

**VISUAL TRANSFORMERS FOR IMAGE  
UNDERSTANDING**

**Thesis Submitted  
in Partial Fulfillment of the  
Requirements for the Degree of**

**MASTER OF TECHNOLOGY  
in**

**SIGNAL PROCESSING AND DIGITAL  
DESIGN (2022-2024)**

**by**

**Himanshu Bisht**

**2K22/SPD/06**

**Under the Supervision of**

**Dr. Rajesh Rohilla**

**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**



# **DELHI TECHNOLOGICAL UNIVERSITY**

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

## **ACKNOWLEDGEMENT**

I would like to express our deep gratitude to our project guide Dr Rajesh Rohilla Professor of Department of Electronics and Communication Engineering of Delhi Technological University, for his guidance with unsurpassed knowledge and immense encouragement.

I would like 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 Transformer and efficientnet2 architectures, which have provided us with powerful tools for feature extraction and caption generation. Our thanks also go to the authors of the benchmark dataset, Flickr 8k which have played a crucial role in training and evaluating our Transformer model.

I am grateful to the open-source community for developing and maintaining user-friendly deep learning frameworks like TensorFlow, PyTorch, and Keras, simplifying the implementation of our research.

I would like to thank our parents, friends, and classmates for their encouragement throughout our project period. At last, but not the least, we thank everyone for supporting us directly or indirectly in completing this project successfully. Your support and inspiration have been truly invaluable.

May, 2024

**Delhi (India)**

**Himanshu Bisht**

**(2K22/SPD/06)**



# DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

## **CANDIDATE'S DECLARATION**

I am Himanshu Bisht, 2k22/SPD/06. student of MTech. (Signal Processing and Digital Design), hereby declare that the project Dissertation titled "**VISUAL TRANSFORMERS FOR IMAGE UNDERSTANDING**" which is 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

Date: 20/05/2024

**(Himanshu Bisht)**



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

**CERTIFICATE BY THE SUPERVISOR**

Certified that **Himanshu Bisht** (2K22/SPD/06) has carried out his research work presented in this thesis entitled **“VISUAL TRANSFORMERS FOR IMAGE UNDERSTANDING”** for the award of **Master of Technology** from Department of Electronics and communication Engineering, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student himself and to the best of my knowledge and belief, the contents of the thesis 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.

Place: Delhi

Date: 31/05/2024

**Dr Rajesh Rohilla**

**SUPERVISOR**



## **ABSTRACT**

Image captioning is a complex undertaking that combines computer vision and natural language processing, with the goal of producing descriptive text for visual stimuli that mimics human language. In this study, we investigate the symbiotic relationship between the EfficientNetB2 image encoder and a Transformer-based language model in the context of image captioning. The utilization of the EfficientNetB2 model serves to capture intricate features within images, while the Transformer model contributes towards the formulation of well-structured and contextually apt captions. The dataset used for training and evaluation is [Flickr8k]. This dataset consists of a diverse collection of images matched with their respective captions. Extensive preprocessing is conducted on both images and captions to ensure compatibility with the selected model architecture. This process involves refining and preparing the data prior to input into the model, in order to optimize the overall performance and accuracy of the system. The image captioning model integrates the EfficientNetB2 image encoder with a customized Transformer-based language model. The model is trained on the prepared dataset with careful consideration given to hyperparameters such as batch size, learning rate, and the number of training epochs. This ensures that the model is optimized for performance and accuracy. Results from the training and evaluation phases are presented, emphasizing the model's proficiency in producing captions that accurately correspond with the visual information. Training and validation metrics, in conjunction with caption quality scores, play a key role in providing a thorough evaluation of the efficacy of the model. This study makes a significant contribution to the field of image captioning by demonstrating the efficacy of integrating EfficientNetB2 and Transformer models. The findings of this project provide valuable opportunities for future research and optimization within the field of integrating computer vision and natural language processing. These insights offer potential avenues for further exploration and development in this interdisciplinary area of study.

## **TABLE OF CONTENTS**

PARTICULARS	PAGE NO.
ACKNOWLEDGEMENT	iii
CANDIDATE DECLARATION	iii
CERTIFICATE BY THE SUPERVISOR	iv
ABSTRACT	v
TABLE OF CONTENTS	v
_LIST OF FIGURES	viii
ACRONYMS	ix
 <b>CHAPTER 1</b>	 1
<b>INTRODUCTION</b>	1
1.1 INTRODUCTION TO IMAGE CAPTIONING	1
1.2 LEVELS OF IMAGE CAPTIONING	2
1.3 TYPES OF IMAGE CAPTIONING	3
1.4 CHALLENGES	4
1.5 SCOPE OF WORK	6
1.6 DISSERTATION ORGANIZATION	7
 <b>CHAPTER 2</b>	 8
<b>LITERATURE SURVEY</b>	8
2.1 OBJECT DETECTION METHODS	9
2.2 TEMPLATE BASED METHOD	9
2.2.1 ADVANTAGES OF TEMPLATE BASED IMAGE CAPTIONING	10
2.2.2 LIMITATION OF TEMPLATE BASED IMAGE CAPTIONING	10
2.3 VISUAL TRANSFORMERS FOR IMAGE	11
2.4 MULTI-TASK APPROACH	12
 <b>CHAPTER 3</b>	 14
<b>BACKGROUND TECHNIQUES</b>	14

3.1 ENCODER-DECODER ARCHITECTURES	14
3.2 AN OVERVIEW OF HYBRIDNETS	15
3.3 TRANSFORMER	16
3.4 IMAGE SEGMENTATION	17
<b>CHAPTER 4</b>	<b>19</b>
<b>PROPOSED METHODOLOGY</b>	<b>19</b>
4.1 DEEP LEARNING BASED	
4.1.1 CNN FOR IMAGE FEATURE EXTRACTION	19
4.1.2 RNN FOR LANGUAGE MODELING	20
4.1.3 LSTM	21
4.1.4 STEPS FOR IMAGE CAPTIONING USING CNN-LSTM	22
4.2 TRANSFORMER BASED METHOD	23
4.2.1 THE IMAGE CAPTIONING STEPS USING TRANSFORMER	23
<b>CHAPTER 5</b>	<b>27</b>
<b>EXPERIMENTS RESULT</b>	<b>27</b>
5.1 DATASET PREPERATION	27
5.1.1 ARCHITECTURE USED	28
5.2 RESULTS AND DISCUSSION	30
5.3 SYSTEM REQUIREMENTS	32
5.3.1 LIBRARY USED	32
<b>CHAPTER 6</b>	<b>33</b>
<b>CONCLUSION AND FUTURE SCOPE</b>	<b>33</b>
REFERENCES	34

## LIST OF FIGURES

Figure 1.1: An example of image captioning .....	04
Figure 2.3: Visual Transformer Architecture .....	12
Figure 3.1: An Overview of YOLOv7 Architecture .....	14
Figure 3.2: HybridNets Architecture .....	16
Figure 3.3: Transformers Architecture .....	17
Figure 3.4: Segmentation Example of Road and Pothole .....	18
Figure 4.1: An example of Deep Learning .....	19
Figure 4.2 : CNN-LSTM architecture.....	20
Figure 4.3: System Architecture of Image Caption Generator .....	23
Figure 4.4: EfficientNetv2 Architecture. ....	25
Figure 5.1: A black and white dog is running through a field .....	30
Figure 5.2: A black and white dog is running on a paved road .....	31
Figure 5.3: A man in a yellow kayak .....	31
Figure 5.4: A girl in a yellow shirt is playing cricket .....	32
Figure 5.5: Epoch -loss characteristics .....	32

## ACRONYMS

<b>CNN</b>	:	Convolution Neural Network
<b>LSTM</b>	:	Long Short-Term Memory
<b>GRU</b>	:	Gated Recurrent Unit
<b>RNN</b>	:	Recurrent Neural Network
<b>VGG</b>	:	Visual Geometry Group
<b>BLEU</b>	:	Bilingual Evaluation Understudy
<b>SGD</b>	:	Stochastic gradient descent
<b>BERT</b>	:	Bidirectional encoder representations from transformer
<b>COCO</b>	:	Common Object in Context

# CHAPTER 1

## INTRODUCTION

### 1.1 INTRODUCTION TO IMAGE CAPTIONING

Image captioning is the process of generating a text description or image caption. It combines computer vision techniques with natural language processing to enable machines to understand and describe visual content. The goal of image captions is to automatically generate a human-like textual description that accurately represents the content of the image. This process requires a deep understanding of the visual information present in the image as well as the ability to generate logical and contextually relevant sentences.

