

Analysis of Multimodal Representation Learning Across
Medical Images and Reports Generation Using Multiple
Vision and Language Pre-Trained Models



Author

Ahmad Hassan

00000318671

Supervisor

Dr. Muhammad Usman Akram

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD
APRIL, 2022

Analysis of Multimodal Representation Learning Across
Medical Images and Reports Generation Using Multiple
Vision and Language Pre-Trained Models

Author

Ahmad Hassan

00000318671

A thesis submitted in partial fulfillment of the requirements for the degree of
MS Computer Engineering

Thesis Supervisor

Dr. Muhammad Usman Akram

Thesis Supervisor's Signature: _____

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,
ISLAMABAD
APRIL, 2022

Declaration

I certify that this research work titled “*Analysis of Multimodal Representation Learning Across Medical Images and Reports Generation Using Multiple Vision and Language Pre-Trained Models*” is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged/referred.

Signature of Student

Ahmad Hassan

00000318671

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student

Ahmad Hassan

00000318671

Signature of Supervisor

Dr. Muhammad Usman Akram

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST College of E&ME. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of E&ME, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the College of E&ME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of E&ME, Rawalpindi.

Acknowledgements

All praise and glory to Almighty Allah (the most glorified, the highest) who gave me the courage, patience, knowledge, and ability to carry out this work and to persevere and complete it satisfactorily. Undoubtedly, HE eased my way and without HIS blessings I can achieve nothing.

I would like to express my sincere gratitude to my advisor Dr. Muhammad Usman Akram for boosting my morale and for his continual assistance, motivation, dedication, and invaluable guidance in my quest for knowledge. I am blessed to have such a co-operative advisor and kind mentor for my research.

Along with my advisor, I would like to acknowledge my entire thesis committee: Dr. Arslan Shaukat and Dr. Sajid Gul Khawaja for their cooperation and prudent suggestions.

My acknowledgment would be incomplete without thanking the biggest source of my strength, my family. I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout in every department of my life. I also want to say thanks to my supportive Sisters and Brother who were with me through my thick and thin.

Finally, I would like to express my gratitude to all my friends and the individuals who have encouraged and supported me through this entire period.

*Dedicated to my exceptional parents: **Sajid Hussain & Ishrat Parveen**, supportive Sisters and Brother whose tremendous support and cooperation led me to this accomplishment. At the end, this thesis is dedicated to all those who believe in the richness of learning*

Abstract

In the medical field, medical images are the visual representation of the organs and their functions. The medical images are used for finding the medical problems present in human being and highly trained professionals interpret these images into reports with a lot of time required in a day, as each report take around 4-6 minutes. Other than that, the reports were used to highlight the diseases, and writing a summarized report nowadays is very important for other inexperienced persons in understanding so that it can help them in better treatment. The automated summarization of radiology reports has tremendous potential. This operationally improve the diagnosis process of diseases. An image-text joint embedding extraction from chest x-rays and radiology reports, in producing summarized reports along with findings/tags will significantly reduce the workload of doctors and help them in treating patients. Because of the sensitivity of the process, the existing methods/techniques are not adequately accurate and limitation of data effects in training the models. Therefore, the generation of a summarization radiology report is an exceedingly difficult task. A novel approach is proposed to address this issue. In this approach, use pre-trained vision-and-language models like VisualBERT, UNITER, and LXMERT to learn multimodal representation from chest x-rays or radiographs and reports. The pre-trained model classified the findings/tags in the chest x-rays (CXR) using Gated Recurrent Units as a decoder to generate a summarized report based on them. The Chest X-rays images and reports data are publicly available Indiana University dataset. There are also different methods for automatic report summarization and findings/tags classification from CNN-RNN-based models but mostly based on text or image only with less accuracy. The image-text joint embedding using the pre-trained models helps in more accurate report generation and improve performance in thoracic findings and summarized report generation task. Experimental results obtained by utilizing Indiana University (IU) CXR dataset showed that the suggested model attains the current state-of-the-art efficiency as compared to other existing solutions to the baseline. As evaluation metrics, BLEU and ROUGE have been applied along with AUC for findings/tags. The experiments are performed in multiple ways and the accuracy achieved in diseases findings is about 98%, BLEU score of 0.35 and ROUGE score of 0.65 for the summarized radiology report.

Key Words: *Bottom-Up Top-Down (BUTD), Chest X-rays Radiology (CXR), Convolution Neural Network, Deep Learning, Gated Recurrent Units (GRU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Visual Question Answering (VQA).*

Table of Contents

DECLARATION	I
LANGUAGE CORRECTNESS CERTIFICATE	II
COPYRIGHT STATEMENT	III
ACKNOWLEDGEMENTS	IV
ABSTRACT	VI
TABLE OF CONTENTS	VII
LIST OF FIGURES	IX
LIST OF TABLES	X
CHAPTER 1: INTRODUCTION	1
1.1 MOTIVATION	3
1.2 PROBLEM STATEMENT	3
1.3 AIMS AND OBJECTIVES	3
1.4 STRUCTURE OF THESIS	4
CHAPTER 2: CHEST ORGANS ANATOMY & THORACIC FINDINGS	5
2.3 STRUCTURE OF LUNGS	7
2.4 IMAGING TECHNIQUES TO ANALYZE LUNGS	8
2.5 THORACIC FINDINGS	10
2.5.1 Atelectasis	10
2.5.2 Infiltration & Consolidation	11
2.5.3 Cardiomegaly	12
2.5.4 Nodules & Masses	13
2.5.5 Pleural Effusion	14
2.5.6 Emphysema	16
CHAPTER 3: LITERATURE REVIEW	17
CHAPTER 4: METHODOLOGY	25
4.1 MATHEMATICAL MODEL	25
4.1.1 Visual Feature Embedding	25
4.1.2 Text Embedding	26
4.1.3 Joint Embedding	26
4.2 BOTTOM-UP TOP-DOWN APPROACH	26
4.3 VISUALBERT	27
4.3.1 Training	28
4.4 LXMERT	29
4.4.1 Embedding	30
4.4.1.1 Word Embedding	31
4.4.1.2 Image Embedding	31
4.4.2 Encoders	32
4.4.2.1 Language and Object Relation Encoder	32
4.4.2.2 Cross Modality Encoder	32

4.4.3 <i>Pre-Training & Fine Tuning</i>	32
4.5 UNITER.....	33
4.5.1 <i>Masked Language Modeling</i>	34
4.5.2 <i>Image Text Matching</i>	34
4.5.3 <i>Masked Region Modeling</i>	35
4.6 COMPARISON OF PRE-TRAINED MODELS.....	35
4.7 VISUAL QUESTION ANSWERING.....	36
4.7 GRU DECODER.....	36
4.8 TRAINING.....	37
CHAPTER 5: EXPERIMENTAL RESULTS	39
5.1 EVALUATION METRICS.....	39
5.1.1 <i>BLEU Score</i>	39
5.2 ROUGE SCORE.....	40
5.3 DATASET.....	41
5.2 QUANTITATIVE RESULTS.....	42
CHAPTER 6: CONCLUSION & FUTURE WORK	48
6.1 CONCLUSION.....	48
6.2 CONTRIBUTION.....	48
6.3 FUTURE WORK.....	49
REFERENCES	50

List of Figures

Figure 1.1: Examples of X-Ray reports by a radiologist	1
Figure 2.1: Structure of Heart	6
Figure 2.2: Medical Imaging of Heart	7
Figure 2.3: Structure of the lungs	8
Figure 2.4: Medical imaging techniques.....	10
Figure 2.5: Atelectasis	11
Figure 2.6: Consolidation.	12
Figure 2.7: Cardiomegaly	13
Figure 2.8: Nodule and Mass.....	14
Figure 2.9: Pleural Effusion.....	15
Figure 2.10: Emphysema	16
Figure 4.1: VisualBERT Architecture	28
Figure 4.2: LXMERT Architecture	30
Figure 4.3: LXMERT Embedding	31
Figure 4.4: LXMERT Encoders.....	32
Figure 4.5: UNITER Architecture	34
Figure 4.6: Gated Recurrent Units (GRU) Network Architecture.....	37
Figure 4.7: A complete and combined model of the pre-trained encoder followed by Classifier and GRU to get required outputs.....	37
Figure 5.1: Training Loss for Proposed Methodology	43
Figure 5.2: Test Accuracy of Model.....	44

List of Tables

Table 3.1: Chest Radiology Dataset Description	21
Table 3.2: Literature Review of Medical Images Captioning.....	23
Table 4.1: Comparison of Pre-Trained Models	35
Table 5.1: AUCs for 13 thoracic findings on IU CXR from Pre-Trained and TieNet Model	42
Table 5.2: Results of ROUGE and BLEU score for summarized medical report on the IU CXR dataset using proposed methodology	44
Table 5.3: Some qualitative results using multiple pre-trained models.....	45
Table 5.4: Best, Worst, and Average case ROUGE score of different pre-trained models for summarization.....	46
Table 5.5: Best, Worst, and Average case BLEU score of different pre-trained models for summarization.....	46

CHAPTER 1: INTRODUCTION

In human life, the chest diseases are fatal and many people face common chest diseases like pneumonia, cardiomegaly, pneumothorax, effusion, etc. [1]. Chest X-rays (CXR) and Computed Tomography (CT) scans are used to diagnose the chest related diseases. The chest abnormalities are captured through these images and after a proper pathological process, the experts get the results in form of reports. An analytical examination conducted by a radiologist through a process finds the presence of abnormalities in the images. A radiologist concludes all the abnormalities in the form of a detailed report which is a textual representation of these abnormalities in the X-ray image of the patient. In the radiology report, the details about the condition of the chest, diseases, and other findings are written. The manual reports are shown in Figure 1.1. Writing a detailed report based on the finding from the images is a very difficult and time-consuming task. Along with that, the young doctors are not skilled enough to get understand reports. This can cause serious errors, so the radiologist converts the reports into summarized form for their understanding and also highlights the disease captured. The number of patients is much more than number of radiologists available. Especially, in the developing countries where large populations and fewer resources are more common. The problem is increasing day by day and countries like Pakistan have no better solution. The patients wait for their turn to get a report from a radiologist and this routine becomes a huge problem for both radiologist and patient.

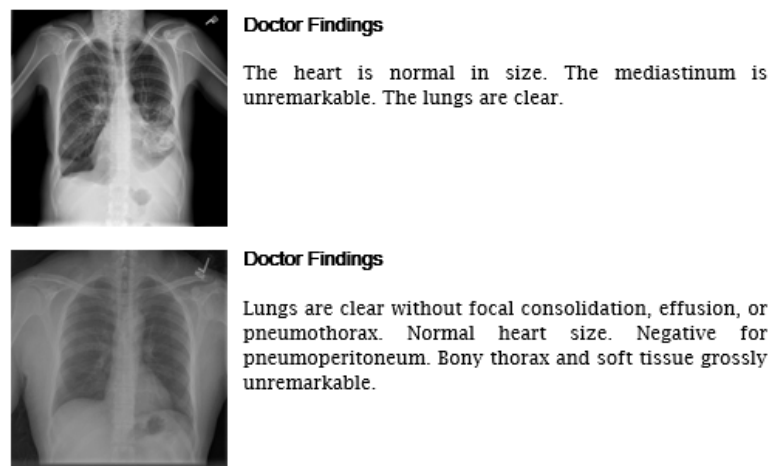


Figure 1.1 Examples of X-Ray reports by a radiologist

The interpretation of CXR in the form of text or radiology report is not efficient. Even for a specialist of respective field the interpretation is not efficient. The shortage of staff in

hospitals or excessive load of work in the hospitals will also cause errors in radiology reports [2]. Also, being less experienced in their fields and excessive workload will result in errors in reports. Many are working in rural areas with few health care facilities and are unable to read and understand a radiograph and describe it in form of text. To better understand and describe a report following skills must need [1].

- (i) The complete information and knowledge related to the basic physiology of chest diseases along with the required information of normality or abnormality of thorax anatomy.
- (ii) The skills and ability to find the relation with other diseases (respiratory function tests, test results and electrocardiograms).
- (iii) The ability to identify the variations in the radiographs with period of time.
- (iv) The knowledge of patient clinical background (complete medical history).
- (v) The ability to study and analyze the fixed pattern of radiograph.

Briefly, the summarization of medical reports along with highlighting the abnormalities is an unpleasant task especially for inexperienced medical professional while in some cases for experienced professionals as well. The proposed methodology is, therefore, derived from the motivation to improvise the diagnostic process of medical system. Automatic radiology report summarization improvise this system and helps medical professionals. The existing radiology report summarization approaches suffering from different problems. Similarly, the identification of thoracic findings from CXR and reports is not much focused. These limitations must be addressed to complete this task and give a better solution to tackle the problems. One of the limitations is a proper knowledge of the diseases which may appear as the white projections of some well-understood patterns on chest X-rays. After the understanding, the patterns of the language semantics for expressing it in a natural language for the layperson. The training of models in this kind of task is exceedingly difficult and time-consuming and the existing models do not focus on image-text joint embedding. Therefore, besides the challenge of the visual and language understanding, a model is required which can understand these patterns and translate them in form of a summarized radiology report and the similar models used for generating the thoracic findings from them.

In comparison to the available methods, the proposed methodology presents a model to solve the problems of the visual and language representation. As a first step, the proposed model takes chest x-ray images and radiology reports as inputs. In the second step, the radiology reports are converted in form of findings or diseases with similar pre-processing use by TieNet

[36] and the feature extracted by the process of Bottom-Up and Top-Down approach (BUTD) [25]. In the third step, the pre-trained vision and language models like VisualBERT [9], LXMERT [10] and UNITER [11] used for the joint embedding and self-attention mechanism help in highlighting the most relevant parts of text and image. From that, the pre-trained models are implemented in two ways one for generating the findings as a Visual Question Answering (VQA) [12] and secondly, the desired summarized report is generated by using these pre-trained as encoder and applying gated recurrent units as a decoder. The proposed methodology work on the self-attention mechanism. The suggested research methodology is motivated by the recent advancements in the vision and language joint embedding techniques, where the goal is to use image and text as joint input and generate results from them using multiple pre-trained models.

1.1 Motivation

In the medical field, doctors use mostly Chest Radiographs for the diagnosis and treatment of Lungs, Heart, and other chest-related diseases. According to many reports, the number of patients increasing day by day in our country and this alarming situation encourage us to find an automated way of generating findings/tags from reports and images, along with a summarized medical report that will not only save a lot of time of professionals as well as a milestone in the field of medical science. Also, some less experienced professionals can understand those detailed reports to get the exact disease and finding from the report, so this research helps to improve healthcare and is the key method for getting better results at lower costs. Therefore the proposed method is derived from the motivation to improve the performance of the system, to get more accurate and fast summarized radiology reports to correctly identify the disease based upon the patient chest x-ray images and reports.

1.2 Problem Statement

The main purpose of this research is to automatically generate the summarized content of a medical image using a pre-trained model which not only captures the diseases in an image but also must express how these diseases relate to each other. The existing solution does not implement the image and text joint using the pre-trained models which results in a slow process and lower accuracies.

1.3 Aims and Objectives

The major objectives of the research are as follows:

- To utilize pre-trained models for correct findings/tags and generate summarized reports.
- To use pre-trained model VisualBERT, UNITER, LXMERT along with bottom-up approach as a common encoder.
- To add classification and text generation head-on pre-trained models for tags and summarized report generation.
- To compare our system with state-of-the-art Methods.

1.4 Structure of Thesis

This work is structured as follows:

Chapter 2 covers the importance of the lungs, heart, and chest wall in the human body and their brief anatomy. It further discusses some possible diseases in them.

Chapter 3 gives a review of the literature and the significant work done by researchers in the past few years for the automated generation of radiology reports and the summarization of reports along with tags/findings.

Chapter 4 consists of the proposed methodology in detail. It includes pre-processing of images and using multiple pre-trained vision and language models for findings/tags and GRU for report summarization.

Chapter 5 introduces the databases used for evaluation purposes. All the experimental results are discussed in detail with all desired figures and tables.

Chapter 6 concludes the thesis and reveals the future scope of this research.

