

# **ADVANCING VISUAL NARRATION THROUGH TRANSFORMERS**

**A Thesis Submitted  
In Partial Fulfillment of the Requirements  
for the Degree of**

**MASTER OF TECHNOLOGY  
in  
Artificial Intelligence  
by**

**SHUBHAM THAKUR  
(Roll No. 2K22/AFI/26)**

**Under the Supervision of  
Dr ARUNA BHAT  
(Associate Professor, Dept of Computer Science & Engineering)**



**To the  
Department of Computer Science and Engineering  
  
DELHI TECHNOLOGICAL UNIVERSITY  
(Formerly Delhi College of Engineering)  
Shahbad Daultpur, Main Bawana Road, Delhi-110042. India**

**May, 2024**

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

**ACKNOWLEDGEMENTS**

I have taken efforts in this survey paper. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to **Dr Aruna Bhat** for her guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing this review paper. I would like to express my gratitude towards the **Head of the Department (Computer Science and Engineering, Delhi Technological University)** for their kind cooperation and encouragement which helped me in the completion of this research survey. I would like to express my special gratitude and thanks to all the Computer Science and Engineering staff for giving me such attention and time.

My thanks and appreciation also go to my colleague in writing the survey paper and the people who have willingly helped me out with their abilities.

*Shubham Thakur*

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

**CANDIDATE'S DECLARATION**

I, **Shubham Thakur**, Roll No. 2K22/AFI/26 student of M.Tech (Artificial Intelligence), hereby certify that the work which is being presented in the thesis entitled “**Advancing Visual Narration Through Transformers**” in partial fulfillment of the requirements for the award of the Degree of Master of Technology in Artificial Intelligence in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from August 2022 to Jun 2024 under the supervision of Dr Aruna Bhat , Associate Professor, Dept of Computer Science and Engineering. The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

Place: Delhi

**Candidate's Signature**

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

**Signature of Supervisor**

**Signature of External Examiner**

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

**CERTIFICATE**

Certified that **Shubham Thakur** (Roll No. 2K22/AFI/26) has carried out the research work presented in the thesis titled “**Advancing Visual Narration Through Transformers**”, for the award of Degree of Master of Technology from Department of Computer Science and Engineering, Delhi Technological University, Delhi under my supervision. The thesis embodies result of original work and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree for the candidate or submit else from the any other University /Institution.

Date:

Dr Aruna Bhat  
(Supervisor)  
Department of CSE  
Delhi Technological University

# **Advancing Visual Narration Through Transformers**

Shubham Thakur

## **ABSTRACT**

Image captioning has experienced significant advancements over the past decade, transitioning from traditional semantic approaches to sophisticated neural network models. This thesis explores the enhancement of image captions using a novel architecture combining MobileNet and Transformers. Traditional models, such as the Encoder-Decoder framework with CNN-RNN architectures, have laid the groundwork for image description generation. However, these models often face limitations in capturing complex image contexts and generating coherent, contextually rich descriptions. The proposed model leverages MobileNet for efficient feature extraction and Transformers for superior sequence generation, addressing the limitations of earlier methods. By incorporating attention mechanisms, the model enhances the understanding of intricate image details, resulting in more accurate and descriptive captions. Experimental results demonstrate a 12.5% improvement in captioning performance compared to standard methods, showcasing the potential of this approach. This work contributes to the ongoing innovation in image captioning, paving the way for more advanced techniques and applications in the field.

## **LIST OF PUBLICATIONS**

1. Paper title: 'Textualizing Images: A Comprehensive Review of Image Captioning Using Deep Learning' is accepted in '2<sup>nd</sup> International Conferences on Optimization Techniques in Engineering and Technology (ICOTET 2024)' SCIE in May 2024.
2. Paper title: 'Enhancing Image Captioning with Transformers and EfficientNet for Fine-Grained Detail Inclusion' is accepted in '1<sup>st</sup> International Conference on Applied Artificial Intelligence and Machine learning (IC2024)' SCIE in May 2024.

## **TABLE OF CONTENTS**

<b>Title</b>	<b>Page No.</b>
<b>Acknowledgements</b>	<b>ii</b>
<b>Candidate's Declaration</b>	<b>iii</b>
<b>Certificate</b>	<b>iv</b>
<b>Abstract</b>	<b>v</b>
<b>List of Publications</b>	<b>vi</b>
<b>Table of Contents</b>	<b>vii</b>
<b>List of Tables</b>	<b>ix</b>
<b>List of Figures</b>	<b>ix</b>
<b>CHAPTER -1 INTRODUCTION</b>	<b>1-4</b>
1.1    IMPORTANCE OF VISUAL NARRATION	1
1.2    APPLICATIONS OF VISUAL NARRATION	1
1.3    CHALLENGES IN VISUAL NARRATION	2
1.4    TRANSFORMERS AND THEIR POTENTIAL	3
1.5    RESEARCH GOAL AND OBJECTIVE	3
<b>CHAPTER – 2 LITERATURE REVIEW</b>	<b>5-17</b>
2.1    DIFFERENT DATASETS	5
2.2    DIFFERENT EVALUATION METHOD	8
2.3    TRADITIONAL METHODS OF VISUAL NARRATION	12
<b>CHAPTER – 3 RELATED WORKS</b>	<b>18-21</b>
3.1    CONVOLUTIONAL NUERAL NETWORK	18
3.2    RECURRENT NUERAL NETWORK	19
3.3    LONG SHORT TERM MEMORY	19
3.3    ATTENTION	20

<b>CHAPTER – 4 DATASET</b>	<b>22-24</b>
4.1    DATA COLLECTION	22
4.2    DATA PRE-PROCESSING	24
<b>CHAPTER – 5 METHODOLOGY</b>	<b>25-32</b>
5.1    TRANSFORMER MODEL ARCHITECTURE	25
5.2    BENEFIT OF ATTENTION MECHANISM	28
5.3    LOSS FUNCTION	29
5.4    STRUCTURE OF THE MODEL	31
<b>CHAPTER – 6. EXPERIMENTAL SETUP AND RESULT ANALYSIS</b>	<b>33-37</b>
5.1    EXPERIMENTAL SETUP	33
5.2    PERFORMANCE EVALUATION	33
5.3    QUANTITATIVE ANALYSIS	34
5.4    QUALITATIVE ANALYSIS	35
5.4    SOME RESULTS	36
5.    DISSCUSSION AND FINDINGS	37
<b>CHAPTER – 7. CONCLUSION, FUTURE WORK &amp; SOCIAL IMPACT</b>	<b>38-39</b>
5.1    CONCLUSION	38
5.2    OBJECTIVE ACHIEVED	39
<b>References</b>	<b>40</b>
<b>List of Publications and their proofs</b>	<b>46</b>
<b>Plagiarism Report</b>	<b>48</b>



## List of Tables

<b>Table Number</b>	<b>Table Name</b>	<b>Page Number</b>
1	Use Case Of Image Narration	2
2	Research Quaestions	5
3	Different Evaluation Methods	8
4	Evaluation Scores	9-11
5	Different Statistical Based Methods For Image Narration	13
6	Different Types Of Loss Function Used In Image Narration	29
7	Quantitative valuation Of Proposed Model	34
8	Qualitative Evaluation Of Proposed Model	35
9	Some Generated Results	36

## List of Figures

<b>Figure Number</b>	<b>Figure Name</b>	<b>Page Number</b>
1	Pie Diagram Representation Of Share Of Papers Using A Particular Dataset	7
2	Different Types Of Methods Used For Image Narration	13
3	Encoder-Decoder Architecture	15
4	Graph Based Image Narration	16
5	Architecture Of CNN	18
6	Architecture Of RNN	19
7	Architecture OF LSTM	20
8	Architecture Of Attention Based Method	21
9	Flickr8K Dataset	23
10	Transformer Architecture	25
11	Attention Block	27
12	Process Of Calculating Attention Score	28
13	Calculating Cross Entropy Loss	31
14	Calculating Average Loss	31

## **CHAPTER 1: INTRODUCTION**

### **1.1 The Importance of Visual Narration**

One of the most important applications that is used in many different fields is visual narration. With the advancement of artificial intelligence, this discipline has made great progress. In essence, visual narration is the process of describing an image's depiction of a situation using a model. This makes it possible for the model to gather environmental data more accurately.

This approach is used in many sectors these days. For example, the education sector uses visual narrative to give thorough subject descriptions. Comparably, visual narrative is used in the entertainment industry to generate accurate descriptions for scenes, improving convenience and user experience.

Apps that narrate images can tell users about different locations in the tourism industry. Image narration models are used in the healthcare industry to provide thorough explanations of medical reports. There are many more processes that use image narration; these are only a few common use cases.

This thesis's primary objective is to determine the best process for creating picture narratives. Image narratives can be made in a variety of traditional ways. For the sake of this thesis, we have thoroughly considered each strategy and noted these models' drawbacks.

### **1.2 Application In Various Industries**

Image narration is becoming increasingly popular these days, with many industries utilizing it to enhance their work efficiency and provide a better user experience. Specifically, the e-commerce and healthcare industries are using image narration extensively to improve their operations. In the e-commerce sector, image narration is widely adopted by commerce sites to enhance user experience by providing detailed information about objects.

