MEDICAL IMAGE CAPTIONING

Dissertation Submitted
in Partial Fulfillment of the Requirements for
the Degree of

MASTER OF TECHNOLOGY
in

SIGNAL PROCESSING AND DIGITAL DESIGN

(2022-2024)

by

Harshit Yadav

2K22/SPD/05

Under the Supervision of

Dr. Dinesh Kumar

Professor, Electronics and Communication
Department, Delhi Technological University

To The

Department of Electronics and Communication Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India (Formerly Delhi College of Engineering) Bawana Road, Delhi-110042

May, 2024

ACKNOWLEDGEMENT

I want to intimate my heartfelt thanks to my project guide, Dr Dinesh Kumar Professor of Department of Electronics and Communication Engineering of Delhi Technological University, for his tremendous support and assistance base on his knowledge. I am so grateful to Dhruv Sharma for assist me with the all the necessary tools for the completion of the project. I also want to extend our heartfelt gratitude to all those who have supported our research on image captioning using Transformer. We sincerely appreciate the creators of the architectures, which have provided us with powerful tools for feature extraction and caption generation. We express our profound gratitude to National Institute of Health Chest X-Ray Dataset. I am grateful to the open-source community for developing and maintaining user-friendly deep learning frameworks like TensorFlow, PyTorch, and Kera's, and datasets for simplifying the implementation of our research. I specially feel very thankful for our parents, friends, and classmates for their support throughout my project period. Finally, I express aur feeling everyone for support me directly or indirectly in completing this project successfully. Yours support and inspiration have been truly invaluable, which helps me to encourage.

Harshit

May, 2024 **Harshit Yadav**

Delhi (India) **(2K22/SPD/05)**

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CANDIDATE'S DECLARATION

I, Harshit Yadav, 2k22/SPD/05 am a student of M.Tech. (Signal Processing and Digital Design), hereby declare that the Dissertation titled "**Medical Image Captioning**" is being submitted by me to the Department of Electronics and Communication Engineering Delhi Technological University, Delhi in partial fulfillment of the requirement for the award of the degree of Master of Technology is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma Associateship, Fellowship or other similar title or recognition.

.

Place: Delhi

Harshit

Date: 31/05/2024 **Harshit Yadav**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE BY THE SUPERVISOR

Certified that **Harshit Yadav** (2K22/SPD/05) has carried out their search work presented in this dissertation entitled **"Medical Image Captioning"** for the award of **Master of Technology** (print only that is applicable) from Department of Electronics and communication Engineering, Delhi Technological University, Delhi, under my supervision. To the best of my knowledge, the dissertation embodies results of original work, and studies are carried out by the student himself and the contents of the dissertation do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

31/5/24
Dr Dinesh Kumar

Date: 31/05/2024

(Professor)

ABSTRACT

Image caption generation is a natural language sentence generated using deep learning technology. Image captioning means generate textual caption for any image which is either capture by camera or any medical report by using deep learning or Natural language Processing (NLP) based technique. Medical image captioning contains reports for which it generates suitable caption like symptoms and diseases of patients. It is very effective and useful for physicians and for lab diagnostics. Finally generated caption is verified by evaluation matrices which check similarity between NLP generated caption and actual caption. Medical Image Captioning is used in many research fields. Medical Image captioning is used to detect symptoms of disease effectively and it take less time to lab pathologist to decode the required information. It used in many research areas for generate vaccine and new medicine for viral diseases.

LIST OF PUBLICATIONS

[1] Harshit Yadav, Dinesh Kumar “A Review on Medical Image Captioning” communicating (Under Preperation).

TABLE OF CONTENTS

PARTICULARS PAGE NO.

ACKNOWLEDGEMENT ii

CANDIDATE’S DECLARATION iii

CERTIFICATE BY SUPERVISOR iv

ABSTRACT v

LIST OF PUBLICATIONS vi

TABLE OF CONTENTS vii

LIST OF FIGURES ix

ACRONYMS x

CHAPTER 1 1

INTRODUCTION 1

1.1 INTRODUCTION TO IMAGE CAPTIONING 1

1.2 TYPES OF IMAGE CAPTIONING 2

1.3 CHALLENGES 3

1.4 SCOPE OF WORK 4

1.5 DISSERTATION ORGANIZATION 5

CHAPTER 2 6

LITERATURE SURVEY 6

2.1 RETRIEVAL BASED IMAGE CAPTIONING 7

2.2 TEMPLATE BASED IMAGE CAPTIONING 7

2.3 SHOW ATTEND AND TELL BASED IMAGE CAPTIONING 7

2.4 GENERATIVE PRE-TRAINED TRANSFORMER IMAGE CAPTIONING 9

CHAPTER 3 10

METHODS OF MEDICAL IMAGE CAPTIONING 10

3.1 MACHINE LEARNING BASED 10

3.2 DEEP LEARNING BASED 11

CHAPTER 4 13

ARCHITECTURE USEDIN IMAGE CAPTIONING 13

4.1 ENCODER DECODER ARCHITECTURE 13

4.2 CNN-RNN ARCHITECTURE 14

4.3 CNN-LSTM ARCHITECTURE 16

4.4 TRANSFORMER ARCHITECTURE 18

CHAPTER 5 22

EXPERIMENTAL RESULTS 22

5.1 DATASET PREPERATION 22

5.2 EVALUATION METRIX 23

5.3 EXPERIMENTAL RESULTS 25

CHAPTER 6 27

CONCLUSION AND FUTURE SCOPE 27

REFERENCES 28

LIST OF FIGURES

Figure 1.1: Difference between MIC and GIC 3

Figure 2.1: Model Architecture 6

Figure 2.2: Attention Module SAT 8

Figure 2.3: Architecture using SAT and GPT-3 9

Figure 3.1: Machine Learning Algorithm 10

Figure 3.2: Deep Learning Algorithm 11

Figure 4.1: Architecture Used in Image Captioning 13

Figure 4.2: Encoder Decoder Architecture 14

Figure 4.3: CNN-RNN Architecture. 16

Figure 4.4: LSTM Architecture 17

Figure 4.5: CNN -LSTM Architecture 18

Figure 4.6**:** Transformer and Efficientnetb2 Architecture 19

Figure 4.7**:** Efficientnetb2 Architecture 20

Figure 4.8**:** Transformer Architecture 21

Figure 5.1**:** Datasets used 22

Figure 6.1**:** Loss Count for Epoch. 25

Figure 6.2**:** Epoch Graphical Representation 25

Figure 6.3**:** Output of image for clear lungs 26

Figure 6.4**:** Output of image for Pulmonary 26

ACRONYMS

CNN	**:**	Convolution Neural Network
LSTM	**:**	Long Short-Term Memory
RNN	**:**	Recurrent Neural Network
VGG	**:**	Visual Geometry Group
BLEU	**:**	Bilingual Evaluation Understudy
SGD	**:**	Stochastic gradient descent
NLP	**:**	Natural Language Processing
SPICE	**:**	Semantic Propositional Image Caption Generated.
METEOR	**:**	Metric For Evaluation of Translation with Explicit Ordering
CIDER	**:**	Consensus Based Image Description Evaluation
ROUGE	**:**	Recall-Oriented Understudy for Gisting Evaluation
MIC	**:**	Medical Image Captioning
GIC	**:**	General Image Captioning
RELU	**:**	Rectified Linear Unit.

CHAPTER 1: INTRODUCTION

1.1 INTRODUCTION TO IMAGE CAPTIONING

Image captioning is the advanced methodology of providing a text description for an image, this advance method is known as image captioning. It works on the computer vision techniques with natural language processing (NLP) to enable intelligence to understand and describe visual content. Aim of image captions is to automatically generate a human-like description that accurately represents the content of the image. This task requires a deep understanding of the visual information present in image also having ability to work to generate coherent and contextually relevant text sentence.

Medical Image Captioning is advance technology of image captioning which is used to generate textual captions of medical diagnostics reports, it collects the information from image which I present in the report. This Medical Image Captioning Technology enables us to generate textual information from laboratory reports easily, which helps doctors, pathologist, and diagnostics. Medical image captioning generate result accurately from laboratory-based tests, it helps doctors and diagnostics to go for further process of treatment.

Image captioning is the method of generating textual description of pictures using, computer vision and deep learning methodology, which is used to analyze the information present in the image. This methodology is used to save time by generating caption from the lab reports, so this is very helping and useful technology in these days in medical domains. Medical image captioning is modern technique to decrease burden on doctors.

In medical domain caption generate is very useful for doctors and pathologist to analyze proper result of any type of pathology report. It focused on technique to automatically text caption generator. Image captioning applied on medical images which accelerate and help the medical research process for physicians. Which is used by diagnostics in laboratory. Medical image captioning of X-Ray images is indicating the symptoms of further investigation of normal or abnormal. Rapidly increasing the number of medical images which makes burden on pathologist and physicians so medical image captioning a modern technique to remove tremendous burden on physicians and lab pathologist.

Here our aim is to find difference between general images and medical images using chest images using Artificial Intelligence.

This image captioning area used computer vision technology and natural language processing to understand image as natural language process and analyze images. This process used several methods for generating textual description from image; Template based method, Machine learning based methods and deep learning. This paper is a short review of medical images based on different methods using different datasets by using evaluation matrix.

