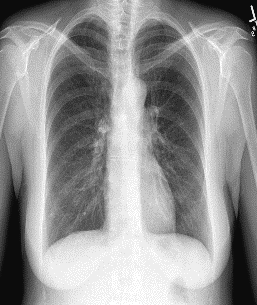
**Table 3-2 Software requirements for the proposed system**

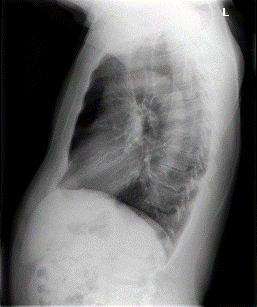
# PROPOSED METHODOLOGY

The study presented in this thesis describes a novel method of image captioning, which is intended to be used for text report generation.

The architecture of the system is described in [4.1](#_bookmark42). A more detailed view of the system is presented in [4.2,](#_bookmark43) followed by a detailed description of the software design in [4.3](#_bookmark50).

Frontal X-Ray image





**Decoder**

**Encoder**

CheXNet

[Feature Extractor]

Lateral X-Ray image

**Figure 4-1 System High level Architecture**

# 4.1 System Architecture

The overall system high level architecture is illustrated in [Figure 4-1.](#_bookmark44) The model architecture is composed of 3 sub-models,

1. Feature Extractor
2. Encoder
3. Decoder

The last 2 sub models represent the transformer. A sequence-to-sequence model is a deep learning model that takes a sequence of items (in our case, features of an image) and outputs another sequence of items (reports).

The encoder processes each item in the input sequence, it compiles the information that captures into a vector called the context. After processing the entire input sequence, the encoder sends the context over to the decoder, which begins producing the output sequence item by item. The steps followed are following:

# 4.2 Add Token in text data:

After creating new data points from existing data points, we will add <start> and <end> token into text data and prepare decoder input and output. The start and End tokens are special tokens added at the Start of the sentence and End of the sentence respectively to let the model learn the beginning and ending of the sentences.

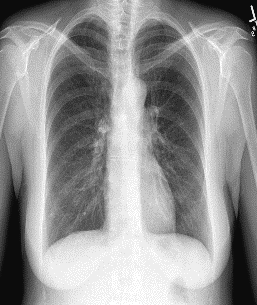
**4.3 Tokenization:**

Machines only understand numerical value and we cannot feed text data into deep learning and machine learning models. We will convert text data into numerical data using Tokenizer. The Tensorflow deep learning library provides tools to perform this operation.

**4.4 Image features:**

The pre-trained CheXNET model is used to extract image features and converted to feature vector using transfer learning technique. CheXNET Model is a Denset121 layered model which is trained on 112,120 number of chest x-ray images for the classification of 14 diseases. We can load the weights of that model and pass the image through that model. The top layer will be ignored.

As there are two X-Rays corresponding to each patient, so each image is pre-processed according to the input of the DenseNet-121 model and the model’s predictions for both the images is concatenated at the end. DenseNet -121 is the preferred pre-trained model for feature extraction (AB Amjoud, 2021), because it does not need a great number of parameters, which means we do not spend a significant time learning redundant feature maps.



a1

a2

.

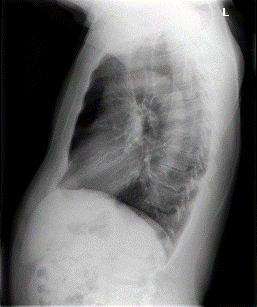
.

.

a1024

CheXNet

[Pre – trained Denset 121 layered model]



**Figure 4-2 CNN Architecture**

**4.5 Encoder-Decoder Architecture:**

The concatenated image tensors are passed to encoder, we feed image features into a dense layer having 512 neurons and then a dropout layer is added for tuning. In the decoder part, there is an embedding layer, dropout layer and LSTM layer. The input sequence i.e. encoder\_output is passed to the embedding layer. Then we pass this layer output to the dropout layer and finally feed to LSTM.