Image captioning, serves as the backbone for extracting intricate features from images. The primary goal of image captioning is to create algorithms that can describe and understand the details of the image and describe and state in in a simple language. The implementation of image description systems involves deep learning techniques, that particularly CNN and RNN, they are widely used for image captioning. CNNs used to extract high level featured for the input image and RNNs commonly use LSTM and GRU variants, that are used to generate captions based on the Image features extracted by CNNs, it also generate features word by word. In image captioning attention mechanism and transformer models are also used in which attention mechanism are often integrated into encoder and decoder framework. It focus on different part of images and generate each word in descriptive language. Transformer based models or mechanism it capture a global dependencies between image and generate more context base image captions. there are some different variants in it such as BERT it utilise self attention mechanism to capture descriptive relationships between words in a sentence.

When we train in image captioning ,model learns to measures dissimilarity between the generated captions, it includes minimizing a loss function includes cross-entropy loss and various reinforcement learning based objectives.

Evaluation of image captioning by several matrices that are used to evaluate quality of the captions that is generated by above methods. There are different matrices that is used to evaluate like BLEU(Bilingual Evaluation Understudy), METEOR(Metric for

Evaluation of Translation with Explicit ordering), CIDEr(Consensus based Image Description Evaluation)and ROUGE-L(Recall Oriented Understudy for Gisting Evaluation - Longest Common Subsequence).

## **1.2 LEVELS OF IMAGE CAPTIONING**

The image captioning can be categorized into different levels based on quality , complexity, edification and sophism of the captions that are generated from images after the feature extractions. There are several levels of image captioning:

**Descriptive captioning:** This level describes that what is visually present in the image, it focus on objects, people, animal, vehicles and different scenes.

**Contextual Captioning:** In this level, captions provide more context by considering their relationships and involvement within the image it also describes the visual elements of an image.

**Creative Captioning :** In this level it embrace generating caption that are beyond demonstration in form of template, in which it generate captions on based on story telling. In this caption it may also involves riddles, different language, humorous things, metaphor and poetic language.

**Narrative captioning :** This level involves creating a story around the image, including together multiple elements to create a coherent and informative description. It also gives demonstration, references, imaginative elements to describe the image.

**Analytical captioning :** In this level it involves analyzing the images in detail often consolidating elements of critical thinking, demonstration and evaluation. In analytical caption cultural significance, symbols and art style techniques that are visible in image can easily be described and generated.

**Interactive captioning :** In this level it involves the viewers by creating interaction , response and feedback, it may suggest questions or suggest alternate prospective and viewers can also imagine themselves in the scene.

Educational Captioning : In this level it provide informative content that can educate viewers about image features. It may includes factual information, historical evidence, scientific relations or cultural activities that can give information to viewers.

### **1.3 TYPES OF IMAGE CAPTIONING**

We can classify image captioning in various different criteria, such as input data required, output format, method used to generate captions. Here are several types of image captioning:

(1) Template based captioning: In this type of captioning it gives an approach that generates image captions by using predefined templates or sentence structures and we extract information from image by using computer algorithms, Computer vision techniques are use to generate predefined template, In this type it is more flexible to arrange the elements of images. A template have slots for subject, action and location, instead of generating captions from start, template-based methods provide a well structured foundation to guide the caption generation process.

(2) Deep learning based captioning: In this type of image captioning particularly it use CNN(conventional neural network) for image processing, RNN(recurrent neural network) for caption generation and transformers work with efficientb2 with proper synchronization to find relation between image and its objects that is, Encoder-Decoder architectures for image captioning Deep learning based techniques mainly based on large datasets of paired images and captions to learn the relationship between images and there captions that is features are learned automatically from training data and they can handle a large and diverse set of images in order to generate captions.

(3) Traditional rule based captioning: this type of captioning relies on pre-defined templates and rules to generate description based on the objects or content in the image, in this it involves humanly crafted rules for identifying the content in image that is objects, scenes and things inside the images a rule based system that have template fills the blanks based on detected objects in images .

(4) Multimodal captioning: this method integrate information from multiple modes such as image, text, audio to generate description. It can combine visual features that are extracted from images with textual features from associated attributes or audio features



from stated descriptions. It can generate captions for videos by just analyzing audio tracks and visual content.

(5) Conditional captioning: In this approach generation of captions is based on condition to generate caption that is based on specific metadata, attributes or requirements. It gives more control over the generated captions, it can generate captions based on specific contents or style and requirements such as caption suitable for focusing on humor.



Fig. 1.1 An example of image captioning

## 1.4 CHALLENGES

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 and 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 VISUAL TRANSFORMER

- Transformers process sequential data, but images have spatial structures. It needs to consciously handle this spatial information to understand the different objects of an image this is known as spatial information handling.
- There are various changes can occur in image captioning because of their different size and complexity but our visual transformer must be able to handle the process of both small and large size images without disrupting their performance and efficiency.
- There is long range dependency, a type of challenge in image caption by visual transformer in which many images contain information like scenes or objects that covers a large area of images so visual transformer must capture long range dependencies to generate accurate captions that describe their entire scene inside the image systematically.
- Visual transformers often require large amounts of required data for training an image, which is hard to be available at specialized domain. Data efficiency is a very important part of visual transformers and a major challenge.
- Visual transformer use many evaluation matrices such as BLEU(Bilingual Evaluation Understudy) and METEOR(Metric for Evaluation of Translation with Explicit ordering), these evaluation matrices may not be able to capture the quality and diversity of generated captions specially for longer, wider and more hard to understand or complex descriptions, to address this challenge it requires more research into models, methods, architectures, training, strategies in evaluation matrices.
- Transformers typically require considerable computational resources, especially as image resolution increases. This can make training and conclusion computationally expensive, this will limit their practical applicability. So transformers are computationally complex.

- In image captioning it involves combination of both visual and textual approach effectively. Transformers need to learn to combine information from both visual and textual modalities while generating textual description effectively. It is necessary to handle multi modal inputs.

## **1.5 SCOPE OF WORK**

Significant objectives and contributions of this work are:

- To develop a model that can productively understand visual content of an image and generate descriptive descriptions or captions.
- To explore transformer based architecture customize for processing visual data, such as visual transformer optimized for image understanding.
- To researching, developing and comparing various neural network models that are customize for image captioning such as CNN(conventional neural network), RNN (recurrent neural network ) and transformer based methods.
- To define and utilize evaluation matrices and benchmarks assess the information from the Image that is quality of generated captions, including matrices like BLEU(Bilingual Evaluation Understudy) and METEOR(Metric for Evaluation of Translation with Explicit ordering), CIDER(Consensus based Image Description Evaluation),ROGUE(Recall Oriented Understudy for Gisting Evaluation ) and SPICE(Semantic Propositional Image Caption Evaluation) these can measure major factors like eloquence, smoothness, connection, variety etc.
- To investing pre-trained models, those trained on large scale datasets like flickr8k and fine-tuning them for image captioning tasks to litigate training and improve performance.
- Major scope of work is to explore the techniques to improve the understanding of visual content and its information with textual captions, including multi-modal combination method and important mechanism that combine bot image nad text modalities.
- Last major scope of work is to use image captioning in real world in form of application that are image search engine, assistive technology for visually distiguish an individual , contents, required information, creation, tools, multimedia contents and understanding the systems.

- By addressing these areas a researcher can advance in the state of image captioning and develop real life solution with different types of applications.

## **1.6 DISSERTATION ORGANIZATION**

The content of the dissertation is organized into six chapters:

- Chapter I INTRODUCTION TO IMAGE CAPTIONING
- Chapter II LITERATURE SURVEY
- Chapter III BACKGROUND TECHNIQUES
- Chapter IV PROPOSED METHODOLOGY
- Chapter V EXPERIMENTAL RESULTS
- Chapter VI CONCLUSION AND FUTURE SCOPE

**Chapter I** – Includes the introduction to image captioning by visual transformers and overview about types and challenges about image captioning..

**Chapter II** – This chapter is literature survey, which gives an insight about the research papers published based on object detection in an image, deep learning and architectural approaches.

**Chapter III** – This chapter gives an insight into the background techniques that are being used in the implementation of visual transformer by using image captioning.