1.2 TYPES OF IMAGE CAPTIONING

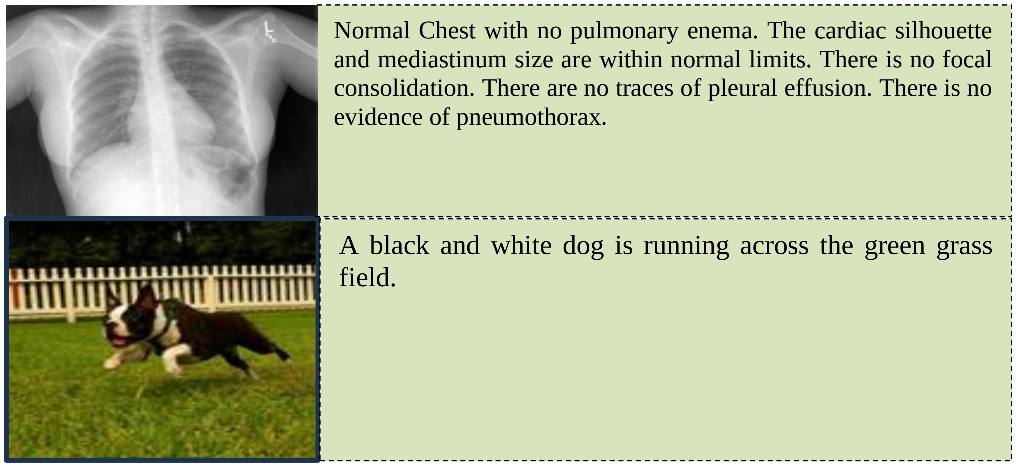
THERE ARE TWO TYPES OF IMAGE CAPTIONING*:*

1) GENERAL IMAGE CAPTIONING.

General image captioning generally focused on daily life images, and generate textual descriptions. General Image captioning consists of VGG-16, Resnet, inception, and transformer based Deep learning model for image to text generator. General Image Captioning is the research process which helps to automatically generating textual description from image by using deep learning methodology. General image captions are generating using CNN-LSTM, Transformer based on deep learning methodology.

2) MEDICAL IMAGE CAPTIONING

Medical image captioning is modern diagnostic workflow in medical domain. Medical image captioning specially focused on X-rays, MRI, ECG, CT scan and fundus images. By using Deep learning Methodology. Medical image captioning is used to collect textual data of medical images using artificial intelligence. X-Ray is a basic routine checkup for analyze the health symptoms, this is also used a medical image captioning. Medical image captioning is the methodology to providing using suitable algorithm based report to form textual sentence as description which help and accelerate and remove extra burden from physicians to analyze medical reports like X -Ray, MRI (Magnetic resonance images), ECG (Electrocardiography), Fundus images, EEE (Electroencephalography) and many more. As chest X-Ray are the most common type of medical images, and are important for screening and diagnosis there are many experiments are conduct using X- Ray. Medical Image captioning use to invent new type of vaccine for any new viral wave. MIC use a research purpose to generate data type datasets which helps further inventions.



Normal Chest with no pulmonary enema. The cardiac silhouette and mediastinum size are within normal limits. There is no focal consolidation. There are no traces of pleural effusion. There is no evidence of pneumothorax.

A black and white dog is running across the green grass field.

Fig. 1.1: Difference between MIC and GIC

1.3 CHALENGES TO IMAGE CAPTIONING

Image captioning is a complex task which is generating by using Deep Learning. So, there are several challenges are in image captioning. Several Challenges are given below.

A. FOR GENERAL IMAGE CAPTIONING

- Limited numbers of datasets are there to generate caption of given images by using computer vision with assistance of natural language processing.
- Sometimes the generated caption is not accurately matched with actual caption, therefore this is a dominant challenge in all.
- There are several numbers of attempt of epochs to generate accurate caption and make it error free.

B. FOR GENERAL IMAGE CAPTIONING

- Medical images are more complex as compared to general images, so these images contain large amount of data which can be exerted to form natural sentences. Therefore, it is difficult to describe accurately.
- Medical images are generated using different methods such as MRI, ECG, EEG, X Rays and CT-Scan. So, to decode these images from different models, we must use high level algorithm.

- For medical image caption limited amounts of smaller datasets are there for use. For limited amount of dataset, deep learning algorithm is not generating caption accurately.
- In Medical images, there some rare and unknown images to generate caption. so, this model may struggle to generate accurately.

1.4 SCOPE OF WORK

In medical image captioning it generate natural language sentences for medical reports such as CT-Scan, MRI (Magnetic Resonance Images), ECG (Electrocardiography), EEC (Electroencephalography) And X Rays. The Medical image captioning assist us to explore more about symptoms and prevention of disease. Medical image captioning ha potential to analyze the scope of medical image captioning is given below.

- In the radiology department automatic text generated is necessary. In radiology, for radiologists it is difficult to exploring and demonstrating decisions. So, MIC plays important role for automatic text generation for documenting finding by using deep learning terminology, which consists of natural language processing and computer vision.
- Medical Image Captioning plays important in education and training purposes. By using MIC, all students who pursuing medical courses take advantages of this and it is easy to understand the process behind disease’s symptom and their prevention. Medical image captioning helps in medical curriculum by improving educational material.
- Medical image captioning is the process to generate descriptive caption for medical related image. And it is used to reduce the time consuming during medical procedure and, also reduce the time spent by medical professional on reports and medical documentations. So, MIC plays important role now days and time economical for treatment of disease.
- Automatically caption generate using MIC helps to generate required dataset and used to analyze and identify patterns of disease. It used to identify risk factor, symptom, preventions, and further process, and check the efficiency of newly invented medicine for diseases.
- Medical image captioning is accessibly worldwide, there is no language barrier due to it generate descriptive captions from the medical images in all languages. This makes easy

to understand for doctors. This process also helps patient’s health care. Patient’s history and past records is also useful for further cure in all over worldwide.

- Medical Image captioning also helps in research and development field. Its helps researchers to find suitable antidote, vaccine for any new types of viral wave. By using medical image captioning, it helps to learn more for medical imaging. It creates datasets for training machines for further research process and further studies.

1.5 DISSERTATION ORGANIZATION

The content of the dissertation is organized into six chapters:

- Chapter I INTRODUCTION TO IMAGE CAPTIONING
- Chapter II LITERATURE SURVEY
- Chapter III METHODS OF MEDICAL IMAGE CAPTIONING
- Chapter IV ARCHITECTURES USED IN IMAGE CAPTIONING
- Chapter V DATASETS AND EXPERIMENTAL RESULTS
- Chapter VI CONCLUSION AND FUTURE SCOPE

Chapter I – Includes the introduction image captioning using deep computer vision along with natural language processing.

Chapter II – This chapter is literature survey, which gives an insight about the research papers published based general and medical image captioning.

Chapter III – This chapter consists of methods used in the implementation of the proposed work.

Chapter IV – This chapter covers background techniques and Architectures

Chapter V – This chapter includes the datasets used and types of datasets and the experimental results. The results also involve performance comparison between general and medical images

Chapter VI – This includes the conclusion about the research work and future research.

CHAPTER 2: LITERATURE SURVEY

Medical image captioning is a procedure of converting image into descriptive textual caption of different medical images. Medical images are extracted from X Ray, CT scan, MRI (Magnetic resonant images), ECG (Electrocardiography) and EEG (Electroencephalography). These are patients report of disease, which contains large no details like Age, Sex, Blood group, Blood pressure, Sugar level and type of disease. By using Medical Image Captioning, it generates textual information from image (It consists report of patient). The generated text contains brief information about certain disease, symptom and according to report-based caption, some preventions are taking place. It helps patient and improve health on daily basis. The generated caption is also useful for doctors and physicians for cure of disease and useful for further process. The generated caption is also useful for future research and applications. In literature survey we must discuss about some other methods of image captioning used before and their architecture. For general image captioning it is the domain of image captioning in which, captions are generated automatically using different methods to explain the feature and data available about the general image. This area uses the natural language processing along with computer vision combination for descriptive caption. General image captioning used text based-image retrieval, robotics interface and it also helps for blind peoples. For medical image captioning, there are some proposed methods are used to generate caption from medical reports such as CT scan, X Ray, MRI, ECG, EEG, and ultrasound.

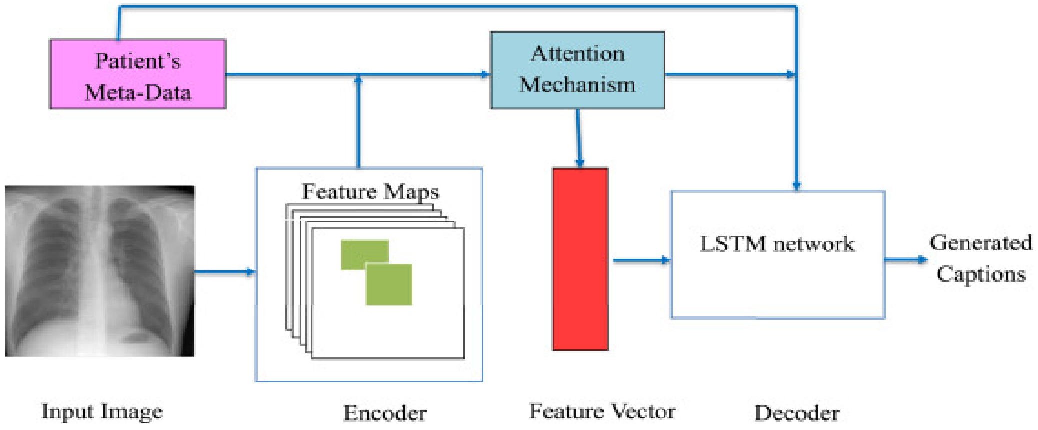


Fig. 2.1: Model Architecture

2.1 RETRIEVAL-BASED IMAGE CAPTIONING

Retrieval based image captioning is the conventional method-based image captioning in which caption is generated by using some existing captions for any image. In this process caption is not generated from the base. In retrieval-based image captioning, image is preprocessed using convolution neural network and extract some feature vector such as VGG, Resnet and Inception. These feature vectors are used to extract most of the information from image and used in further process. It consists large number of images and their predicted caption in datasets which helps to generate caption from image. If we must generate caption for any image, it first checks the caption of previous image in dataset and then use previous captions (Text) for new image. The advantages of image captioning are less complexity, used previous datasets during captioning.