CHAPTER 2: CHEST ORGANS ANATOMY & THORACIC FINDINGS

In the human body, there are many vital organs behind the chest like the lungs, trachea, heart, esophagus, and thoracic diaphragm. The lungs and the heart are major organs. The respiratory system runs through the lungs and vital organs that exchange oxygen (O₂) and carbon dioxide (CO₂) between the atmosphere and body. Heart means the life of a human. It is an organ that transfers blood throughout the body. The trachea (a tube) connect the larynx and bronchi in the body. Esophagus is also a tube but a larger one that has a key role in supplying food (water and food) to the stomach. This chapter will briefly cover the heart and lungs anatomy, imaging techniques, diseases related to them.

2.1 Structure of Heart

A heart in a human body is a four-chambered muscular organ. A pericardial sac covers the heart which is lined with parietal layers. It looks similar to a man's closed fist. It is the center of the circulatory system of the body, which pumps blood. A heart is formed from three layers named epicardium (the outer layer), myocardium (the middle layer) and endocardium (the innermost layer). The heart cavity is divided into four different chambers Left atrium, Right atrium, Left ventricle and Right Ventricle. Each chamber receives a different kind of blood. The deoxygenated blood is received in the right atrium while oxygenated blood comes in the left atrium. The role is quite simple, it takes blood from two atria through the veins and supplies the blood in the body through both ventricles. A valve set is required to pump the fluid in one way. Atrioventricular is a valve between ventricles and atria that keeps the blood flow easier while semilunar valves are at the bases of the ventricles. One atrioventricular is tricuspid and the other one is bicuspid.

A pulmonary semilunar valve is between the pulmonary trunk and ventricle (right). The semilunar valve is between the aorta and left ventricle. In contraction and relaxation of the heart, the backflow is prevented through atrioventricular. To prevent the blood flow back into ventricles the semilunar valves play their role. A double sort of pump work simultaneously means the heart ventricles and atria both contract and relax at the same time. Blood from the right atrium flows to the ventricle (right) which is supplied next to the heart in the lungs to get the oxygen. The lungs supply the blood to the left atrium which then moved into the left ventricle. After the process, the blood is supplied to the body. The supply of oxygen and nutrients is very necessary for a heart to work efficiently. The waste products are removed, and

oxygen is supplied through a vast number of blood vessels. The structure of heart is shown in figure 2.1.

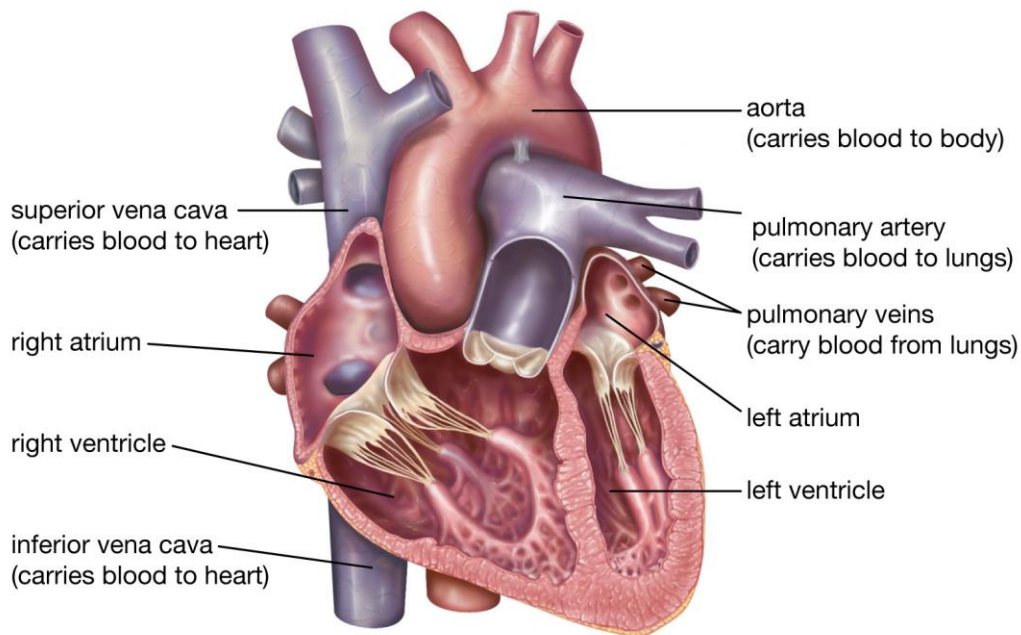


Figure 2.1 Structure of Heart [28]

2.2 Imaging Techniques to Analyze Heart

Heart imaging techniques are used to diagnose heart problems. A radiologist (cardiac specialist) interprets the medical images in the form of reports and diagnose the heart diseases. The imaging techniques can be of diverse types like Computed Tomography (CT) scans, X-rays, and Magnetic Resonance Imaging (MRI) scans. The difference in imagining is shown in figure 2.2. The chest x-rays can also be helpful for heart disorder findings [41]. Normally the x-rays are taken from the front and side. A beam is thrown on the chest of a human using a machine and collects the details in the form of an image (x-ray). Through this, the shape and size of the heart are found easily. The chest x-rays can easily highlight the abnormalities in shape as well as in the heart of a person. The condition can also be identified through an x-ray. X-rays can easily detect any abnormality in heart size. Heart enlargement which is common cause of heart failure can detect through it. Constrictive pericarditis can be diagnosed by x-rays.

The pulmonary arteries narrowing and enlargement can easily tell about the blood pressure. MRI on the other side is a more beneficial tool in medical imaging. High strength magnetic waves use to take images of the body. Without using the ionizing radiation which is harmful

in some other cases, MRI cardiac help in diagnosing the heart-related issues. A small bed like system attached to a tunnel-like machine take scan and give output in form of images.

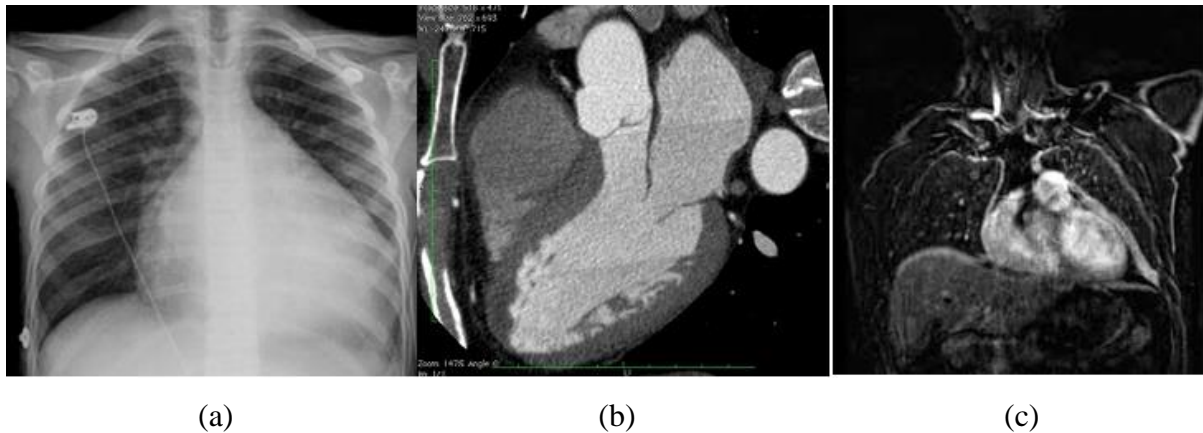


Figure 2.2 Medical Imaging of Heart (a) X-ray (b) CT Scan (c) MRI

2.3 Structure of Lungs

Lungs are made up of sacks of tissues, whose location lies in the chest's thoracic cavity between the rib cage and diaphragm. In organisms, each lung is enclosed in a thin membranous structure known as the pleura, and each of the lobes is linked by its central bronchus, which is a significant passage of air with the trachea also called the windpipe and pulmonary arteries connect it to the heart. On each lung's inner side, the hilum is preset nearly two-thirds of the distance from its base to its apex, and this is the point where bronchi, pulmonary arteries and veins, lymphatic vessels, and nerves are joining the lung. The blood receives oxygen from the lungs; oxygen is extracted from the air and finally delivered from the lungs to every individual cell of the human body.

In the chest, since the left lung shares some space inside the heart; therefore, the left lung is usually considered smaller than the right lung. Air flows through the route of nose or mouth into the respiratory system and goes to the trachea through the pharynx. The air moves through the trachea unless it splits into two bronchi that connect with the lungs. Three lobes are present in the right lung, whereas two lobes are present in the left lung, as the size of the left lung is smaller compared to the right lung. The bronchi break further within each lung into several tinier air passages known as bronchioles, significantly expanding surface area. Growing bronchiole ends with an air sacking cluster termed as veoli. The gas exchange occurs with the bloodstream in the alveoli, which contains various capillary veins in their walls.

Typically lungs contain some air after birth, and they are light, flexible, fibrous, and soft organs. They would float in water upon being squeezed when healthy, and if diseased, they shall sink. Moreover, excluding respiratory functions, the lungs are also involved in performing other tasks of the body, including absorption and excretion of alcohol, water, and other pharmacological factors.

The structure is shown in Figure 2.3.

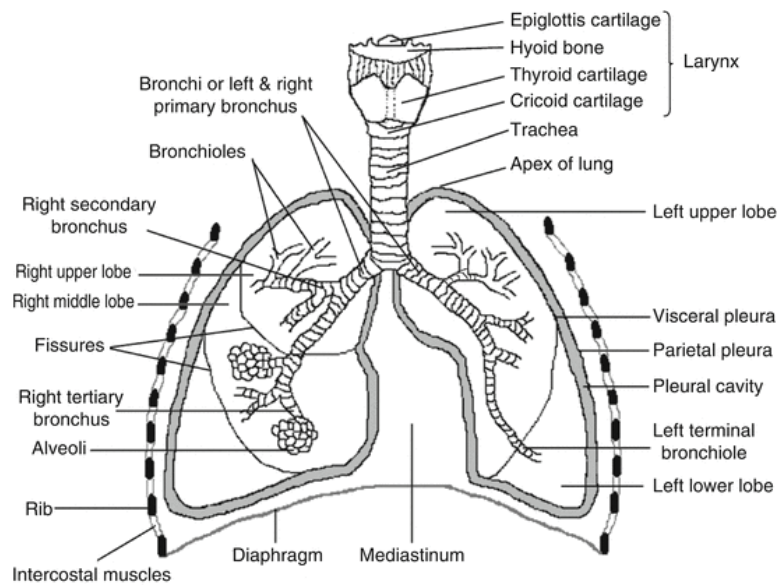


Figure 2.3 Structure of the lungs [29]

2.4 Imaging Techniques to Analyze Lungs

Medical imaging is a well-known technique used for the visual representation of the body for medical analysis. The different types of medical imaging are X-rays, Computed Tomography (CT) scans, Magnetic Resonance Imaging (MRI), and Ultrasound (Figure 2.4). Chest x-rays include a clear overview of the heart: the cardiovascular organ and main blood vessels, which can typically show severe lung disease in neighboring areas, or chest wall, like ribs.

For example, most cases of pneumonia, lung tumors, chronic obstructive pulmonary disease, a collapsed lung (atelectasis), and air (pneumothorax) or fluid (pleural effusion) can be found in pleural space by the help of X-rays. Plain film radiography is a kind of radiations which, when passing through the body, produce an image on the x-ray film according to the density of striking objects.

The radiations that pass through the objects strike the film and burn it that makes the film black. Bones appear whitest; as they are dense enough to absorb almost all the radiations, so very few

radiations pass through them and hit the film. Soft tissues such as muscles appear grey, and air/gas seems black. The bodily images, which are 360° cross-sectional, are provided with the help of plain film radiography. Magnetic Resonance Imaging produces details of body parts without the use of radiations by combining a strong magnetic field with radio waves and advanced computer systems. Ultrasound gives the internal structure of various parts of the body using sound waves of frequencies higher than the human audible range. It is significantly used for the detection of fluid present in pleural space. When a needle is used for fluid removal, ultrasonography can be used as guidance. Pneumothorax is also diagnosed through bedside technique. A material that can be seen on x-rays (called a radiopaque contrasting agent) can be administered intravenously or delivered by mouth during CT to help explain some chest anomalies. More advanced CT procedures are the high-resolution CT and helical (spiral) CT. Mostly CT scans and x-rays images are used for the analysis of lungs. X-ray images are preferred over CT scans because CT scan imaging uses a high dose of radiation, while x-rays use a low quantity of radiations, and it is a fast and reliable method.

The CT angiography makes use of a radiopaque contrasting agent inserted into an arm vein to create a picture of blood vessels, such as the pulmonary artery that carries blood to the lungs from the heart. CT angiography is usually performed to treat blood clots in the pulmonary artery (pulmonary embolism) instead of nuclear lung screening. High detailed images are produced by MRI, which are particularly helpful when the doctors suspect blood vessel abnormalities like an aortic aneurysm in the chest. MRI also takes a longer time to do, however, and is more costly than CT. In a pulmonary artery, injection of contrasting radiopaque agent by a thin, long catheter tube passing through a vein in the heart and after that in the pulmonary artery is done, known as pulmonary artery angiography. Positron emission tomography can be used during the doubt of cancer [14]. Various metabolic rates of malignant as compared to benign tissues are dependent on this imaging radiographic procedure.



(a)

(b)

(c)

(d)

Figure 2.4 Medical imaging techniques (a) CT scan (b) MRI (c) Ultra-sound (d) X-ray

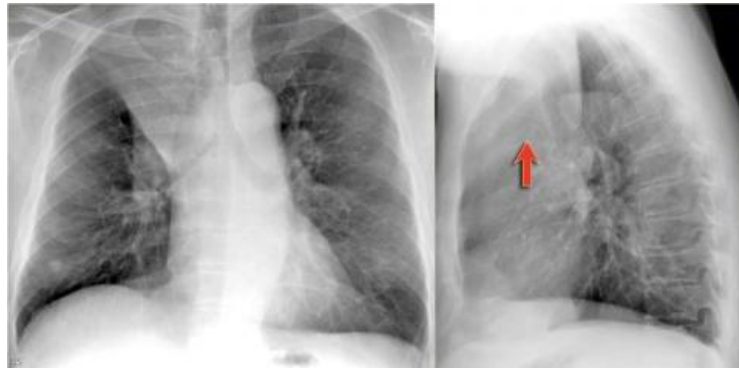
2.5 Thoracic Findings

There are many diseases which can be identified through a chest x-rays. The radiograph of each is different showing a specific pattern. Some of the patterns are shown in x-ray images. On a larger scale the irregular patterns can be recognized as Atelectasis, Infiltration, & Consolidation, Cardiomegaly, Pneumonia, Nodules & Masses, Pleural effusion, Pneumothorax, Edema, Emphysema, Fibrosis, Pleural Thickening and Hernia [59], [18]. The detail of some patterns is described below.

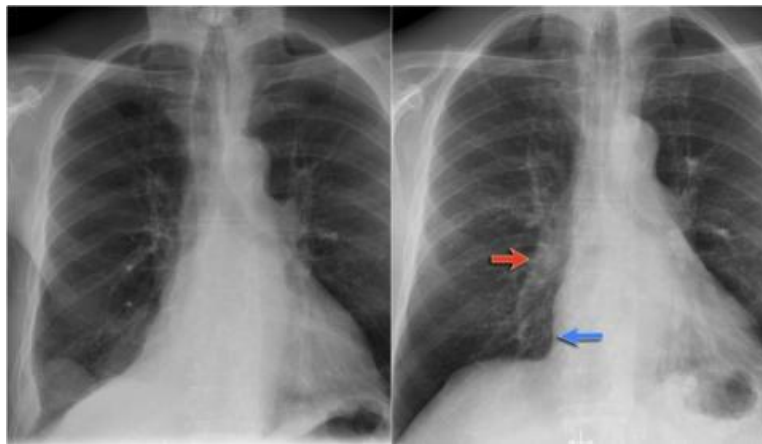
2.5.1 Atelectasis

Atelectasis is commonly known as the reduction in all or some parts of lungs. It is also named as Lobar Atelectasis. When alveoli (tiny air sacs) are deflated in the lungs or filled with a fluid named as alveolar fluid. Mucus plug is one of the reasons for this problem. The mucus plug grows up in the airways. After a person goes through a surgery, has not been able to breathe normally. The main reason is the use of such medicines during surgery while suctioning the lungs makes the lungs clear. The mucus plugs are most common in less age. One of the other reasons behind Atelectasis is inhaling a small object like almond or peanut, most commonly by children. Another reason is the growth of an abnormal tumor inside the airway. Lung diseases like asthma and bronchiectasis or weak breathing capacity are major factors of this disease. Atelectasis is the most common complication that occurs while breathing. This is often associated with abnormal displacement of fissures, vessels, bronchi, heart, and diaphragm. It can directly affect the breathing system. Atelectasis can be of two kinds: one is Linear Atelectasis and the second is Round Atelectasis. Round Atelectasis is fibrotic and thickened interlobular septa which is round and collapsed. Linear Atelectasis is the thickening of a focal area. The thickness may reach to 1

centimeter in size. There are different signs like Cough, difficult in breathing and shallow and rapid breathing.