LSTM layer is Long Short-Term Memory networks — are a special kind of RNN, capable of learning long-term dependencies. The outputs of encoder and decoder are then added using the Add layer of Keras. This output has been passed to a Time distributed dense layer. The time distributed dense layer is applied at the end because the output is a sequential output and it should be applied to every temporal slice of output.

**Decoder**

**Encoder**

Vector

(1624)

Concatenate

(1024 + 600 =1624)

lungs

1 X 1024

clear

are

**<end>>**

LSTM

(256)

LSTM

(256)

LSTM

(256)

Flatten (600)

LSTM

(256)

CheXNet

[Pre – trained model]

3 X 200

Embedding

Layer

Embedding

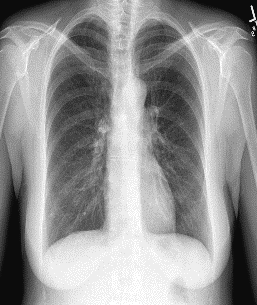
Layer

Embedding

Layer

Embedding

Layer



Embedding layer

<start>

Indication

[Ex: Negative TB Test]

**Figure 4-3 Encoder - Decoder Architecture**

**4.6 Model Inference:**

In the inference stage, the argmax based Greedy search is used to find the output sentence. Greedy search is a vanilla implementation for generation of output which is selecting a single word with maximum probability from the entire vocabulary.

**5 IMPLEMENTATION DETAILS**

This chapter describes the details of the implementation for the X-ray report generation.

The software and tools used are described in [5.1](#_bookmark63). Different modules used in the project are described in [5.2.](#_bookmark65) The procedure used to find the best layer for feature extraction is described in [5.3.](#_bookmark69)

# 5.1 Software and Tools

The project runs on a Laptop that has an Intel Core i3 Processor, and is configured with Windows10, 64bit. The system has 8GB RAM. The entire project has been developed in Python 3.6. The tools and libraries used in python for various tasks are summarized in [Table 5-1.](#_bookmark64)

|  |  |  |
| --- | --- | --- |
| **Task** | **Library/Framework** | **Version** |
| Reading Images, Resizing Images | OpenCV | 4.1 |
| Classifiers for Feature Extraction | OpenCV | 4.1 |
| Text Generation | Nltk | 3.9 |

**Table 5-1 Python libraries and their version for the various tasks used in the project**

# Modules

## Feature Extraction

Multiple choices of models exist for feature extraction. For our report generation architecture we adopt the transformer model introduced by Vaswani et al. (2017) for neural machine translation (NMT). The transformer is an encoder-decoder model where the encoder and decoder both consist of stacked layers of self-attention and position-wise feed-forward neural networks. We refer the reader to Vaswani et al. (2017) for a detailed description of the model. The primary difference between our setting and that of Vaswani et al. (2017) is that instead of translating a source language to a target language, we must translate an image into a corresponding textual annotation. Therefore instead of operating on word embeddings, our encoder operates directly on image features.

The model considered here is CheXNet model. CheXNet, is a 121-layer convolutional neural network trained on ChestX-ray14, currently the largest publicly available chest X-ray dataset, containing over 100,000 frontal-view X-ray images with 14 diseases. Four practicing academic radiologists annotate a test set, on which we compare the performance of CheXNet to that of radiologists. We find that CheXNet exceeds average radiologist performance on the F1 metric. We extend CheXNet to detect all 14 diseases in ChestX-ray14 and achieve state of the art results on all 14 diseases.

## Text Generation

NLTK is a leading platform for building Python programs to work with human language data. It provides easy-to-use interfaces to [over 50 corpora and lexical resources](https://www.nltk.org/nltk_data/) such as WordNet, along with a suite of text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, wrappers for industrial-strength NLP libraries, and an active [discussion forum](https://groups.google.com/group/nltk-users).

NLTK is suitable for linguists, engineers, students, educators, researchers, and industry users alike. NLTK is available for Windows, Mac OS X, and Linux. Best of all, NLTK is a free, open source, community-driven project.

