

True Image Description Using CNN and LSTM

SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS OF THE DEGREE OF
BACHELOR OF ENGINEERING IN
COMPUTER ENGINEERING

BY

RONAK DHINGRA
SOHAIL GIDWANI
VANSHIKA GURBANI
DIVYA PANJWANI

GUIDE: PROF. JUHI JANJUA

DR. TANUJA SARODE
(PROFESSOR,
DEPARTMENT OF COMPUTER ENGINEERING, TSEC)



COMPUTER ENGINEERING DEPARTMENT
THADOMAL SHAHANI ENGINEERING COLLEGE
UNIVERSITY OF MUMBAI

2022-2023

True Image Description Using CNN and LSTM

Submitted in partial of the fulfilment degree
of the requirements of

BACHELOR OF ENGINEERING

In

COMPUTER ENGINEERING

By

Group No: 53

1902037	Ronak Dhingra
1902044	Sohail Gidwani
1902049	Vanshika Gurbani
1902118	Divya Panjwani

Guide: Prof. Juhi Janjua

DR. TANUJA SARODE

(Professor, Department of Computer Engineering, TSEC)



Computer Engineering Department
Thadomal Shahani Engineering College
University of Mumbai
2022-2023

CERTIFICATE

This is to certify that the project entitled “**True Image Description using CNN and LSTM**”
is a bonafide work of:

1902037

Ronak Dhingra

1902044

Sohail Gidwani

1902049

Vanshika Gurbani

1902118

Divya Panjwani

Submitted to the University of Mumbai in partial fulfilment of the requirement for the award
of the degree of “**Bachelor of Engineering**” in “**Computer Engineering**”.

Prof. Juhi Janjua
Guide

Dr. Tanuja Sarode
Head of Department

Dr. G. T. Thampi
Principal

Project Report Approval for B.E

Project report entitled ***True Image Description using CNN and LSTM*** by

1902037

Ronak Dhingra

1902044

Sohail Gidwani

1902049

Vanshika Gurbani

1902118

Divya Panjwani

is approved for the degree of “**BACHELOR OF ENGINEERING**” in
“**COMPUTER ENGINEERING**”.

Examiners

1. _____

2. _____

Date: 18th April, 2023

Place: Mumbai

Declaration

I declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. we also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

- 1) _____
Ronak Dhingra - 1902036
- 2) _____
Sohail Gidwani – 1902044
- 3) _____
Vanshika Gurbani - 1902049
- 4) _____
Divya Panjwani - 1902118

Date: 18th April, 2023

Abstract

Image captioning is oriented towards describing an image with the best possible use of words that can provide a semantic, relatable meaning of the scenario inscribed. Different models can be used to accomplish this arduous task depending on the context and requirement of what needs to be achieved. An encoder–decoder model which uses the image feature vectors as an input to the encoder is often marked as one of the appropriate models to accomplish the captioning process. In the proposed work, a dual-modal transformer has been used which captures the intra- and inter-model interactions in a simultaneous manner within an attention block. The transformer architecture is quantitatively evaluated on a publicly available Microsoft Common Objects in Context (MS COCO) dataset yielding a Bilingual Evaluation Understudy (BLEU)-4 Score of 85.01. The efficacy of the model is evaluated on Flickr 8k, Flickr 30k datasets and MS COCO datasets and results for the same is compared and analyzed with the state-of-the-art methods. The results shows that the proposed model outperformed when compared with conventional models, such as the encoder–decoder model and attention model.

In recent years, with the rapid development of artificial intelligence, image caption has gradually attracted the attention of many researchers in the field of artificial intelligence and has become an interesting and arduous task. Image caption, automatically generating natural language descriptions according to the content observed in an image, is an important part of scene understanding, which combines the knowledge of computer vision and natural language processing. The application of image caption is extensive and significant, for example, the realization of human-computer interaction[1].

Table of Content

List of Figures		iii
List of Tables		iv
Chapter 1	Introduction	1
1.1	Introduction	1
1.2	Problem Statement and Objectives	2
1.3	Scope	2
Chapter 2	Review of Literature	4
2.1	Domain Explanation	4
2.2	Review of Existing Systems	10
2.3	Limitations of Existing Systems/ Research Gaps	10
Chapter 3	Proposed System	12
3.1	Analysis/Framework	12
3.2	Design Details (GUI Design, DFD, Flowchart, Deployment Diagram, etc.)	13
3.3	Methodology	15
Chapter 4	Implementation Details	16
4.1	Experimental Setup	16
4.1.1	Dataset Description/Database Details	16
4.2	Software and Hardware Setup (Description of Libraries Used)	17
Chapter 5	Results and Discussion	19
5.1	Performance Evaluation Parameters	19
5.2	Implementation Results	22
5.3	Results Discussion	23
Chapter 6	Conclusion and Future Work	24