**Chapter IV** – This chapter covers the methodology that includes deep learning and visual transformers.

**Chapter V** – This chapter includes the experimental results. The results also involve performance comparison between CNN and Transformers.

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

## CHAPTER 2

### LITERATURE SURVEY

#### 2.1 OBJECT DETECTION METHODS

Deep learning models specializing in bounding-box-based object detection on two-dimensional image signals have can be particularly useful in predicting obstacles, such as vehicles, on roadways. Several methods have been proposed over the years to improve upon the process of object detection. Deep learning algorithms, notably Convolutional Neural Networks (CNNs) [9], have demonstrated remarkable capabilities in achieving precise object identification. However, they are often associated with much prolonged inference times. The Region-Convolutional Neural Network (R-CNN) generates throughput comprising sets of bounding boxes corresponding to obstacles, along with the corresponding class identification outputs [10]. RCNN relies on a selective search algorithm as its foundation, which aids in predicting objects within designated regions. To accomplish this, RCNN employs linear Support Vector Machines (SVM) as its classification method. The approach exhibits limitations in terms of real-time detection speed primarily stemming from redundant feature computations. Subsequent advancements were introduced through the enhancements made via Fast R-CNN [14] and Faster R-CNN [15]. In contrast to RCNN, Fast R-CNN adopts the approach of utilizing the entire image as input for feature extraction, rather than processing each proposed region individually. However, it still faces challenges in achieving efficient detection speed, primarily due to limitations imposed by proposal detection. Conversely, Faster R-CNN eliminates the need for the selective search algorithm by enabling the network to learn region proposals, thereby positioning itself as one of the pioneering near-real-time deep learning object detectors. Although Faster R-CNN successfully overcomes the speed limitations of Fast R-CNN, it still encounters computational redundancy in subsequent detection stages.

## 2.2 TEMPLATE BASED METHOD

It is one of the most used and oldest methods of image captioning in which it generates image descriptions by filling in pre-established templates with features extracted from the visual content, now let us define how template based methodology works:

- In template based captioning, there are templates that represent a specific structure for generating textual description of an image and these templates are pre-defined. This may vary from complexity and technicality depending on the desired level of details in generated textual description. These templates consist of placeholders or slots that can be filled with information related to the information of the image.
- The very first step in template based captioning is to analyze the information from the input image to extract important content. It uses computer vision techniques such as object detection, feature extraction, scene classification, visual content extraction are used to analyze and understand the image and identify visuals, scenes and other objects in the image.
- After analyzing the content of the image the feature extraction from the image, such as object labels, structural relationships and visual characteristics are drawn to the corresponding positions in the pre-defined templates. These visual features serve as input to the template based captioning system, providing the necessary information to generate textual description of an image.
- After the visual features extracted from the image such as objects, scenes and information and visual attributes, they are used to fill in the surrogate or position in the pre-defined templates. The filled templates represent candidate captions for the input image, with each template capturing a different aspect or interpretation of the visual content.
- In some cases of template methodology multiple templates can be generated for a single input image, each providing different levels of detail and different perspective. So for choosing a correct or accurate template that is most suitable based on the

criteria of coherence, relativity or user preferences a selection mechanism is created. This process is called as template selection.

- After selecting the final template, we have to provide some post-treatment steps that may applied to clarify the generated description. To clarify this may involve grammatical correction, language fluency improvement or context adjustment to ensure the quality and readability of the generated description.
- Now, at last for output in template based method, image captioning is a set of caption generated from the input image and every set confirms that it is a pre defined template structure and these pre defined templates provide a interpretation of the visual content, organized according to the chosen template.

#### **2.2.1 Advantages of template-based image captioning :**

- Template based methodology is used to design structured model, that analyze and synthesize the caption visual caption process. It also decide structured framework of sentences. It has a guided caption generation.
- Templates gives us control over the generated captions, enabling steadiness and sustainable in the output. It give us control and consistency in image captioning.
- Template-based methods can be mathematically effective as they avoid the need for complex language models or serially generated algorithms.

#### **2.2.2 Limitations of template-based image captioning:**

- Templates are restrictive they limit the diversity and creativity of the generated captions they are not flexible.
- Template-based methods have limited adaptability they struggle to handle images that do not fit well within the predefined templates or contain rare or unusual visual content.
- Difficulty handling, they face difficulty in handling complex images.

Template-based approaches may face challenges in generating accurate and detailed captions for images.

## **2.3 VISUAL TRANSFORMERS FOR IMAGE**

Visual transformers identifies itself as a type of advanced learning architecture that relates the transformer model straight away to image data, by dealing with image patches in order of token, very similar to the words we use in a sentence. When modifying for image captioning, Visual Transformers can provide attachment to their powerful self-attention mechanism in order to generate descriptive captions.

In visual transformer architectures, used to process visual feature extracted and textual feature from the caption of an image. Transformer model can be pre-trained on vast of dataset same as natural language processing and computer vision.

A survey has been provided in order to understand how exactly visual transformers perform on image captioning, along with its application, merits and demerits.

Here are some merits of Visual transformer:

The self-attention mechanism allows the model to capture long-range dependencies and global context, which is crucial for understanding complex scenes in an image it is important for Global Context Understanding.

Visual Transformers can be scaled by increasing the number of layers or the size of the embeddings, It can scale very easily.

Here are some Demerits of Visual Transformer:

Visual Transformers require remarkable computational resources, especially for large images and deep models they are computationally intensive.

Training effective, visual Transformers often requires large datasets to capture the diverse visual and textual relationships, they require more data for getting the image content. They are highly data hungry.



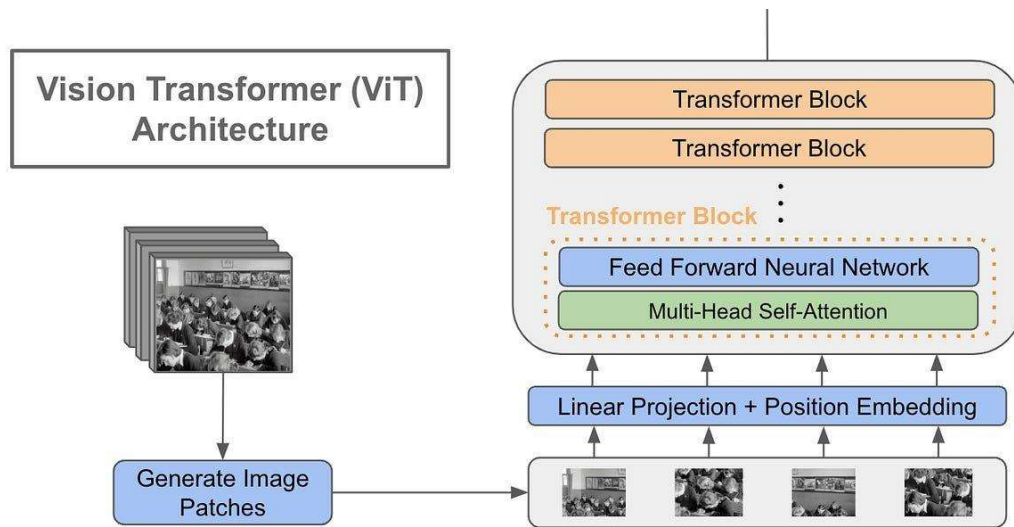


Fig. 2.3: Visual Transformer Architecture

## 2.4 MULTI-TASK APPROACH

Multi task approach of visual transformer methodology is in multiple research areas, it can simultaneously do multiple task that can enhance the efficiency of this method and its performance , we can use visual transformer in multiple areas and it has multiple applications that are following :

- It can be used in Autonomous Vehicles, It can understanding the scene, detecting objects, and segmenting road elements to assist in navigation and safety all at the same time.
- One of the growing department is Robotics, in which it can Multi-task learning for scene understanding, object manipulation, and navigation.
- In medical fields or healthcare department, it can generate image captions for medical images, detecting anomalies, and segmenting regions of interest all at the same time or simultaneously.

- It can moderate the in research fields, It can create automated systems that can caption images, detect inappropriate content, and segment images for further analyzing and researching for different fields.
- It can give us a resource efficiency, It shares the encoder that reduces the computational resources required compared to training separate models for each task, without encoder they are computationally intensive.
- It can improve performance by joint learning that can improve performance on separate tasks by manipulating shared representations and mutual information.
- Ensures consistency across related tasks since they are learned manipulating from the same feature representations, they are highly consistence.
- A single model operating multiple tasks simplifies arrangement and maintenance in production environments.