(a)



(b)

Figure 2.5 Atelectasis of (a) right lobe (b) left lobe [31]

2.5.2 Infiltration & Consolidation

Consolidation (or air-space opacity) represents a condition in which the air inside the alveoli has been substituted with an alternative material. That substance can be blood (pulmonary hemorrhage), pus (pneumonia), edema fluid, or tumor. It is recognized on a chest x-ray when one or more of the following features are seen: (1) vessels obscured by ill-defined or irregular homogeneous opacity, (2) Lung / soft-tissue impairment system, (3) No volume loss, (4) Extends the fissure or pleura but does not cross it, and (5) Air-bronchogram. Consolidation may be present for the diagnosis of pneumonia. Consolidation includes some common signs like:

1. On the affected side, thorax expansion is reduced upon inspiration.
2. Increased vocal fremitus present on an affected side

3. Dull percussion is found on an affected side.
4. Pleural rub and bronchial breath sounds can be present.
5. An increase in vocal resonance

Greater radiopacity is found in consolidated tissue, and it is a late-stage pulmonary complication. Figure 2.6 shows consolidation in different lobes of the lungs.

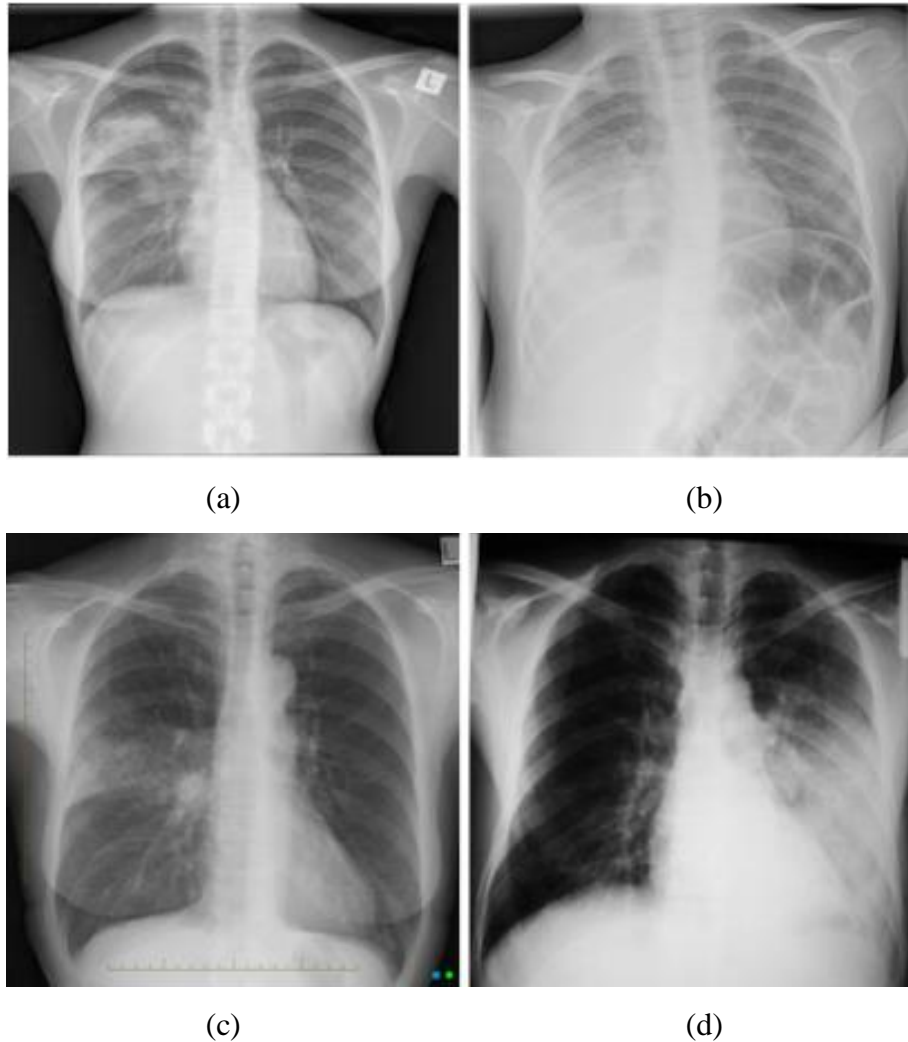


Figure 2.6 Consolidation of (a) right upper lobe (b) right lower lobe (c) anterior segment of right upper lobe (d) left lower lobe [15].

2.5.3 Cardiomegaly

Cardiomegaly is referred as enlarged heart. Normally this is not consider as a disease but a condition which refers to diseases. Through chest x-rays the enlarged hearts are identified and further test diagnose the actual condition which causing the heart enlargement. There are certain conditions that are causing the heart to enlarge, including

- high blood pressure
- heart valve diseases

- pulmonary hypertension
- thyroid disorder
- excessive input of iron in the body

Sometimes the heart have to pump harder for the blood supply which can cause enlargement of muscles and eventually it weaken the heart muscles. The supply of blood is in regular pattern but sometime if the valve condition not well then an irregular heartbeat can cause enlargement in the heart. The supply toward the lungs is harder which can also cause the heart to enlarge. Thyroid disorders and low red cell count in the body can lead towards the enlarge heart. Enlarge heart leads toward many serious complications some of them are heart failure, heart murmur, blood clots and sometime death.

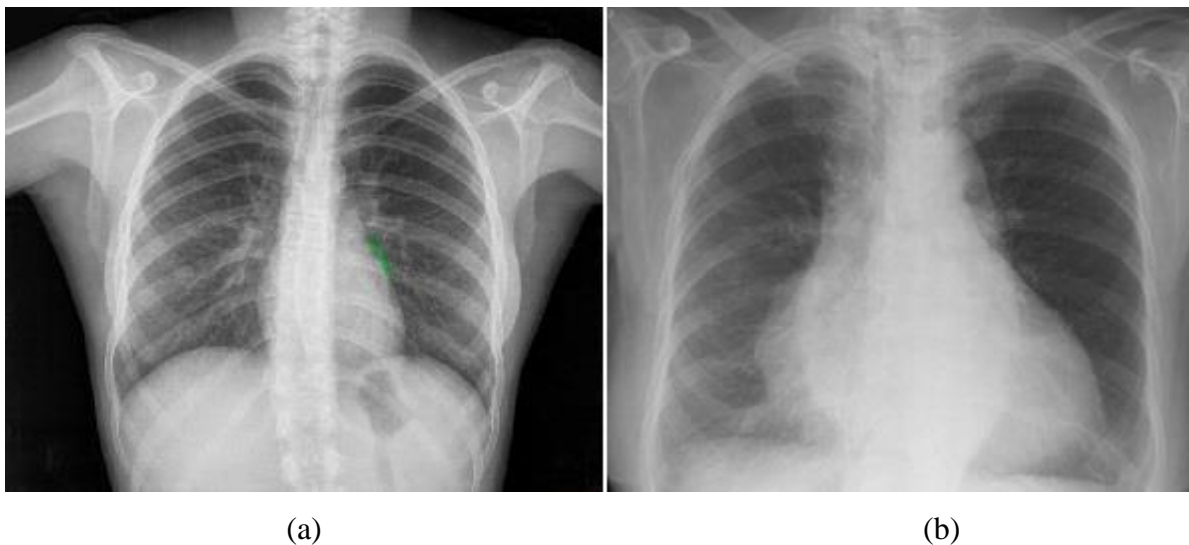


Figure 2.7 Cardiomegaly (a) Normal heart (b) Enlarged heart

2.5.4 Nodules & Masses

Nodules & Masses represent any space-occupying lesion, either solitary or multiple. They are mostly benign. The nodule is also known as "coin lesion" or "spot on the lung." A discrete, nearly circular opacity on a chest x-ray ranging up to 3 cm in size is defined as a nodule. A mass is essentially the same as a nodule but differs in size, as it is greater than 3 cm.

There can be many causes of benign nodules, like many of them occur due to an infectious disease or inflammation in the lungs. Most infections occurring with nodules are not active like Mycobacterium tuberculosis and fungal infections like histoplasmosis, coccidioidomycosis, and aspergillosis. Nodules may be found actively or due to the formation of scar tissue. The non-inflammatory causes include rheumatoid arthritis, granulomatosis, or sarcoidosis. Figure 2.8 points out nodules and masses on chest x-rays.

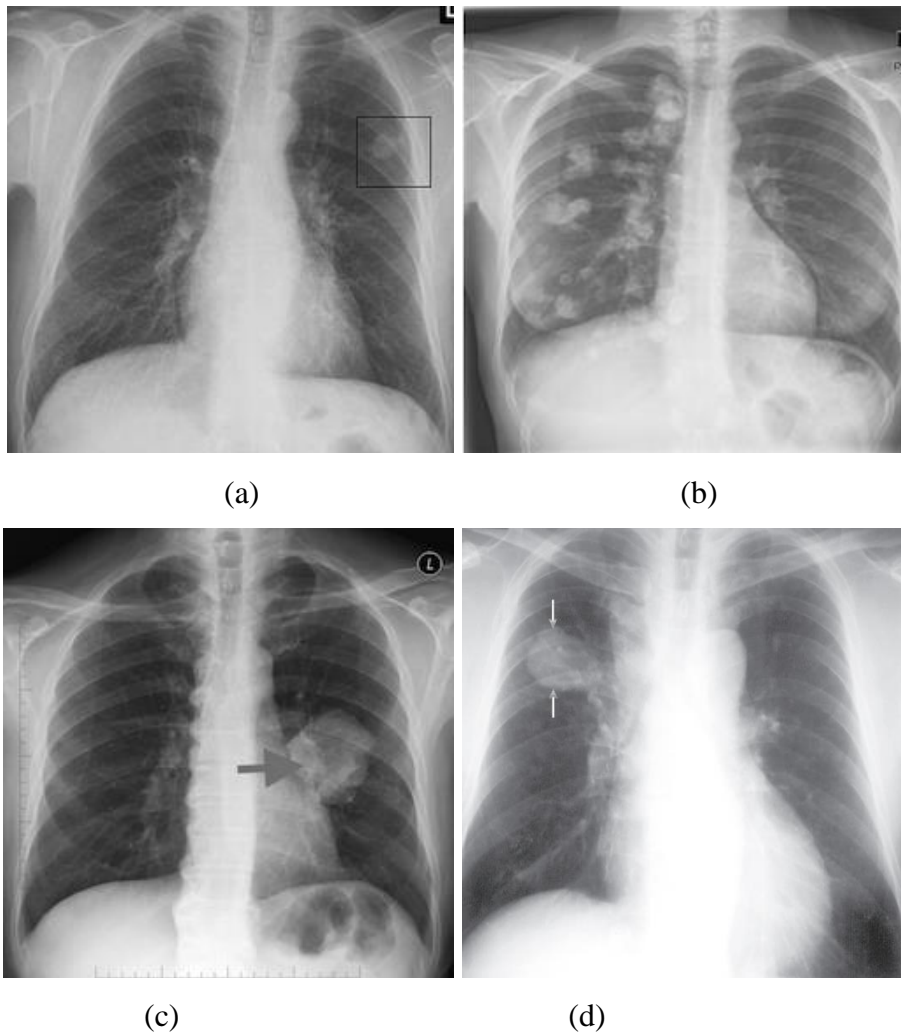


Figure 2.8 Nodule and Mass (a) Nodule in left lung (b) Multiple calcified pulmonary nodules in right lung (c) Mass in left lung (d) Mass in right lung [18], [20], [32].

2.5.5 Pleural Effusion

A thin membrane lines the inner side of the chest cavity called pleura that envelope the lungs. Their function is lubrication and facilitating respiration. Mostly, pleural space contains a small quantity of fluid, i.e., in the space which is present outside of the lungs between the pleural layers. Pleural space abnormalities include pleural effusion, pleural masses, pleural thickening, the air within the pleural space (pneumothorax), and pleural calcification. The most common among these is pleural effusion, which is often referred to as "water on the lungs." It is the existence of a substantial amount of liquid in pleural space. The effusion can range from minimal blunting to massive with a complete whiteout of the one side of the chest (hemithorax). Its symptoms usually include shortness of breath, dry cough, chest pain, pyrexia, and orthopnea.

Pleural effusion can be drained through the pleural drain, or chest tube thoracostomy can be done for making a small cut in the chest wall, and a plastic tube is placed in pleural space. A substance like doxycycline is injected in pleural space through a chest tube, and this substance helps the chest wall and pleura bind tightly to each other upon healing. This procedure can prevent come back from effusions in a lot of cases. The effusion may be transudative and exudative, depending on the protein content. The most commonly found causes of transudative (protein deficient) are congestive heart failure, cirrhosis, pulmonary embolism, and exudative (protein-rich) is lung cancer, pneumonia, kidney, or inflammatory diseases. Pleural effusion can also occur due to tuberculosis, pleural effusion, ovarian hyperstimulation syndrome, Autoimmune diseases and Chylothorax.

Abdominal surgery, radiation therapy, and certain medications can also cause pleural effusion. It may also occur with certain kinds of cancers like lymphoma, breast, and lung cancer. Sometimes, the fluid itself can be a result of chemotherapy, or it may be malignant. Figure 2.9 illustrates different examples of pleural effusion.

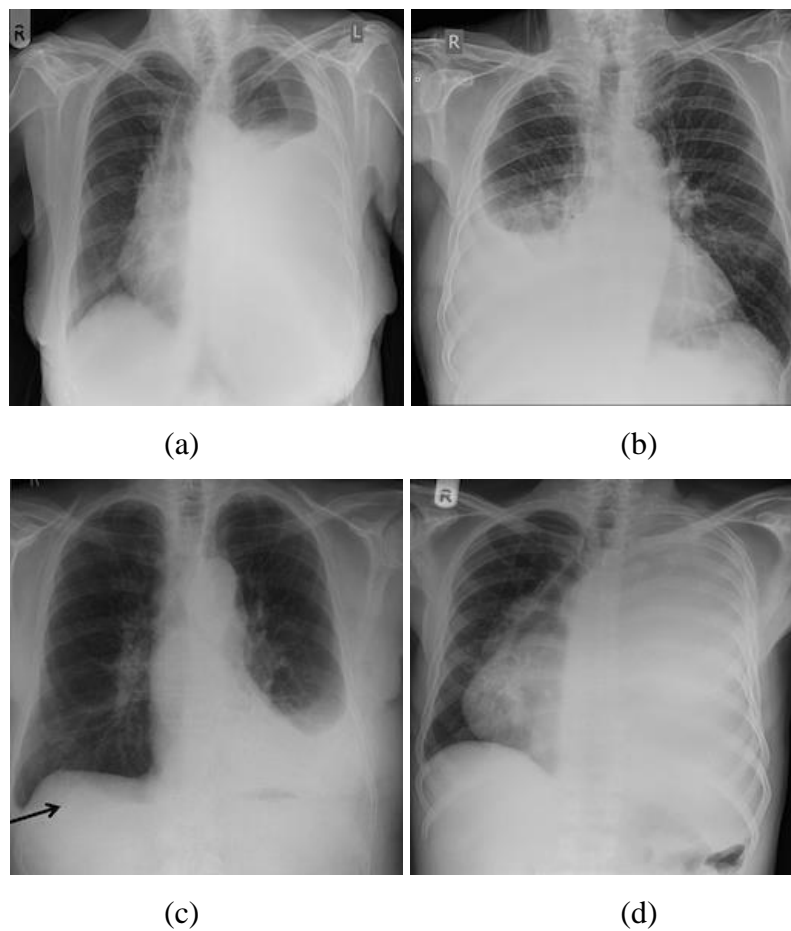


Figure 2.9 Pleural Effusion (a) Left pleural effusion (b) Right pleural effusion (c) moderate left pleural effusion and subpulmonic effusion on the right (d) Massive Pleural effusion with mediastinal shift [15], [24].