Appendix

References

25

List of Publication

List of Figures

Figure No.	Description	Page No.
Figure 2.1	Deep Learning	6
Figure 2.2	CNN Architecture	7
Figure 2.3	Architecture of Inception V3	8
Figure 2.4	Architecture of Transformer	9
Figure 3.1	Predicted Results	13
Figure 3.2	Another Predicted result on website	14
Figure 3.3	Flowchart of True Image Description	14
Figure 4.1	COCO 2017 Dataset Sample	17
Figure 5.1	Results Using Transformer	20
Figure 5.2	Results Using LSTM	21
Figure 5.3	LSTM and Transformer Comparision	24

List of Tables		
Table No.	Description	Page No.
Table 5.1	Implementation Results	7

Chapter 1

Introduction

1.1 Introduction

Image Caption Generation has branched out into several applications ranging from assistive technology to agriculture and manufacturing sectors. The existing solutions use this deep learning application to aid visually impaired people by helping them understand their environment, in robotic and industrial applications by automating processes and limiting human intervention. We present a detailed analysis of different uses of image captioning, technologies implemented, their limitations and the scope to extend utilization in other domains.

In the past few years, computer vision in the image processing area has made significant progress, like image classification and object detection. Benefiting from the advances of image classification and object detection, it becomes possible to automatically generate one or more sentences to understand the visual content of an image, which is the problem known as Image Captioning. Generating complete and natural image descriptions automatically has large potential effects, such as titles attached to news images, descriptions associated with medical images, text-based image retrieval, information accessed for blind users, human- robot interaction. These applications in image captioning have important theoretical and practical research value. Image captioning is a more complicated but meaningful task in the age of artificial intelligence. Given a new image, an image captioning algorithm should output a description about this image at a semantic level. In this an Image caption generator, basis on our provided or uploaded image file. It will generate the caption from a trained model which is trained using algorithms and on a large dataset.

1.2 Problem Statement and Objectives

The principal aim of this project is to design and develop an image caption generator. The image caption generator must be capable of identifying features in a picture and combining those features to form a statement that makes sense. A secondary aim of the project takes into account the difficult nature of the task. The project has been worked on for a number of years, and it has proved a considerable task to fully develop a fully functioning image caption generator. Thus, in the interest of future projects, the work carried out during this project will be guided towards producing a system that can readily be further developed by subsequent project groups. To make this a reality the project will be documented as closely and as accurately as possible, thus providing subsequent groups with all the knowledge required continuing the development of this system.

1.3 Scope

The overall workflow can be divided into these main steps:

1. **Read Captions File:** Reading the text and token flickr8k file, finding the length of the file and splitting it.
2. **Data Cleaning:** Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct.
3. **Loading Training Testing Data:** The process includes training Images File, testing it and creating a train description dictionary that adds starting and ending sequence
4. **Data Preprocessing – Images :** Loading the image, preprocessing and encoding it and testing it.
5. **Data Preprocessing – Captions_:** Loading the captions, appending the start and the end sequence, finding the maximum length of the caption.
6. **Data Preparation using Generator:** Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making corrections to data and the combining of data sets to enrich data.
7. **Word Embedding:** Converting words into Vectors (Embedding Layer Output)
8. **Model Architecture:** Making an image feature extractor model, partial caption sequence model and merging the two networks.
9. **Train Our Model:** A training model is a dataset that is used to train an ML algorithm. It consists of the sample output data and the corresponding sets of input data that have an influence on the output.

10. **Predictions:** Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome here predicting Caption for a photo.

Chapter 2

Review of Literature

2.1 Domain Explanation

Artificial intelligence (AI) has been a rapidly growing field over the past several years, with a wide range of applications across industries. Within this field, Machine Learning (ML) and Deep Learning (DL) have emerged as key subfields, offering powerful tools for processing and analyzing data. ML algorithms are designed to learn from data, allowing them to make predictions or decisions without being explicitly programmed. DL goes beyond traditional ML algorithms by processing large amounts of data with complex neural networks. Both ML and DL are highly dependent on data, and as the availability of big data has grown, so too has the potential for these algorithms to make an impact in a variety of fields. As such, the study of ML and DL is of great significance in the field of AI, with the potential to make significant contributions to society in areas ranging from healthcare to finance to manufacturing.