## CHAPTER 3

### BACKGROUND TECHNIQUES

#### 3.1 ENCODER-DECODER ARCHITECTURES

Encoder-Decoder architectures are simple structures that perform feature extraction by generation of feature maps in the first part and detection/segmentation in the next part. In other words, the proposed architecture consists of an encoder component designed to process input sequences of varying lengths, and a decoder component that functions as a conditional language model. The encoder receives the input sequence and performs encoding operations to capture its underlying features. On the other hand, the decoder utilizes the encoded input as well as the preceding context of the target sequence to predict the next token in the target sequence. This dynamic interaction between the encoder and decoder enables the model to effectively generate accurate predictions for subsequent tokens based on the contextual information. YOLOv7 is also an encoder-decoder architecture, with its backbone serving as the encoder and head as its decoder.

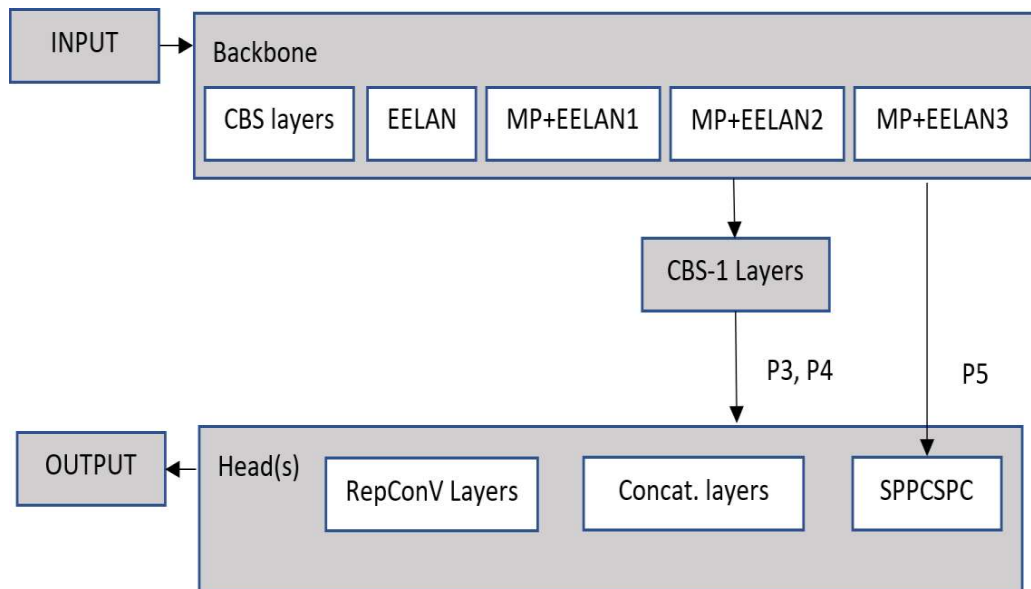


Fig. 3.1 An Overview of YOLOv7 Architecture

### 3.2 AN OVERVIEW OF HYBRIDNETS

The Hybrid Nets model, similar to the proposed methodology, adopts an encoder-decoder network architecture. It integrates both semantic segmentation and object detection techniques, allowing for the comprehensive detection of various elements within the scene. By incorporating these two decoders, namely object detection and caption generation, the model is capable of simultaneously addressing multiple aspects of interest. This comprehensive approach enables the HybridNets model to effectively analyze and interpret the complex visual information captured on the image. EfficientNet-B2, an influential component of the architecture, assumes the role of the encoder-backbone. EfficientNet-B2 represents a class of neural network architecture designed to balance performance and efficiency across different processes, including image captioning, image classification and object detection. This architecture is designed to achieve better accuracy and efficiency by optimizing the model's depth, width and resolution. It is a one of the specific variant of EfficientNet family distinguish by its average size and computational efficiency. It is larger and more powerful than smaller variants like EfficientNet BO, but more lightweight compared to larger variants like EfficientNet-B3 and EfficientNet-B7. It makes a stability between model complexity and computational difficulties, making it acceptable for wide range of assignments. It can scale where the width, depth and resolution of the architecture uniformly. It consist a pile of convolutional neural network layers, followed by batch organization, activation function and down sampling operation such as max pooling. The architecture also compromise bottleneck layers and compress excitation blocks to intensify representative capacity and method articulateness. The essential goal of Efficient Net-B2 is to increase accuracy while decreasing computational complexity resources and memory impression. By carefully stabilizing architecture size and computational complexity, It can achieve state of the performance on various computer vision tasks, including image classification, object detection and image text generation. In transfer learning scenarios, this algorithm used as feature extraction method where trained weights learned on large scale datasets are fined turn on task specific datasets. Transfer learning with Efficient Net-B2 gives researchers and interpreter to attach the representation power of the algorithm for specific assignments with limited computational complexity.

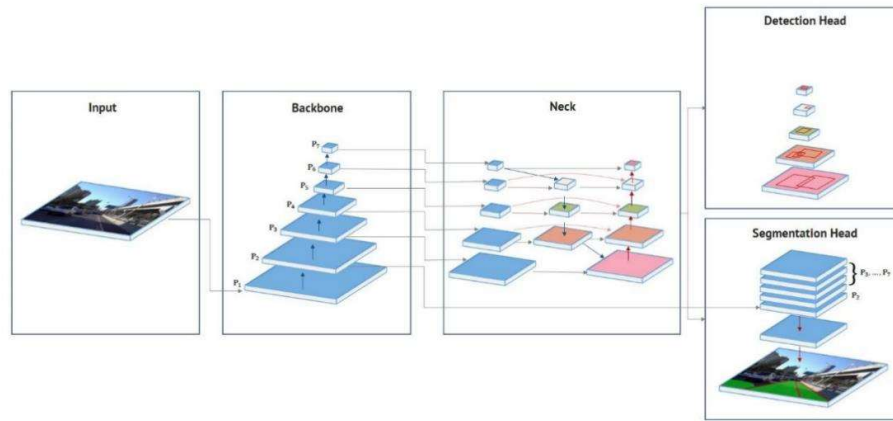


Fig. 3.2 HybridNets Architecture [30]

### 3.3 TRANSFORMER

This architecture in image captioning used to develop a natural language processing task, in generating descriptive text for an images. This process is approach that includes adaptive transformers models to process visual data effectively and generate logical and provisional relevant description. There are some major aspects of using transformer in image captioning that are model architecture, pre-training, fine tuning, mechanism, multimodal combination, Evaluation architecture etc.

Transformer is a important tool which is a type of a neural network based on self attention mechanism to extract important features and provide relationship between different elements serially. In visual transformer architectures, used to process visual feature extracted and textual feature from the caption of an image. Transformer model can be pre-trained on vast of dataset same as natural language processing and computer vision. The dataset can be imageNet and COCO(Common Object in Context) which is improve the performance of visual transformer, This type of model based on attention mechanism which focused on important parts of image and extract important characteristics, attention mechanism enable the model to scan all parts of an image while generating textual information. Transformer consist both visual and textual algorithm using multimodal combination techniques, this model combine accurately visual with textual caption. This model is evaluated by using different types of evaluation metrix such as ROGUE(Recall Oriented Understudy for Gisting Evaluation), SPICE(Semantic Propositional Image

Caption Evaluation ), METEOR(Metric for Evaluation of Translation with Explicit ordering), BLEU(Bilingual Evaluation Understudy)and COCO(Common Object in Context).

Transformer based approach for image captioning having capability to capture long range dependency and contextual information to improve visual transformer .