# 6 INTERMEDIATE RESULTS AND DISCUSSION

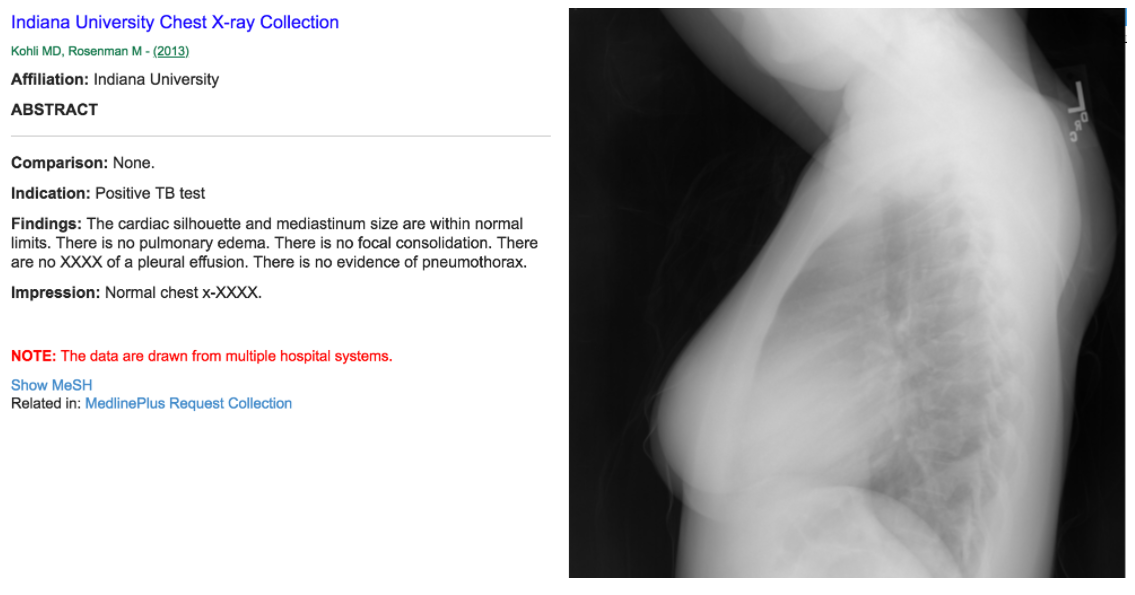
This chapter presents the intermediate results found for the procedures elaborated in the previous section, and also describes the datasets used for the project, exploratory data analysis and the evaluation metrics considered. The results presented here are intended only as a baseline, and will be further worked upon to improve them in the next phase of the project.

# Datasets

The dataset used in this project is from public and open source database. The dataset is widely used by researchers and it is “Indiana University X-rays” dataset that contains both X-Ray image and radiology text reports. It comprises of 7471 images in .png format which contains front view and lateral view of each patient’s chest and there are about 3955 patient’s text reports available in XML format.

Each X-ray image has been provided with four fields such as Comparison, Indication, Findings and Impressions. Impressions provide clear descriptions of the salient entities and events. The finding field gives maximum information present in the images. The goal of this project is to predict the findings of the medical report attached to the images.

Sample Data point is shown in the below figure.