Industry	Use Case
Education	Image narration is used to enhance the learning environment by providing visual and textual information about various subjects.
HealthCare	Image narration can be used in healthcare industry to aid the doctors in the process of reviewing the reports
Tourism	Image narration application are one of the most popular apps now days , which provides user etc the detail information combine with text and visual .
Social Media	Social media platform utilises image narration in there content to provide user with great user experience
Ecommerce	Ecommerce company widely uses image narration in their cataloging process .
Real Estate	In this sector image caption is widely used to catalog the different properties . Which provide the user a great ease in understanding the different amenities present in the property

Table 1 : Use Cases Of Image Narration

Similarly, in the tourism sector, many mobile applications use image narration to offer tourists a new way to learn and discover things.

[Table 1] represent the applications of image captioning across different industries

### 1.3 Challenges in Visual Narration

Although there are various models for the image narration process, each capable of generating accurate captions, one problem they all share is the inability to generate detailed captions about the image. As a result, they miss a lot of important information in the image and fail to understand the semantic relationships between the objects present, causing a significant amount of useful information to be lost in the image narration generation.

## **1.4 Transformers and their Potential for Visual Narration Enhancement**

Transformers are among the technologies that have revolutionized various fields of AI. The field of image narration has also greatly benefited from the introduction of transformers. These transformers are capable of understanding and generating captions for images effectively, thereby enhancing visual narration. They can comprehend the semantic relationships between the objects in an image, generating more information-rich and detailed narrations about the scene. Popular uses for attention-based methods:

1. Interactive storytelling and educational materials can be developed using these methods.
2. Compared to conventional models, these models are more accurate.
3. These models can withstand words that have many interpretations inside a single sentence, or ambiguous terms.
4. It is simple to integrate these models with a variety of activities, particularly those that use NLP.
5. These models are flexible and may adjust to the preferences of the user.

## **1.5 Research Goals and Objectives**

This thesis primary goal is to examine the many models for the visual narration process that are now in use, pinpoint their shortcomings, and create a new model that can get over these restrictions.

Objectives:

1. Studying the many models available for picture narration is one of the goals
2. To research the assessment techniques applied to ascertain these model's accuracy.
3. To study the databases that are available for our suggested model's training.

4. To develop a new model that is capable of overcoming every drawback of the versions that are now in use.
5. To compare the suggested model's performance to that of conventional ones.

These objectives will help us first understand the area of image narration, including the different models available, the widely used databases available on the public internet, the best evaluation methods, and more. To perform this study, we will first conduct a literature survey on the topic of image narration.

## CHAPTER 2: LITERATURE REVIEW

To begin the literature survey for this topic, we will first formulate some research questions that we aim to answer through our literature review.

Research Question 1	What are the different datasets used in the papers related to image textual description generation
Research Question 2	What are the different evaluation methods that are extensively used in research papers related to image textual description generation
Research Question 3	Finding the different methods that are used in various research to generate image textual descriptions
Research Question 4	What are different challenges and difficulties in the field of the image description

Table 2 : Research Questions

All of the questions in **[Table-2]** will form the foundation of the literature survey, covering important techniques, methods, and advancements in the field, allowing for an in-depth exploration of research on image description generation.

### 2.1 Different Datasets

Datasets are one of the most important factors in determining the accuracy of a model. If the dataset is not diverse enough, the model may be biased toward certain cases and will not yield accurate results in the real world. Therefore, selecting the right dataset for model development is a critical step in this thesis. Some popular datasets that are widely used for image narration tasks are publicly available.

### PASCAL Dataset

Pattern Analysis, Statistical Modeling, and Computational Learning are the acronyms for the PASCAL dataset. It is a benchmark dataset for categorization and object identification. Twenty object types are present in the annotated photos in this dataset. Bounding boxes surrounding the interesting objects are included in the annotations.

### ABSTRACT Dataset

The photos in the ABSTRACT 50S dataset are frequently used in studies to assess image recognition and classification methods. There are 5,000 photos in all within the collection, which is composed of 50 categories of things, each with 100 photographs.

### Stanford Dataset

The Stanford Dogs dataset, which comprises of dog photos for object identification and classification, is one of the datasets in the Stanford Dataset collection.

### Visual Genome Dataset

The Visual Genome dataset offers thorough explanations of the information contained in photos. It has around 100,000 photos in total, with an average of 30 area descriptions and 5–6 questions on the image's content each image.

### Flickr Dataset

Images gathered from the photo-sharing website Flickr make up the Flickr dataset. Each of its more than 100 million photos has user-generated metadata tagged on it.

### MS-COCO Dataset

The MS COCO stands for Microsoft Common Objects in Context dataset. It is a large-scale image recognition, segmentation, and captioning dataset. It contains more than 330,000 images, each annotated with object-bounding boxes, segmentation masks, and at least five captions. It has been used as a benchmark dataset for several of the research papers studied so far in this systematic literature review. The dataset has also been used to evaluate the performance of image captioning models, leading to significant advances in the field.



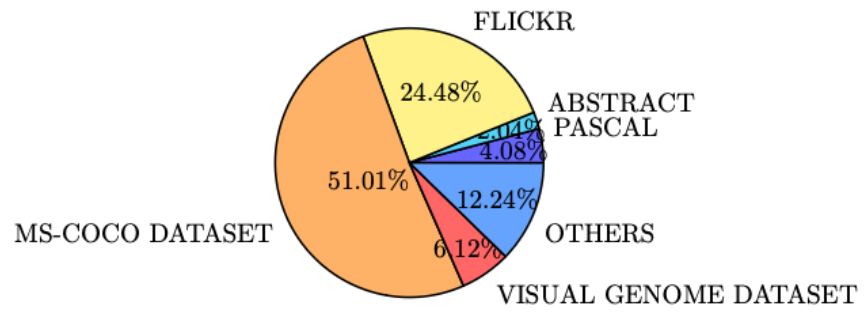


Figure 1 : Pie Diagram Representation Of Share Of Paper Uses A Particular Dataset

## 2.2 Different Evaluation Methods

Evaluation methods are used to verify the results obtained by the model whether they are correct or not.

There are multiple evaluation techniques that are widely used in image narration related models. Some of which may tackle the problem of ambiguity, some may not, so selection of evaluation methods depends on the use case of our model.

Some of the popular evaluation techniques are:

BLEU	BLEU stands for Bilingual Evaluation Understudy. It is used to evaluate the quality of the machine-translated text. It measures the overlap between the machine-generated text and a set of reference text by calculating the number of n-gram in the machine-generated text that is also present in reference text
METEOR	The Metric for Evaluation of Translation with Explicit Ordering (METEOR) is a widely used metric for evaluating the quality of machine-translated output. Unlike other metrics that rely solely on word overlap or exact match, METEOR takes into account the fluency, adequacy, and ordering of the translation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is a set of metrics used for evaluating the quality of automatic summarization and text generation, also known as gisting. It is designed to evaluate the quality of summarization outputs but can be used for other gisting tasks as well

Table 3 : Different Types Of Evaluation Method

<b>Paper</b>	<b>Dataset</b>	<b>Evaluation-Score</b>
[51]	STANFORD	CIDEr: 17.4
		BLEU 1: 41.60
		METEOR: 15.60
[28]	COCO	CIDEr: 1.042
		BLEU 1: 74.80
		METEOR: 26.60
		ROUGE: 55.00
	MS COCO	CIDEr: 104.2
		BLEU 1: 70.80
		METEOR: 24.30
	FLICKR	BLEU 1: 64.70
		METEOR: 18.90
[30]	MS COCO	CIDEr: 150.5
		BLEU 1: 97.70
		METEOR: 34.10
		ROUGE: 67.30
	COCO	CIDEr: 104.2
		BLEU1: 93.70
		METEOR: 35.40
		ROUGE: 70.50
[38]	COCO	CIDEr: 118.00
		CIDEr: 117.10
		BLEU 1: 80.10
		METEOR: 27.40
	FLICKR	ROUGE: 57.00
		CIDEr: 65.0
[31]	MS-COCO	ROUGE: 49.90
		BLEU 4: 30.10
		METEOR: 24.70
		ROUGE: 51.50

Paper	Dataset	Evaluation-Score
[42]	FLICKR	CIDEr: 97.00
		BLEU 4: 29.40
		METEOR: 23.00
	COCO	BLEU 4: 37.50
		METEOR: 28.50
		ROUGE: 58.20
		CIDEr: 125.5
[47]	MS-COCO	BLEU 4: 35.40
		METEOR: 26.40
		ROUGE: 36.40
[27]	COCO	CIDEr:116.00
		BLEU 4: 33.20
		METEOR: 25.70
		ROUGE: 55.00
[23]	MS-COCO	CIDEr: 100
		BLEU 4: 28.51
		METEOR: 25.32
		ROUGE: 56.53
[7]	PASCAL	BLEU1: 0.398
		Rouge: 0.21
[53]	FLICKR	BLEU: 0.398
[14]	FLICKR	BLEU 4: 27.70
		METEOR: 23.70
		CIDEr: 85.50
[26]	MS-COCO	METEOR: 15.95
		CIDEr: 13.52
		BLEU 4: 8.69