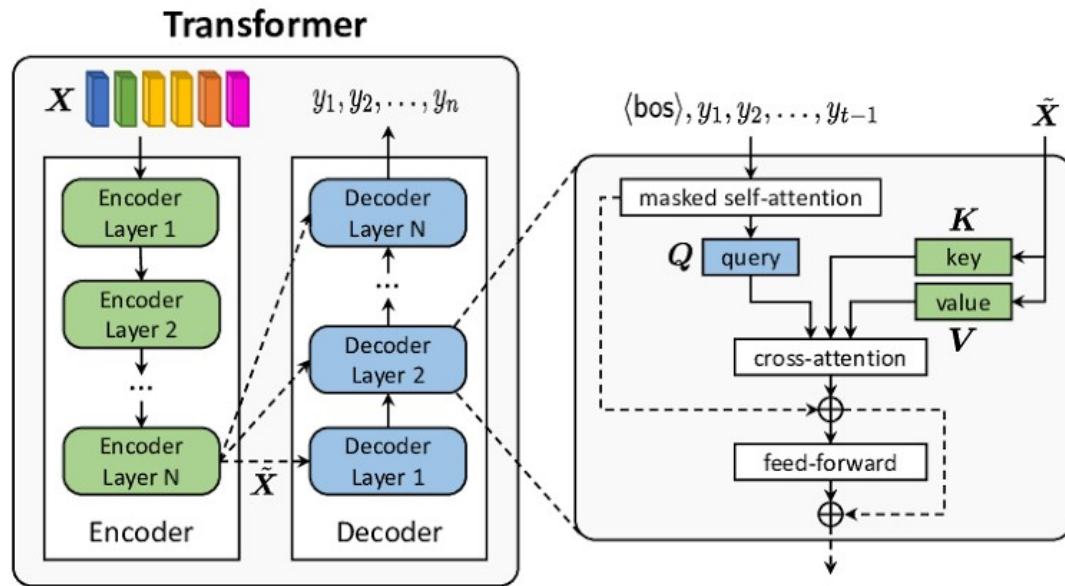


Fig 3.3 Transformers Architecture

### 3.4 IMAGE SEGMENTATION

Image segmentation is a sophisticated technique employed in the realm of computer vision to partition a digital image into discrete and meaningful subgroups known as image segments. This process serves the purpose of simplifying the complexity of the image, thereby facilitating subsequent processing or analysis on each individual segment. From a technical standpoint, segmentation entails the intricate task of assigning labels to individual pixels, thereby discerning and identifying objects, people, or other significant elements present within the image. In order to optimize computational resources and improve the efficiency of object detection, a prevalent approach involves

the utilization of an image segmentation algorithm to identify and extract objects of interest within the image. By employing this strategy, the subsequent object detector can focus exclusively on the predefined bounding boxes derived from the segmentation algorithm's output. This targeted approach eliminates the need for the detector to process the entire image, resulting in enhanced accuracy and reduced inference time.



Fig 3.4 Segmentation Example of Road and Pothole

## CHAPTER 4

### PROPOSED METHODOLOGY

#### 4.1 DEEP LEARNING BASED

One of the most excellent strategy is profound machine learning based strategies, highlights, substance and data are learned consequently from preparing information and they can handle a huge and different sets of pictures in arrange to create captions. Here are a few illustration of profound learning strategies,.

- Convolutional Neural Networks (CNNs) for image feature extraction
- Recurrent Neural Networks (RNNs) for language modeling
- Encoder-Decoder architectures for image captioning

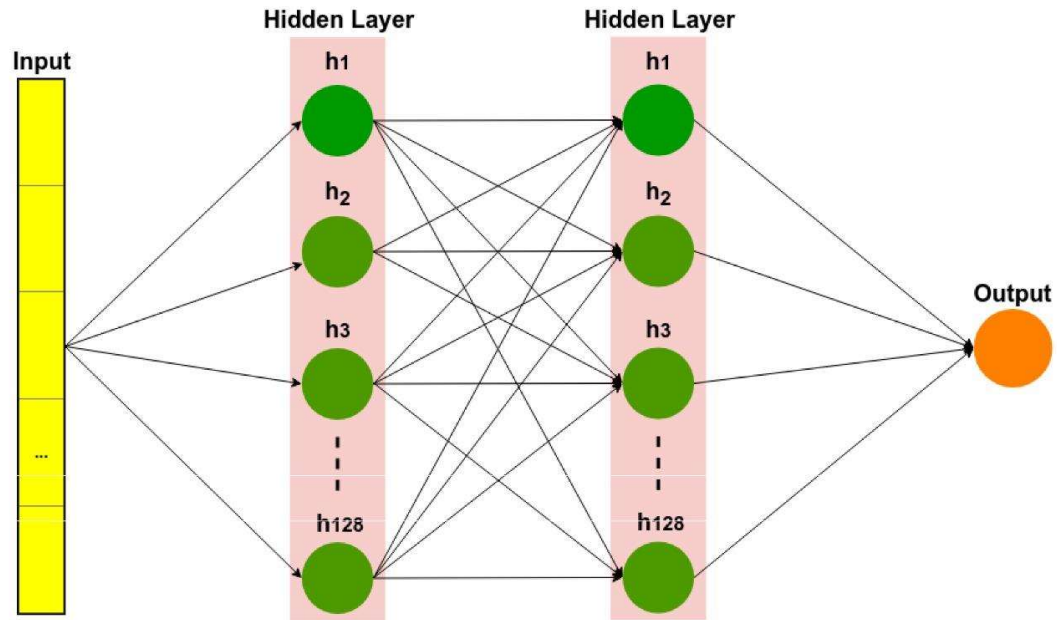


Fig 4.1 An example of Deep Learning

##### 4.1.1 Convolutional Neural Networks (CNNs) for image feature extraction

###### *CNN-LSTM*

CNN-LSTM is a popular favoured architecture for image captioning that combines Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks. This architecture holds the strengths of CNNs in image feature extraction and LSTMs in customizing sequence to generate descriptive textual captions for images. Here's an overview of the CNN-LSTM architecture for image captioning:



$$\theta^* = \arg \max_{\theta} \sum_{(I,S)}^n \log p(S|I; \theta)$$

$$\log p(S|I) = \sum_{t=0}^N \log p(s_t|I, s_0, \dots, s_{t-1})$$

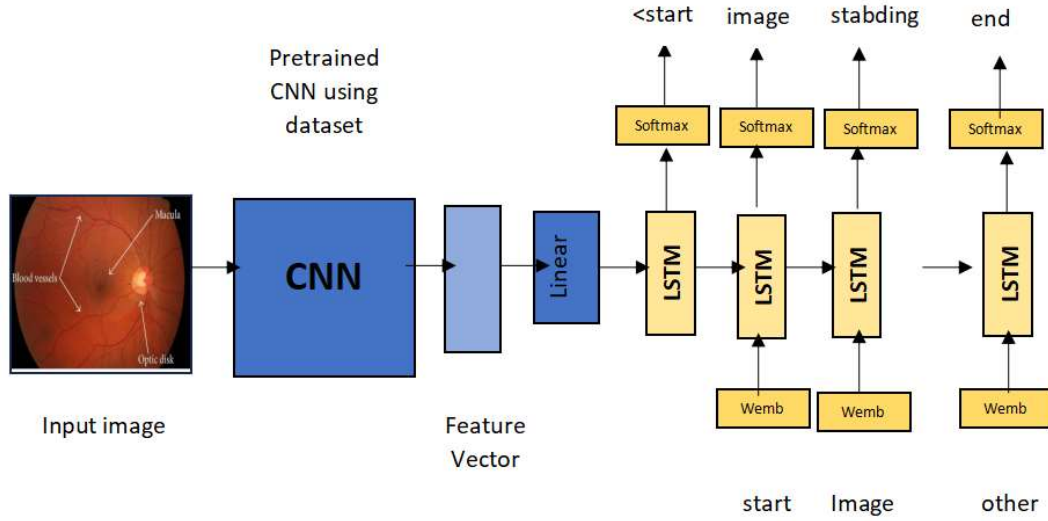


Fig 4.2 CNN-LSTM architecture

In CNN-CNN based show where CNN is utilized for both encoding and translating groundswatch that CNN-CNN demonstrate has tall misfortune which isn't passable as the createdcaptions won't be precise and the captions created here will be adjacent to the point to thegiven test picture.CNN was planned to form picture information to an yield variable. The benefit of utilizing CNN istheir capacity to create and speak to the two dimensional capacity. This permitted the demonstrateto memorize position and scale consistent structure within the information on working with pictures. Itworked well with the information which has computative connections.

#### 4.1.2 Recurrent Neural Networks (RNNs) for language modeling

There is CNN-RNN in which there is less loss while image captioning as compared to CNN-CNN based model and the training time is more in CNN-RNN, but training time affects the whole efficiency of the method, here is another limitation in RNN and that problem is vanishing gradient problem. In which gradient update the weights during

backpropagation that become very small, effectively causing the learning process to wait in RNN method.