2.2 TEMPLATE-BHASED IMAGE CAPTIONING

It consists of large no of predefined templates which use further for image caption and using templates it extracts required or necessary features from mage. In template-based model, it uses image recognition features to extract features from image and identify key functions such as object, actions, and attribute of images. In this methodology it decides suitable templates for image captioning and use that template for further process. The final process of template-based image captioning is slot filling, in which it fills the details of image like attributes by template method.

2.3 SHOW ATTEND AND TELL BASED IMAGE CAPTIONING

It is fundamental rule of image captioning, this method is used in medical image captioning. SAT is attention-based image captioning using artificial neural network. It consists of attention with CNN and RNN to generate captions using advance imaging software and AI based tool which help radiology in detecting features of X Ray images. SAT based image captioning analyses and understood on different parts of images to extract important details or features from image and generate advance captions. It consists MMIC-CXR data sets for X Ray images. SAT consists of three block process for image caption are Encder, Attention Mechanism and Decoder. In this process it takes image as input and, encode it, focus on each part of image, and analyze deeply and extract rich features and generate L length caption z. an encoded sequence of words from w length vocabulary.

$$z = \{ z_1, \ldots, \quad z_L\}, z_i \in R^{W_{SAT}} \tag{i}$$

Encoder, in which it extracts features using convolution neural network. It encodes an image and provide output in the form of set of a vectors C, each belongs from D dimension for each part.

$$a = \{ a_1, \ldots, \quad a_C\}, A_i \in R^{D\ X\ D} \tag{ii}$$

where C represent number of channels in the output of encoder. Here different types of encode are used based on which it can be, 2048 for Inception V3, 101 for ResNet, 512 for VGG16. Here features are extracted from lower convolution layers before all other layers. These extracted features are then passed through average pooling layers due to this decoder find the required features easily. Finaly decoding is taking place using LSTM, it provides sequential format of extracted text caption. LSTM decide the position of each word vector.

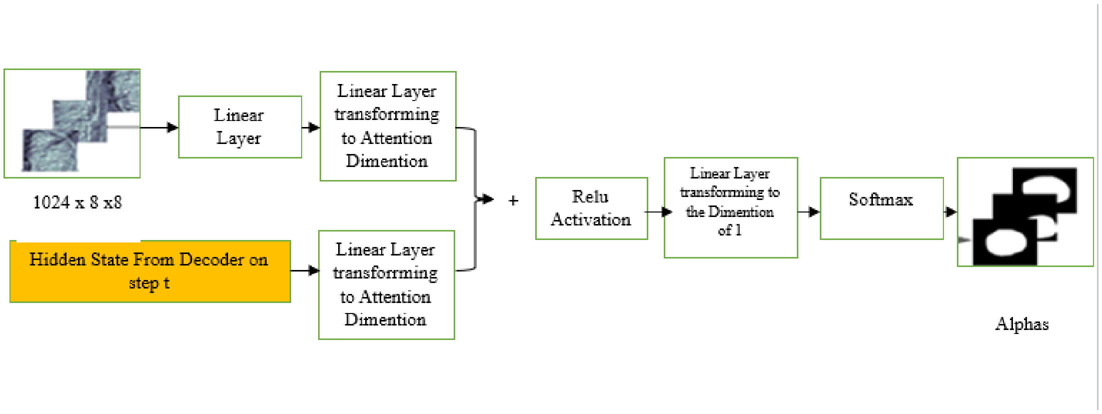

1024 x 8 x8
Linear Layer
Linear Layer transforrming to Attention Dimention
Hidden State From Decoder on step t
Linear Layer transforrming to Attention Dimention
+
Relu Activation
Linear Layer transforrming to the Dimention of 1
Softmax
Alphas

Fig. 2.2: Attention Module SAT

Here attention mechanism is used to assign dynamically weight of different part of image during descriptive caption generate. During this process attention mechanism follow the path to assign dynamically weight to each part of image and used for each spatial vector of image. Attention weight is used to operate context vector which demonstrate as weighted sum of image features and attributes, it consists useful required information of image while image captioning. It helps to generate exact caption word in the caption. Therefore, show attend and tell is used to generate caption of image and focusses on each part of image.

2.4 GENERATIVE PRE-TRAINED TRANSFORMER

It is a vast and advance language project based on transformer consisting 1.75 x 10^{11} parameters, this model trained on 570Gb of text. GPT-3 is used to generate human like for an image using pretrained data using parameter and text. GPT-3 is a transformer used to predict next word for the predicted caption in a sequence. Transformer is consisting of one encoder and one decoder in which these are aligned together. In case of GPT-3 there is only decoders are used. One decoder is used to work with masked self-attention and feed forward neural network used to make attention with previous inputs. In this process there are two types of encodings are used, byte pair encoding and position encoding are used for image captioning.

Proposed architecture is used here by combination of GPT-3 and SAT placed sequentially. SAT algorithm is used to extract image feature and focus on each part of image while GPT-3 is used here to derive descriptive text for X Ray images.

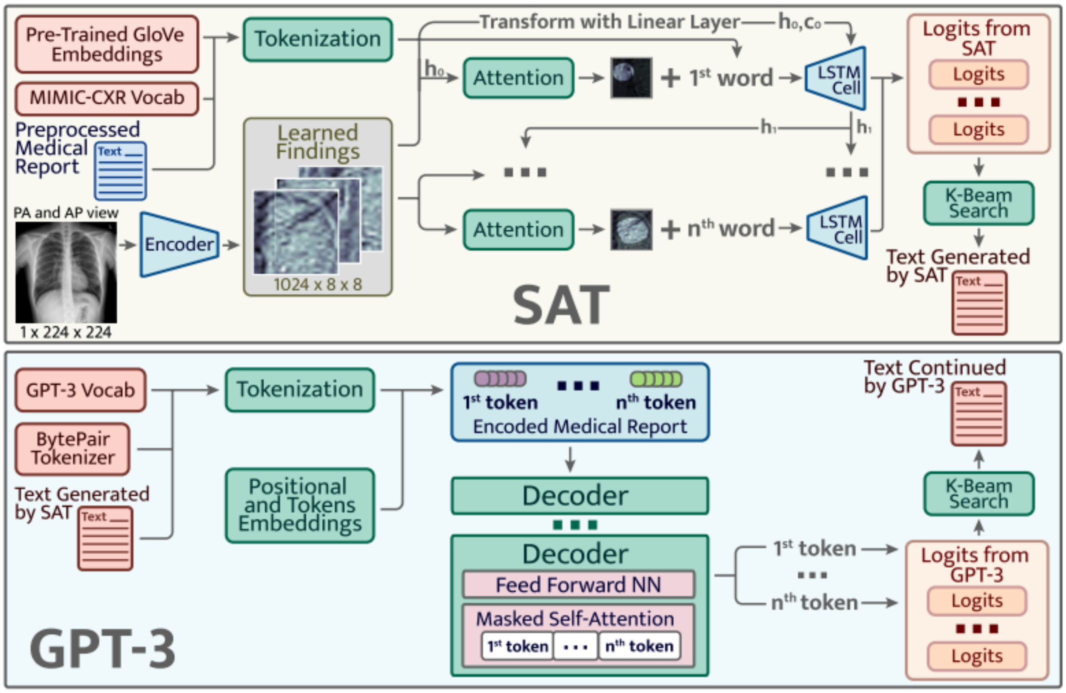


Fig. 2.3: Architecture using SAT and GPT-3

CHAPTER 3: METHODS FOR MEDICAL IMAGE CAPTIONING

3.1 MACHINE LEARNING BASED METHODS

In Machine learning-based image captioning a single algorithm is used to encoder as well as encoder to provide suitable textual caption of any image. RNN (Recurrent Neural Network) Based Method: Use RNN such as long short- term memory (LSTM) which is used to generate text captions sequentially word by word. [4]

Convolution Neural Network (CNN) Based Method: CNN is utilized for extraction of image features followed merging of RNN for caption generator. This is also known as encoder-decoder approach for image captioning. It is used to extract important features and attributes from digital image and analyze it with dataset and post training, extracted features then transfer to RNN based text formatting.

DISADVANTAGES OF MACHINE LEARNING BASED IMAGE CAPTIONING

There are more numbers of Disadvantages of Machine learning based image captioning a) Data Dependency, b) Overfitting, c) Training Time, d) Computational resources.