2.5.6 Emphysema

Lungs are major part of respiratory system and a human can breath with the help of lungs. The air sacs sometime can be damaged due to excessive smoking of tobacco, air pollution and through chemical dust which can cause the shortness of breath (Emphysema). The weakness in the inner walls of the air sacs and ruptures create the bigger spaces which reduces the lungs surface area and it reduces the oxygen amount. The lungs takes oxygen and releases the carbon dioxide but the damage in alveoli can cause the air trapped and leaving no space for the fresh air. Emphysema patients also faces chronic bronchitis (inflammation of air carrying tubes in the lungs). Both can damage the lung tissues. The complications which may face the emphysema patients are

- Pneumothorax
- Heart Diseases
- Bullae (Empty space in lungs)

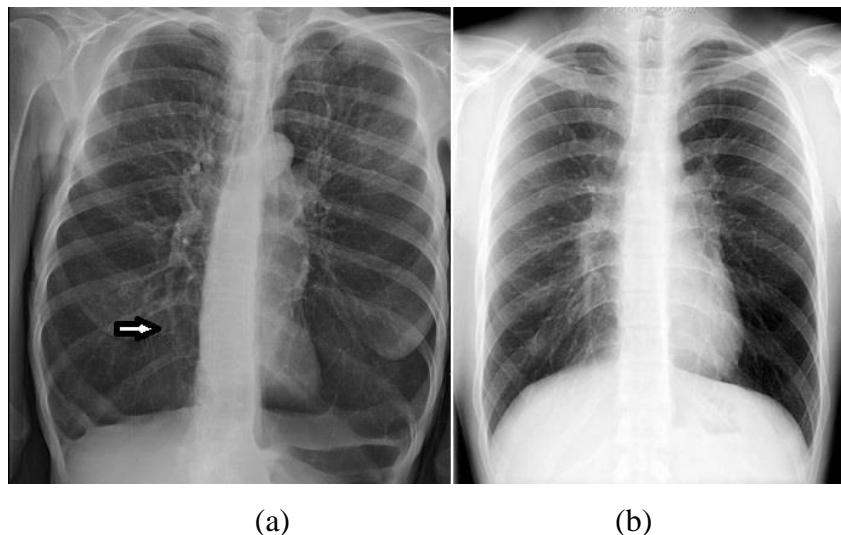


Figure 2.10 Emphysema (a) Mild Stage (b) Severe Stage Emphysema

The rise of chest related diseases especially coronary heart diseases enforce the researches to talk more and more about the diagnosis of those diseases. The importance of different parts of chest described in this chapter. The normal and abnormal conditions of different parts showing the serious consequences if someone failed to identify the effects or changes in the chest. These diseases are fatal for human life. The only way to save the life is to get identify these thoracic diseases and conditions of different parts to the professional.

CHAPTER 3: LITERATURE REVIEW

In research, radiographs (specially the chest) are always an interesting field for researchers in medical imaging. Multiple imaging techniques are utilized nowadays but traditional radiography is still the initial modality for medical report generation. . Computed Tomography (CT) scans and Magnetic Resonance Imaging are the primary techniques for assessing and diagnosing particular parts of the body. In the last few years, several groundbreaking research types have been published on computer-aided diagnosis of heart and lung diseases. These diagnosis can used generate the automated medical report. Multiple approaches were proposed for the role of image captioning. We classified them in three major types/groups. The formation of template description is the first group which that on exploring and identifying the artifact results of attributes. This group also ensure the basic grammatical rules while exploring. The captions produced by these methods are considerably basic type and appeared correct in case of the grammar. But the major drawback of these methods are dependence on hard coded visual notions. This reduce the efficiency and variation. The second way casts the challenge as a problem of retrieval predicted on the presumption of providing similar caption on the same images.

The main purpose is to get a collection of similar titled images from a huge database. The other purpose is to either transfer titles of similar images to a new paragraph directly or generate a new caption by joining candidate's caption pieces based upon specific rule. The last one is completely based on deep learning methods. The last two groups and approaches have fallen out in favor of dominant method. In the past the research work only focused on neural networks and similar kind of models but no one focused on multimodal architecture. In recent years, a tremendous success has been made by self-attention models like transformers. The pre-trained models like VisualBERT, BERT, LXMERT and UNITER are used to achieve targets in classification and image and text combine embedding techniques improve the results. In this chapter, all valuable researches in this domain are summarized and the datasets used by these researches are in focus.

Litjens et al. [22] methods of preference for analyzing the medical images highly accepted in learning algorithms deeply. The author reviewed the main detailed learning ideas which apply to medical picture analysis. The survey is based upon classification of images, segmentation of images, object detection and other roles on deep learning methods. Reports on each application area are given with thoroughly overview of pulmonary, cardiac, neurological, breast, ocular,

abdominal and musculoskeletal. The rightly reading of chest x-ray is an exasperating activity and the main reason is variation, variability and complexity in diseases and their treatment. One such example is CT scans available in Rubin et al. [30]. The chest x-rays dataset of MIMIC-CXR deep convolution neural network were trained to recognize several prevalent thorax disorders. Convolution neural network (CNN) models are analyzed and described for lateral and frontal chest x-rays. This gained much lesser consideration from previous programs. The researcher develop a clear prototype for viewing posteroanterior and anteroposterior (frontal). The frontal and lateral chest x-rays of a patient were introduced through a DualNet architecture which demonstrates the utility in improvising the efficiency against standard classifiers. According to Q. You et al. [33] Image captioning is in focus because of its importance in practical applications. The major challenge in image captioning is creating a meaningful description of the image. The previous approaches are not as good as required, the images converted in words or words used to caption images. But these models lacked in giving fine details and coherent description of images. The proposed method is top-down and bottom-up approaches are combined and a new captioning model proposed in this paper. The Recurrent Neural Network use top-down features for getting global information and gets attribute feedback from bottom-up using attention mechanism. The feedback system in the algorithm predict more accurate words and fill the semantic gap between prediction and image objects. The CNN used to create global visual description and the input node give an overview of image content. A list of visual attributes or concepts get from attribute detector which corresponds to entry in dictionary. The features are used for caption generation in RNN [66],[67]. The input attention model assign a score to each attribute vector. The output attention model give attention score and designed to attend certain cognitive cues in feature vector, the extracted information from image most relevant to parsing existing words and predict words. The results shows that the proposed method of image captioning give better performance by combining the two approaches and extract more fine details of image with RNN. The model is able to give semantically important regions and switch more weights on those features which are more important according to task and gives better caption to the image. One of the major challenging task is Visual Question Answering, it requires more effective semantic embedding and fine grained visual understanding and the models neglect spatial context and high level image semantics. More effective semantic embedding and fine visual understanding is required D.Yu et al. [34]. A spatial encoding approach in which the context aware visual features are extracted from image using bidirectional RNN model. Joint learning in multi-level attention which

reduce the semantic gap from vision to language. Semantic Attention find the question related concepts from image. Using deep Convolution Neural Network train a concept detector, which define the concept as word. Context-aware Visual Attention fill the semantic gaps between images and questions, the question related region in image are find using fine tuned CNN model and get the relation between region and question on the basis of attention score. Joint Attention Learning add question vector into attended image features extracted from layers and use element wise multiplication to combine two types of attention together then joint feature into softmax layer to predict probability of answer. The highest probability candidate is the final answer. A novel method which combine visual and semantic attention to get automatic question answering. In past the textual attention is more focused but the proposed model exploit more concepts from the image. The proposed model implemented on different datasets and comparing the attention model with high level concepts without attention mechanism, the result shows that the model achieved high performance. In another work, A. Zaman et al. [40] work on Tuberculosis (TB), which is 2nd largest death disease after HIV [42], the main reasons of the death are late detection of this disease. The deadly disease effect the lungs. The detection of TB is through different process like skin test, DNA based test (very expensive) and using the x-rays analysis which is time consuming because of shortage of radiologist. Indiana University Chest X-ray dataset used in this research in which lungs are segmented through Random walker segmentation and selected features are extracted, then classified using SVM. Semi supervised segmentation using user based seed points and the segment the organ in the volumetric medical image. The color contrast based features extracted from the segmented parts and used for classification. Support Vector Machine is used for better classification. The results shows that the deadly disease can be easily detected using the chest x-ray and methodology shows 73% accuracy of the results which is decent percentage for the reliability of method proposed. The method reduced the workload on the medical system and through easy and early detection of disease save lives and early treatment.

According to [21],[23], training representative computational models from data on medical images need vast sets of instructing data. For large quantities of data, voxel-level annotation is often impracticable. The substitute for manual annotation uses the immense quantities of information stored during the clinical procedure in image data and accompanying records. He first suggested a poorly supervised learning method to use semantic explanations as labels in reports to better classify tissue patterns in OCT imagery. They specify how accurate voxel-level classifiers would be and how these details improve the performance of intraretinal SRF,

IRC, and regular retinal tissue classification. They suggest using a semantic representation of clinical results as a lesson objective, which is anticipated by a convolutionary neural network from imaging data. It is demonstrated how detailed voxel-level classifiers based on inadequate volume-level semantic definitions based on a collection of 157 volumes of optical coherence tomography (OCT) can be obtained. They illustrated how semantic information improves intraretinal cystoid fluid (IRC), subretinal fluid (SRF) and normal retinal tissue classification precision, and also how the learning algorithm relates semantic principles to imaging features and calculation.

According to Tanti et.al. [8] the generative models are classified into inject and merge architecture. In the first type, the input given tokenized captions and images into a vector (RNN block) while in second type the captions are only input in the RNN block which merge the image and output. This learns effective mathematical models and procedures by strengthen the knowledge in medical field. This type of combination in which both image and text data used for report generation can leverage the model performance. According to S. Candemir et al. [3] The chest radiography are used in many disease diagnosing process like lungs and heart pathologies. Cardiomegaly an abnormal heart enlargement disease which increase the cause of heart attack. Cardiomegaly detection depends upon the limited number of radiologist and later detection have serious effects on life. Fine tune deep CNN architecture for cardiomegaly detection. Pre-trained model (CXR-based) to overcome the limited annotated data. The end-to-end training to CNN architecture on NIH-CXR dataset which is then fine-tune the final layers of this model with limited cardiomegaly CXR. Low level features learn in earlier layers and more specific features to cardiomegaly from images. The disease with different level of severity detected through training a multi class classification system at different level. The Indiana dataset used to classify cardiomegaly images from all based upon the severity using a softmax probability. In this paper deep CNNs used for automatic detection of cardiomegaly. CXR based pre-trained models for pulmonary classification in CXR and tested for cardiomegaly classification. The results shows that the system confidence increases with severity in distribution of each severity class and average probabilities. In another research work by S. Singh et al. [4] Without radiologist correct interpretation and complete summarization of medical images is very difficult task. The radiology practice is error prone due to limited number of radiologist and increasing number of patients. Automated system required which gives correct report based upon the medical image [44]-[48], provided. Encoder Decoder based multimodal machine learning model proposed which can automatically

generate radiology reports from given images. The system identify the salient findings from the Chest X-rays using CNN and then fed into LSTM for generating a report. Chest X-rays image provided to convolution neural network which encodes the information in vector based image representation form. The encoded information passed through Recurrent Neural Network which act as decoder and generate a detail report based upon the vectored image. Transfer learning used to overcome the less annotated medical data which initialize pre-trained weights of the model that trained on large scale image classification dataset. The encoder-decoder framework proposed give better results. The promising results achieved in experiments which generate the reports from images and reduce the workload of radiologists. The automated generated radiology reports from medical images show the effectiveness of proposed model.

Table 3.1: Chest Radiology Dataset Description

Dataset	Source Institution	Disease Labeling	Images	Reports	Patients w.r.t Reports
IU Chest X-Ray (Demner-Fushman et al. [17])	Indiana Network for Patient Care	Expert	8,121	3,996	3,996
MIMIC-CXR (Johnson et al. [39])	Beth Israel Deacones Medical Center	Automatic (CheXpert labeler)	4,73,057	2,06,563	63,478
Chest-XRay8 (Wang et al. [26])	National Institutes of Health	Automatic (DNorm + MetaMap)	1,08,948	-	32,717
PadChest (Bustos et al. [38])	Hospital Universitario de San Juan	Expert + Automatic (Neural network)	1,60,868	2,06,222	67,625
CheXpert (Irvin et al. [37])	Stanford Hospital	Automatic (CheXpert labeler)	2,24,316	-	65,240

M. Sajad et. al. [6] Teeth have greater importance in human body but once they are infected diseases it gave a lot of pain. For dentists, the correct prediction of lesion and its type using a radiograph is very difficult and time consuming. Lesions are of different types and dental x-rays are opaque, which required an expert and a lot of time to classify them. The treatment become useless if the lesion is not correctly diagnose. The research proposed a method to

automatically predict and make the diagnosis easier for dentist. Endo-perio lesions in periapical x-rays with two different experiments. First, the features extracted using Alexnet and trained on conventional classifiers like K-Nearest Neighbor and Support Vector Machine. In second the previous Alexnet model was fine-tuned on training data both with and without augmented data and used as feature extractor to classify using Softmax function. The fully connected layer which used for feature extractor and K-NN & SVM were trained. The results showed that the process made diagnosis process easier and faster in identifying type of lesions using periapical x-rays. The transfer learning and data augmentation technique used for classification. By performing two experiments the results showed that the SVM classifier trained on features extracted by retained model performed well. X. Huang et. Al. [7] Chest X-rays are now widely used to diagnose the diseases, but due to lack of expert radiologist the misdiagnosis is increasing which effects in correct treatment of patients. The radiology reports are generated based on radiology image but still there exist some problems in extracting features and generating report. The background information neglected in diagnosis process which also effects the correct report generation. Multi-attention encoder decoder model which receives images and generate its feature representation using a CNN and multi-attention module by paying attention on channel and spatial information. The background information fusion module encodes the patient's background information. The decoding process generate the text using LSTM structure, first generate a series of high-level topic vectors representing the sentence and then generate a sentence based on each topic. The background information is fuse with word embedding. The incorporation of background information in model along with attention mechanism gives more effective results as compared to others. The model achieves state of the art performance by incorporating background information and enhancing the mapping between text and image entity position in report generation. Zhang et al. [5] presented a generalized framework to improve the factual correctness of summarization models. The framework evaluates the factual correctness of generating a summary by the fact-checking method. The model is implemented on radiology report datasets. The neural model training strategy helps in giving optimized and summarized results along with factual correctness. Factual correctness is a key requirement applied to two different datasets collected from hospitals. The improved performance was clearly shown via both human and automatic evaluation using factual correctness in summarized results.

Machine learning has already been performed for many years by defining a sequence of many activities, including aligning phrases, reordering, and independently translating terms. But with

passage of time the advancement in using pre-trained model in image-text join embedding give state of the art performance. The medical images also perform better in these modeling techniques as compared to previous encoder decoder-based methods.