[32]	MS-COCO	BLEU 1: 77.20
		BLEU 4: 36.20
		METEOR: 27.00
		ROUGE: 56.40
		CIDER: 113.50
[34]	STANFORD	BLEU 1: 42.38
		BLEU 4: 9.43
		METEOR: 18.62
		CIDEr: 20.93
[37]	AMAZON DATA	BLEU 1: 46.32
		BLEU 4: 18.01
		METEOR: 17.64
		CIDEr: 17.17
[49]	MS-COCO	BLEU 1: 80.80
		BLEU 4: 38.40
		METEOR: 26.40
		ROUGE: 58.60
		CIDEr: 127.80
[52]	MS-COCO	BLEU 1: 81
		BLEU 4: 38.80
		METEOR: 28.80
		ROUGE: 58.80
		CIDEr: 129.60
[43]	MS-COCO	BLEU 1: 77.40
		BLEU 4: 37.80
		METEOR: 28.40
		ROUGE: 57.50
[36]	MS-COCO	CIDEr: 119.80

Table 4 : Evaluation Score Of Different Paper Study

Some of the frequently used evaluation metrics used in the various models are:

BLEU [1]

METEOR [3]

CIDEr [13]

ROUGE [2]

So this research study focusing on identifying the frequently used evaluation metrics for comparing image captioning methods. Among all the metrics, the analysis reveals that BLEU is the most frequently used evaluation method across the considered papers with more than 2/3 of papers use BLEU as one of the evaluation metrics. Followed by METEOR. By utilizing BLEU along with other widely recognized evaluation metrics, this study aims to establish a consistent and standardized approach for comparing various methods of image description generation. This approach ensures a reliable and objective assessment of the performance and effectiveness of different techniques in generating image descriptions

### **2.3 Traditional Methods for Visual Narration**

While transformers are a powerful new tool, it is important to explore how other methods have traditionally been used to enhance visual narration. Here's a breakdown of some common techniques.

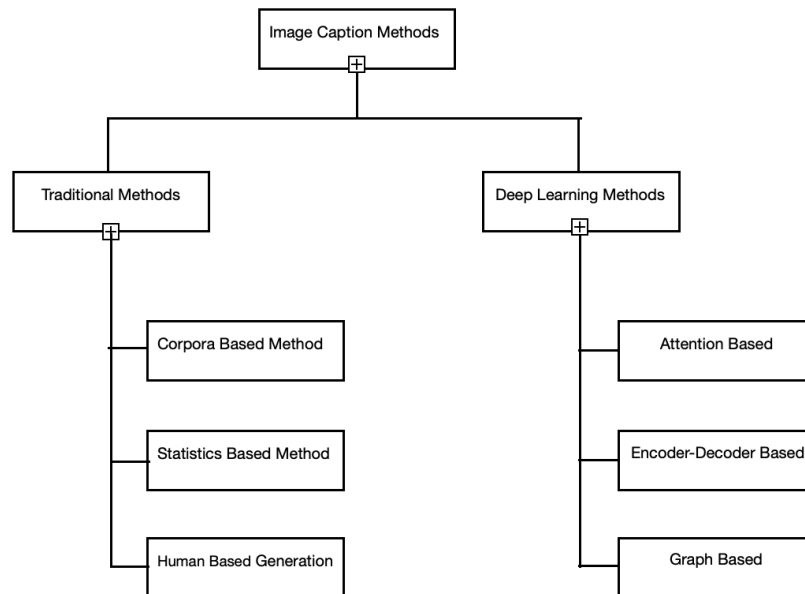


Figure 2 : Different Types Of Methods Used For Image Narration

### 2.3.1 Corpora Based Method

It is among the oldest techniques for creating visual storytelling. This approach creates a new narrative for the image by using a prepared corpus of text sentences. Basically, the model uses several object identification techniques to first identify the various objects that are present in the photos. Following object classification, it searches redefined corpora for sentences produced by combining those types.

These techniques have the advantage of producing narratives relatively quickly, but a disadvantage is that they consistently produce narratives of the same kind for a given word count.

### 2.3.2 Statistical Based Method

These models used statical method to determine the captions for the images .they use the probability distribution of the words in the document to determine the best narration possible for the give set of the objects. Once the model learn to generate detailed caption it then uses its previous learned information to generate the new narration .

Although these method are very fast , they may not able to give accurate narration . As they only look at the frequency of the words in the document and also does not take consider of any relationship between those words .

Different types of method using statistics :

<b>N-gram Models</b>	These models look at how often groups of N words appear together in the text. They predict the next word based on the ones before it to choose the best sequence of words for captions.
<b>Bag-of-Words Models</b>	These models treat text as a bunch of individual words without caring about their order. They use methods like TF-IDF to figure out which words are most important for making captions.
<b>Topic Modeling</b>	Techniques like LDA find topics in the text and pick words related to those topics for captions.
<b>Markov Models</b>	These models look at the chances of one word coming after another in the text to pick words for captions based on these chances.

Table 5 : Different Types Of Statistical Method



### 2.3.3 Deep Learning based methods

#### 1. Encoder -Decoder based Model

An encoder-decoder model for image captions consist of two parts encoder and decoder. The encoder part looks at the picture and turns it into numbers. Then, the decoder part takes those numbers and turns them into words to describe the picture. It's like translating from picture language to word language. First, the encoder takes the picture convert them into latent space . Then, the decoder takes those latent space embedding and turns them into words to describe what the picture shows.

One benefit of using this model is that it can understand pictures and describe them in words. For example, if the picture shows a beach, the model might say, "A sunny beach with people swimming and sunbathing on the sand."

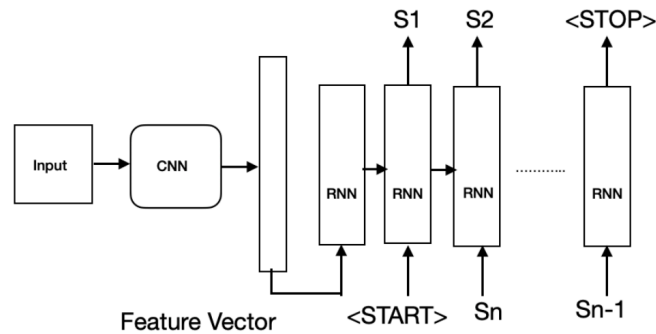


Figure 3 : Encoder - Decoder Architecture

From [Figure 3], it is clearly observed that the encoder-decoder is divided into two modules. One is the encoder module, which mainly utilizes a CNN architecture to extract important features from the image, and the other is the decoder module, which uses those extracted features to generate a caption.

## 2. Graph based Model

Graph-based methods are used for image generation tasks as they are able to detect the relationship between different objects and entities in the input image with the help of which they are able to generate more accurate and structured textual descriptions of the image data. This method is also able to keep the caption generated in the semantic order due to the availability of relationship exist between the different objects in the input image. Information passing takes place within the graph, allowing for the exchange of knowledge and context between different elements. Feature aggregation combines relevant information from different nodes in the graph to form a comprehensive representation. Using these aggregated features, a description is generated for the image. Finally, the generated caption is evaluated to assess its quality and relevance. This technique leverages the power of graph structures to improve the accuracy and contextuality of image captions.

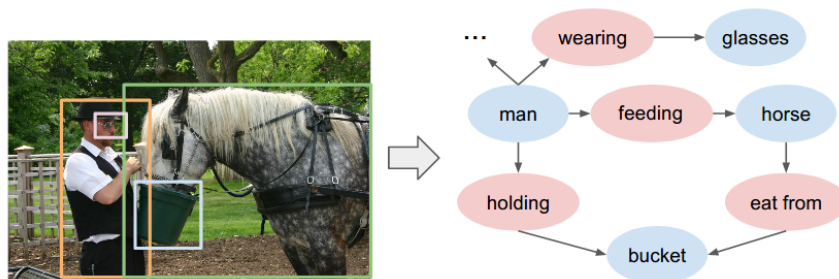


Figure 4 : Graph Based Image Narration

[Figure 4] illustrates how the graph-based model breaks down different components of an image to form a graph, with nodes depicting the objects and edges representing the relationships between these objects. Consequently, one can obtain a fully semantic relationship graph, which can be used to generate more semantically based captions.

### 2.3.4 Drawback of Previous Method

Even though the models we have so far studied are all capable of writing good captions, they can only write short ones. The brevity results in

overlooking tiny but pivotal elements that exist within the images. An instance can be seen when for example a model gives a precise description as “a dog playing in a park” but fails to provide other contexts such as “a red ball near the dog” or “a group of children playing in the background.” These missing parts are very important in analyzing an image wholly. Hence, while these captions are generally accurate and coherent, there is room for improvement with regard to capturing and describing finer details which would give a more complete picture of it.