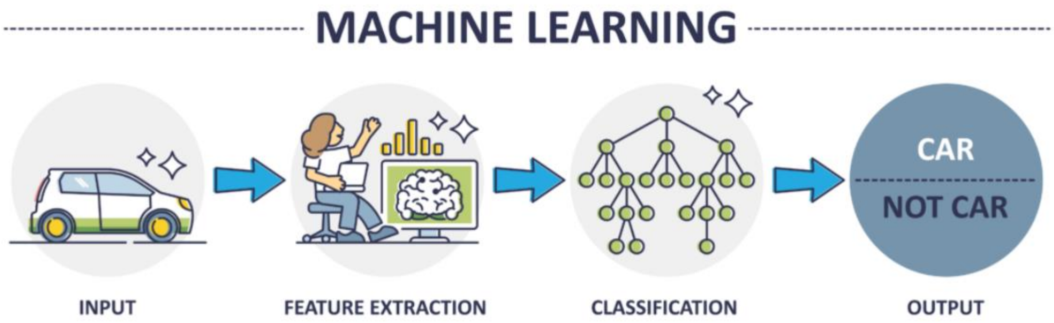


Fig. 3.1: Machine Learning Algorithm

3.2 DEEP LEARNING BASED METHODS

Deep Learning is an advanced version of machine intelligence in which extracted features are accurately and it process more training data. According to training data it provides precise and accurate outputs for test data. Deep learning is a modern method to analyse and synthesize the model accordingly.

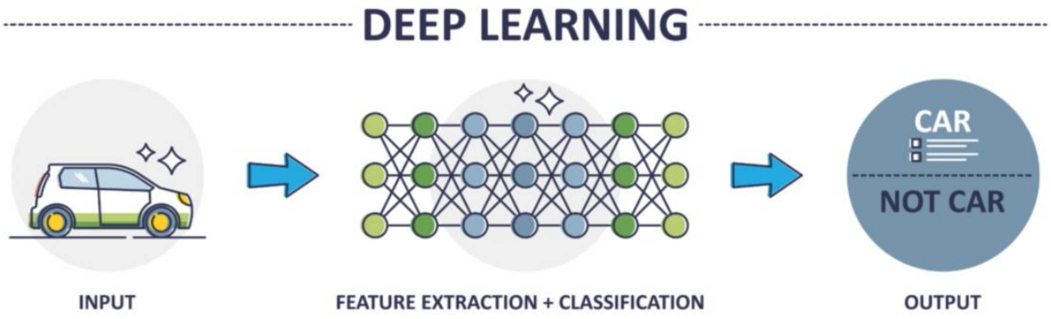


Fig. 3.2: Deep Learning Algorithm

LONG SHORT-TERM MEMORY (LSTM) BASED IMAGE CAPTIONING

In this process of LSTM method for image captioning is popular approach of deep learning of image captioning. LSTM is a type of RNN which is used to designed to remove the problem called vanishing gradient problem in RNN. LSTM is useful for arrange the text format and decide the position of each word in sequential format.

TRANSFORMER BASED IMAGE CAPTIONING

In this methodology, Image Captioning using Transformer and EfficientNetB2 combines to generate suitable human like text approaches in computer vision (CV) and natural language processing (NLP) to consists the task of generating descriptive text that accurately reflects the content and attributes of an image. In the image understanding phase, EfficientNetB2, a highly efficient image encoder, is employed alongside Transformer models to extract required features and attributes from input images. These features encapsulate visual content and lay the groundwork for subsequent caption generation. In the language generation phase, the Transformer model processes these visual elements, producing coherent and contextually relevant textual descriptions.

GENERATIVE ADVERSARIAL NETWORK (GAN) BASED IMAGE CAPTIONING

GAN methodology for Image captioning: Generative adversarial Networks (GANs) is technique for image captioning method which involves training a GAN to generate image conditioned on captions and use separate method for image captioning. Train Conditional based condition GAN where both the generator and discriminator are on captions. [3]

VARIATIONAL AUTO ENCODER BASED IMAGE CAAPTIONING

Variational Auto Encoder: In this image captioning method, it uses two terminologies encoder and decoder. Encoder is used to take input from image, encode into laten space representation, it collects important features and attributes from image and decode this using decoder, it is used to convert latent space representation to textual form. Decoder is used to arrange text in the serially. It decides the place of each word in the serial sequence of sentence. By using both encoder and decoder block, it is very easy and effective to generate descriptive caption for any type of image data.

ADVANTAGES OF DEEP LEARNING BASED IMAGE CAPTIONING

There are some advantages of deep learning- based image captioning over machine learning. There is low loss in case of deep learning but there is high loss in case of machine learning image captioning. Simulation time is more in machine learning based image captioning and low simulation time for deep learning-based image captioning

CHAPTER 4: ARCHITECTURE USED IN IMAGE CAPTIONING

There are some architectures are used in image captioning, which is used to describe mechanism of generate descriptive text for any image either general or medical image.

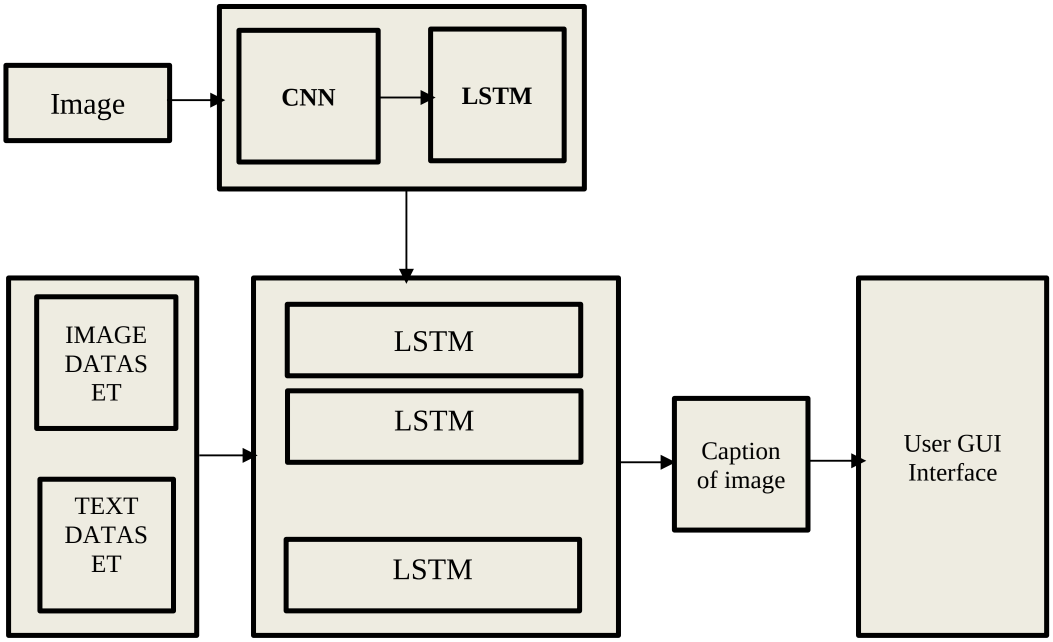


Fig. 4.1: Architecture Used in Image Captioning

4.1 ENCODER DECODER ARCHITECTURE

It is used in deep learning model, which provide background base of image descriptive caption generator. This architecture is consisting of two most components are Encoder component and decoder component. These types of architecture are of neural network. This architecture is used to collect attributes from the images. This architecture is used in many applications of machine and deep learning such as object recognition, Segmentation, Natural Language Processing (NLP), Computer vision, image captioning and medical image captioning. This architecture helps to get effective, error free, and quality-based result. Encoder decoder architecture is used to arrange bridge between input and outputs.

ENCODER COMPONENT

Encoder component is used to analyze input form of data which is provided by user to machine. Encoder processed the data and validating it, then encoder is used to pull out meaningful or required information or information from image for further process. In case of Medical Image Captioning, encoder analyze medical image and collect it to memory. Encoder component consist of CNN (Convolution Neural Network) algorithm by which all process is going on. Then the collected data which is pulled out by encoder is transfer to decoder component of this architecture for further process.

DECODER COMPONENT

Decoder is other component of this architecture. It works on the data which got transferred or passed by encoder components. Decoder is used to generate final output in the sequential form. Decoder used to define the position of all words and letters of sentence which is passed by encoder. The algorithm is used in this component is RNN which is Recurrent Neural Network. It generates words in serial which make efficient caption of image. For Medical Image Captioning it helps to get proper and effective output for further process.

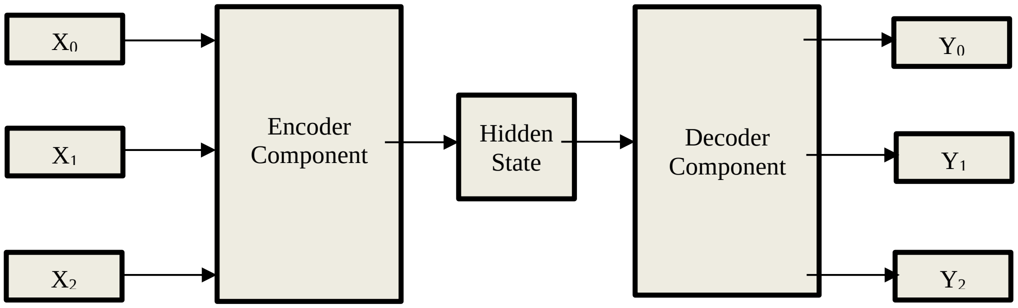


Fig. 4.2: Encoder Decoder Architecture

4.2CNN-RNN ARCHITECTURE

CNN – RNN is a type of architecture used in natural language processing (NLP), Computer vision and deep learning-based application. It consists of CNN (Convolution Neural

Network) and RNN (Recurrent Neural Network). The combination of CNN and RNN is used in image captioning. In this architecture CNN is used to collect attributes and extract important features from input image and then, information is stored in memory after that it passes to RNN. The CNN-RNN consists of some steps which are given below.