Table 3.2: Literature Review on Image Captioning

Author	Year	Model	Type	Dataset	Metrics
Baoyu Jing et al. [50]	2018	CNN + MLC + Hierarchal LSTM	Medical Report Generation	IU X-Ray and PEIR Gross	BLEU, METEOR, CIDER and ROUGE
Liu G. et al. [65]	2018	CNN + Retrieval Policy Module	Medical Report Generation	CX-CHR and IU X-Ray	BLEU and ROUGE
Z. Chen et al. [43]	2020	Transformer Encoder + Decoder	Detailed Medical Report Generation	IU and MIMIC CXR	BLEU, METEOR, and ROUGE
MacAvaney et al. [49]	2019	Pointer-generator network	Report Summarization in the medical domain	MedStar Georgetown University Hospital	ROUGE
Zhang et al. [5]	2020	Pointer Generator + Reward	Report Summarization with factual correctness	Stanford University Hospital and RIH	ROUGE, Factual F1

The proposed model has applied multiple pre-trained models (UNITER, LXMERT, VisualBERT) for learning multimodal representation from Indiana CXR radiograph and its associated report. The pre-trained model used in two ways to show the importance of using them 1) generate the thoracic findings as a classification task, the entire process is based on self-attention mechanism. 2) The joint image text embedding input in pre-trained models and the results used for generation of automatic summarized report using GRU based model as decoder which is shown in next chapter. The main contributions of the proposed research in the summarized medical report generations are:

- A novel model that provides solution for the problems with more advanced and fast modelling techniques using pre-trained model in learning and classification of thoracic findings and GRU for automatic report summarization.
- The proposed method uses the joint image-text from text only embedding, gives a better performance by using the pre-trained models as compared to train models from scratch.
- Finally, substantial experiments on Indiana University Chest X-rays dataset have been demonstrated the significance of our proposed method.

CHAPTER 4: METHODOLOGY

The prior chapter has thoroughly discussed the research conducted in the relevant subject. In this chapter, the methods are explained clearly. This chapter presents a detailed methodology for automatic generation of findings/tags and report summarization using pre-trained models and also include the detailed mathematical equations and their association with each other and how they are working and applied in model. Starting from the mathematical model involved in the entire process and later discussed the detailed methodology of the proposed architecture.

4.1 Mathematical Model

A probabilistic model is proposed to produce the summarized text report by using joint embedding. Recent achievements in the domain of machine learning shows that the strong model of text summarization give state-of-the-art results by explicitly optimizing the probability of successful translation, provided by joint embedding sequence both for training and inference. Such models use pre-trained models that converts the variable size input of encoder to fixed size vector. The fixed size vector is then used as input to decoder part that converts this into a meaningful sequence of words. Thus, in our proposed model the variable size input is text and CXR which combined and use as joint embedding, the pre-trained model is used as encoder and GRU used as decoder for summarization part. While the same process repeats but the classifier head added on top of encoder part output the thoracic findings. The main objective is to directly maximize the likelihood of accuracy of summarized medical report as described initially by the radiologist. This is achieved by mathematical formulation represented in equation 1.

$$\mathcal{L}(\theta) = - \sum_{t=1}^n \log p_{\theta}(y_t | y_{<t}, H) \quad (1)$$

In the above equation, θ considered as the model parameter, and θ is learned by maximizing the likelihood of observed sequence.

4.1.1 Visual Feature Embedding

As we discussed above the input CXR image pre-process through bottom-up feature extractor and get the 36 objects. Given an image v , the visual feature obtained from bottom-up attention layer as:

$$\mathbf{v} = \{v_1, v_2, v_3, \dots, v_k\}, \quad v_i \in R^c \quad (2)$$

The location or position of features are also important factors extracted as:

$$\mathbf{x} = \{x_1, x_2, x_3, \dots, x_k\}, x_i \in R^c \quad (3)$$

here k indicates the number of visual features and c the hidden dimension size. The final visual features

$$\tilde{\mathbf{v}} = \{\tilde{v}_1, \tilde{v}_2, \tilde{v}_3, \dots, \tilde{v}_k\} \quad (4)$$

$$\tilde{v}_i = v_i + x_i + T_v \quad (5)$$

here T_v is a semantic vector shared by all visual features to differentiate them from language.

4.1.2 Text Embedding

Language or Text feature embedding is performed through similar processing as used by BERT model to encode the textual information. A given text is split into sequence of tokens using WordPiece tokenizer [55]. The tokens are then converted into a vector which is represented as:

$$\mathbf{w} = \{w_1, w_2, w_3, \dots, w_N\}, w_i \in R^d \quad (6)$$

here d is embedding dimension size. The position vector p is represented as:

$$\mathbf{p} = \{p_1, p_2, p_3, \dots, p_N\}, p_i \in R^d \quad (7)$$

The final language feature embedding is obtained as follows:

$$\tilde{\mathbf{w}} = \{\tilde{w}_1, \tilde{w}_2, \tilde{w}_3, \dots, \tilde{w}_N\} \quad (8)$$

$$\tilde{w}_i = w_i + p_i + T_L \quad (9)$$

Here T_L is a semantic vector shared by all text features to differentiate them from visual features.

4.1.3 Joint Embedding

The joint-embedding implemented after obtaining the visual and language embedding, the input sequence is constructed by concatenating them. The joint embedding uses some tokens [SEP] and [CLS] to the joint embedding block as:

$$\mathbf{H} = \{[CLS], \tilde{v}_1, \tilde{v}_2, \dots, \tilde{v}_k, [SEP]v, \tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_N, [SEP]L\} \quad (10)$$

The joint-embedding vector is used by the model to get the output results.

4.2 Bottom-Up Top-Down Approach

In modern computer vision systems, visual attention mechanisms are known as key components. These attention mechanisms are inherent part of expectational achievements in almost every field like image captioning, object detection and many more fields. The bottom-up and top-down approach takes the visual attention one step further and achieved

expectational performance on different datasets for image captioning and visual question answering. The attention mechanism is generally divided in two major groups

Detection Proposals: Faster R-CNN proposals commonly known as detection proposals. The region of interest pooling (utilizing single feature map) operation works on attention mechanism which enables the detector second stage to work only with the relevant feature. The main drawback of this approach is that these proposals are not able to use any information other than that, which may be extremely useful for better classification.

Global Attention: The entire feature map re-weight in global attention which learned according to attention heat maps. But the major drawback in this approach is the no usage of information of the object in attention map generation in the image.

Bottom-up and top-down approach consider these approaches and overcome their drawbacks. This approach neglects the global feature mapping and generate the attention maps by region proposed network. This difference in detection mechanism is shown in diagram.

The bottom-up and top-down approach Faster R-CNN generate the 36 top proposals and 2048 feature map generated by region of interest pooling. The pooled feature map averaged in a single feature map. The single feature map input into an LSTM attention and gets and output a vector of size 36. The feature map is calculated in next step after summing all the feature maps in pool. The feature maps used an input for next network which perform the task. In this approach the LSTM generate the image captions for an input. This method of attention is extremely useful for many domains. The other models using only top-down attention only aggregates the features from all layers in the image and use average weight for weighting. The bottom-up attention use more modified version of Faster R-CNN to predict the object class and attribute class from dataset.

4.3 VisualBERT

Bidirectional Encoder Representations from Transformer (Devlin et al) is a form of transformer with subwords and objective of language modelling. Subwords (input) mapped to a set (E) taken from three parts token embedding, position embedding and segment embedding. The position embedding highlights the position of token and segment embedding specifies the segment pair. This input passed through transformers layers which is build up with contextualized representation of subwords. The model trained in two step one is pre-training and other is fine tuning. Pre-training segment include masked language modelling (MLM) and next sentence prediction. In the MLM, some parts are randomly replaced with special tokens

[MASK] while others remains same, in next sentence prediction two consecutive sentences identified through model. The model is fine tuned for a specific task from pre-trained parameters and task specific objectives gained.

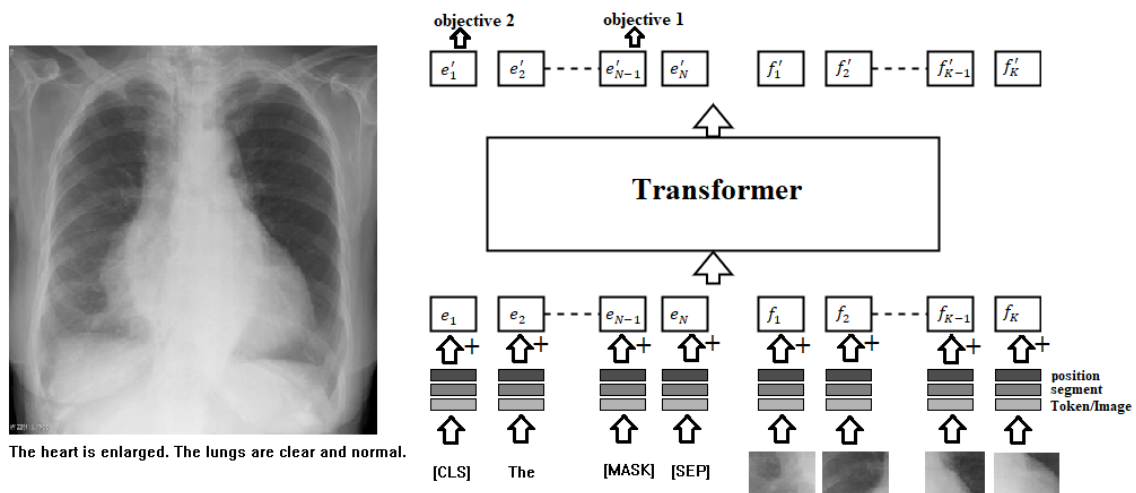


Figure 4.1 VisualBERT Architecture

A simple model for a vision and language task with more flexible framework. VisualBERT consist of transformers which align elements of an input image regions and text through self attention. The mechanism is based on self attention within the transformer. A visual embedding set introduced in this model along with all the components of BERT model for image modelling. For every $f \in F$ there is a bounding region in the image, taken from object detector. Each visual embedding computed from multiple embeddings firstly from a visual feature representation of bounding region which is computed by CNN represented as f_0 , secondly a segment embedding of image represented as f_s and thirdly a position embedding for alignment between the bounding regions and words which are input and set for the sum in position embedding according to align words (f_p). The multi-layer transformer input with visual embedding and set of text embedding for model to find useful information and relation among the sets and new representation is developed.

4.3.1 Training

VisualBERT model training is very much similar to BERT, but a difference is visual and language combine learning. The training procedure has three phases

- **Task-Agnostic:** VisualBERT train with two objective one with MLM with the image and some text parts as input, to predict the mask word keeping their respective image region without masking. There are multiple captions associated with a single image.

There may be one caption which represent the image well then other. The model trained to identify such captions which are more associated with the image and neglect the others.

- **Task-Specific:** A downstream task to train model using the data with MLM for image objective. A new domain to target as compared to previous one.
- **Fine Tuning:** VisualBERT and BERT model has same fine-tuning step, based on specific task along with its inputs and outputs and the objectives all introduced and trained to get maximum performance from the transformer.

4.4 LXMERT

Vision and text reasoning demands an understanding of language semantics and visual concepts and relationship among them. Learning Cross-Modality Encoder Representations from Transformers commonly known as LXMERT a framework which learn the relations between these vision and language modalities. The model is based on three encoders:

- (1) A language encoder
- (2) An object relation encoder
- (3) A cross modal encoder

In this framework, a large-scale transformer is build using these encoders. The model has the capability to connect vision and language semantics, and pretrain the model with huge dataset containing the image and sentence pairs. In making model ability to link the semantics (vision-language) based on five pre-training assignments to learn the connection not only among them but also in cross modal relations:

- (1) Predicting mask object (label classification)
- (2) Predicting mask object (feature regression)
- (3) Masking Language model
- (4) Cross modal matching
- (5) Question Answering (image based)

Many researchers work on developing a model for visual understanding and show their effectiveness on large vision datasets. The models designed are backbone for understanding the different task using pretraining and fine tuning of models. LXMERT model is develop and inspired from bidirectional encoder representation transformer popularity. Cross modal alignment between language and vision learned through LXMERT which is different from BERT single modality masked language modelling. The masked features predicted through multi-modality pre-training.

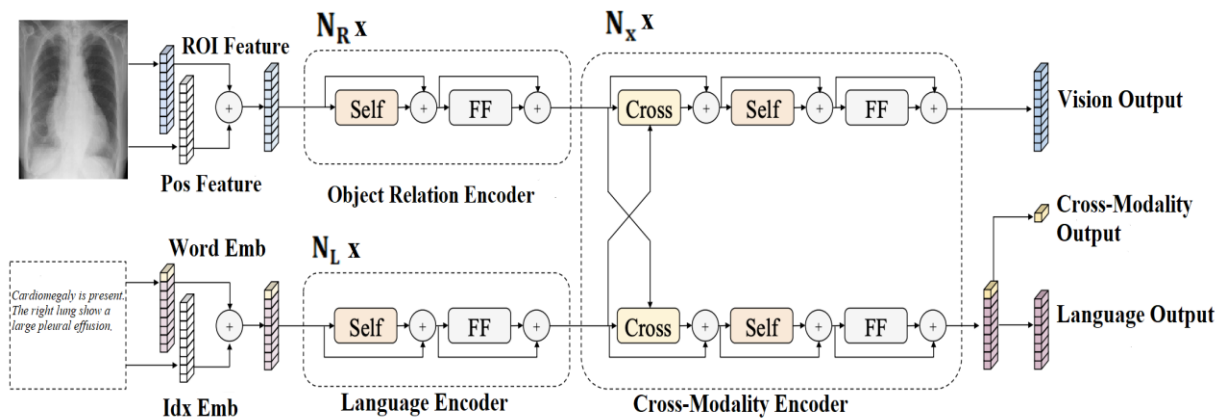


Figure 4.2 LXMERT Architecture

A cross- modality model with both cross and self attention mechanism work for natural language processing. The image and text both input into the model and pass through the attention layers to answer the question and representations from the given input. The LXMERT architecture is shown in figure 4.2. The architecture is based upon embedding (input), object relation encoder, language encoder, cross-modality encoder, and outputs.

4.4.1 Embedding

The image and text input into the model which pass through embedding layers and converts the inputs into a sequence of features. The embeddings features are of two type one is word-level embedding of sentence and second is object level embedding of image. After getting these embedding features then passed to next encoding layers.

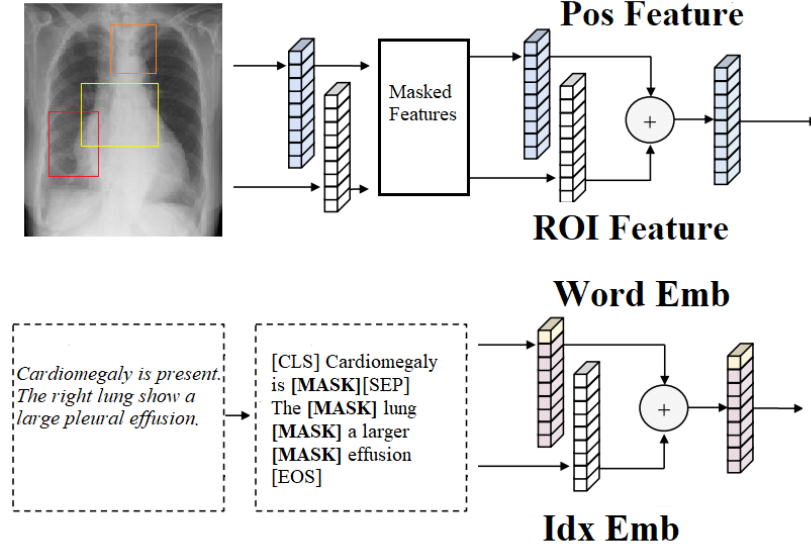


Figure 4.3 LXMERT Embedding

4.4.1.1 Word Embedding

The language or text input are sentences which are divided or split into words. The words are split in length ‘n’ of words x_1, x_2, \dots, x_n through word tokenizer. The words along with index are projected to vectors. These words are projected by embedding layer which then added to index aware embedding.

$$\hat{x}_i = WEmbed(x_i) \quad (11)$$

$$\hat{y}_i = IndexEmbed(i) \quad (12)$$

$$z_i = LayerN(\hat{x}_i + \hat{y}_i) \quad (13)$$

Here \hat{x}_i is the word and i is the index and \hat{y}_i represent the index embedding.