For example an image of boy playing cricket , the traditional model may give captions like “ Boys are playing cricket” but in this they lack various fine grained details such as description of the ground , description of boy going to bowled the bowl etc.

## CHAPTER 3 : RELATED TERMS

Before going deep into our proposed model , let us first understand some basic terms which are very useful during the building of our model .

### 3.1 Convolutional Neural Network

Convolution Neural Network(CNN) is a deep neural network that is primarily used in analyzing visually imaginary. This process involves applying a series of filters to the input image so that the relevant information from the image such as the edge can be extracted. As the dimension of the input in this method is reducing, so it makes the computation fast and ideal for large input images.

The structure of a CNN layer is comprised of a stack of multiple convolution layers each outputting a filter image and followed by the activation, pooling operation. With the help of multiple those layers a CNN network can work on complex input data

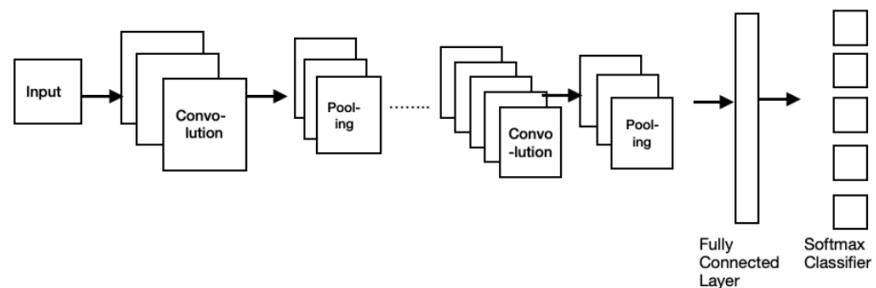


Figure 5 : Architecture Of CNN

The architecture consists of three main layers:

- Convolutional layer: This layer extracts features from the input image. These features can be edges, lines, and shapes.
- Pooling layer: This layer reduces the dimensionality of the data by summarizing the information from the convolutional layer.
- Fully connected layer: This layer classifies the input data based on the features extracted from the convolutional layers.

### 3.2 Recurrent Neural Network

It is a type of neural network that is used to process sequential information (speech, text etc), This neural network has hidden states that can store data related to previous output for a longer time, which is useful in finding the pattern in sequential data .

The Recurrent Neural Network is comprised of mainly three layers

1. Input Layer: This layers accepts the input .
2. Recurrent Layer: This layer is used for processing the input sequence and generating the sequence of the hidden state .
3. Output layer: This layer takes sequence of hidden state and produce the final output

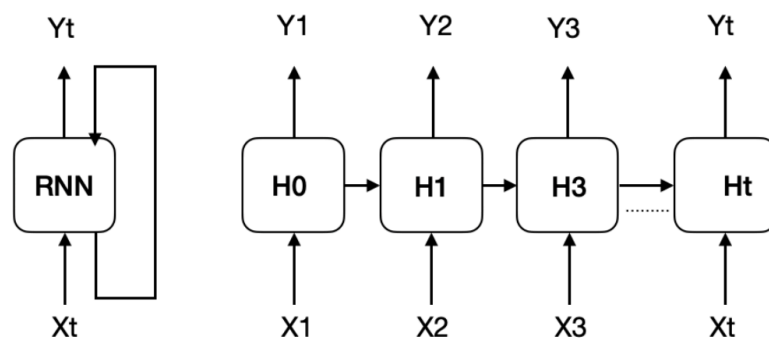


Figure 6 : Architecture Of RNN

### 3.3 Long Short-Term Memory

LSTMs are a type of neural network that can overcome the drawbacks of RNN models, such as their inability to capture long sequences and the problem of gradient descent.

The LSTM architecture consists of three gates:

**Cell State:** This is the part that has a lengthy history of information storage.

**Input Gate:** This gate selects the data to be stored in the cell state from the current input.

**Output Gate:** This gate selects which cell state data should be used to produce the output.

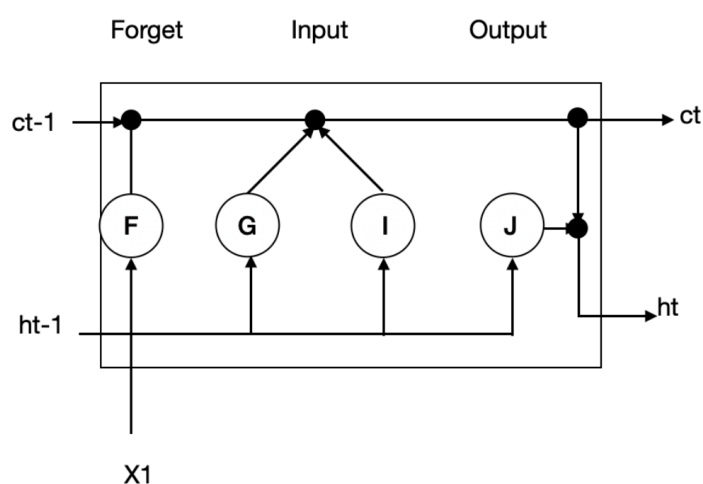


Figure 7 : Architecture Of LSTM

### 3.4 Attention

By using this technique, the system is able to ignore non-relevant information for a given word and concentrate solely on the pertinent elements associated with the caption it is creating. With the use of the attention approach, the model is better able to extract pertinent information from each word in the caption, producing descriptions that are more precise and thorough.

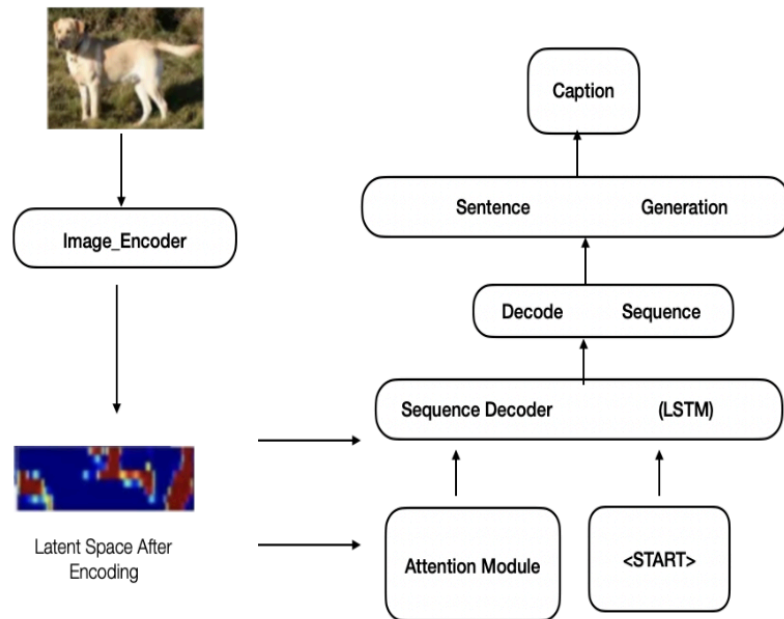


Figure 8 : Architecture Of Attention Mechanism

The design of an attention-based network is shown in [Figure 8]. Within an encoder-decoder structure, the attention mechanism functions. To extract visual information, the picture is first run through an encoder, usually a convolutional neural network (CNN). The attention mechanism uses these traits to determine attention weights for various sections of the picture. The significance of every location in producing the current word is determined by the attention weights. In order to build the caption sequentially, the decoder—typically a recurrent neural network (RNN) or Long Short-Term Memory (LSTM) network—considers the attended visual areas in addition to previously generated words.

## CHAPTER 4: DATASET

### 4.1 Data Collection and Preprocessing

We require a dataset that has both basic and comprehensive descriptions for every image in order to improve the model's accuracy. We must build a new dataset because there isn't one in the public domain that we can use to train our model.

We made the decision to build our unique dataset using the FLICKR 8K dataset. There are pictures in the FLICKR 8K dataset, and each picture has five distinct captions of varied lengths. Using two captions from the FLICKR dataset for each image—one for the basic caption and another for the detailed caption—is our novel approach.

The length of the sentences serves as the basis for determining which caption is the most extensive and which is the least detailed. Whichever caption has the fewest words is chosen as the basic caption, and whichever caption has the most words is chosen as the detailed caption.

#### 4.1.1 Visual Data

Number Of Images: Graphic Information 3000 photos from the FLICKR 8K dataset are included in the custom dataset.

Data variety: To guarantee diversity in our dataset, we have hand-picked photos from various classifications.

#### 4.1.2 Textual data

The dataset contains pairs of sentences as the textual data:

Sentence 1: The caption with the fewest details

Sentence 2: The caption with the greatest details



#### 4.1.3 Point to consider while creating dataset

Quality of dataset: The caliber of the data points in our dataset has been taken into account. We have made an effort to include only low-noise photos in our custom dataset.

Diversity of dataset: We have aimed to include data from a diverse set of classes to ensure that our model does not become underfit while training on the dataset.