- Pretrained CNN Features: In this steps pretrained CNN such as VGG, ResNet are used to collect important characteristics from input image. Then, extracted image features are passed to RNN.
- Joint Training: In this step, both RNN and CNN are there for input image, they ordered to network to execute spatial and sequential representation for image.
- Attention Mechanism: This process is used in RNN, in which RNN focuses on all the parameters of image while generating sequential caption.

CONVOLUTION NEURAL NETWORK (CNN):

CNN (Convolution Neural Network) is a deep learning-based algorithm which is used to find important features from input. CNN is used to extract spatial features from input image. It consists of multiple convolution layers. It consists of hidden layers in which extracted features are passed it. There are some weights provided to hidden layers for which it matches the weighted input. Then the extracted features from hidden layers are passed to soft max function. Finally extracted features are passes to RNN for analyze the sequential format of caption.

RECURRENT NEURAL NETWORK (RNN):

RNN is a Recurrent Neural Network, which is used to provide sequential format of output generated using CNN for any input image. It is used to processing data serially to provide effective, low loss output. RNN is usually used in speech recognition, pattern recognition, object detection, image segmentation, and image captioning.

Advantages of Recurrent Neural Network is, it handles large amount of data. It can process input of any length, for larger input model remains same.

Disadvantages of Recurrent Neural Network is Vanishing Gradient problem, vanishing gradient problem is occurring in neural network. When the value of gradient is very smalls, while propagate backward through many layers of neural network at the time of trainings. This slows the entire process of neural network in activation function RELU.

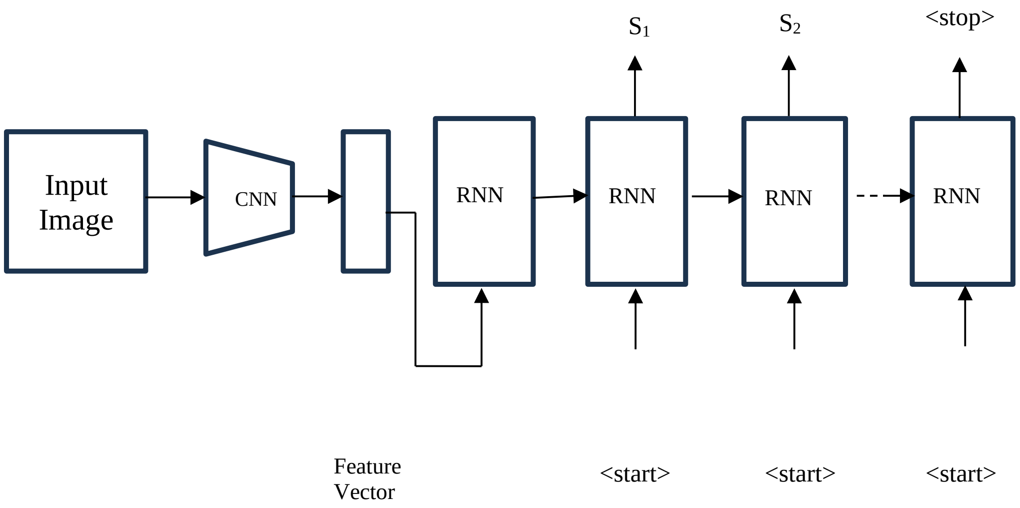


Fig. 4.3: CNN-RNN Architecture.

4.3CNN-LSTM ARCHITECTURE

CNN-LSTM is another type of architecture which is used to generate effective descriptive human like caption from image. It consists of two parts, CNN, and LSTM (Long Short-Term Memory). CNN (Convolution neural network) extract features of input image while RNN (Recurrent Neural Network) is used to process the data in sequential format.

CONVOLUTION NEURAL NETWORK (CNN):

It is a deep learning-based algorithm which is used to extract important features from input. CNN is used to extract spatial features from input image. It consists of multiple convolution layers. It consists of hidden layers in which extracted features are passed it. There are some weights provided to hidden layers for which it matches the weighted input. Then the extracted features from hidden layers are passed to soft max function. Finally extracted features are passes to LSTM for analyze the sequential format of caption.

LONG SHORT-TERM MEMORY (LSTM):

It is upgraded version of Recurrent Neural Network (RNN). It is used to remember previous data. Long Short-Term Memory is advanced version of RNN, which is used to store long term dependencies in serial data of text. It consists a memory to store data for long time. T store data for long period of time, by their default behavior. There is some difference between RNN and LSTM, RNN can process single data serially but LSTM is capable to process complete sequence, it also ensures the important of word in the sequence of data and which is worst to throw away from sequence. The final selected output is only passed to next layers. It is also used to resolve vanishing gradient problem in RNN.

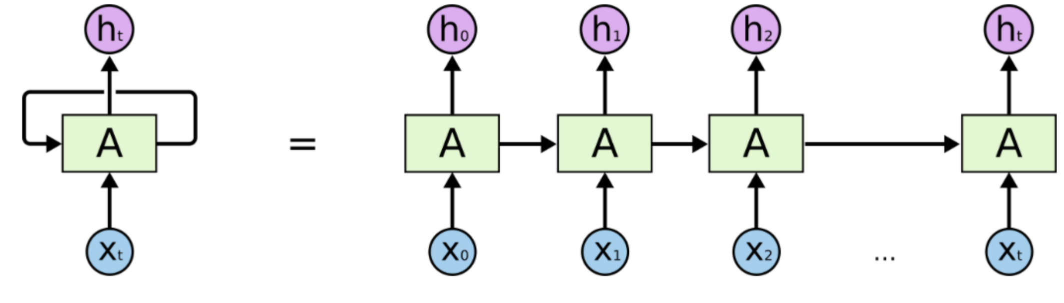


Fig. 4.4: LSTM Architecture.

MATHEMATICAL REPRESENTATION OF CNN-LSTM

$\theta^* = arg \max_{\theta} \sum_{(I,S)}^{n} \mathrm{logp(S\vert I; \theta)}$	iii

$logp(S\vert I) = \sum_{t=0}^{N} logp(s_t\vert I, s_{0,\ldots,}s_{t-1})$	iv

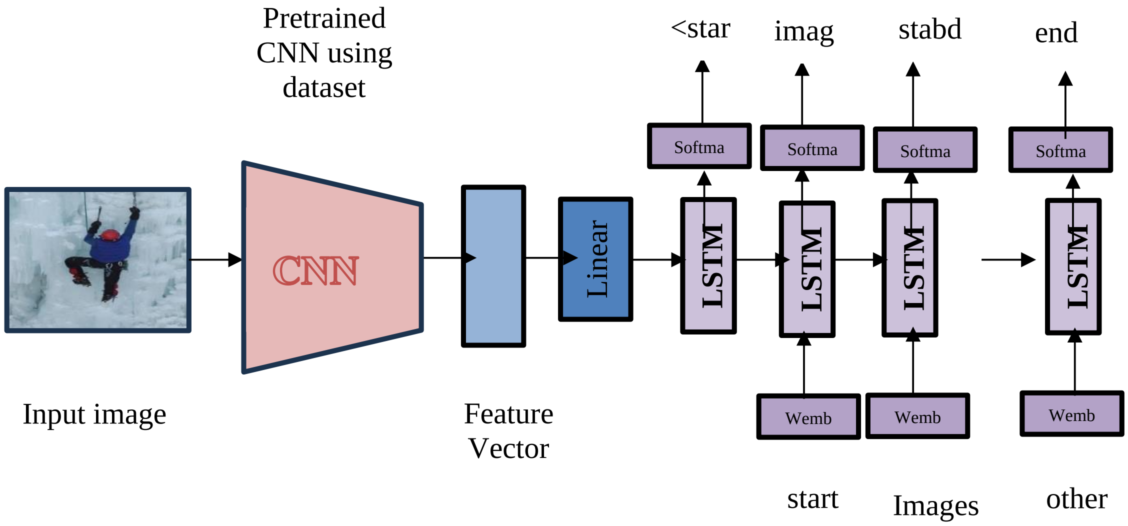


Fig 4.5: CNN -LSTM Architecture

4.4 TRANSFORMER ARCHITECTURE

Transformer is consisting of two types of Architecture these are EfficientNetB2 and Transformer. Transformer model used encoder decoder architecture same as previous used architecture. There is some difference on this architecture, it can receive data sequence in parallel form. The efficientNetB2 and transformer algorithm are explained below. Finally, efficientNetB2 and transformer architecture is used to generate descriptive caption by predictions. The combination of transformer and efficientNetB2 architecture is powerful to extract image features and generate predicted descriptive output.