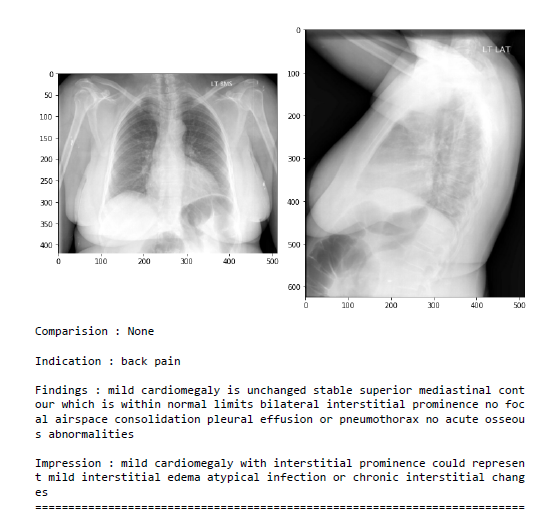
**Figure 6-1 Sample Data Point**



**Figure 6-2 Radiology Text Report tree view in XML format**

The above figure shows the radiology text report tree view in XML format. This XML file with raw data is parsed and structured as data points. Then these data points are stored as csv files for future model requirements. This XML file has lot of information related to patients such as image\_id, text captions like — comparison, indication, findings, impression etc. The findings feature is extracted from these files and consider them as reports because they are more useful for the medical report. Image\_id also needs to be extracted from these files to get the corresponding X-rays to each report.

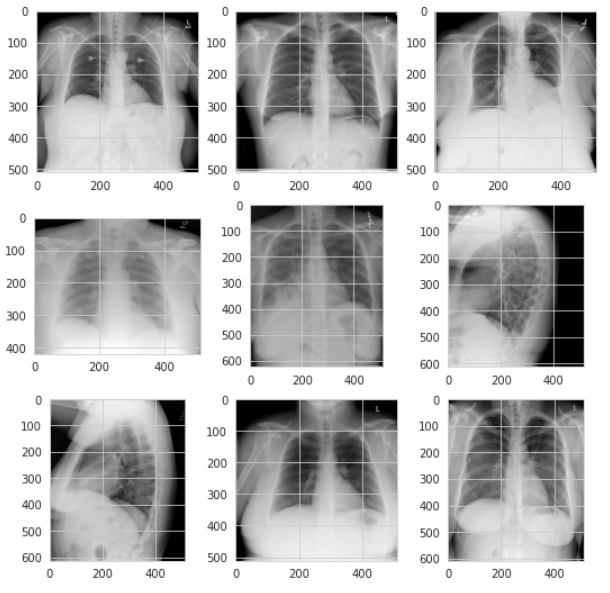
The below figure shows a structured sample data point after extracting data from XML:



**Figure 6-3 Structured Sample Data Point**

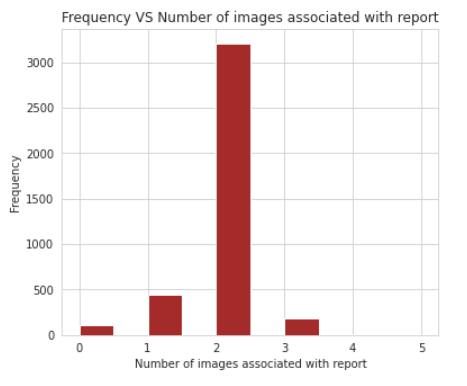
# Exploratory data analysis:

## Radiology Image analysis:

Below figure shows the 9 sample X-rays of different size and they are in both front and lateral view.

**Figure 6-4 Sample X-Ray Images**

The below figure shows the total images present per data point or report. Here more than 3000 reports have 2 images, minimum image count per report is 1 and maximum image count per report is 5.



**Figure 6-5 Total Images Present Per Data Point or Report**

There are only two image types — Front and Lateral, but each patient has multiple x-rays associated with them. The maximum number of images associated with a report can be 5 while the minimum is 0. The highest frequency of being associated with a report are 2 images.

We have more than 2 images and in some cases less than 2 images are associated with each data point. The data points which are having 1,3,4,5 images needs to be structured using the following approach that could help in this case.

Limiting the data point to 2 images per data point, if we have 5 images, it’s 4+1 (all image + last image) so make it as 4 data points, Here last image should be Lateral if we have frontal as remaining images. Converting multiple images into two images. For that following implementation has been done using following steps:

1. If we have 5 images then total 4 data points created

1st image + 5th image

2nd image + 5th image

3rd image + 5th image

4th image + 5th image

2. If we have 4 images then total 3 data points created

1st image + 4th image

2nd image + 4th image

3rd image + 4th image

3. If we have 3 images then total 2 data points created

1st image + 3rd image

2nd image + 3rd image

4. If we have 2 images then it is according to our requirement

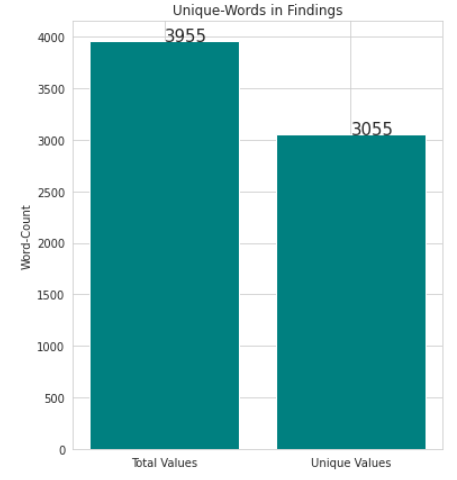
5. At last, If we have 1 image, we just replicate it and make it 2.

Code for the above explained data structuring.

## Text features analysis:

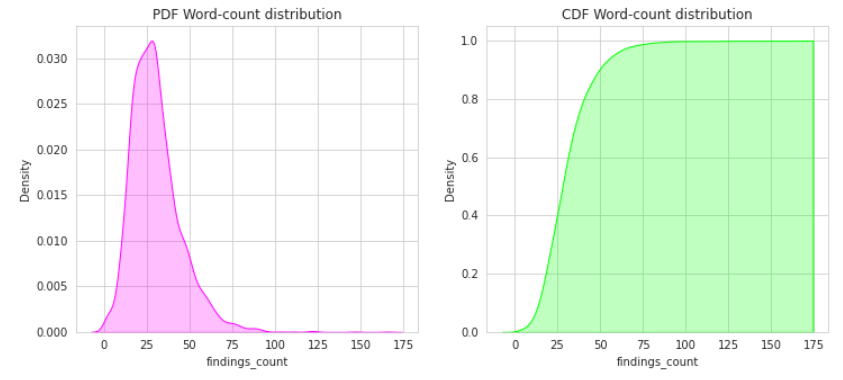
For text analysis, finding feature is considered as it is the target feature and it provides maximum information of x-ray images.

The below figure shows the unique words count, there are total of 3955 entries and 3055 are unique in all of them. They are never repeated.



**Figure 6-6 Unique word counts in Target Feature**

The below figure shows the PDF and CDF for word count distribution of Findings feature. 50% data have less than 20 words per findings, 99% data have less than 50 words per findings.



**Figure 6-7 PDF and CDF Word Count Distribution of Target Feature**

The below figure shows the wordcloud: pleural, effusion, limits, within, normal, lungs, cardiomediastinal silhouette are the highlighted words i.e. these are important words.

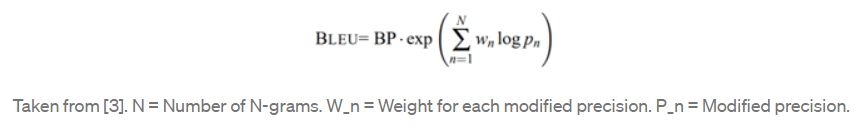


**Figure 6-8 WordCloud**

# Performance Metrics

The performance metrics used to evaluate the effectiveness of the system proposed in this thesis is – BLEU score. It is a conventional natural language generation (NLG) metrics scores used for evaluation metrics and quantitative evaluation. This metric is commonly used in the field of Image Captioning. They enable to assess the quality of the generated reports by comparing N-gram’s occurrence with the ground truth. We note that the use of these metrics in the medical and clinical field is slightly limited. For the BLEU metric, we employ BLEU-1, BLEU-2, BLEU-3 and BLEU-4.

BLEU stands for Bilingual Evaluation Understudy Score. It can be calculated using the NLTK Python library, which provides a function sentence\_bleu() that allows for evaluation of generated text against a reference. Essentially, BLEU scoring is performed by using sampled reference translations. The algorithm looks at typical sentences that contain the same words as those in the real caption. Then, it calculates how many of those words are in the predicted caption.