#### 4.1.4 Sample Dataset



Figure 9 : FLICKR8K Dataset

## 4.2 Preprocessing: Preparing Data for Transformers

After the creation of the dataset, the next step is to preprocess it. As working directly on the unprocessed dataset can hinder the model's ability to learn effectively and reduce its accuracy. So different types of preprocessing we apply to our dataset are :

### 4.2.1 Visual Data Preprocessing

**Resizing:** We resize all the images to a standard size.

**Normalization:** We normalize the pixel values from the range of -255 to 255 to a new range of -1 to 1.

**Data Augmentation:** To increase our dataset, we apply data augmentation strategies such as flipping, cropping, and rotating.

### 4.2.2 Textual Preprocessing

**Tokenization:** We break down the sentences into small tokens that can be easily understood by the model.

**Text Cleaning:** We remove unnecessary elements such as punctuation and stop words.

**Vocabulary:** We import a prebuilt vocabulary of 50,000 English words.

### 4.2.3 Data Embedding

Before passing the data to the transformer architecture, we need to embed the data, as transformers can only understand data that is in the same embedding form.

Additionally, we also add positional embeddings since transformers have no inherent way to understand the positions of tokens. Therefore, we include positional embeddings at the start of the encoder module of the transformers.

## CHAPTER 5 : METHODOLOGY

In contrast to conventional techniques that only addressed object recognition, this method takes image-text relationships into account. The system learns to map visual elements to their corresponding textual descriptions by feeding an image and its existing caption into a transformer model. The transformer can now recognize objects beyond simple shapes thanks to this learning process. It can comprehend the overall composition of the picture and produce more engaging, natural-sounding captions that capture the scene's narrative elements. The quality and usefulness of image descriptions could be greatly enhanced by using this method.

### 5.1 Transformer Model Architecture

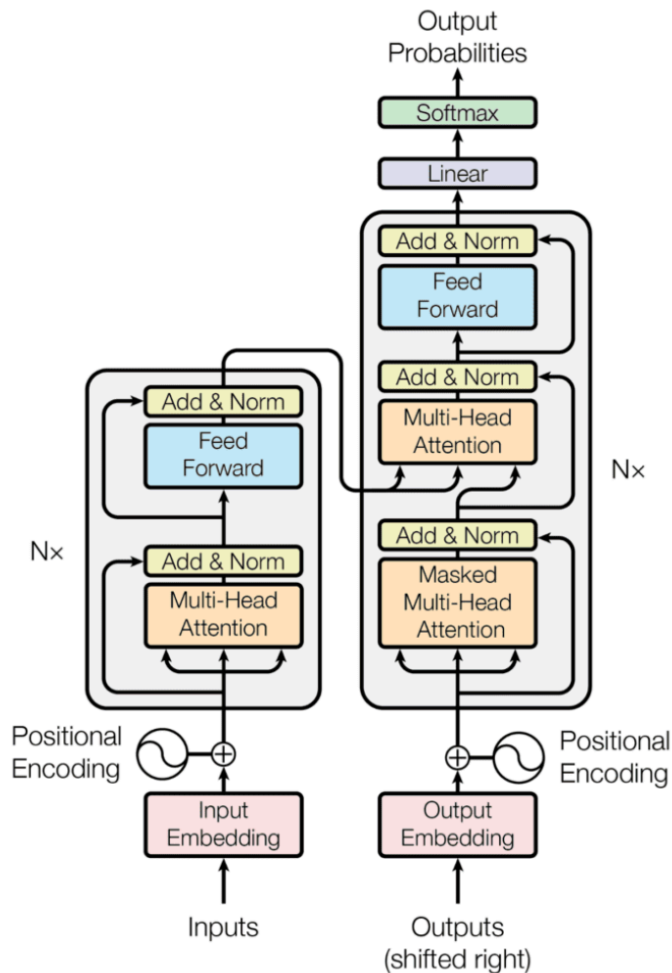


Figure 10 : Transformer Architecture

From [Figure 10] it is observed that the transformer consists of two modules: the encoder module and the decoder module. Both of these modules are used in various different large language models (LLMs). For instance, GPT uses the decoder side of the transformers, while models such as BERT use the encoder side of the transformers.

### 5.1.1 Encoder-Decoder Structure

**Encoder:** The encoder acts as the visual interpreter for the transformer. It converts the input data into a form that can be understood by the decoder side of the transformer. It is capable of outputting the visual representation while keeping in mind the different relationships between the data points, which it discovers using the attention mechanism.

The encoder normally consists of two components:

**Self-Attention:** This layer is responsible for analyzing the relationships between different data points in the sequence. For example, in the case of a sentence, it finds the relationships between the different words present in the sentence.

**FeedForward Network:** This layer is responsible for introducing non-linearity into the model.

**Decoder:** The decoder takes the input from the encoder and generates the narration word by word or token by token.

There are mainly three components of the decoder side of the transformers:

**Masked Self-Attention:** This layer is responsible for allowing the model to focus on previously generated regions of the image narration while attuning the current visual encoder part.

**FeedForward Network:** The non-linearity in the model is brought about by this layer.

### 5.1.2 Attention Mechanisms

The attention mechanism is one of the transformer's primary functions. It is the heart of the transformer, making it more successful in generating tasks compared to its predecessors. The attention mechanism allows the transformer to shift its attention to different parts of the image when needed. This capability helps the model understand the relationships between different objects in the image and generate information-rich narrations about the image.

#### Attention

The attention mechanism is based on three different vectors: Query, Key, and Value.

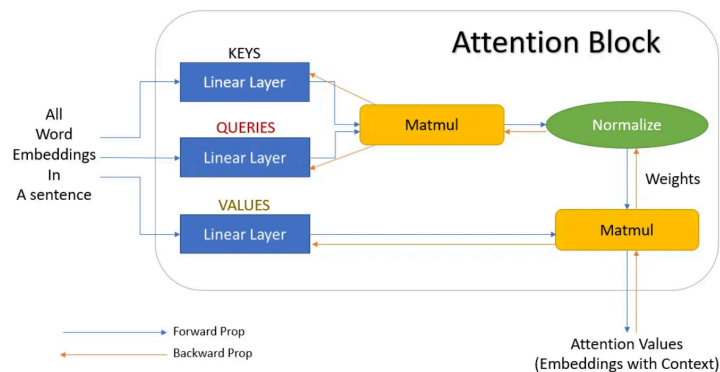


Figure 11 : Attention Block

Three Components of the Attention Mechanism:

Query Vectors: Represent the current focus of attention.

Key Vectors: Represent different parts of the input data.

Value Vectors: Contain the actual information associated with each key vector.

## Calculating Attention Scores

The model calculates a score for each key vector based on its similarity to the query vector. This score indicates how relevant each part of the input data is to the current focus. [Figure 12] is representing how the attention score is calculated.

## Weighted Sum

The model takes a weighted sum of the value vectors, using the attention scores as weights.

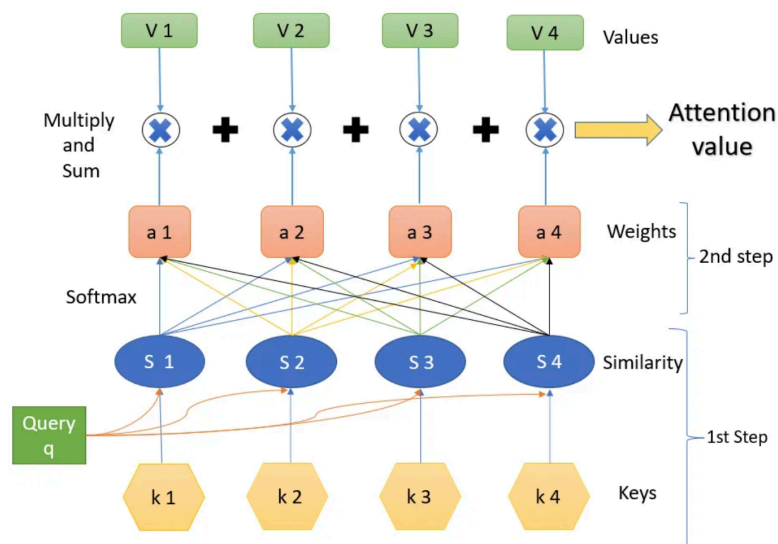


Figure 12 : Process Of Calculating Attention Score

## 5.2 Benefits of Attention Mechanisms

Improved Understanding of Visual Relationships

Context-Aware Narration Generation.

Dynamic Focus

### 5.3 Training Objectives and Loss Functions

Trying objective is to train the proposed model such that it is able to generate the image narration closely similar to a human generated narration and are accurate and coherent . Loss Function is used to quantify the difference between the model generated narration and the human generated narration and on the basis of that loss we again fine tunes our model accordingly. Our goal is to minimise the loss function to such a low range that there are no big difference between the model generated narration and between the human generated narration .

Some Of the common loss function which are widely used for transformers are :

Teacher-Forcing	During training, the decoder might be exposed to the entire ground truth narration at each step.This can be helpful for initial training but is not ideal for real-world scenarios
Cross-Entropy Loss	This measures the difference between the predicted probability distribution of the next word in the narration and the actual distribution from the ground truth. It encourages the model to generate words with higher probabilities according to the desired narrative.
BLEU Score	While not a direct loss function, BLEU score is a metric used during evaluation to assess how similar the generated narration is to the ground truth in terms of n-gram precision (n-grams being sequences of n words).