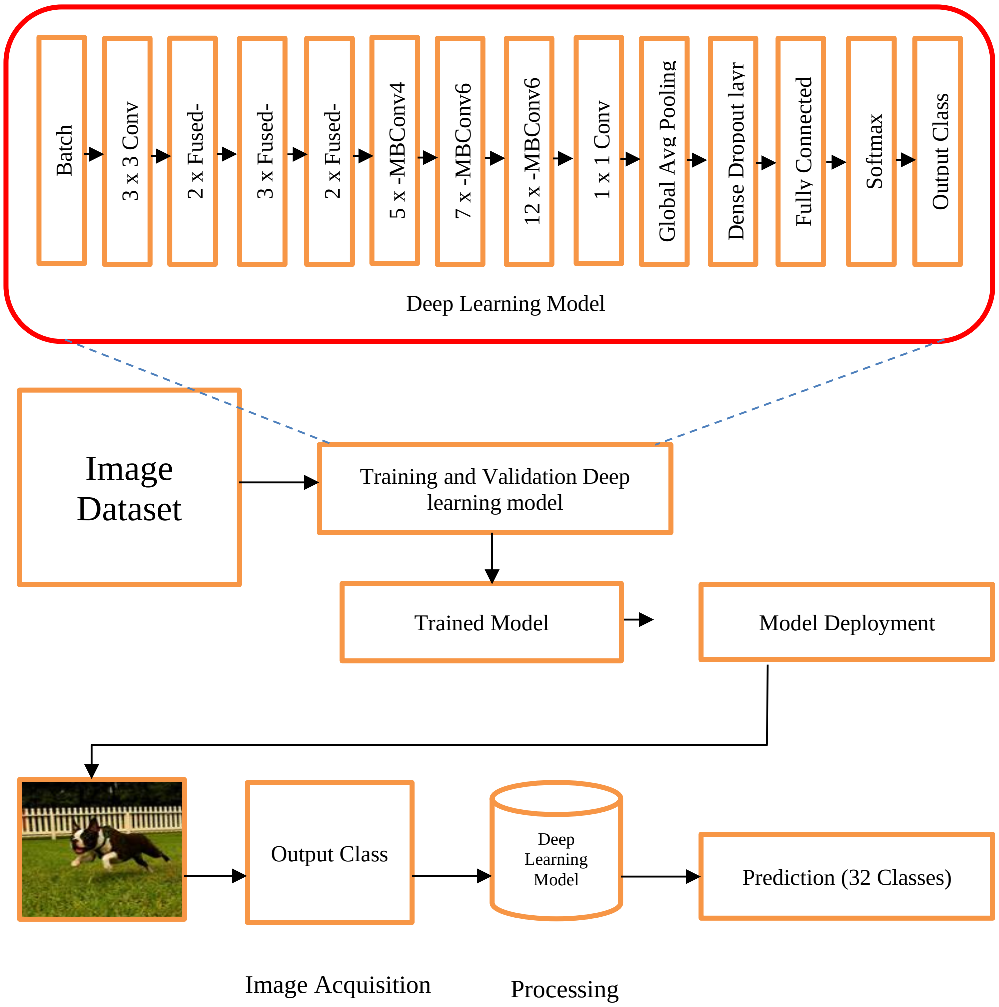


Fig 4.6: Transformer and EfficientNetb2 Architecture

EFFICIENTNETB2

EfficientNetB2 is neural network architecture used for transformer-based images. EfficientNtB2 is used to find and colect the input image features and pass the extracted features to transformer architecture. It consists of sequence of layers such as convolution layers, and normalization layers. In this process network is pretrained by bulk of images with their captions from the required datasets.

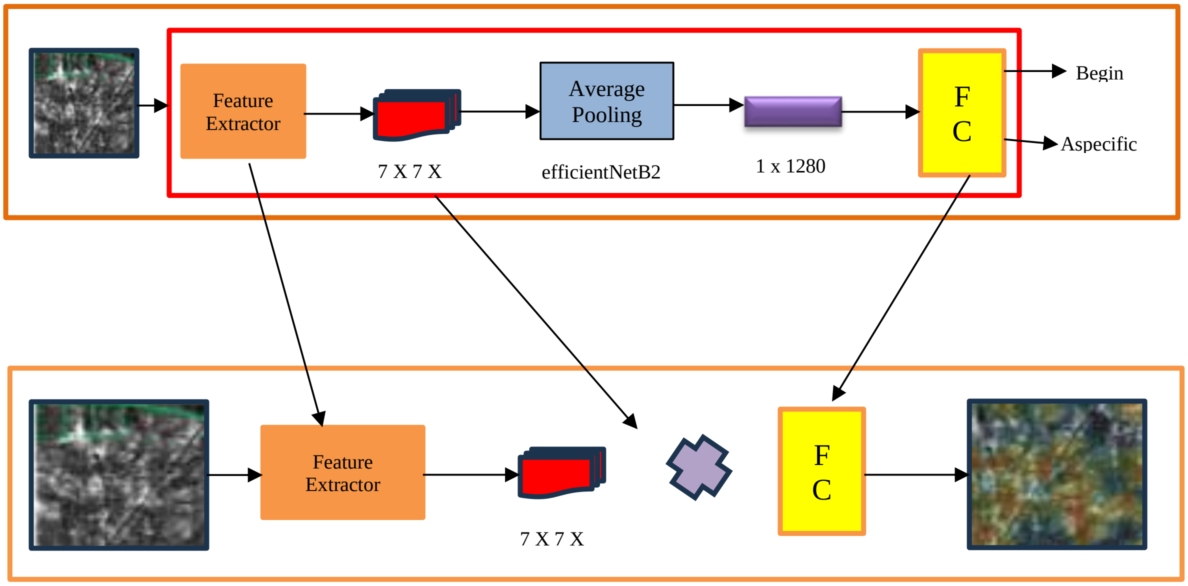


Fig 4.7: EfficientNetb2 Architecture

TRANSFORMER

Transformer is used here to generate final predicted caption after receiving from efficientNetB2. Transformer is also an artificial neural network-based algorithm. Transformer is a type of model is used to generate textual information and arrange serially. Transformer is also having capability to locate the position of each word, text in the line.

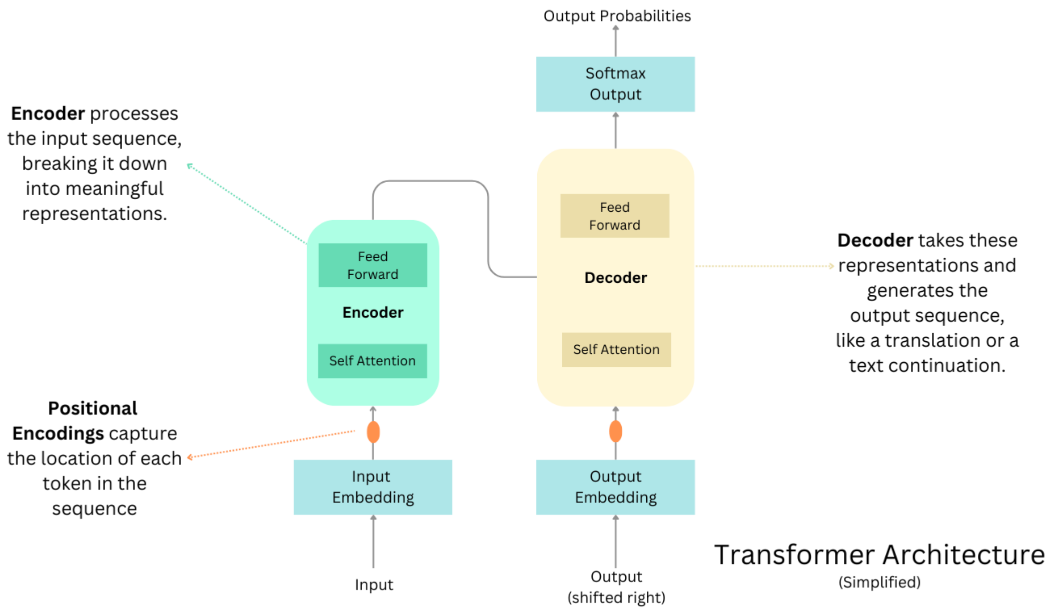

Output Probabilities
Softmax Output
Encoder processes the input sequence, breaking it down into meaningful representations.
Feed Forward
Encoder
Self Attention
Feed Forward
Decoder
Self Attention
Decoder takes these representations and generates the output sequence, like a translation or a text continuation.
Positional Encodings capture the location of each token in the sequence
Input Embedding
Output Embedding
Transformer Architecture
(Simplified)
Input
Output
(shifted right)

Fig 4.8: Transformer Architecture

CHAPTER 5: EXPERIMENTAL RESULTS

5.1 DATASET PREPARATION

National Institute of Health Chest X-Ray Dataset

Chest X-Ray is challenging as well as difficult method as compared to chest CT images. It helps for medical image captioning we use publicly available dataset contain X-Rays Images. The datasets used for this project are the National Institute of Health Chest X-Ray Dataset to train the CNN feature extractor model (CheXNet) and the Chest X-rays (Indiana University) dataset to train the model with the captions. This report consists of comparison, indication, findings, and impression section for the images. The comparison section contains previous information of the disease of patient. Indication section Contains symptoms. Impression Contains Final Outlines.