4.4.1.2 Image Embedding

The image objects are represented by the position features named as bounding box and its 2048-d dimensional region of interest feature. The convolutional network feature map is not used in finding the features of objects in image embedding. The object detection finds n objects y_1, y_2, \dots, y_n , region of interest features \hat{a}_j and bounding box \hat{p}_j from the image, e_j represents the addition of two fully connected layer.

$$\hat{a}_j = LayerNorm(W_F a_j + b_F) \quad (14)$$

$$\hat{p}_j = LayerNorm(W_P p_j + b_P) \quad (15)$$

$$e_j = (\hat{a}_j + \hat{p}_j) \quad (16)$$

For masking object prediction task both spatial and positional information is necessary. The image embedding and attention layer both are agnostic to absolute indices of inputs, so the order of object is not specified.

4.4.2 Encoders

The encoder used in LXMERT are three, one is language encoder and the second one is object relationship encoder and the last is the cross-modality encoder. These encoders are built on self attention and cross attention layers.

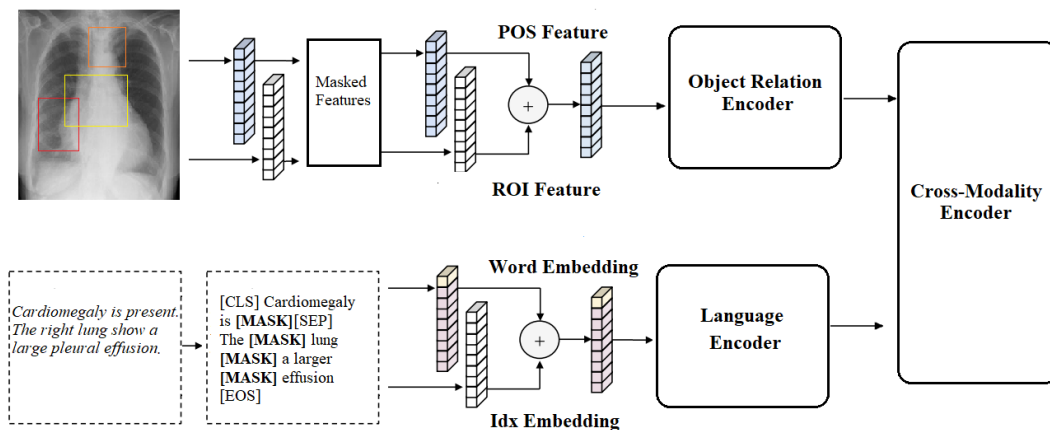


Figure 4.4 LXMERT Encoders

4.4.2.1 Language and Object Relation Encoder

The inputs passing through embedding layers passed through transformer encoders, both work on single modality. In comparison to BERT model which only use language transformer encoders, LXMERT applied to both vision and language task.

4.4.2.2 Cross Modality Encoder

In the cross modality encoder, each layer consist of one bi-direction cross attention, two self attention and two feed forward sub-layers. In the last layer the bi-direction cross attention sub layer is applied first containing the two unidirectional (vision to language and language to vision) cross attention sublayer. The context vectors and query are outputs of the layer. In cross attention the information is exchange and two modalities are aligned in order to learn joint cross modality.

4.4.3 Pre-Training & Fine Tuning

LXMERT model pre-train on different modality (cross) pre-training tasks for better initialization and vision and language connections. The masked language model is similar to

BERT model which randomly masked words with probability of 0.15 and the model predict the masked words. As compared to BERT the model predicts the masked words from text and image using cross modality model architecture. This can help in building the relations and connections from language to vision and vision to language modality. The masked object prediction pre-train by masking (ROI features) objects randomly. The model predicts the properties of masked objects and work similarly to the masked language model, predict from image to text and text to image (cross modality). The region of interest feature regression and label classification task are performed for cross modality specification. The regression of object features and detected label masked the objects with cross entropy loss.

Fine tuning is robust, and few changes required in model for specific task. A learning rate of $1e^{-5}/5e^{-5}$, a batch size of 32 and 4 epochs to fine-tune the model.

4.5 UNITER

In human life multimodal learning is generally present. The human being analyses different content in different ways, whether through text, audio, or images (visualization) etc. The learning through multiple modes give human an experience. Multiple modes combination for a learning is simple example to explain multiple modalities interaction. The researcher inspired from this, started work on machine learning for understanding multiple modes on datasets and task. In the start, multimodal tasking just works for image and audio task but later on focus different modalities for task like Visual question answering, Image text retrieval and reasoning for text, images, and audio. Before multimodal tasking, unimodal tasking was explored and good representation extracted from modalities (Word2Vec and GloVe used to generate embeddings for text and for image features extraction convolutional neural network). The major problem is using unimodal features in multimodal tasking. Many researchers tried to resolve this problem and multiple solutions of join representation also available but they all work for a specific task and no one focus on generalized model for different task.

UNITER model came with all the solution and more generalized joint embedding for images and text. It works on self-supervised learning similar to BERT which ensured the embedding is more generic. The image and text jointly trained in this model with millions of pairs of datasets. The model architecture is shown in figure 4.5. The model using WordPiece tokenized sentence along with word position information to extract all text features and Faster R-CNN along location information to extract the image features. After extracting all the features, passed through multiple layers of transformers to learn the join embedding.

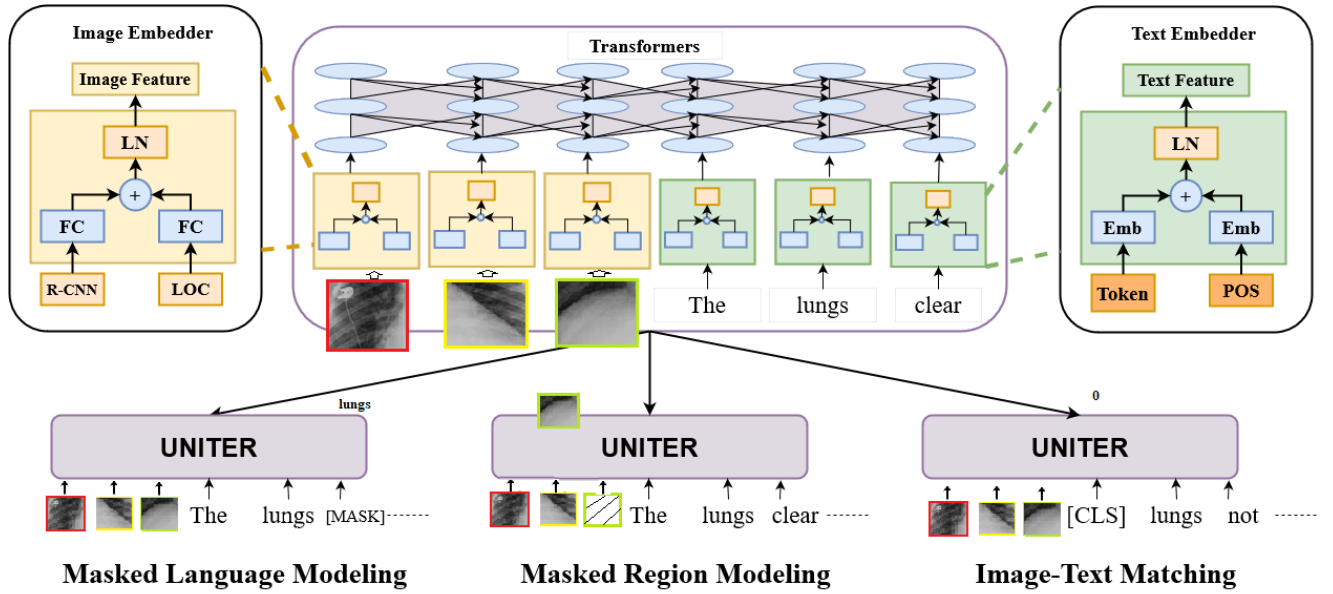


Figure 4.5 UNITER Architecture

In self-supervised training, the main objectives. These task have input (x, r) where x is the paired text and r is image region and a trainable parameter θ .

4.5.1 Masked Language Modeling

In masked language modeling the input text is masked with a token [MASK], and the probability of masking is about 15%. The main objective is to predict the masked words based upon the image regions and sentence. The masking is done by reducing the negative log likelihood as given in equation where x_m is the masked word and $x \setminus m$ are surroundings words.

$$\mathcal{L}_{MLM}(\theta) = -E_{(x,r) \sim D} \text{Log} P_{\theta}(x_m | x \setminus m, r) \quad (17)$$

4.5.2 Image Text Matching

In Image-Text matching the token [CLS] is appended in the beginning of the text input and it captures the joint context of input text and image, similar to BERT model [CLS]. This token is used to measure the score of the image and text matching level. A binary cross entropy loss used to optimize this positive negative input pairs.

$$\mathcal{L}_{ILM}(\theta) = -E_{(x,r) \sim D} [y \log s_{\theta}(r, x) + (1 - y) \log(1 - s_{\theta}(r, x))] \quad (18)$$

Here the value of y is set to be 1 for positive pairs and zero for negative pairs.

4.5.3 Masked Region Modeling

In masked region modeling the regions are filled with zeros instead of [MASK] and the image features are continuous making impossible to max the log likelihood. Therefore, in UNITER masked region modeling has three different objective functions which fit in normal equation.

$$\mathcal{L}_{MRM}(\theta) = E_{(x,r) \sim D} f_{\theta}(r_m | r \setminus m, x) \quad (19)$$

Similar to masked language modeling here r_m is the masked image regions and $r \setminus m$ are surrounding regions.

4.6 Comparison of Pre-trained Models

As we discussed and implement multiple pre-trained models to show the importance of using them as compared to CNN-RNN based models. Here is just a brief comparison of all 3 pre-trained models. Initial processing is similar in all, but the only difference is number of transformer (BERT) layers. All pre-trained model uses the same input embedding scheme but the way they use in layer to get the image text joint embedding to generate the required outputs.

Table 4.1: Comparison of Pre-Trained Models

Model	Pre-Training Task	Text Embedding	Visual Embedding	Transformer Streams	BERT Layers
VisualBERT	MLM, ITM, Task specific Pre-Training	WordPiece	BUTD	Single	12
LXMERT	MRM, MLM, IQA, ITM	WordPiece	BUTD	Double	9 (language) 5 (vision) 5 (cross)
UNITER	WRA, ITM, MLM, MRM	WordPiece	BUTD	Single	12

4.7 Visual Question Answering

In natural language processing and computer vision one of the most interested problems is visual question answering (VQA) which gathered huge amount of interest from natural language processing, computer vision and deep learning. In visual question answering the algorithm is to answer the text-based question about the image provided. The questions can be of any kind depending on the problem in computer vision.

- Attribute classification
- Object detection & recognition
- Object Count
- Scene based classification

The complex problems like relationship between image objects and text can be asked through this algorithm. The information extracted from the image based upon the text provided are in details from the whole image. VQA model is capable of solving wide range of problems in computer vision.

4.7 GRU Decoder

GRU is very famous variant of Recurrent Neural Network which is commonly known as improved LSTM. The GRU decoder perform similarly to LSTM model, but GRU is much cheaper in computation as compared to LSTM. The GRU decoder architecture is shown in figure 4.6. At every time step a decoder output the probability over the target based upon the previous generated word. The probabilities are generated by encoder annotation H , previous generated word y_{t-1} and internal state s_{t-1} .

$$y_t = f_{dec}(y_{t-1}, s_{t-1}, H) \quad (20)$$

Decoder function f_{dec} consist of two things first is the conditional GRU and second is the bottleneck function.

$$\acute{s}_t = f_{gru}(y_t, s_{t-1}) \quad (21)$$

$$c_t = f_{att}(\acute{s}_t, H) \quad (22)$$

$$s_t = f_{gru}(\acute{s}_t, c_t) \quad (23)$$

The bottleneck function project the conditional GRU output into probabilities over the target vocabulary.

$$b_t = \tanh(W^{bot}[s_t, c_t]) \quad (24)$$

$$y_t = \text{softmax}(W^{proj}b_t) \quad (25)$$

Here W is the mapping matrix. The y_t project the probabilities of words in sequence. GRU connections are shown in [13]. The proposed model used transfer learning approach to train architecture, to efficiently extract the features from the input images (CXR).

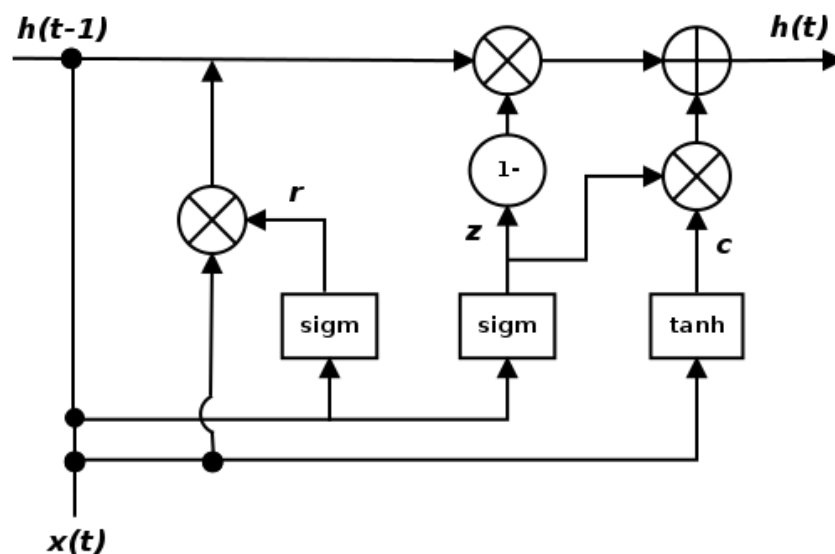


Figure 4.6 Gated Recurrent Units (GRU) Network Architecture

4.8 Training

The model is trained for two different results to show the importance of using transformer based pre-trained models. The initial processing is same for all the models, but the end results are different because two different ending gave different outputs. The classifier head added on encoded part of pre-trained model used to get the thoracic findings from the model while the GRU based decoder added to get the summarized textual report from the inputs. The step-by-step working is shown in chapter 5 while the complete processing is shown in figure 4.7. The model fine-tuned for 6 epochs and weights are added, including a batch size of 16 while the learning rate is 5e-4.

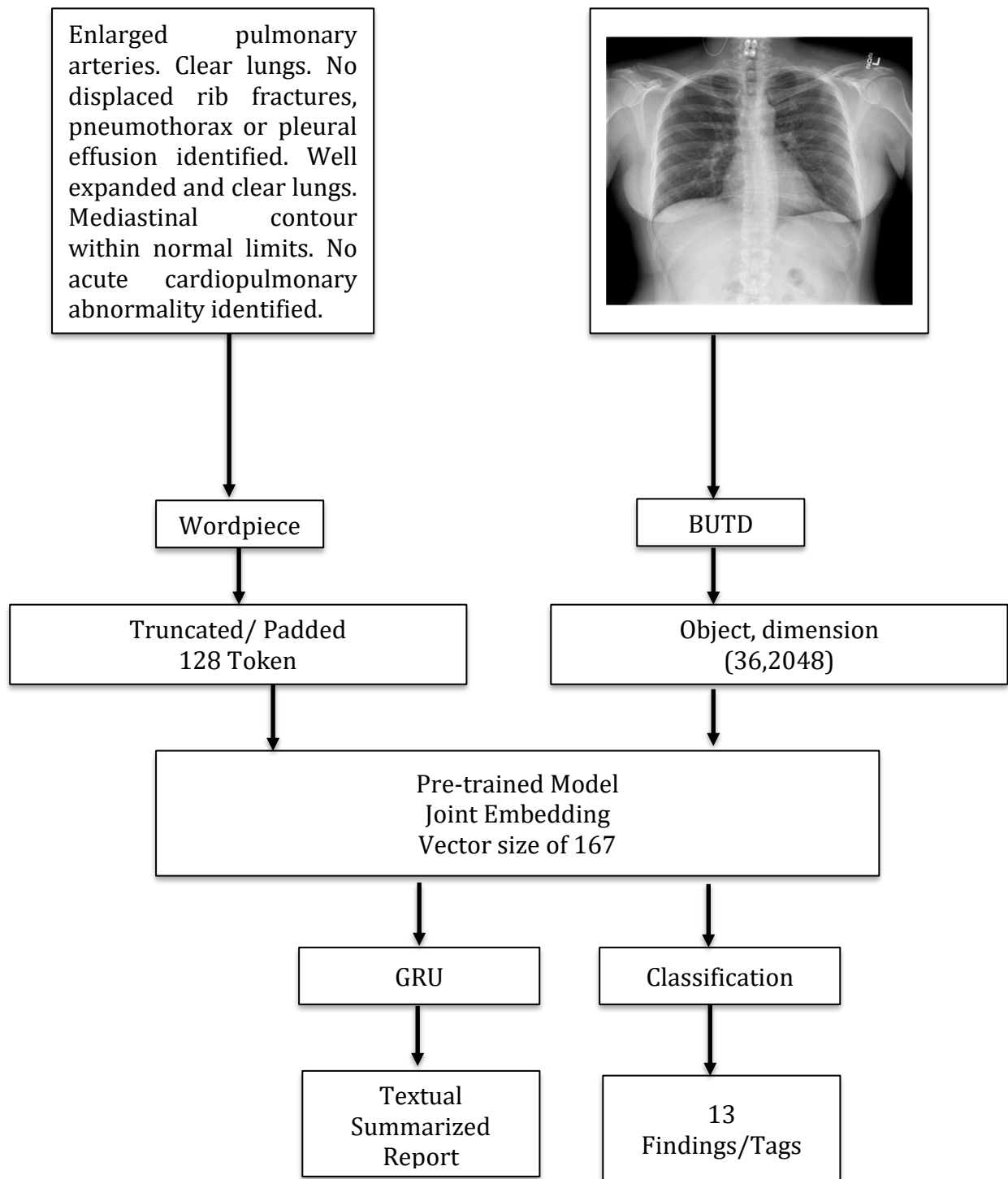


Figure 4.7 A complete and combined model of pre-trained encoder and followed by Classifier and GRU to get required outputs.