Gradient is the parameter which is used to calculate the rate of loss per the given input parameter comparing both inputs and outputs. If RNN and LSTMs in general particularly have received the most success when working with sequence of word and paragraph, this is called natural language processing. This include both sequences of text and sequence of language

RNN used for:

- 1) Text data
- 2) Speech data
- 3) Generative word

In Repetitive neural systems it has the capacity to work as comparable as human brain it can make sense of a word and keep this in intellect and create the another word, essential Neural systems don't have the capacity to do this. As we know the human brain is advanced in such a way so as to form sense of past words, and keeping these in intellect create the following words, in this way shaping a idealized sentence. In any case, headways in Repetitive neural systems address this issue. They are systems in RNN with circles in them, permitting data to preserve for a whereas within the circle, by making utilize of their inner states, in this way making a criticism circle.

#### **4.1.3 Long Short Term Memory networks – usually just called “LSTMs” –**

LSTM may be a long brief term memory which are an extraordinary kind of RNN, have the potential of learning long-term dependence. They contain memory cell which is utilized to store information for long time period and keep in mind data for long periods and for all intents and purposes its their default behavior and this behavior is controlled with the assistance of “Gates”. as we know RNNs handle single information focuses but LSTMs can handle whole groupings. They have capacity to hold vital information for visual change and they can toss absent

pointless information. Thus, the as it were significant data is passed on to the another layer. LSTM units utilize a memory cell that can keep up data in memory for long periods of time. Presently days, most of the assignment that utilized in arrangement to arrangement learning assignment are LSTMs based models..

#### **4.1.4 Steps For image captioning by using CNN-LSTM**

1. The input image is provided to CNN vector, such as VGG16, ResNet, or Inception, to analyse and detect high-level visual features. The CNN processes the image in a ranked manner, capturing both low-level and high-level visual representations this process is called as CNN Feature Extraction.

2. We can perform sequence modeling with LSTM by extracting the visual features and then fed as input to this network. This models the sequential nature of captions by processing the features word by word serially. This method takes the visual features along with the previously generated words as input at each time step.

3. Each word within the caption is spoken to as a word inserting vector. These word embeddings capture the satisfactory meaning of the words and offer assistance the demonstrate get it the connections between distinctive words.

4. The LSTM creates captions by anticipating the following word within the grouping based on the Input visual highlights and the already produced words. This handle proceeds until an end-of-sentence token is created or a predefined most extreme caption length is come to, this lead to caption description.

5. The CNN-LSTM show is prepared employing a datasets with matched image-caption cases. The model is optimized to play down the irregularity between the anticipated captions and the ground truth captions. This includes backpropagating the slopes and overhauling the demonstrate parameters.

6. To form interface, the prepared CNN-LSTM demonstrate is utilized to produce captions for unused pictures. The picture is handled through the CNN to extricate visual highlights, which are at that point passed to this strategy for caption era. They can predicts each word within the caption serially, going to to the visual highlights and already created words of an picture.

The CNN-LSTM engineering benefits from the CNN's capacity to extricate important visual highlights from pictures and the LSTM's capability to show successive conditions

in content. By combining these components, the demonstrate can produce captions that are relevantly significant to the picture substance and phonetically coherent.

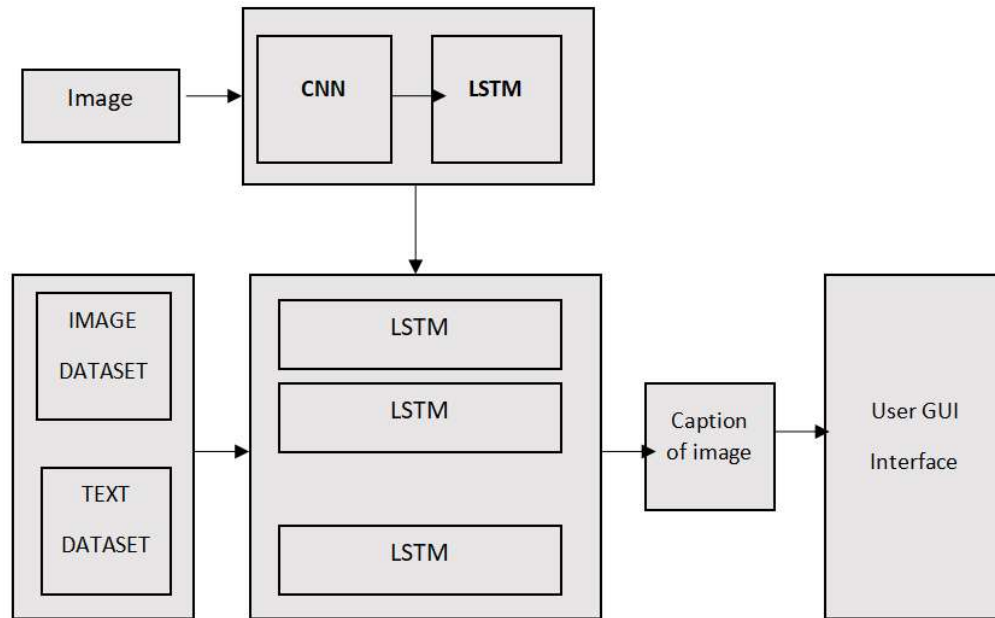


Fig 4.3 System Architecture of Image Caption Generator

## 4.2 TRANSFORMER BASED METHOD

The transformer network plays an important role in an encoder-decoder architecture which is very similar to an RNN. Transformers being able to receive input sentences or sequences in parallel is the most important difference, as there is no time step connected with the input and simultaneously all the words in the sentence are passed.

### 4.2.1 The image captioning steps using transformers are followed below :-

Let's begin with comprehension the input to the transformer

#### 1. EfficientNetB2 for Image Feature Extraction:

Architecture :- the efficient image categorization is designed through convolutional neural network (CNN), which is also known as EfficientNetB2.it carries a series of layers which involves convolution layers, normalization layers and squeeze-and-excitation

blocks.

Extraction characterisation :- the network on large scale image characterization task is pre-trained. You'll have to remove the final classification layer, in order to use it for image captioning and from the last convolutional layer as image features use the output. It's the same layer which captures high level abstract features.

## **2. Language Modeling by Transformer Model :-**

Architecture :- neural network architecture comes under the category of Transformer Model which is introduced for natural language processing tasks. The mechanism used in it is self-attention, permitting weighing of different words in separate sequences by the model.

Attention Mechanism: Attention mechanisms help the model to focus on other parts of the input sequence while generating output sequence of each element. language modeling and sequence-to-sequence tasks attain most benefit from it.

Positional Encoding: the order of the input sequence has not been understood by Transformers. In order provide information about the positions of tokens in the sequence, the addition of Positional encoding is done to the input embeddings.

## **3. Combining EfficientNetB2 and Transformer for Image Captioning:**

Image Features Integration: The image features extracted by EfficientNetB2 are combined with the tokenized and embedded captions before feeding them into the Transformer model.

Dual Input :- the image features as well as the tokenized captions are the two inputs taken by model during training and inference. The context about the visual content is provided by the image features, and the language modeling aspect is provided by the captions.

Concatenation :- Before the process of passing through the final layer to predict the

upcoming word in the sequence two things are carried out, i.e the image features and text embeddings are Concatenation.

**Training Objective :-** To minimize the dissimilarity among the predicted captions and the ground truth captions, the model is been trained. Optimizing the model parameters using gradient descent and using a suitable loss function (for example sparse categorical cross entropy) are involved in it. The model has been trained to generate captions that are semantically and syntactically just like ground truth.

#### 4. Training and Inference:

**Training Data :-** training data occurs by pairs of images consisting of a dataset and corresponding captions. Captions that are semantically and syntactically alike to each other to the ground truth are known as trained models.

**Inference :-** throughout inference, for a given new image, the caption is generated by the model through repetitively predicting the upcoming word in the sequence. until an end token is predicted or maximum length sequence is reached this process is repeated.