Table 6 : Different Kind Of Loss Function Used In Image Narration

Additional Considerations:

Early Stopping: To prevent overfitting, early stopping used In which the training is halted if the model's performance on a validation set fails to improve for a certain number of epochs

Regularization: Techniques like dropout can be used to prevent the model from overfitting to the training data and improve generalization to unseen visuals.

### 5.3.1 Implementing Loss Function

We use the cross-entropy loss function for this model, as it effectively captures the difference between the probability distribution of the next word for the generated and ground truth sentences.

$$H(t,p) = - \sum_{s \in S} t(s) \cdot \log(p(s))$$

#### Implementing Cross-Entropy Loss for Image Captioning

##### 1. Ground Truth Preparation:

We take the image and its corresponding token, converting the caption into a sequence of numerical tokens, each representing a word in the vocabulary.

##### 2. Model Prediction:

The input image is fed to the model, which predicts the probability distribution of the next word in the caption sequence at each position. This prediction provides the likelihood of the next possible word.

##### 3. Loss Calculation for Each Word:

The loss is calculated using the embedded form of the ground truth and the probability distribution of the next predicted word in the sequence. We compare the one-hot encoding with the probability distribution at that position.

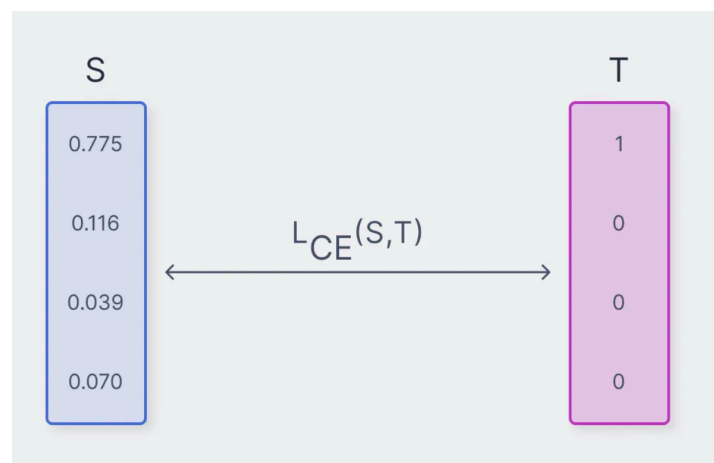


Figure 13 : Applying Cross Entropy Loss



#### 4. Overall Loss Calculation:

We calculate the average of the losses at each location in the sequence to determine the total loss.

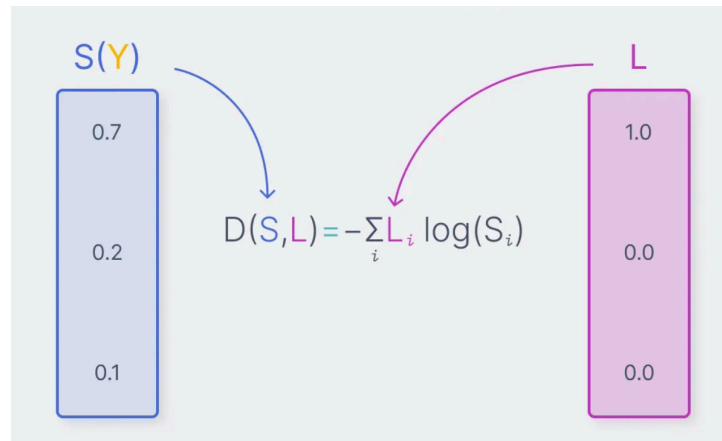


Figure 14 : Calculating Average Loss

### 5.4 Structure Of The Model

Our approach combines the Flickr8K dataset, which has many images and captions, with MobileNet, a basic image analyst. Our model's goal is to provide comprehensive and logically contextualized captions for every image by utilizing transformers, which are renowned for their ability to comprehend language.

First, we break down the images into smaller parts using MobileNet. This method helps our model understand what's in each picture without feeding it too much information. Then, we incorporate transformers to aid improve our model's comprehension of the words and sentences in the captions. Transformers are superior to traditional neural networks at understanding long sentences and deciphering their underlying meaning. Our methodology can provide captions that are both contextually meaningful and factual by integrating transformers. The Flickr8K dataset was chosen because it contains a wide variety of images with captions written by humans. This dataset is a useful tool for our model's training, which helps it create organic relationships between words and images.

Our goal is to create subtitles that go beyond just summarizing the substance of the images. Our goal is for our model to capture the meaning and feelings that the photographs portray. By combining MobileNet, transformers, and the Flickr8K dataset, we hope to make image captions more engaging and entertaining for all users.

## **CHAPTER 6: EXPERIMENTAL SETUP & RESULT ANALYSIS**

### **6.1 Experimental Setup**

In this chapter we outline the experimental steps needed to create this models and to evaluate its performance on real word dataset.

#### 6.1.1 Software Requirements

Python Programming Environment:

1. Frameworks: TensorFlow and PyTorch
2. Libraries: NumPy, Pandas, and Scikit-learn
3. Text Processing Libraries: NLTK , spaCy

#### 6.1.2 Hardware Requirements

1. GPU Acceleration: High-performance GPUs
2. Memory: Sufficient RAM
3. Storage: SSD storage

### **6.2 Performance Evaluation**

#### 6.2.1 Experimental Protocols

We follow strict rules when testing our model to make sure our results are accurate and reliable. Here's how we set up our experiments:

1. Splitting the Dataset: The Flickr8K dataset is split into three sections: test sets, training sets, and validation sets. This aids in the training, adjustment, and performance assessment of our model.
2. Cross-Validation: To ensure the accuracy of our results, we employ a procedure known as cross-validation. This entails evaluating the model's performance across various scenarios using various dataset segments.

3. Adjusting Settings: In order to train the model, we conduct experiments to determine the ideal hyperparameters. to assist the model in maximizing performance and determining the ideal configuration for a given job.

### 6.3 Quantitative Analysis

In this section , we evaluate the result generated for our proposed model on various standard evaluation metrics. [Table 8] represent the BLEU , ROUGE evaluation score for our proposed model and the baseline model (Encoder-Decoder).

Image		BLEU	ROUGE
	Base Model	62.01	0.48
	Our Model	80.02	0.73
	Base Model	58.05	0.53
	Our Model	80.29	0.78
	Base Model	50.36	0.40
	Our Model	76.45	0.69

Table 7 : Quantitive Evaluation Of Proposed Model

From [Table 7], we can clearly observe that our proposed model achieves greater accuracy in generating human-like, detailed sentences compared to the previous baseline model.

In our testing phase, for each batch of 32 images, we were able to achieve an average BLEU score of 80.34. This represents a significant jump from the traditional method, which yielded a BLEU score of 71.29 on the same batch of images. Consequently, we achieved a significant increase of 9.05 points, or a 12.69% improvement, over the previous baseline models.

## 6.4 Qualitative Analysis

In this section, we'll compare the descriptions generated by our new model to those created by the baseline model. This will help us understand how our model performs compared to the existing approach.[**Table 8**]




Image	Output Generated	
	Base Model	People sitting on bench
	Ground Truth	Three people sit on wooden Bench set on white orange tile.
	Our Model	Three people are sitting on benches made of wood. These benches are placed on a floor with white and orange tiles.
	Base Model	Children playing in water
	Ground Truth	Children playing in the water fountain
	Our Model	Young children, filled with excitement and laughter, splash and play amidst the refreshing jets of water from the fountain
	Base Model	Men standing near bicycle
	Ground Truth	A group of man standing and drinking water near the bike
	Our Model	A group of men gather around a bike, sharing water and laughter, enjoying each other's company

Table 8 : Qualitative Analysis Of Proposed Model

[**Table 8**] clearly demonstrates that our proposed model generates sentences that are more similar to the actual descriptions (ground truth) compared to the baseline model's outputs. In some instances, our model even surpasses the ground truth by providing even more detailed descriptions.

## 6.5 Some Results

Image	Enhanced Visual Narration
	<p>A cat is lying down on the couch. Next to it, there's a remote control. The cat seems relaxed and comfortable, enjoying its time on the couch</p>
	<p>A person in a bright red outfit confidently rides a motorcycle on a rough dirt track. The bumpy terrain makes it exciting as they navigate sharp turns, showing off their skill and control on the bike</p>
	<p>A dog with black and white fur is happily swimming in a pool. Its paws paddle through the water as it moves gracefully, enjoying the coolness of the pool on a sunny day.</p>

Table 9 : Some Results

[Table 9] represents some of the generated result from our proposed model.

## 6.6 Discussion & Findings

### 6.6.1 Strengths of the Proposed Methodology

1. Detailed Caption Generation: Our approach is effective in producing captions that accurately explain photos.
2. Diversity Robust :Our proposed model performs well with a variety of image types and datasets. It scores highly in tests, representing its robustness to the diverse datasets.
3. Works Across Various Devices: Our proposed model implement using Mobilenet to optimize efficiency across various devices.