CHAPTER 5: EXPERIMENTAL RESULTS

We performed a systematic series of studies to test proposed model's efficacy by comparing previously developed models as well as with the help of metrics like ROUGE and BLEU score.

5.1 Evaluation Metrics

As discussed earlier that, describing a chest x-ray is a major challenge and without an expert radiologist the description is almost impossible. The radiologist is too busy in writing the reports so they have less time to convert these radiographs and reports into a summarized results for new doctors. There are some skills which are needed to interpret an x-ray. One need to acquire knowledge about basic physiology of chest diseases, natural structure of thorax, understanding of the changes in radiographs over time and analyzing the radiographs through fixed pattern. Also, the writing may contain some errors that can be harmful for patients. Therefore, a novel method for generating a summarized results which are similar to experts generated results. The best evaluation for text generation or summarization part is human evaluation because it is more efficient. The existing research used many different evaluation matrices to evaluate the performance of proposed model. To measure the performance one of the techniques is human evaluation. But this is impossible in our case as we discussed the issues. There are basically two methods used for evaluation. The first one is widely used metric on sentence generation evaluation is Bilingual Evaluation Understudy (BLEU) score [16]. The second method used for evaluation are best for summarized results named as ROUGE score [19]. The AUC evaluation metric used for thoracic findings/tags.

5.1.1 BLEU Score

Bilingual Evaluation Understudy score is one of the ways to find good sentence translation. In this technique the original and auto-generated sentence compared to find a relation. The original sentence is also named as reference sentence while generated sentence is termed as candidate sentence. This process used to evaluate the text generated after processing. With a perfect match between the candidate and reference sentence the max score (1.0) is given by BLEU and when a complete mismatch happen then min score (0.0) is given by sentence evaluation metric. The BLEU score lies between 0.0 to 1.0 in natural language processing tasks. The algorithm is not completely beneficial but there are some advantages of using it which are as follow:

- The algorithm is simple, fast and inexpensive.
- The algorithm does not depend on language.
- This is adopted most widely
- Algorithm is much similar to human evaluation.

This technique works by counting the n number of grams in the reference sentence to the n number of grams in the candidate sentence, where 1 gram is said to be unigram that is actually a token or word and same as it is bigram comparison is the comparison of each word pair. This comparison is without considering the order of words. Matching n-grams counting is changed to confirm that it takes the no of occurrences of the words in reference sentence taking into account, by not rewarding a candidate translation that generates a plenty of reasonable words. This thing is called the modified n-grams precision.

This algorithm is for comparing candidate sentences and reference sentences. But at the same time modified versions that regularize n-grams by their number of occurrences is also proposed for more and more better results using multiple sentences. In practice, a perfect score is not possible because in this case the automatic generated text and reference text must be same which is difficult. This is even not possible when humans are evaluating in comparison of reference sentences. The quantity and performance of the references used to measure the BLEU score ensures it may be difficult to evaluate ratings through datasets. In addition to translation, we may use the BLEU score using deep learning approaches for other language generation problems.

- Generation of one language to another
- Generation of image Captioning
- Speech recognition.
- Summarization of text

The model accuracy can be measured by BLEU score with four types: BLEU-1, BLEU-2, BLEU-3 and BLEU-4. The only difference is the use of grams words comparison. As we found best results for our proposed model on BLEU-2. Other than that cumulative weights have been used since they give better outputs. Adam optimizer [39] is used for parameter learning. Researchers are focusing on this subject have identified other metrics that are considered more relevant for medical report assessment. So far, we focused on one metric but hoping other may come in future debate.

5.1.2 ROUGE Score

Recall-Oriented Understudy for Gisting Evaluation is a very famous metric for evaluating the summarized results generated automatically using machine translation. In natural language processing, the metrics compare the proposed summary with the translation against the human-generated summary. ROUGE-N score evaluate the summary by overlapping the n-grams between the reference and candidate summary. The n-grams can be 1,2,3 means unigram, bigram etc. ROUGE-L score is used for evaluating longest common sequence in the results.

The main reasons of using ROUGE score in our methodology evaluation are as following:

- ROUGE score is the most widely used set of metrics for evaluating automatic summarization of texts.
- It works by comparing an automatically produced summary or translation against a set of reference summaries.
- Candidate sentence is the autogenerated summary.
- Reference sentence is the ground truth or the original impression

5.3 Dataset

The Indiana University (IU) Chest X-Ray Collection of team lead by Demner-Fushman is used to evaluate proposed model that is also publicly available. In this dataset the chest x-rays and their corresponding reports both are available. In some cases, there are more than one report associated with a image. The IU dataset contain 3,684 radiology reports pairs from different health networks within the archive of the Indiana Network for Medical Care. The IU dataset contain a total of 7,470 related chest x-rays [17]. This means that for almost each report relate to two X-Rays of patient. These CXR are both frontal and lateral views of chest. There are multiple areas termed as impression, findings, comparison, and indication in the dataset. In this research, we used the impression or summarized reports of doctor as the target summarized reports to be generated.

Starting from pre-processing the data and discard all the images and associated reports which has no frontal view. The pre-processing step also converts the data of long findings of doctors to a short report. There is now more the one report associated with each X-Ray. Beside this all the tokens in the reports to coveted to lowercases, removing all the tokens which are not alphabet those results in 765 unique words. In the next step the findings are converted in the form of labels or annotations of thoracic findings based on MeSH indexing. The conversion of reports into labels are successfully implemented in [35]. The raw data is converted into a

dataframe, and the idea was taken from [36] where in TieNet model used the annotations for classification task.

5.2 Quantitative Results

We report the results of thoracic findings/tags using the Accuracy under the curve (AUC). The task has two parts one is based on the prediction or classification from visual question answering using pre-trained model which classified the 13 thoracic findings in the image. The other part is based on GRU based decoder model to generate the summarized report part from the input data and evaluate on BLEU and ROUGE scores. The performance is shown in Table 5.1 and Table 5.2 which is much better than all other mentioned networks. The contrast between these models indicates explicitly how powerful proposed modeling technique from pre-trained models. Through multiple pre-trained models the prediction analysis is also different, VisualBERT performance is much better than LXMERT and UNITER and the BLEU and ROUGE score for each pre-trained findings are also shown.

Table 5.1: AUCs for 13 thoracic findings on IU CXR from Pre-Trained and TieNet Model

Findings	Pre-Trained Model (Image+Report)			Model (Image+Report)
	VisualBERT	LXMERT	UNITER	TieNet [5]
Atelectasis	0.988	0.989	0.982	0.976
Cardiomegaly	0.991	0.980	0.978	0.962
Effusion	0.993	0.983	0.992	0.977
Infiltration	0.973	0.972	0.972	0.984
Mass	0.961	0.996	0.966	0.903
Nodule	0.967	0.966	0.966	0.960
Pneumonia	0.987	0.990	0.986	0.994
Pneumothorax	0.992	0.958	0.983	0.960
Consolidation	0.989	0.993	0.998	0.989
Edema	0.991	0.995	0.989	0.995
Emphysema	0.968	0.960	0.968	0.868
Fibrosis	0.993	0.990	0.994	0.960
Pleural Thickening	0.980	0.01	0.979	0.953
Average	0.985	0.984	0.981	0.965

They're proposed model work on image-text joint embedding which also used by TieNet but the difference is the pre-trained models used for captioning. The results for thoracic findings

for pre-trained models are much better than the TieNet model which using CNN based encoder technique which was discussed in chapter 3. The results clearly shows that the findings from image-text joint embedding are much better with pre-trained models and especially in VisualBERT model. Model was trained on Google Colab which provides the 1x NIVIDIA Tesla K80 GPU with 12GB GDDR5 VRAM.

Below figure 5.1 and 5.2 showing the accuracy of our best model VisualBERT training loss and Test accuracy.

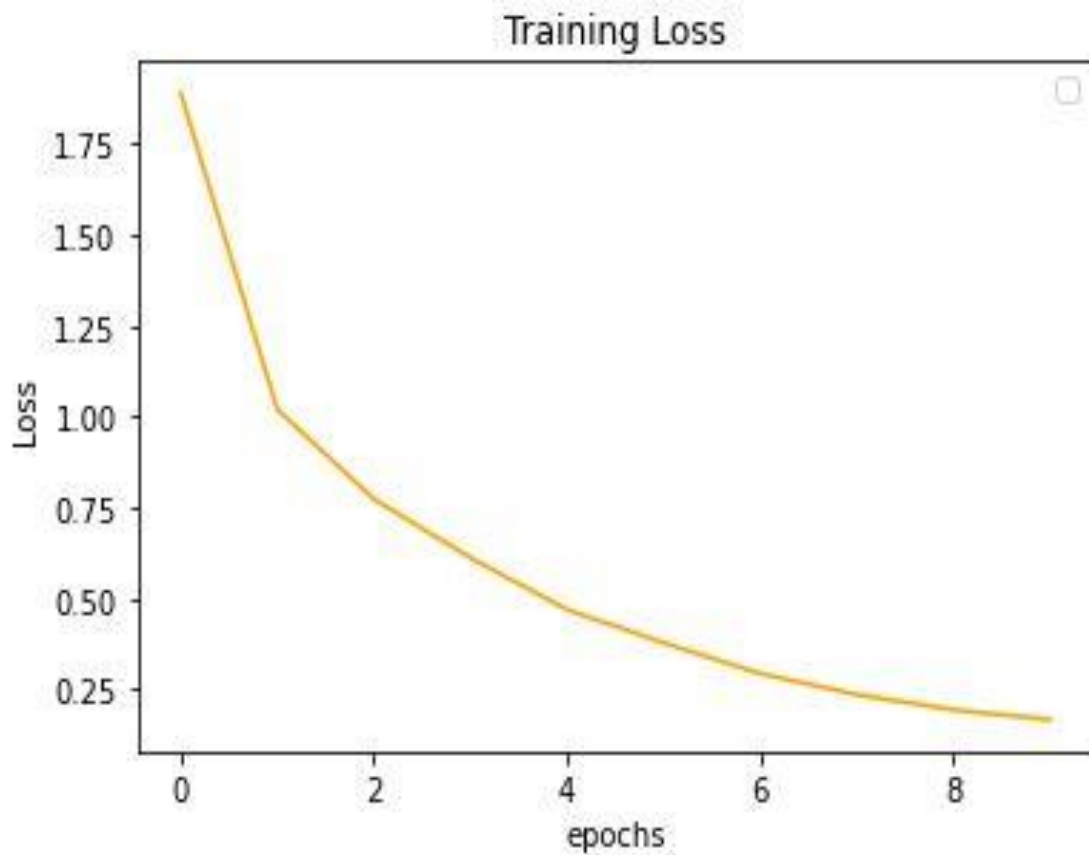


Figure 5.1 Training Loss for Proposed Methodology

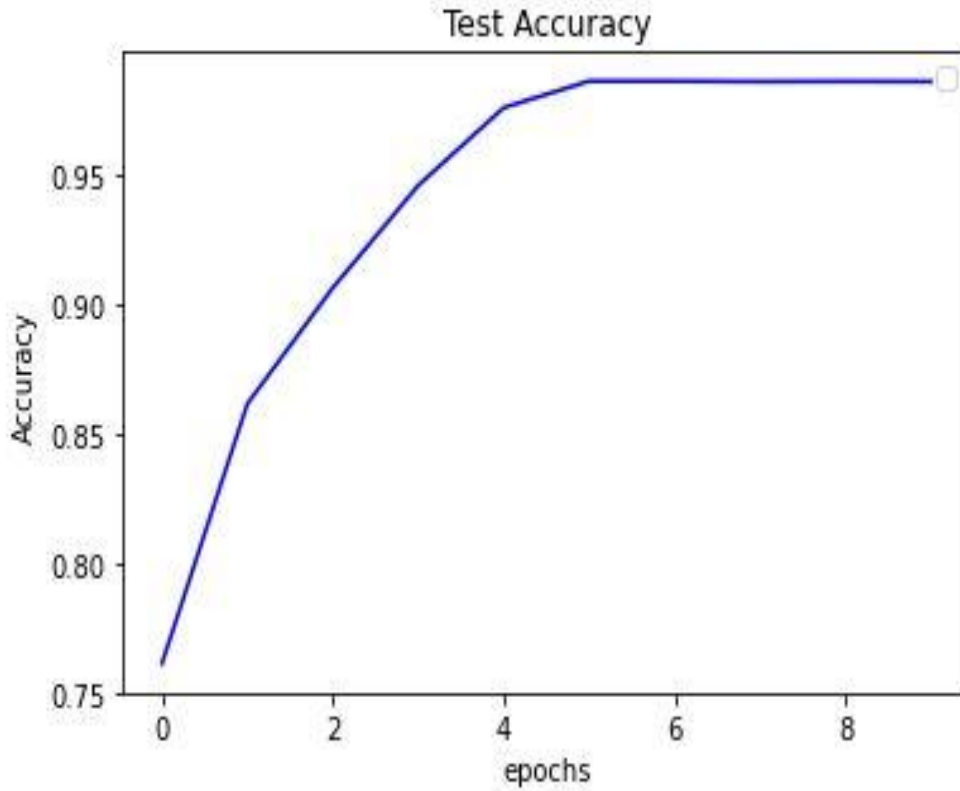


Figure 5.2 Test Accuracy of Model



The summarization parts results are shown in table 5.2 with comparison to multiple models which are used for summarization of radiology reports.

Table 5.2 Results of ROUGE and BLEU score for summarized medical report on the IU CXR dataset using proposed methodology

Methods	ROUGE	BLEU
Pointer Generator [51]	0.33	0.226
Pointer Generator + RL [52]	0.58	0.229
BiGRU [56]	0.54	0.29
CNN-GRU [54]	0.32	0.243
VisualBERT+ GRU (proposed)	0.65	0.35
LXMERT+ GRU (proposed)	0.53	0.33
UNITER + GRU (proposed)	0.58	0.30

Some qualitative results of medical report generation using proposed methodology is shown in Table 5.3. First column contains the Radiographs, second contain the ground truth or findings by doctors and 3rd one contains the results generated by proposed pre-trained model.