BLEU scoring is a unique method for this that uses modified n-gram precision. “An n-gram is a sequence of words occurring within a given window where n represents the window size” [3]. In order to calculate the BLEU score we need 2 pieces: the precision score and the Brevity Penalty (BP). The following equation shows the calculation of BLEU score:

A perfect match results in a score of 1.0, whereas a perfect mismatch results in a score of 0.0.

The score was developed for evaluating the predictions made by automatic machine translation systems. It is not perfect, but does offer 5 compelling benefits:

* It is quick and inexpensive to calculate.
* It is easy to understand.
* It is language independent.
* It correlates highly with human evaluation.
* It has been widely adopted.

The BLEU score was proposed by (Kishore Papineni, 2002). in their 2002 paper “[BLEU: a Method for Automatic Evaluation of Machine Translation](http://www.aclweb.org/anthology/P02-1040.pdf)“.

# Results:

The below table shows the results of our proposed model as well as other existing model.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **BLEU-1** | **BLEU-2** | **BLEU-3** | **BLEU-4** |
| Our proposed Encoder-decoder model | 0.541 | 0.392 | 0.281 | 0.208 |
| (AB Amjoud, 2021) proposed model | 0.479 | 0.359 | 0.219 | 0.160 |
| (Lovelace J, 2020) | 0.415 | 0.272 | 0.193 | 0.146 |
| (Boag W, 2020) | 0.305 | 0.201 | 0.137 | 0.092 |

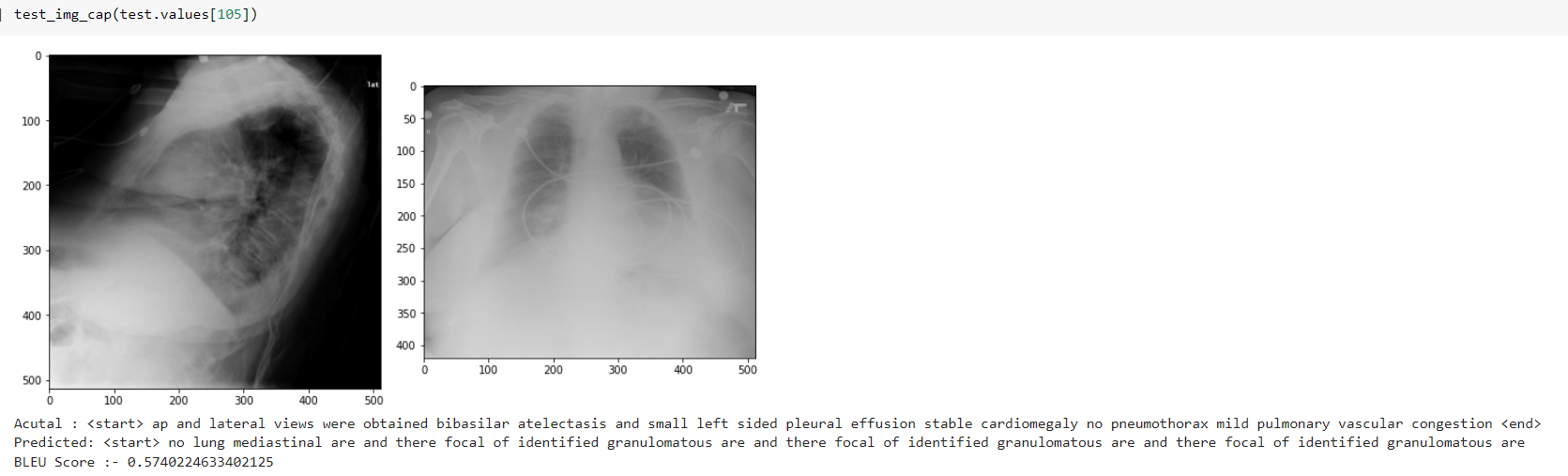
**Table 6-1. Quantitative results and comparison of our model with other existing models on the IU X-Ray dataset test set**