### 6.6.2 Limitations and Challenges

1. Complex Meanings: Occasionally, our proposed model has problem of generating complex captions where there are a lot of complex relationship between the objects are exists
2. Data Might Not Be Perfect: There may be a chance that data we are using to train our model may not contain objects from all the classes , hence our model may be in some rare cases can generate wrong narration.

### 6.6.3 Areas for Improvement

1. Making Captions More Consistent: We can improve in this area , as thee are still lot of drawback in the existing model..
2. Better Integration of Pictures and Words: We can try to find out ways to integrate picture and words more efficiently .

### 6.6.4 Recommendations for Future Research

1. Understanding Images in More Detail: Creating a multimodal approach for generating the image narration
2. Integrating LLMs for Better Efficiency : LLMs can be used to make the model more efficient in generating captions and extracting important features form the image

## **CHAPTER 7: CONCLUSION, FUTURE WORK & SOCIAL IMPACT**

### **7.1 Summary of Key Findings and Contributions**

The purpose of this work was to improve computers' ability to interpret and understand images. We tried a new strategy by combining two powerful technologies, MobileNet and transformers. Transformers are known for their exceptional abilities in language processing and understanding. They make it easier for computers to produce text that seems human. Nevertheless, MobileNet is a type of neural network that is particularly good at detecting and interpreting images. Our goal was to find out if integrating these two methods will enable computers to describe images more effectively.

We did several experiments with the help of this transformer and MobileNet. We found that this approach significantly improved computers' ability to accurately and comprehensively characterize visual content. For the field of artificial intelligence (AI), this finding has important ramifications. Particularly useful for jobs like picture comprehension—where AI must process visual data—and human-like discourse production—where AI must generate coherent, natural-sounding language—are artificial intelligence's specialties.

Numerous exciting prospects arise as a result of this evolution. It might lead to the development of more user-friendly and intuitive technology, for example. Imagine artificial intelligence (AI) systems that can describe images to the blind or visually impaired or that can provide comprehensive captions for images in real time. Such technology could improve living standards for a great number of people and be very beneficial.

More work needs to be done despite these hopeful outcomes. We are simply beginning our research. Computers need to be made much more capable of understanding complex concepts and minute contextual factors. A computer should be able to recognize objects in an image, for instance, and also understand how those objects relate to one another and work together.

It is also necessary to address important issues including ethical considerations and data biases. Data bias can occur when the training set used by AI systems is not representative of all possible results. AI systems that perform well in some circumstances but poorly in others could arise from this, which could have unjust or unanticipated consequences. Ethical



considerations are also critical to ensuring that AI is used responsibly and does not harm individuals or society.

Our research concludes by showing the revolutionary potential of combining transformers and MobileNet to fundamentally alter how computers see and comprehend images. This tactic may increase the AI systems' intelligence, usability, and practicality for humans. However, much work needs to be done to strengthen these systems and address major challenges in order to ensure their dependability and fairness.

## **7.2 Objective Achieved**

1. We came up with a creative way to create comprehensive and logical image descriptions by utilizing transformers and MobileNet.
2. Our approach proved successful with a variety of picture formats, demonstrating its capacity to process a broad spectrum of visual data.
3. We made sure that our approach runs quickly and works with many devices, which makes it useful for a wide range of practical uses.

## CHAPTER 8 REFERENCE

- [1] Kishore Papineni et al. “BLEU: a Method for Automatic Evaluation of Machine Translation”. In: (Oct. 2002).doi: 10.3115/1073083.1073135.
- [2] Chin-Yew Lin. “ROUGE: A Package for Automatic Evaluation of summaries”.In: Proceedings of the ACL Workshop: Text Summarization Braches Out 2004 (Jan. 2004), p. 10.
- [3] Alon Lavie and Abhaya Agarwal. “METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments”. In: (July 2007), pp. 228–231.
- [4] B. Z. Yao et al. “I2T: Image Parsing to Text Description”. In: Proceedings of the IEEE 98.8 (Aug. 2010), pp. 1485–1508.
- [5] E. Hoque, O. Hoerber, and M. Gong. “CIDER: Concept-based image diversification, exploration, and retrieval”. In: Information Processing Management 49.5 (Sept. 2013), pp. 1122–1138.
- [6] G. Kulkarni et al. “BabyTalk: Understanding and Generating Simple Image Descriptions”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (Dec. 2013), pp. 2891–2903.
- [7] G. Kulkarni et al. “BabyTalk: Understanding and Generating Simple Image Descriptions”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 35.12 (Dec. 2013), pp. 2891–290.
- [8]Y. Verma et al. “Generating Image Descriptions Using Semantic Similarities in the Output Space”. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (June 2013).
- [9] J. Donahue et al. “Long-term Recurrent Convolutional Networks for Visual Recognition and Description”. In: (Nov. 2014).
- [10] X. Chen and C. L. Zitnick. “Mind’s eye: A recurrent visual representation for image caption generation”. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015).
- [11] X. Jia et al. “Guiding the Long-Short Term Memory Model for Image Cap-tion Generation”. In: 2015 IEEE International Conference on Computer Vision (ICCV) (Dec. 2015).
- [12] A. Karpathy and L. Fei-Fei. “Deep visual-semantic alignments for generating image descriptions”. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015).

- [13] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. “CIDEr: Consensus-based Image Description Evaluation”. In: (2015). arXiv: 1411.5726 [cs.CV]
- [14] O. Vinyals et al. “Show and tell: A neural image caption generator”. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2015).
- [15] Oriol Vinyals et al. “Show and Tell: A Neural Image Caption Generator”. In: (2015). arXiv: 1411.4555 [cs.CV]
- [16] Peter Anderson et al. “SPICE: Semantic Propositional Image Caption Evaluation”. In: (2016). arXiv: 1607.08822 [cs.CV]
- [17] D. J. Kim et al. “Sentence learning on deep convolutional networks for image Caption Generation”. In: 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI) (Aug. 2016).
- [18] K. Tran et al. “Rich Image Captioning in the Wild”. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (June 2016).
- [19] Kelvin Xu et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention”. In: (2016). arXiv: 1502.03044 [cs.LG]
- [20] Q. You et al. “Image Captioning with Semantic Attention”. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016).
- [21] Quanzeng You et al. “Image Captioning with Semantic Attention”. In: (2016). arXiv: 1603.03925 [cs.CV]
- [22] Quanzeng You et al. “Image Captioning with Semantic Attention”. In: (2016). arXiv: 1603.03925 [cs.CV]
- [23] J. Gu et al. “An Empirical Study of Language CNN for Image Captioning”. In: 2017 IEEE International Conference on Computer Vision (ICCV) (Oct. 2017).
- [24] Jiuxiang Gu et al. “An Empirical Study of Language CNN for Image Captioning”. In: (2017). arXiv: 1612.07086 [cs.CV]
- [25] A. Karpathy and L. Fei-Fei. “Deep Visual-Semantic Alignments for Generating Image Descriptions”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence 39.4 (Apr. 2017), pp. 664–676.

- [26] J. Krause et al. “A Hierarchical Approach for Generating Descriptive Image Paragraphs”. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(July 2017).
- [27] S. Liu et al. “Improved Image Captioning via Policy Gradient optimization of SPIDeR”. In: 2017 IEEE International Conference on Computer Vision (ICCV) (Oct. 2017).
- [28] J. Lu et al. “Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning”. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(July 2017).
- [29] Jiasen Lu et al. “Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning”. In: (2017). arXiv: 1612.01887[cs.CV]
- [30] T. Yao et al. “Boosting Image Captioning with Attributes”. In: 2017 IEEE International Conference on Computer Vision (ICCV)(Oct. 2017).
- [31] L. Zhou et al. “Watch What You Just Said”. In: Proceedings of the on Thematic Workshops of ACM Multimedia 2017(Oct. 2017).
- [32] P. Anderson et al. “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (June 2018).
- [33] Peter Anderson et al. “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”. In: (2018). arXiv: 1707.07998[cs.CV]
- [34] M Chatterjee and A. G. Schwing. “Diverse and coherent paragraph generation from images.” In: Computer Vision – ECCV(2018), pp. 747–763.
- [35] Moitrey Chatterjee and Alexander G. Schwing. “Diverse and Coherent Paragraph Generation from Images”. In: (2018). arXiv: 1809.00681[cs.CV]
- [36] F. Fang, H. Wang, and P. Tang. “Image Captioning with Word Level Attention”. In: 2018 25th IEEE International Conference on Image Processing (ICIP)(Oct. 2018).
- [37] S. Liu et al. “Image Captioning Based on Deep Neural Networks”. In: MATEC Web of Conferences 232 (2018). Ed. by Y. Wang, p. 01052.
- [38] X Liu et al. “Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data.” In: Computer Vision – ECCV(2018).