Table 5.3: Some qualitative results using multiple pre-trained model

Chest X-RAY		Results	Findings/Tags
			Normal
Radiology Report	Heart size and mediastinal contour are within normal limits. There is no focal airspace consolidation or suspicious pulmonary opacity. No pneumothorax or large pleural effusion. Mild degenerative change of the thoracic spine.	Summarized Report	No acute cardiopulmonary abnormality.
Chest X-RAY		Results	Findings/Tags
			Atelectasis, Cardiomegaly
Radiology Report	The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, pneumothorax, or pleural effusion. Cardio mediastinal silhouette is unremarkable. There is cardiopulmonary abnormality.	Summarized Report	There is a calcified opacity in the left lung

As we learned from the proposed methodology the accuracy score of ROUGE is much better than BLEU because ROUGE are more helpful in summary evaluation but just for the sake of comparison, we implemented both method. The AUC of some findings are more accurate in

different models but the overall average results shown by VisualBERT is clearly demonstrating that its task specific pre-training capability helps in getting more accurate results. The Table 5.3 showing the different performance output of proposed methodology. In each case the input are the text and image which work as joint embedding framework and out the required results based upon the header/decoders. This, clearly showing that the joint embedding is much more helpful as compared to text and image only input and using a BERT type pre-trained model helps in improving the accuracy of model in both cases.

In Table 5.4, some best case, worst case and average results of different pre-trained models are shown, as we can see that the VisualBERT best case gave us much more accurate results as compared to LXMERT or UNITER base model implemented for report summarization part. But their accuracy is again good for only ROUGE score while in BLEU is results are not much satisfactory.

Table 5.4: Best, Worst and Average case ROUGE score of different pre-trained model for summarization

Model	ROUGE Score		
	Best	Worst	Average
VisualBERT+ GRU (proposed)	0.98	0.30	0.65
LXMERT+ GRU (proposed)	0.93	0.19	0.53
UNITER + GRU (proposed)	0.95	0.23	0.58

Table 5.5: Best, Worst and Average case BLEU score of different pre-trained model for summarization

Model	BLEU Score		
	Best	Worst	Average
VisualBERT+ GRU (proposed)	0.62	0.18	0.35
LXMERT+ GRU (proposed)	0.50	0.10	0.33
UNITER + GRU (proposed)	0.52	0.09	0.30

From the results, one thing is clear that using of joint embedding for text summarization and classification of diseases are much better as compared to previous text or image only methods. Also, the use of transformers based pre-trained models which are initially trained on larger dataset can also be reliable on smaller dataset performance. The results of best case gained from text summarization part clearly representing that someone may get more better results

using a different transformer based pre-trained model. The results proving our stance that using of pre-trained models along with joint-embedding technique give a great performance imporvemnts.

CHAPTER 6: CONCLUSION & FUTURE WORK

6.1 Conclusion

The proposed model of joint embedding is to create summarized reports for chest x-rays and original detailed reports, to assist medical field in writing summarized reports more efficiently and effectively for inexperienced doctors along with thoracic findings/tags. It is based upon a pre-trained model like UNITER, LXMERT, VisualBERT and bottom-up attention using detectron2 implemented for feature segment extraction converts an image into a segments of objects vector along with the positional information of the most likely objects, and a text part pass through the wordpiece tokenizer to extract the text segments and positions, then followed by these pre-trained models that generate corresponding findings/tags from images and sentences adding a classifier head on top while the summarized results generate from adding a GRU based decoder on these pre-trained model encoders. The model performance and effectiveness are analyzed both quantitatively and qualitatively on the Indiana University Dataset. For the comparison multiple methods has been presented to check the influence of different components on the medical report summarization. As the multiple pre-trained models are analyzed to demonstrate the different cases of proposed method. The results show that pre-trained models generally work slightly better than modeling from zero like pointer generator models also taking less time for training as well as for sentence summarization. The performance may also increase when more advanced datasets used which are bigger and by training on a greater number of images. The analysis of multiple pre-trained models experiments on IU Dataset validate the effectiveness of the proposed architecture using joint embedding.

6.2 Contribution

- An automatic summarized report generation system from vision and language joint embedding using pre-trained models.
- The complete model implementation using pre-trained models to extract findings/tags and generate summarized reports with findings.
- Experiments on Indiana University dataset using multiple pre-trained models to demonstrate the significance of the proposed methodology.
- Review & comparison of recent developments in automated summarization of reports

6.3 Future Work

In this research we used the Indiana University (IU) Chest X-Ray to classify the diseases present in the image and text. There are many other datasets available. We can extend this research by carrying out experiments using many other datasets. We perform our experiments using the architecture that is the combination of LXMERT, VisualBERT and UNITER for findings along with LSTM for report generations. Someone can extend this research by experimenting using different architectures.

REFERENCES

- [1] Louke Delrue, Robert Gosselin, Bart Ilsen, An Van Landeghem, Johan de Mey, and Philippe Duyck. Difficulties in the interpretation of chest radiography. In *Comparative Interpretation of CT and Standard Radiography of the Chest*, pages 27–49. Springer, 2011
- [2] Brady, A., Laoide, R.O., McCarthy, P., McDermott, R.: Discrepancy and error in radiology: concepts, causes and consequences. *Ulster Med. J.* **81**(1), 3 (2012).
- [3] S. Candemir, S. Rajaraman, G. Thoma and S. Antani, "Deep Learning for Grading Cardiomegaly Severity in Chest X-Rays: An Investigation," 2018 IEEE Life Sciences Conference (LSC), 2018, pp. 109-113, doi: 10.1109/LSC.2018.8572113.
- [4] S. Singh, S. Karimi, K. Ho-Shon and L. Hamey, "From Chest X-Rays to Radiology Reports: A Multimodal Machine Learning Approach," 2019 Digital Image Computing: Techniques and Applications (DICTA), 2019, pp. 1-8, doi: 10.1109/DICTA47822.2019.8945819.
- [5] Y. Zhang, D. Merck, E. Tsai, C. D. Manning, and C. Langlotz, "Optimizing the factual correctness of a summary: A study of summarizing radiology reports," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
- [6] M. Sajad, I. Shafi and J. Ahmad, "Automatic Lesion Detection in Periapical X-rays," 2019 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), 2019, pp. 1-6, doi: 10.1109/ICECCE47252.2019.8940661.
- [7] X. Huang, F. Yan, W. Xu and M. Li, "Multi-Attention and Incorporating Background Information Model for Chest X-Ray Image Report Generation," in IEEE Access, vol. 7, pp. 154808-154817, 2019, doi: 10.1109/ACCESS.2019.2947134.
- [8] Bista, N., Shrestha, P., & Maskey, S. (2019). A Study on Morphological Variations of Fissures and Lobes of Human Lungs with its Clinical Significance. *Journal of Nobel Medical College*, 8(2), 21-25.
- [9] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A Simple and Performant Baseline for Vision and Language," *arXiv [cs.CV]*, 2019.
- [10] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019
- [11] Y.-C. Chen *et al.*, "UNITER: UNiversal Image-TExt Representation Learning," in *Computer Vision – ECCV 2020*, Cham: Springer International Publishing, 2020, pp. 104–120.
- [12] A. Agrawal *et al.*, "VQA: Visual question answering: Www.Visualqa.Org," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 4–31, 2017.

- [13] D. Bahdanau, K. Cho, and Y. Ben-Gio, "Association for Computational Linguistics. Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediq 2021 shared task on summarization in the medical domain," in Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021. Association for Computational Linguistics. Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, Ann Arbor, Michigan; Doha, Qatar: Association for Computational Linguistics, 2005, pp. 1724–1734.
- [14] Yabluchansky, M., Bogun, L., Martymianova, L., Bychkova, O., Lysenko, N., & Brynza, M. (2017). Signs and symptoms of respiratory system diseases: Syndrome of Diffuse and Focal Consolidation of the Lungs.
- [15] Radiopaedia. [Online], www.radiopaedia.org, accessed on December, 2017
- [16] Ebner, L., Tall, M., Choudhury, K. R., Ly, D. L., Roos, J. E., Napel, S., & Rubin, G. D. (2017). Variations in the functional visual field for detection of lung nodules on chest computed tomography: impact of nodule size, distance, and local lung complexity. *Medical physics*, 44(7), 3483-3490.
- [17] Parkar, A. P., & Kandiah, P. (2016). Differential diagnosis of cavitory lung lesions. *Journal of the Belgian Society of Radiology*, 100(1).
- [18] Charles L. Daley, Michael B. Gotway, and Robert M. Jasmer. Curry International Tuberculosis Center.[Online],http://www.currytbcenter.ucsf.edu/sites/default/files/radiographic_complete_2nded.pdf, accessed on March, 2020
- [19] C.-Y. Lin, "ROUGE: A package for automatic evaluation summaries," in Proceedings of the Workshop on Text Summarization Branches Out, Barcelona, Spain, 2004, pp. 25–26.
- [20] S Knop et al., "A new case of Carney triad: Gastrointestinal stromal tumours and leiomyoma of the oesophagus do not show activating mutations of KIT and platelet-derived growth factor receptor alpha," in *Journal of Clinical Pathology* 59(10):1097-9, November 2006.
- [21] Ferreiro, L., Toubes, M. E., San José, M. E., Suárez-Antelo, J., Golpe, A., & Valdés, L. (2020). Advances in pleural effusion diagnostics. *Expert review of respiratory medicine*, 14(1), 51-66.
- [22] Muruganandan, S., Azzopardi, M., Thomas, R., Fitzgerald, D. B., Kuok, Y. J., Cheah, H. M., ... & Singh, B. (2020). The Pleural Effusion And Symptom Evaluation (PLEASE) study of breathlessness in patients with a symptomatic pleural effusion. *European Respiratory Journal*, 55(5).
- [23] Jany, B., & Welte, T. (2019). Pleural effusion in adults etiology, diagnosis, and treatment. *Deutsches Ärzteblatt International*, 116(21), 377.

- [24] Clare Hooper, Y C Gary Lee, and Nick Maskell, "Investigation of a unilateral pleural effusion in adults: British Thoracic Society pleural disease guideline 2010," in *Thorax* 2010;65(Suppl 2):ii4eii17, 2010
- [25] P. Anderson *et al.*, "Bottom-up and top-down attention for image captioning and visual question answering," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
- [26] Sheard, S., Moser, J., Sayer, C., Stefanidis, K., Devaraj, A., & Vlahos, I. (2018). Lung cancers associated with cystic airspaces: underrecognized features of early disease. *Radiographics*,38(3), 704-717.
- [27] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.
- [28] Encyclopædia Britannica. [Online], <https://www.britannica.com/science/heart#/media/1/258344/121131>, accessed on December 2021.
- [29] Baker D.J. (2016) The Structure of the Airways and Lungs. In: Artificial Ventilation. Springer, Cham. https://doi.org/10.1007/978-3-319-32501-9_2
- [30] Park, S., Lee, S. M., Lee, K. H., Jung, K. H., Bae, W., Choe, J., & Seo, J. B. (2020). Deep learning-based detection system for multiclass lesions on chest radiographs: comparison with observer readings. *European Radiology*, 30(3), 1359-1368.
- [31] HealthTap. [Online], https://edc2.healthtap.com/htstaging/user_answer/avatars/965837/large/open-uri20130324-3652-1dpby9f.jpeg?1386647275, accessed on January, 2020.
- [32] Colah's Blog. [Online], <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>, accessed on July, 2020
- [33] D. Yu, J. Fu, X. Tian, and T. Mei, "Multi-source multi-level attention networks for visual question answering," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 15, no. 2s, pp. 4709–4717, 2019, doi: 10.1145/3316767.
- [34] C. Gao, Q. Zhu, P. Wang, and Q. Wu, "Chop chop BERT: Visual Question Answering by chopping VisualBERT's heads," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, 2021.
- [35] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [36] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-rays," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

- [37] wikiRadiography. [Online], <http://www.wikiradiography.net/page/Interstitial+vs+Alveolar+Lung+Patterns>, accessed on July, 2020
- [38] Toyokawa, G., Yamada, Y., Tagawa, T., Kinoshita, F., Kozuma, Y., Matsubara, T., ... & Oda, Y. (2018). Significance of spread through air spaces in resected lung adenocarcinomas with lymph node metastasis. *Clinical Lung Cancer*, 19(5), 395-400.
- [39] Al-Githmi, I. (2017). Mediastinoscopy in Assessing Mediastinal Lymphadenopathy and Lung Disease. *Open Journal of Thoracic Surgery*, 7(4), 55-61.
- [40] Collu, C., Fois, A., Crivelli, P., Tidore, G., Fozza, C., Sotgiu, G., & Pirina, P. (2018). A case-report of a pulmonary tuberculosis with lymphadenopathy mimicking a lymphoma. *International Journal of Infectious Diseases*, 70, 38-41.
- [41] Radiology Key. [Online], <https://radiologykey.com/pediatric-chest/>, accessed on December, 2017
- [42] Christina C. Chang et al., "HIV and co-infections Authors," in *Immunological Reviews* 254(1):114-42 · July 2013.
- [43] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020
- [44] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghahfoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I S´anchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
- [45] Thomas Schlegl, Sebastian M Waldstein, Wolf-Dieter Vogl, Ursula Schmidt-Erfurth, and Georg Langs. Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer, 2015
- [46] Hoo-Chang Shin, Le Lu, Lauren Kim, Ari Se, Jianhua Yao, and Ronald M Summers. Interleaved Text/image deep mining on a very large-scale radiology database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1090–1099, 2015.
- [47] Xiaosong Wang, Le Lu, Hoo-chang Shin, Lauren Kim, Isabella Nogues, Jianhua Yao, and Ronald Summers. Unsupervised category discovery via looped deep pseudo-task optimization using a large scale radiology image database. *arXiv preprint arXiv:1603.07965*, 2016.
- [48] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chest-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In the *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471. IEEE, 2017.

- [49] S. MacAvaney, S. Sotudeh, A. Cohan, N. Goharian, I. Talati, and R. W. Filice, “Ontology-Aware Clinical Abstractive Summarization,” in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019.
- [50] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017
- [51] Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In The 2017 Annual Meeting of the Association of Computational Linguistics (ACL 2017).
- [52] S. MacAvaney, S. Sotudeh, A. Cohan, N. Goharian, I. Talati, and R. W. Filice, “Ontology-Aware Clinical Abstractive Summarization,” in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019.
- [53] Y. Zhang, D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz, “Learning to summarize radiology findings,” in Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, 2018.
- [54] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, “Hybrid Retrieval-Generation Reinforced Agent for medical image report generation,” arXiv [cs.CV], 2018.
- [55] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, “Fast WordPiece Tokenization,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.
- [56] A. Zaki, “Multilayer Bidirectional LSTM/GRU for text summarization made easy (tutorial 4),” HackerNoon.com, 31-Mar-2019. [Online]. Available: <https://medium.com/hackernoon/multilayer-bidirectional-lstm-gru-for-text-summarization-made-easy-tutorial-4-a63db108b44f>. [Accessed: 10-September-2021]
- [57] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225, 2017.
- [58] Qingji Guan, Yaping Huang, Zhun Zhong, Zhedong Zheng, Liang Zheng, and Yi Yang. Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification. *arXiv preprint arXiv:1801.09927*, 2018.
- [59] Pulkit Kumar, Monika Grewal, and Muktabh Mayank Srivastava. Boosted cascaded convnets for multilabel classification of thoracic diseases in chest radiographs. *arXiv preprint arXiv:1711.08760*, 2017.
- [60] Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *arXiv preprint arXiv:1803.02315*, 2018.
- [61] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology

- examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310, 2015
- [62] Alistair EW Johnson, Tom J Pollard, Seth Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-Ying Deng, Roger G Mark, and Steven Horng. Mimic cxr: A large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.
- [63] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vay' a. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*, 2019.
- [64] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *arXiv preprint arXiv:1901.07031*, 2019.
- [65] Liu, G., Hsu, T. M. H., McDermott, M., Boag, W., Weng, W. H., Szolovits, P., & Ghassemi, M. (2019). Clinically accurate chest x-ray report generation. *arXiv preprint arXiv:1904.02633*.
- [66] Imane Allaouzi, M. Ben Ahmed, B. Benamrou, and M. Ouardouz. 2018. Automatic Caption Generation for Medical Images. In *Proceedings of the 3rd International Conference on Smart City Applications (SCA '18)*. Association for Computing Machinery, New York, NY, USA, Article 86, 1–6. DOI:<https://doi.org/10.1145/3286606.3286863>
- [67] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).