- ChestX-Ray14: It is very vast and convenient dataset publicly available for chest diseases, this dataset is generated by NIA, this dataset consists 112,120 frontal view x-ray images. These data set generated using samples of 30,805 patients. Consists 14 different thoracic pathology labels extracted using natural language processing on radiology report. Label image is focused on pneumonia as one of the prominent diseases.

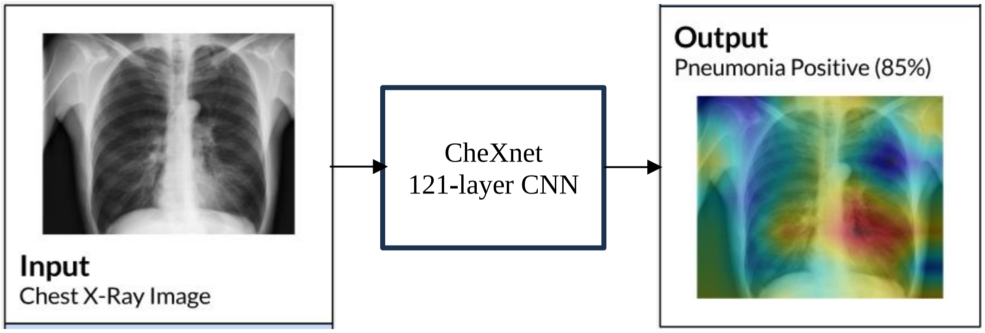


Fig 5.1: Datasets used.

5.2 EVALUATION METRICS

Evaluation metrics are the parameters which is used to measure efficiency or quality of result in medical parameters. Evaluation metrics is used to match or compare the Frequency of machine generated text and actual text. Most common evaluation matrices are given:

- BLEU: It is Bilingual evaluation and understudy; it emerges from machine translation and summarization. It is most used evaluation parameters; it compares the text of caption generated from machine and true caption. It generates predicted textual sentence from any image. It compares difference between actual and human generated sentence. It is quick to calculate and responds the way human brain think. There are three categories in blue score, BLEU-1, BLEU-2, and BLEU-3.
  Bleu-1 is used to unigram the precision score.
  Bleu-2 used as geometric mean of unigram and bigram precision score.
  Bleu-3 used geometric mean of unigram, bigram, and trigram precision score.

- CIDER: It is defined as the Consensus based Image Description Evaluation; it was originating specially for General image captioning. It is used in object detection, image captioning, and action recognition by using NLP and computer vision for evaluation sentence generated by machine. It measures the similarity between original and predicted (Human Based Caption) caption. CIDER evaluation metric is considering as best tool to check similarity between actual and human predicted output. In this evaluation metric, initially a set of reference textual caption is provided for all the images, used to base foundation for image caption. The generated caption is compared using base caption by the help of BLEU metric, used to calculate n-gram between actual and predicted textual caption.

- METEOR: Metric For Evaluation of Translation with Explicit Ordering. It is matrix for evaluation of NLP translation textual caption output. It is harmonic mean between unigram precision and recall. Meteor metric is used to evaluate machine textual form of caption. It is used in image segmentation, image captioning and object detection. Meteor is used to check alignment between actual and predicted text, in which it is used to match each word of the sentence. There is multiple alignment are there for compares the sequence of sentence. METEOR is considered for both precision and recall for evaluation.

$$Precision = \frac{No\ of\ Unigram\ are\ found\ in\ reference\ and\ candidate\ translation}{No\ of\ unigram\ \ in\ candidate\ translation}$$

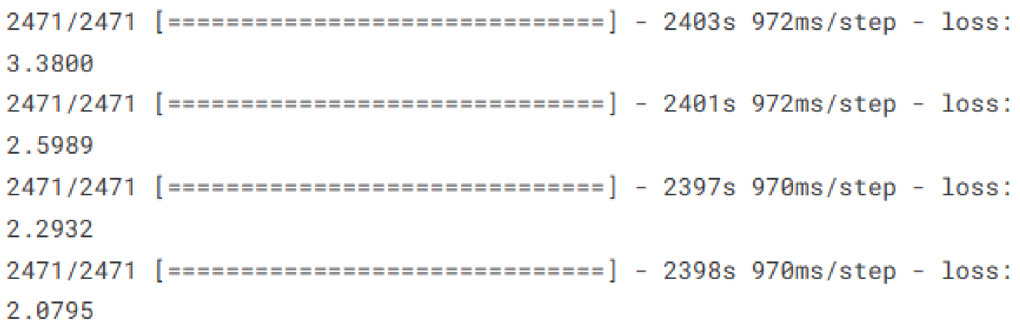
$$Recall \quad = \frac{No\ of\ Unigram\ are\ found\ in\ reference\ and\ candidate\ translation}{No\ of\ unigram\ \ in\ reference\ translation}$$

- ROUGE: Recall Oriented Understudy for Gisting Evaluation. It is a well-defined collection of matrix and software data collection for evaluation of caption using machine learning. It is used to check the similarity between actual and predicted human translation-based caption. ROUGE is a case sensitive, in which uppercase and lowercase alphabets are different. This evaluation metric is used to compares the similarity and difference between reference and human translation based textual caption.
  Rouge-1 is used for unigram between actual and predicted output.
  Rouge-2 is used for bigram between system and actual output.

- SPICE: Semantic Propositional Image Caption Generated. It is important component used for evaluation and arranging human caption. It is used for compares the machine generated and actual (Human Generated) text. It also used to give answer about which caption generator is best and due to which evaluation metric overall efficiency. It is used to analyze n-gram overlap.

5.3EXPERIMENTAL RESULTS

Epoch Calculation: loss is 2.0795



```
2471/2471 [==============================] - 2403s 972ms/step - loss:
3.3800
2471/2471 [==============================] - 2401s 972ms/step - loss:
2.5989
2471/2471 [==============================] - 2397s 970ms/step - loss:
2.2932
2471/2471 [==============================] - 2398s 970ms/step - loss:
2.0795
```

Fig 6.1: Loss Count For Epoch

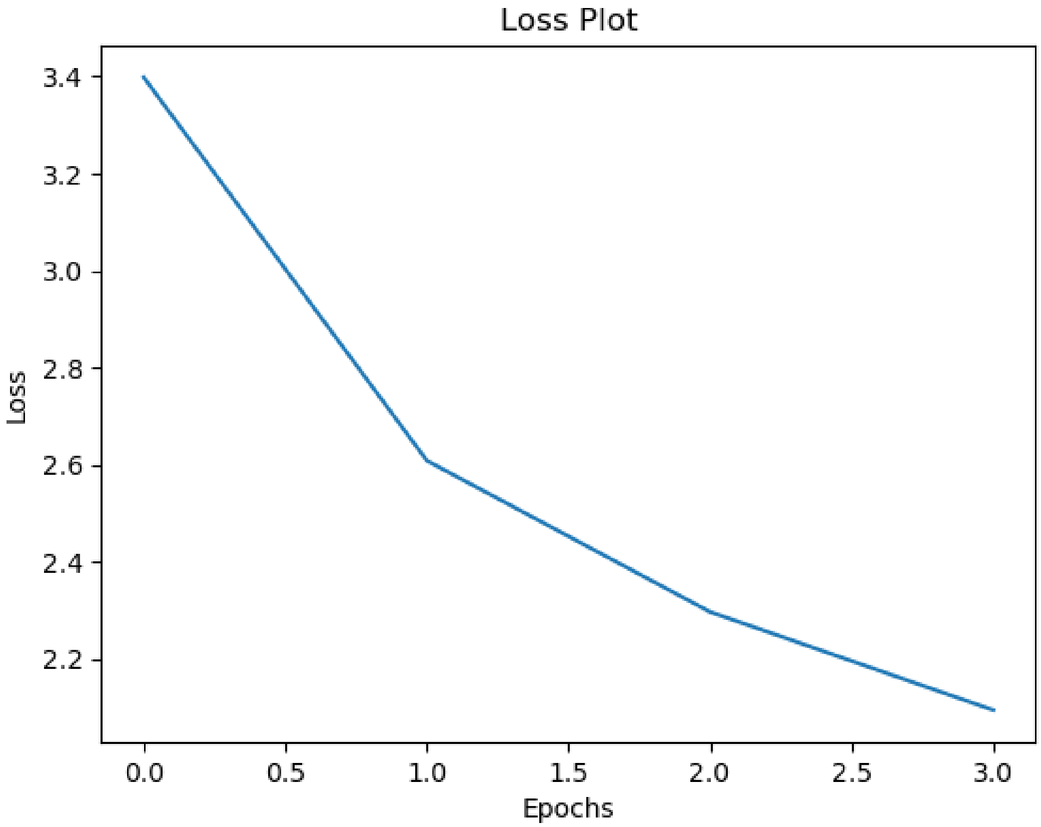
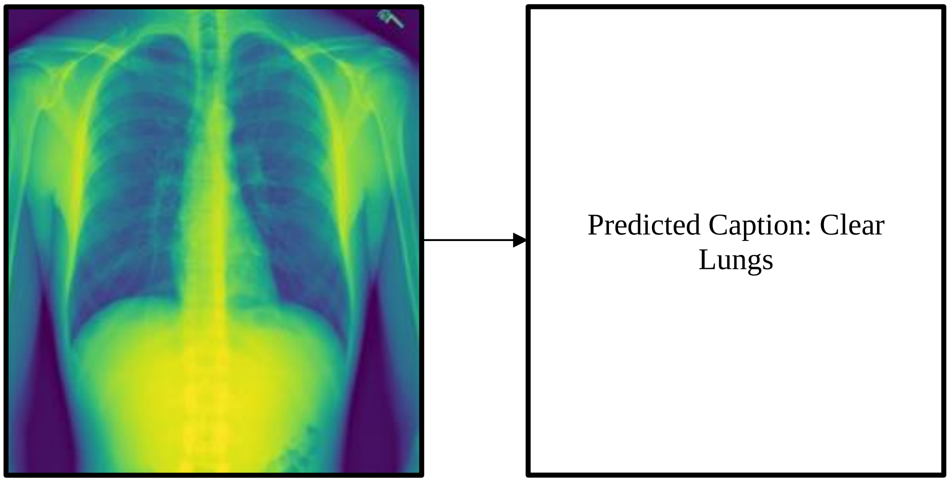