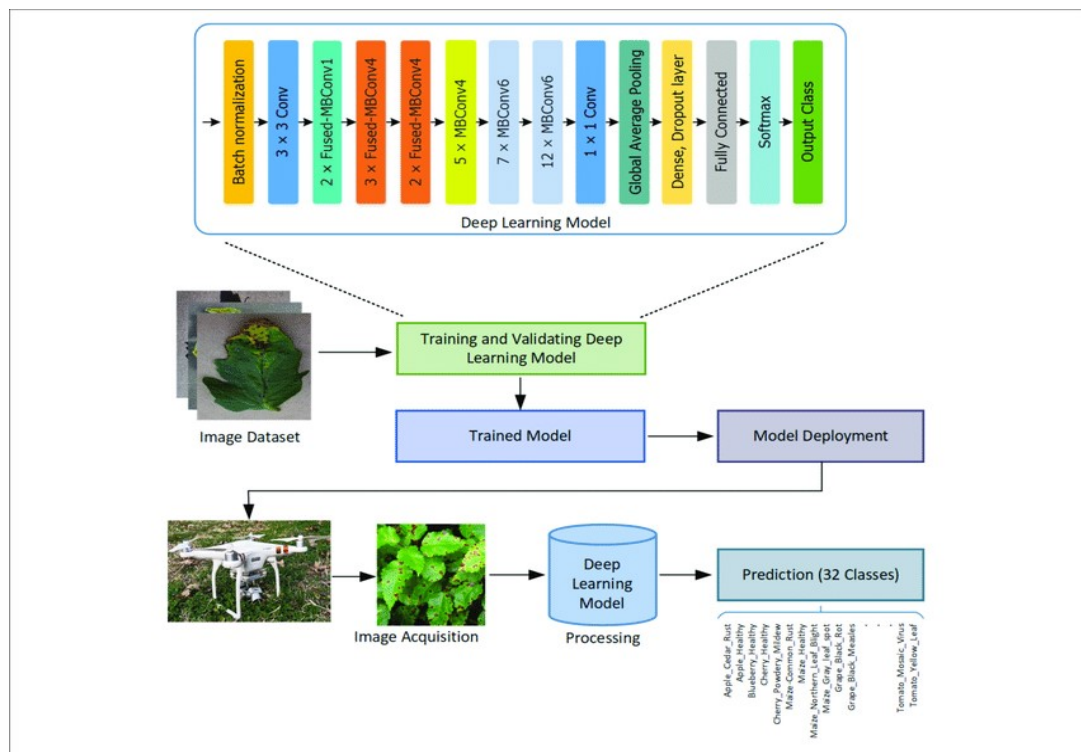


Fig. 4.4 EfficientNetb2 Architecture

### **5. Hyper parameter Tuning and Optimization:**

Hyper parameters is a type of parameters such as learning rates, and dropout rates need to be carefully tuned for superlative performance. Embedding dimensions, the number of attention heads are also tuned so they can give highest performance.

They can manipulate pre-trained models for both EfficientNetB2 and the transformer, when dealing with limited labeled data and that can significantly improve performance, this is called transfer learning. Implementing and fine-tuning in such a method requires a deep understanding of computer vision, natural language processing, and the specific characteristics of dataset. It's often a constant process involving experimentation and tuning to achieve the best results while text generation.

## CHAPTER 5

### EXPERIMENTAL RESULTS

#### 5.1 DATASET PREPARATION

The Flickr 8K, that was presented by Hodosh et al dataset. It may be a broadly utilized benchmark dataset for inquire about in picture captioning. Information, Models, and Evaluation Measurements are in this dataset. Within the dataset, it comprises of 8,000 pictures in which each went with by five human-generated printed portrayal, coming about in a add up to of 40,000 captions.

Key features of the Flickr 8K dataset are as follows:

1. It includes a expansive no. of diverse estimate and assortment in pictures, The dataset contains a blended and a huge collection of pictures from Flickr, covering a wide run of scenes, objects, and exercises. The pictures speak to different real-world scenarios, making it appropriate for preparing and assessing picture captioning models on general-purpose picture understanding..
2. Human-Annotated Captions: Each picture within the Flickr8K dataset is related with five distinctive captions, giving different portrayals for each picture. These captions were collected through a crowdsourcing prepare, guaranteeing a assorted set of printed portrayals for the pictures.
3. Dialect Characteristics: The captions within the dataset display common dialect characteristics, counting varieties in lexicon, sentence structure, and composing fashion. This makes a difference in preparing picture captioning models to produce etymologically assorted and relevantly important captions.
4. Part into Prepare, Approval, and Test Sets: The Flickr8K dataset is separated into three sets for preparing, approval, and testing purposes. The preparing set comprises of 6,000 pictures, the approval set contains 1,000 pictures, and the remaining 1,000 pictures are portion of the test set. This part permits analysts to assess the execution of their models



on concealed pictures amid testing.

5. Standardized Assessment: The Flickr8K dataset gives a standard benchmark for assessing picture captioning models. Common assessment measurements such as BLEU (Bilingual Assessment Understudy), METEOR (Metric for Assessment of Interpretation with Unequivocal Requesting), CIDEr (Consensus-based Picture Depiction Assessment), and ROUGE (Recall-Oriented Understudy for Gisting Assessment) are frequently utilized to survey the quality of created captions against the ground truth.

The Flickr8K dataset has been broadly received within the picture captioning inquire about community and has served as the premise for creating and assessing different captioning models. It has encouraged progressions within the field and contributed to the improvement of novel strategies for producing exact and relevantly significant picture captions.

#### **5.1.1 Architecture Used**

**EfficientNetB2** : EfficientNetB2 could be a variation of the EfficientNet family of neural organize designs, which are planned to supply a great adjust between demonstrate exactness and computational efficiency. EfficientNet was presented by Mingxing Tan and Quoc V. Le within the paper "EfficientNet: Reconsidering Demonstrate Scaling for Convolutional Neural Systems."

Key characteristics of the VGG16 architecture are as follows:

##### **1. Architecture:**

- EfficientNetB2 could be a neural organize design known for its productivity in terms of demonstrate measure and computational assets.□
- It takes after a compound scaling methodology, adjusting width, profundity, and determination to optimize execution.□
- The essential building piece is the Versatile Modified Bottleneck Convolution (MBConv) square, comprising of depthwise distinct convolutions and straight bottlenecks.□
- The design incorporates worldwide normal pooling and a last thick layer for classification

## 2. Deep Stacking:

- EfficientNetB2 is outlined with profound stacking, highlighting numerous layers stacked to capture various leveled highlights.□
- The profundity of the arrange permits it to memorize complex and unique representations from input pictures.

## 3. ImageNet Pretraining:

- □EfficientNetB2 is laid out with significant stacking, highlighting various layers stacked to capture different leveled highlights.□
- The significance of the orchestrate grants it to memorize complex and special representations from input pictures.

## 4. Feature Extraction:

- EfficientNetB2 serves as an compelling highlight extractor in picture preparing errands.□
- The organize forms input pictures through its layers, creating highlight maps with dynamically higher-level representations.
- Deeper layers capture more theoretical and semantic highlights, whereas shallower layers center on low-level visual designs.□
- These include maps can be utilized as input for consequent layers or models, encouraging assignments like picture classification or caption era..

## 5. Transfer Learning:

- □Leveraging its solid highlight extraction capabilities, EfficientNetB2 is well-suited for exchange learning.□
- Transfer learning includes utilizing pretrained weights and include representations from EfficientNetB2 as a beginning point for a particular assignment.□
- Fine-tuning the organize on a littler, task-specific dataset empowers adjustment to unused challenges, such as picture captioning.

EfficientNetB2's effective plan and capable include extraction make it a important choice in different computer vision applications. Within the setting of picture captioning, the organize can be utilized to extricate significant highlights from pictures, which are at that point utilized by ensuing models, such as transformers, to produce expressive captions for the input pictures.

## 5.2 RESULTS AND DISCUSSION

It is simple to see that all of these caption era models can produce to some degree important sentences, whereas the CNN-based models can foresee more high-level words by together abusing history words and picture representations.