**using NLG metrics.**

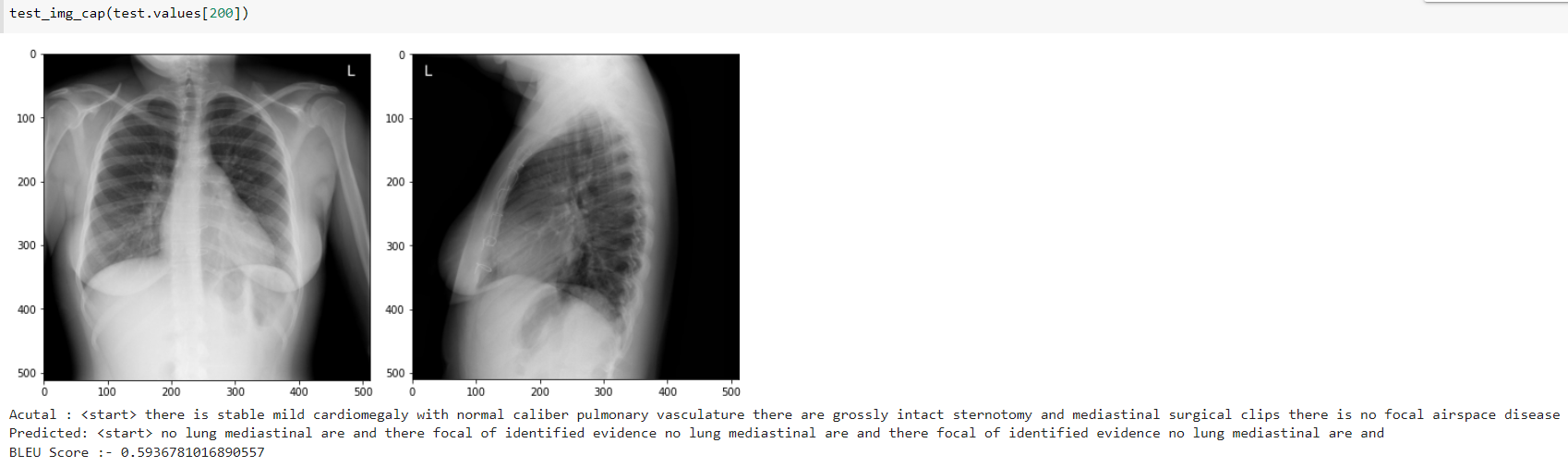
Through table 6-1, we observe that our Encoder-decoder (transformer) model is able to outperform the state of the art models for generating detailed findings chest x-rays reports on UI X-ray test set. Particularly for the BLEU-1, BLEU-2, where the scores were 0.541 and 0.392 respectively. We have also noticed that our improvement of the model has yielded interesting results. On the other hand, we believe that more work needs to be undertaken to ensure better quality reports are generated. One of the challenges we encountered during the model’s development is the small size of the dataset. This has to lead to creating a small corpus, causing a scattering during the generation of new sentences not seen during the training.

We consider that expanding the dataset size by adding proper samples and setting the transform parameters by adding or removing some modules can enhance the learning process and lead to the best results.

The below figure shows the actual text and predicted text for the given X-Ray images with BLEU score.



**Figure 6-9 Sample output - 1**



**Figure 6-10 Sample output - 2**

# CONCLUSIONS AND FUTURE WORK

# Conclusions

Although the project is still in its initial stage, some conclusions can be drawn from the work carried out so far. The overall system architecture, along with each component, the modules and their functionalities are clearly defined. Each have a specific role to play in the process, and have been designed to be optimal in their functionalities.

This study has proposed a chest x-ray image captioning model that generates draft reports of images. We see from our results that it is feasible to train a model to a certain degree of accuracy to write captions for X-rays. By using a variety of methods such as tokenization, recurrent neural networks, we converge on these captions. In this project we have stepped through the processes required to obtain these results, from data collection and exploratory analysis, to NLP and computer vision techniques. To improve on the process within computational constraints, we could increase the size of our dataset, tune our hyper parameters better, or find unique ways of incorporating meta features into our captioning. Finally, we could have considered other ways to test the accuracy of our results other than just the BLEU score. For example, perplexity scoring is also another method used in language modelling for intrinsic evaluation.