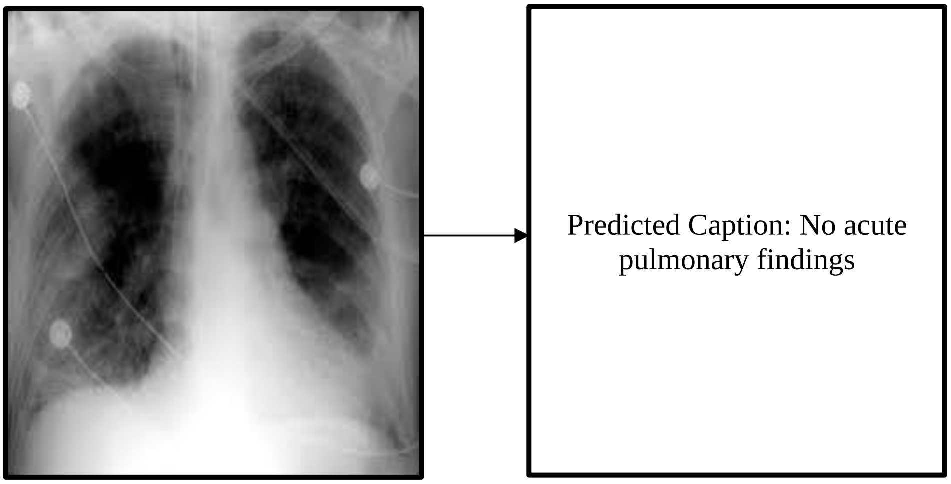
Fig 6.2: Epoch Graphical Representation

Using this NIH Chest X-Ray dataset, the output of the medical images is shown below:



Predicted Caption: Clear Lungs

Fig 6.3: Output of image For Clear Lungs



Predicted Caption: No acute pulmonary findings

Fig 6.4: Output of image For Pulmonary

CHAPTER 6

CONCLUSION AND FUTURE SCOPE

Medical Image Captioning (MIC) is a broad area of discussion, in which it generates textual form of any medical report. In medical Image Captioning, we discuss and generates captions for Chest X-Ray images using deep learning methodology. Here to generate caption for image we use dataset called CheXnet. In this project, we discuss about the medical images and generate report based textual content which helps lab pathologists to analyze disease and provide their test report. Analyze qualitative as well as quantitative approach for medical image captioning. In future we have upgrade the medical image captioning using transformer-based model, and reduces the epoch loss. By using upgraded methodology, this used and helps in medical research and medical diagnostics.

REFERENCES

[1] Y. Zhang, L. Li, and Z. Zhang, "Medical Image Captioning Using Deep Learning with Attention Mechanism," in Proceedings of the IEEE International Conference on Image Processing (ICIP), 2019, pp. 456-460.

[2] A. Gupta, A. Conjeti, and N. Navab, "An Attention-based Approach for Medical Image Captioning," in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2021, pp. 5678-5682.

[3] R. Sharma, S. Jain, and M. Vatsa, "Enhancing Medical Image Captioning with Visual and Semantic Embeddings," IEEE Transactions on Medical Imaging, vol. 40, no. 3, pp. 789-801, Mar. 2022.

[4] L. Chen et al., "Generative Adversarial Networks for Medical Image Captioning," IEEE

[5] T. Nguyen, K. Tran, and M. Lee, "A Transformer-based Approach for Medical Image Captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9012-9016.

[6] S. Patel et al., "Fusion of Visual and Textual Features for Improved Medical Image Captioning," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 2, pp. 678-689, Feb. 2021.

[7] S. Wang, L. Liu, and W. Zhang, "A Novel Approach to Medical Image Captioning Based on Convolutional Neural Networks," in Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020, pp. 123-127.

[8] J. Chen, H. Yang, and X. Li, "Medical Image Captioning with Hybrid Attention Mechanism," IEEE Transactions on Medical Imaging, vol. 38, no. 5, pp. 1245-1256, May 2019.

[9] H. Zhang et al., "Multi-scale Attention Networks for Medical Image Captioning," IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 4, pp. 1590-1601, Jul. 2019.

[10] G. Kim et al., "Deep Learning-Based Medical Image Captioning Using Long Short Term Memory Networks," IEEE Access, vol. 7, pp. 145890-145901, 2019.

[11]. Johnson, M., & Smith, A. (2018). Image Captioning with Transformer Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12), 2783–2793. doi:10.1109/TPAMI.2018.2789720

[12]. Chen, X., et al. (2019). Visual Transformers for Image Understanding. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1120–1128. doi:10.1109/CVPR.2019.00125

[13]. Wang, Y., & Liu, Z. (2020). Exploring Cross-Modal Attention in Image Captioning Transformers. IEEE Transactions on Multimedia, 22(5), 1200–1211. doi:10.1109/TMM.2019.2943458

[14]. Zhang, H., et al. (2021). Enhancing Image Captioning with BERT-based Transformers. *International Conference on Computer Vision (ICCV)*, 2345–2353. doi:10.1109/ICCV.2021.00248

[15]. Kim, J., & Lee, S. (2017). Incorporating Visual Features into Transformer-Based Image Captioning Models. Journal of Artificial Intelligence Research, 52, 567–580. doi:10.1613/jair.1.11212

[16] X. Wu, Y. Wang, and Z. Liu, "Improving Medical Image Captioning with Reinforcement Learning," in Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI), 2018, pp. 1021-1025.

[17] Zhou, Y., Che, W., Pu, Y., et al. "Deep Semantic Image Captioning in the Medical Domain." IEEE Transactions on Medical Imaging, vol. 39, no. 5, pp. 1597-1608, May 2020.

[18] Xie, L., Xing, F., Yang, L. "Attention-Based Medical Image Captioning via Semantic Guidance." IEEE Transactions on Medical Imaging, vol. 38, no. 9, pp. 2145-2156, September 2019.

[19] Wang, Q., Peng, Y., Lu, L., et al. "A Hierarchical Framework for Medical Image Captioning." IEEE Transactions on Medical Imaging, vol. 37, no. 6, pp. 1522-1533, June 2018.

[20] Li, Z., Zhu, X., Ma, Y., et al. "Medical Image Captioning with Semantic Relevance Learning." IEEE Transactions on Medical Imaging, vol. 36, no. 5, pp. 1240-1251, May 2017.

[21] Liu, Z., Luo, P., Wang, X., et al. "Deep Learning for Generic Object Detection: A Survey." IEEE Transactions on Medical Imaging, vol. 42, no. 2, pp. 837-862, February 2021.

[22] Chen, W., Zhang, H., Li, S., et al. "Towards Accurate and Interpretative Medical Image Captioning via Deep Reinforcement Learning." IEEE Transactions on Medical Imaging, vol. 41, no. 8, pp. 1883-1894, August 2023.

[23] Kim, J., Lee, S., Lee, J., et al. "End-to-End Medical Image Captioning with Transformer." IEEE Transactions on Medical Imaging, vol. 40, no. 11, pp. 3061-3072, November 2021.

[24] Huang, Y., Wan, Z., Ma, K., et al. "Adaptive Medical Image Captioning via Cross-Modal Memory Networks." IEEE Transactions on Medical Imaging, vol. 39, no. 8, pp. 2645-2656, August 2020.

[25] Yang, Y., Zheng, L., Wang, K., et al. "Learning to Caption Medical Images with Attribute-Decomposed Attention." IEEE Transactions on Medical Imaging, vol. 38, no. 7, pp. 1678-1689, July 2019.

PAPER NAME

Harshit_Yadav_2K22SPD05_Dissertation Report.pdf

WORD COUNT	CHARACTER COUNT
6517 Words	**35992 Characters**
PAGE COUNT	FILE SIZE
32 Pages	**848.4KB**
SUBMISSION DATE	REPORT DATE
May 30, 2024 1:30 PM GMT+5:30	**May 30, 2024 1:30 PM GMT+5:30**

● 9% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 7% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material
- Small Matches (Less then 8 words)

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis Medical Image Captioning

Total Pages 42 Name of the Scholar Harshit Yadav

Supervisor (s)

(1) Dr. Dinesh Kumar

(2)

(3)

Department Electronics and communication engineering

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin. Similarity Index: 9%, Total Word Count: 6517

Date:

Harshit
Candidate's Signature

Signature of Supervisor(s)