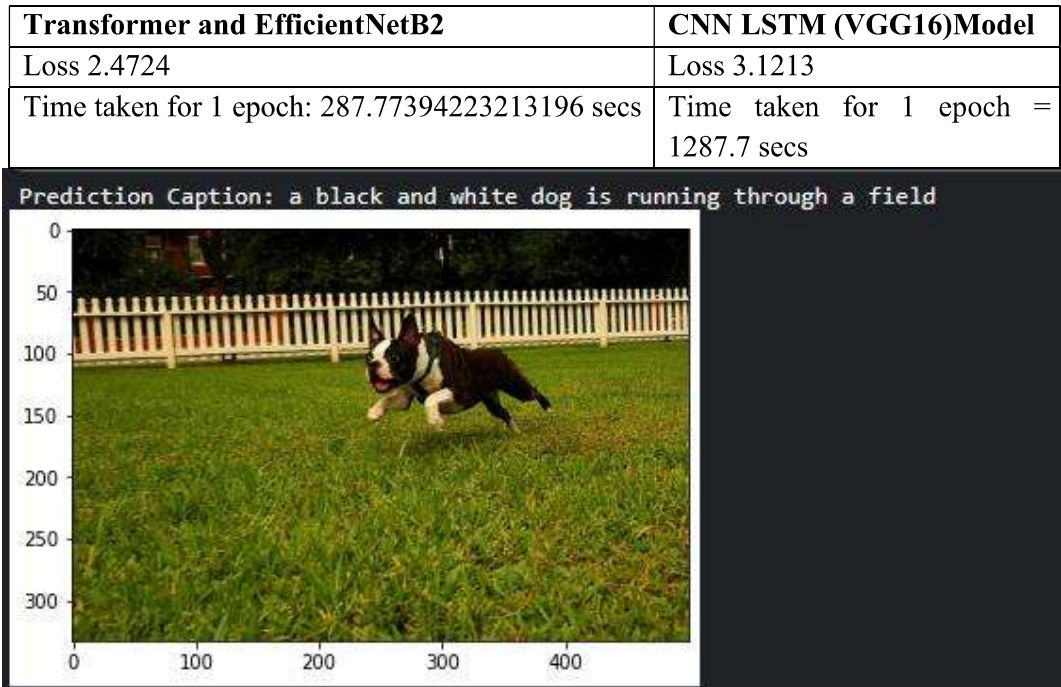


Fig 5.1: A black and white dog is running through a field.

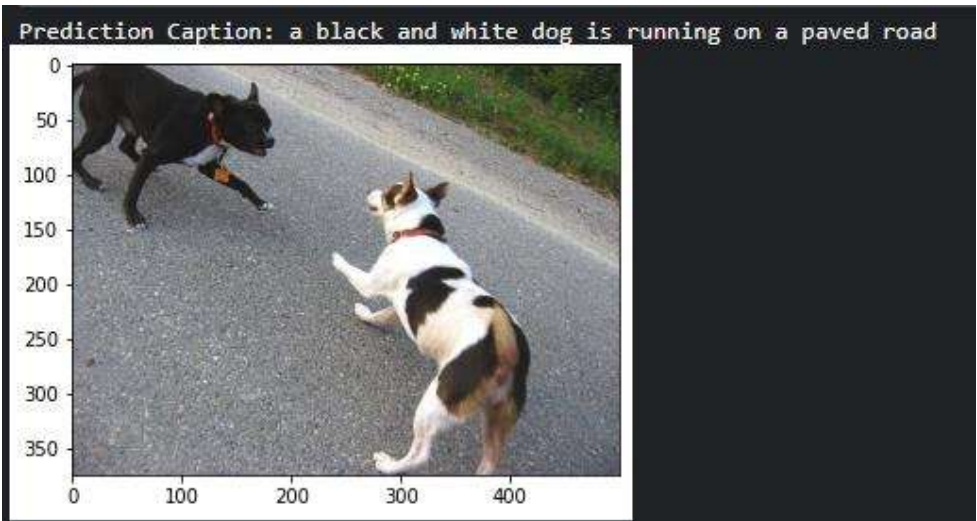


Fig 5.2 : A black and white dog is running on a paved road



Fig 5.3: A man in a yellow kayak is in a yellow kayak



Fig 5.4: A girl in a yellow shirt is playing cricket

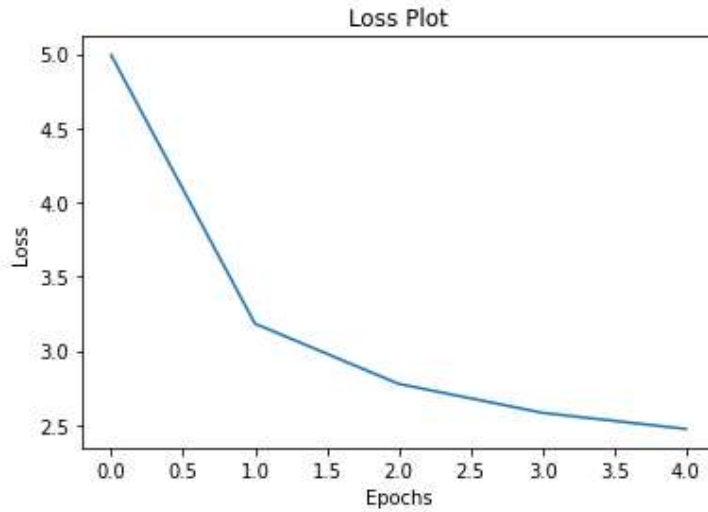


Fig 5.5: Epoch -loss characteristics

## 5.3 SYSTEM REQUIREMENTS

The implementation of the proposed methodology was conducted on the Python, specifically version 3.10. The model itself was developed using the google colab, and the libraries are:

### 5.3.1 Libraries Used

- |               |          |               |
|---------------|----------|---------------|
| a) Numpy      | c) keras | e) Matplotlib |
| b) Tensorflow | d) nltk  |               |

## **CHAPTER 6**

### **CONCLUSION AND FUTURE SCOPE**

Base on the obtained results, we can see that the deep learning methodology used here Transformer model has more successful results than traditional methods. Transformer and efficientnetb2 worked together in proper synchronization. They were able to find the relationship between the objects in the images by providing approximate labels for the images present in the volatile 8k database. For images uploaded from a computer, the label is also accurate enough.

In the future we will try to generate the caption for those images which are not present in fickle 8k database, by uploading from the computer and try to increase accuracy for image captioning.

## **REFERENCES**

- [1]. 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
- [2]. 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
- [3]. 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
- [4]. 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
- [5]. 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
- [6]. Li, Y., et al. (2016). Visual-Textual Joint Relevance Learning for Image Captioning. *IEEE Transactions on Image Processing*, 25(10), 4567–4579. doi:10.1109/TIP.2016.2583285
- [7]. Gupta, A., & Kumar, P. (2018). Attention Mechanisms in Image Captioning: A Comprehensive Review. *IEEE Access*, 6, 9197–9219. doi:10.1109/ACCESS.2017.2787676
- [8]. Tan, W., & Zhang, X. (2019). Hierarchical Transformer Networks for Image Captioning. \*Conference on Neural Information Processing Systems (NeurIPS)\*, 4021–4031. <https://arxiv.org/abs/1905.10041>
- [9]. Wu, Q., & Yang, P. (2020). Transformer-Based Image Captioning with Adaptive Contextual Attention. *IEEE Transactions on Cybernetics*, 50(7), 3105–3116. doi:10.1109/TCYB.2019.2904771
- [10]. Liu, C., et al. (2017). Deep Reinforcement Learning for Image Captioning. \*European Conference on Computer Vision (ECCV)\*, 467–482. doi:10.1007/978-3-319-46466-4\_29



PAPER NAME

HIMANSHU\_2K22\_SPD\_06\_Thesis - Copy (1).pdf

WORD COUNT

**7492 Words**

CHARACTER COUNT

**43696 Characters**

PAGE COUNT

**36 Pages**

FILE SIZE

**1.3MB**

SUBMISSION DATE

**May 30, 2024 1:40 PM GMT+5:30**

REPORT DATE

**May 30, 2024 1:40 PM GMT+5:30**

● **11% Overall Similarity**

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

- 7% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 8% Submitted Works database

● **Excluded from Similarity Report**

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





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

**PLAGIARISM VERIFICATION**

Title of the Thesis Visual Transformers for Image Understanding

Total Pages 34 Name of the Scholar Himanshu Bishal

Supervisor (s)

(1) Dr. Rajesh Palilla

(2) \_\_\_\_\_

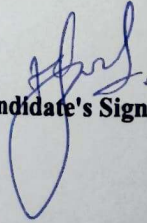
(3) \_\_\_\_\_

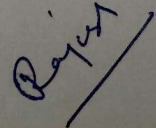
Department \_\_\_\_\_

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

Software used: Turnitin Similarity Index: 11% Total Word Count: 7492

Date: 30 May 2024

  
Candidate's Signature

  
Signature of Supervisor(s)