# Future Work

Although there are still some open questions, a few of the targeted goals have been met. The model implementation is nearing completion, and the work for next phase i.e. to implement model using attention mechanism can begin. The following items are being targeted for the subsequent stage:

7.2.1 Use of Bahdanau additive attention mechanism

7.2.2 Use of Bidirectional GRU at decoder stage

7.2.3 Improving BLEU score

7.2.4 Beam search method to find the output sentences

# ACKNOWLEDGEMENTS

I would like to appreciate Dr. Milan Joshi and Mr. Naga Pavan Kumar, PES University Bengaluru, for their time, patience and valuable inputs which have helped make the project better.

# References

AB Amjoud, M. A. (2021). Automatic Generation of Chest X-ray Reports. *2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS)* (pp. 1-5). Fez, Morocco: IEEE.

Chen Z, S. Y. (2020). Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.

Dzmitry Bahdanau, K. C. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR* (p. 1409.0473). CoRR.

Hoo-Chang Shin, K. R.-F. (2019). Learning to Read Chest X-Rays: Recurrent Neural Cascade Model for Automated Image Annotation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2497-2506). Las Vegas, NV, USA: IEEE Computer Society.

Jacob Devlin, M.-W. C. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: human Language technologies, NAACL-HLT* (pp. 4171–86). Minneapolis, MN, USA: ). Association for Computational Linguistics.

Jeffrey Donahue, L. A. (2015). Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2625-2634). IEEE.

Jing B, X. P. (2017). On the automatic generation of medical imaging reports. *arXiv.* arXiv.

Jonathan Krause, J. J.-F. (2017). A Hierarchical Approach for Generating Descriptive Image Paragraphs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 317-325). Honolulu, HI,USA: IEEE Computer Society.

Kishore Papineni, S. R.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. *40th Annual Meeting of the Association for Computational Linguistics (ACL),*, (pp. 311-318). Philadelphia.

Ni J, H. C. (2020). Learning visual-semantic embeddings for reporting abnormal findings on chest x-rays. *arXiv.* arXiv.

Quanzeng You, H. J. (2016). Image Captioning With Semantic Attention. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4651-4659). Las Vegas, NV, USA: IEEE.

S. Hochreiter and J. Schmidhuber. (1997). Long Short-Term Memory. *Neural Computation* (pp. 1735 - 1780). MIT Press.

Syeda-Mahmood T, W. K. (2020). Chest X-ray report generation through fine-grained label learning. *InInternational Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 561-571). Springer, Cham: InInternational Conference on Medical Image Computing and Computer-Assisted Intervention.

Xiaosong Wang, Y. P. (2017). ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2097-2106). IEEE.

Xiaosong Wang, Y. P. (2018). TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9049-9058). IEEE.

Xiong Y, D. B. (2019). Reinforced transformer for medical image captioning. *InInternational Workshop on Machine Learning in Medical Imaging*, (pp. 673-680). Springer, Cham.

Xu, K. a. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *Proceedings of the 32nd International Conference on Machine Learning* (pp. 2048--2057). Lille, France: PMLR.

Yuan, J. L. (2019). Automatic Radiology Report Generation Based on Multi-view Image Fusion and Medical Concept Enrichment. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019.* Medical Image Computing and Computer Assisted Intervention – MICCAI 2019.

Zhang, Y. W. (2020). When Radiology Report Generation Meets Knowledge Graph. *Proceedings of the AAAI Conference on Artificial Intelligence.* Proceedings of the AAAI Conference on Artificial Intelligence.

Ziegler ZM, M.-K. L. (2019). Encoder-agnostic adaptation for conditional language generation. (p. 1908.06938). arXiv.

Zizhao Zhang, Y. X. (2017). MDNet: A Semantically and Visually Interpretable Medical Image Diagnosis Network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6428-6436). IEEE.