- [39] J. Song et al. “From Pixels to Objects: Cubic Visual Attention for Visual Question Answering”. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (July 2018).
- [40] P. Cao et al. “Image Captioning with Bidirectional Semantic Attention-Based Guiding of Long Short-Term Memory”. In: Neural Processing Letters 50.1 (Jan. 2019), pp. 103–119.
- [41] P. Dognin et al. “Adversarial Semantic Alignment for Improved Image Captions”. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019).
- [42] L. Gao et al. “Deliberate Attention Networks for Image Captioning”. In: Proceedings of the AAAI Conference on Artificial Intelligence 33.01 (July 2019), pp. 8320–8327.
- [43] L. Huang et al. “Attention on attention for image captioning.” In: IEEE/CVF International Conference on Computer Vision (ICCV) (2019).
- [44] Lun Huang et al. “Attention on Attention for Image Captioning”. In: (2019). arXiv: 1908.06954 [cs.CV]
- [45] J. Song et al. “From Deterministic to Generative: Multimodal Stochastic RNNs for Video Captioning”. In: IEEE Transactions on Neural Networks and Learning Systems 30.10 (Oct. 2019), pp. 3047–3058.
- [46] J. Wang et al. “Convolutional Auto-encoding of Sentence Topics for Image Paragraph Generation”. In: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (Aug. 2019).
- [47] “Image Captioning Based On Sentence-Level And Word-Level Attention”. In: 2019 International Joint Conference on Neural Networks (IJCNN) (July 2019).
- [48] Shiyang Yan et al. “Image Captioning Based on a Hierarchical Attention Mechanism and Policy Gradient Optimization”. In: (2019). arXiv: 1811.05253 [cs.CV]
- [49] X. Yang et al. “Auto-Encoding Scene Graphs for Image Captioning”. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019).
- [50] X. Yang et al. “Auto-Encoding Scene Graphs for Image Captioning”. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019).

- [51] R. Li et al. “Dual-CNN: A Convolutional language decoder for paragraph image captioning”. In *Neurocomputing* 396 (July 2020), pp. 92–101.
- [52] X. Yang, H. Zhang, and J. Cai. “Auto-encoding and Distilling Scene Graphs formImage Captioning”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), pp. 1–1.
- [53] V.Julakanti.“Image Caption In: International Journal for Research in Applied Science and Engineering Technology9 (June 2021), pp. 2968–2974.
- [54] K. Nguyen et al. “In Defense of Scene Graphs for Image Captioning”. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV)(Oct. 2021).
- [55] G. Oliveira dos Santos, E. L. Colombini, and S. Avila. “CIDER-R: RobustConsensus-based Image Description Evaluation”. In: *Proceedings of the SeventhWorkshop on Noisy User-generated Text (W-NUT 2021)* (2021).
- [56] T. Vo.“FuzzSemNIC: A Deep Fuzzy Neural Network Semantic-enhanced Approach of Neural Image Captioning”. In: (July 2021)
- [57] T.Vo.“FuzzSemNIC: A Deep Fuzzy Neural Network Semantic-enhanced Approach of Neural Image Captioning”. In: (July 2021).
- [58] Z. Zohourianshahzadi and J. K. Kalita. “Neural attention for image captioning:review of outstanding methods”. In: *Artificial Intelligence Review* 55.5 (Nov 2021), pp. 3833–3862.
- [59] J. H. Tan et al. “ACORT: A compact object relation transformer for parameter efficient image captioning”. In: *Neurocomputing* 482 (Apr. 2022), pp. 60–72.
- [60] K. Yamazaki et al. “VLCAP: Vision-Language with Contrastive Learning for Coherent Video Paragraph Captioning”. In: *2022 IEEE International Conference on Image Processing (ICIP)* (Oct. 2022).
- [61] Yuanen Zhou et al. “Compact Bidirectional Transformer for Image Captioning”. In: (2022). arXiv: 2201.01984 [cs.CV]

## **LIST OF PUBLICATIONS**

1. Paper title: ‘Enhancing Image Captioning with Transformers and EfficientNet for Fine-Grained Detail Inclusion’ is accepted in ‘1<sup>st</sup> International Conference on Applied Artificial Intelligence and Machine learning (IC2024)’ SCIE in May 2024.

2. Paper title: ‘Textualizing Images: A Comprehensive Review of Image Captioning Using Deep Learning’ is accepted in ‘2<sup>nd</sup> International Conferences on Optimization Techniques in Engineering and Technology (ICOTET 2024)’ SCIE in May 2024.

# Proof Of Publications

## Paper 1 : Textualizing Images: A Comprehensive Review of Image Captioning Using

---

### Acceptance : ICOTET 2024

1 message

---

ICOTET2024 <icotetdgi2024@gmail.com>  
To: Shubham Thakur <shubhamt311@gmail.com>  
Cc: icotet@gnindia.dronacharya.info

Fri, May 10, 2024 at 2:45 PM

Greetings from ICOTET 2024!

**Dear Author (s)**

We are pleased to inform you that **Paper ID 2512** entitled “ Textualizing Images: A Comprehensive Review of Image Captioning Using Deep Learning” submitted by you has been accepted by the 2nd International Conference on Optimization Techniques in Engineering and Technology Engineering (ICOTET 2024).

You are advised to register for the conference by 16<sup>th</sup> of May, 2024! Payment details for registration can be found at the bottom of this email.

You are requested to fill out the following Google form for the registration and payment information etc.:

<https://forms.gle/mqyRFhx45hqkJ6cx7>

All the registered and presented papers for the 2nd ICOTET 2024 will be published in the AIP Conference Proceedings (Scopus Index) and Springer Nature Conference Proceedings (Scopus Index). Please note that the plagiarism level of the paper should not exceed 15%.

For further details, please visit the official website: <https://www.icotet.in/registration>

Thanks & Regards

Organizing Committee

**ICOTET 2024.**



## Paper 2 : Enhancing Image Captioning with Transformers and EfficientNet for Fine-Grained Detail Inclusion



Shubham Thakur <shubhamt311@gmail.com>

---

### Acceptance Mail -- ICAAIML

1 message

iccae VIGNT <iccae@vignanits.ac.in>  
To: Shubham Thakur <shubhamt311@gmail.com>

Wed, May 29, 2024 at 12:05 PM

Dear Shubham Thakur

It is our pleasure to inform you that your papers entitled **Enhancing Image Captioning with Transformers and EfficientNet for Fine-Grained Detail Inclusion** (Paper Id: ICAAIML-29) has been provisionally accepted for Virtual oral paper presentation at ICAAIML-2024 on 30th and 31st August 2024, and also your paper has been accepted to publish in **AIP conference proceeding (SCOPUS)**

We request you to complete the early bird conference registration fee and publication charges i.e Rs 3000+ publication charges Rs 8500- 11,500 **If you don't want AIP publication then just pay Rs 3000 only**. After payment send the payment proof along with full manuscript.

PAPER NAME

**Thesis final copy 8.pdf**

WORD COUNT

**7432 Words**

CHARACTER COUNT

**40330 Characters**

PAGE COUNT

**48 Pages**

FILE SIZE

**6.1MB**

SUBMISSION DATE

**May 31, 2024 10:23 AM GMT+5:30**

REPORT DATE

**May 31, 2024 10:23 AM GMT+5:30****● 4% Overall Similarity**

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

- 2% Internet database
- 1% Publications database
- Crossref database
- Crossref Posted Content database
- 3% Submitted Works database

**● Excluded from Similarity Report**

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less than 10 words)

← Sent Successfully

[Share](#) [Help](#)

Amount

₹10,000 

Rupees Ten Thousand Only

To

**Dronacharya Group Of Institutions** 

DI


Canara Bank - 0239 

[Pay Again](#)

From

**Anshu Thakur .**

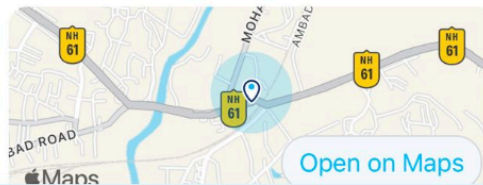
A.

State Bank Of India - 1967 

Paid at 05:41 PM, 16 May 2024

UPI Ref No: 450323451483 [Copy](#)

**View Payment Location**



Powered by 

## Paper 2 : Enhancing Image Captioning with Transformers and EfficientNet for Fine-Grained Detail Inclusion



Shubham Thakur <shubhamt311@gmail.com>

### Acceptance Mail -- ICAAIML

1 message

iccae VONT <iccae@vignanits.ac.in>  
To: Shubham Thakur <shubhamt311@gmail.com>

Wed, May 29, 2024 at 12:05 PM

Dear Shubham Thakur

It is our pleasure to inform you that your papers entitled **Enhancing Image Captioning with Transformers and EfficientNet for Fine-Grained Detail Inclusion** (Paper Id: ICAAIML-29) has been provisionally accepted for Virtual oral paper presentation at ICAAIML-2024 on 30th and 31st August 2024, and also your paper has been accepted to publish in **AIP conference proceeding (SCOPUS)**

We request you to complete the early bird conference registration fee and publication charges i.e Rs 3000+ publication charges Rs 8500= 11,500 **If you don't want AIP publication then just pay Rs 3000 only**. After payment send the payment proof along with full manuscript.