2.1.1 Machine Learning

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects [2].

Types of machine learning:

Machine learning models fall into three primary categories.

Supervised machine learning:

Supervised learning, also known as supervised machine learning, is defined by its use of labeled datasets to train algorithms to classify data or predict outcomes accurately. As input data is fed into the model, the model adjusts its weights until it has been fitted appropriately. This occurs as part of the cross validation process to ensure that the model avoids overfitting or underfitting. Supervised learning helps organizations solve a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox. Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, and support vector machine (SVM).

Unsupervised machine learning:

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. This method's ability to discover similarities and differences in information make it ideal for exploratory data analysis, cross-selling strategies, customer segmentation, and image and pattern recognition. It's also used to reduce the number of features in a model through the process of dimensionality reduction. Principal component analysis (PCA) and singular value decomposition (SVD) are two common approaches for this. Other algorithms used in unsupervised learning include neural networks, k-means clustering, and probabilistic clustering methods.

Semi-supervised learning:

Semi-supervised learning offers a happy medium between supervised and unsupervised learning. During training, it uses a smaller labeled data set to guide classification and feature extraction from a larger, unlabeled data set. Semi-supervised learning can solve the problem of not having enough labeled data for a supervised learning algorithm. It also helps if it's too costly to label enough data [3].

2.1.2 Deep Learning

Deep Learning (DL) is an AI technique that trains PCs to do what easily falls into place for people: Advance as a visual demonstration. Deep learning is a vital innovation behind driverless vehicles, empowering them to perceive a stop sign or to recognize a passerby from a light post.

It is the way to voice control in buyer gadgets like telephones, tablets, televisions, and sans hand speakers. Deep learning is standing out recently and for good explanation. It's accomplishing results that were impractical previously.

In deep learning, a PC model figures out how to perform characterization undertakings straightforwardly from pictures, text, or sound. Deep learning models can accomplish cutting edge precision, at times surpassing human-level execution. Models are prepared by utilizing a huge arrangement of marked information and brain network designs that contain many layers.

Figure 2.1 shows the working of deep learning model and how it takes image as an input and classifies it.

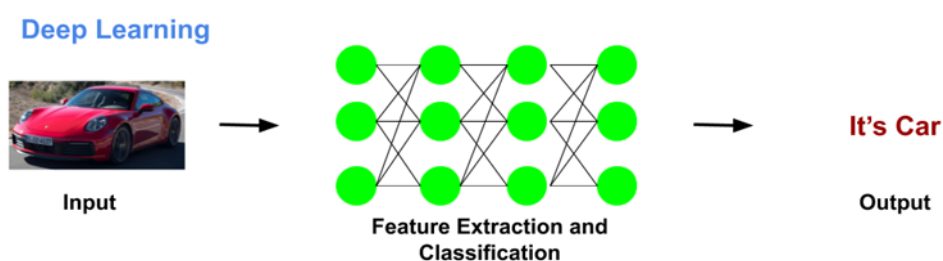


Fig 2.1 Object Detection Example

2.1.3 Convolutional Neural Networks(CNN)

A convolutional neural network, or CNN, is a deep learning neural network designed for processing structured arrays of data such as images. Convolutional neural networks are widely used in computer vision and have become the state of the art for many visual applications such as image classification, and have also found success in natural language processing for text classification. A convolutional neural network is a feed-forward neural network, often with up to 20 or 30 layers. The power of a convolutional neural network comes from a special kind of layer called the convolutional layer.

Convolutional neural networks contain many convolutional layers stacked on top of each other, each one capable of recognizing more sophisticated shapes. With three or four convolutional layers it is possible to recognize handwritten digits and with 25 layers it is possible to distinguish human faces.[4] The usage of convolutional layers in a convolutional neural network mirrors the structure of the human visual cortex, where a series of layers process an incoming image and identify progressively more complex features.

The layers of the neural network are mentioned below:

Convolutional Layer

Pooling Layer

Rectified Long Measure Layer

Fully Connected Layer

Loss Layer

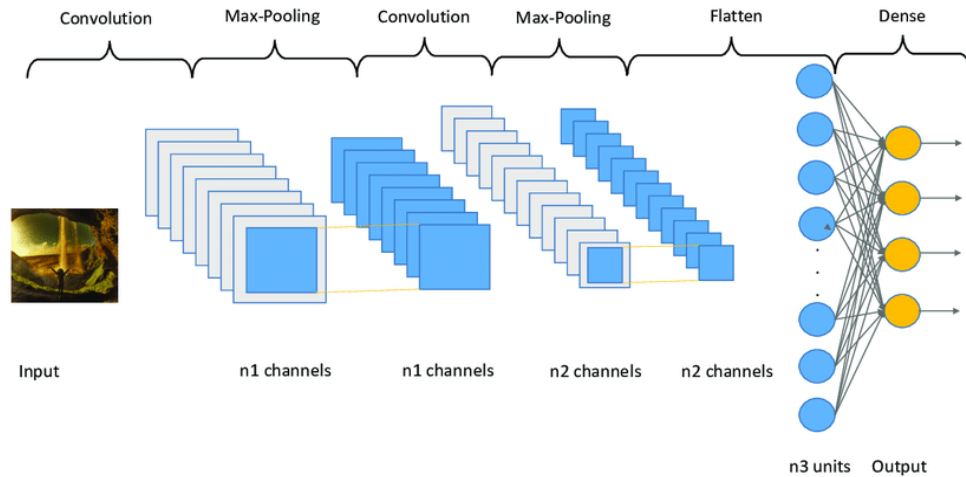


Fig 2.2 CNN Architecture

2.1.4 Inception V3

Inception v3 is a deep neural network architecture designed for image classification tasks. It was introduced by Google researchers in 2015 as an improvement over the previous Inception v1 and v2 models. Inception v3 is characterized by its use of a highly optimized convolutional neural network (CNN) with multiple Inception modules that enable it to capture features at various spatial scales. It also incorporates batch normalization, a technique that normalizes the inputs to each layer to improve the speed and stability of training. Another key feature of Inception v3 is its use of factorization into smaller convolutions, which helps to reduce the computational complexity of the model while maintaining its accuracy. Inception v3 has achieved state-of-the-art results on several image classification benchmarks, demonstrating its effectiveness and utility in real-world applications.

The Architecture of Inception v3 is shown in figure 2.3.

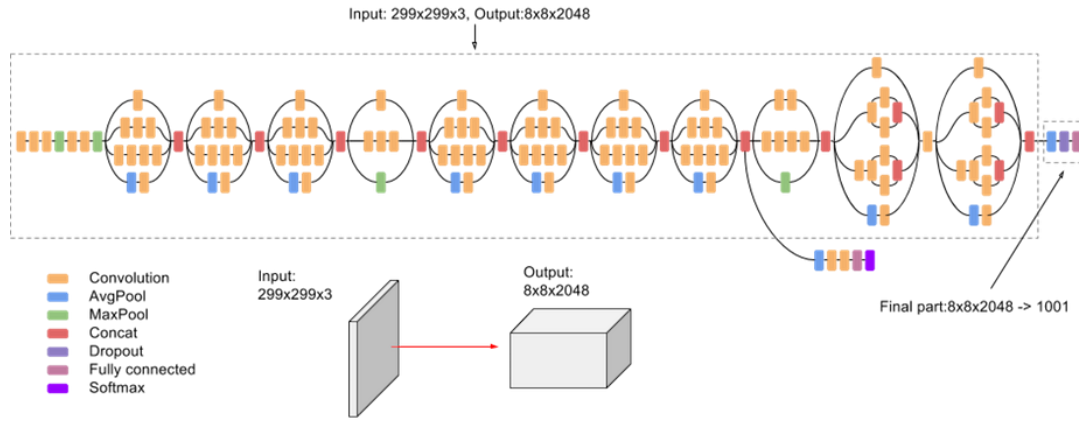


Fig. 2.3 Architecture of Inception v3

2.1.5 Transformer

Transformers, originally developed for natural language processing, have recently been applied to image captioning tasks with impressive results. The transformer-based image captioning models are able to capture complex image features and generate accurate and detailed captions for the input images. In a transformer-based image captioning model, the image features are extracted using a convolutional neural network (CNN) and are then fed into the transformer encoder. The transformer encoder processes the image features and produces a sequence of hidden states, which are then fed into the transformer decoder along with a start token. The decoder then generates a sequence of words one by one, with each word being conditioned on the previous words and the image features. The probability distribution of the next word is computed using the attention mechanism, which calculates the relevance of each word in the previous sequence to the current word being generated.

The formula for the attention mechanism can be expressed as shown in equation i:

$$\text{Attention}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_k})V \dots(i)$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimension of the key matrix. The attention mechanism computes a weighted sum of the values based on the similarity between the query and key matrices. This allows the model to focus on relevant image features and generate captions that are more accurate and descriptive. Overall, transformer-based image captioning models have shown great promise in generating captions that are both accurate and informative. Figure 2.4 shows us the flow of a transformer.

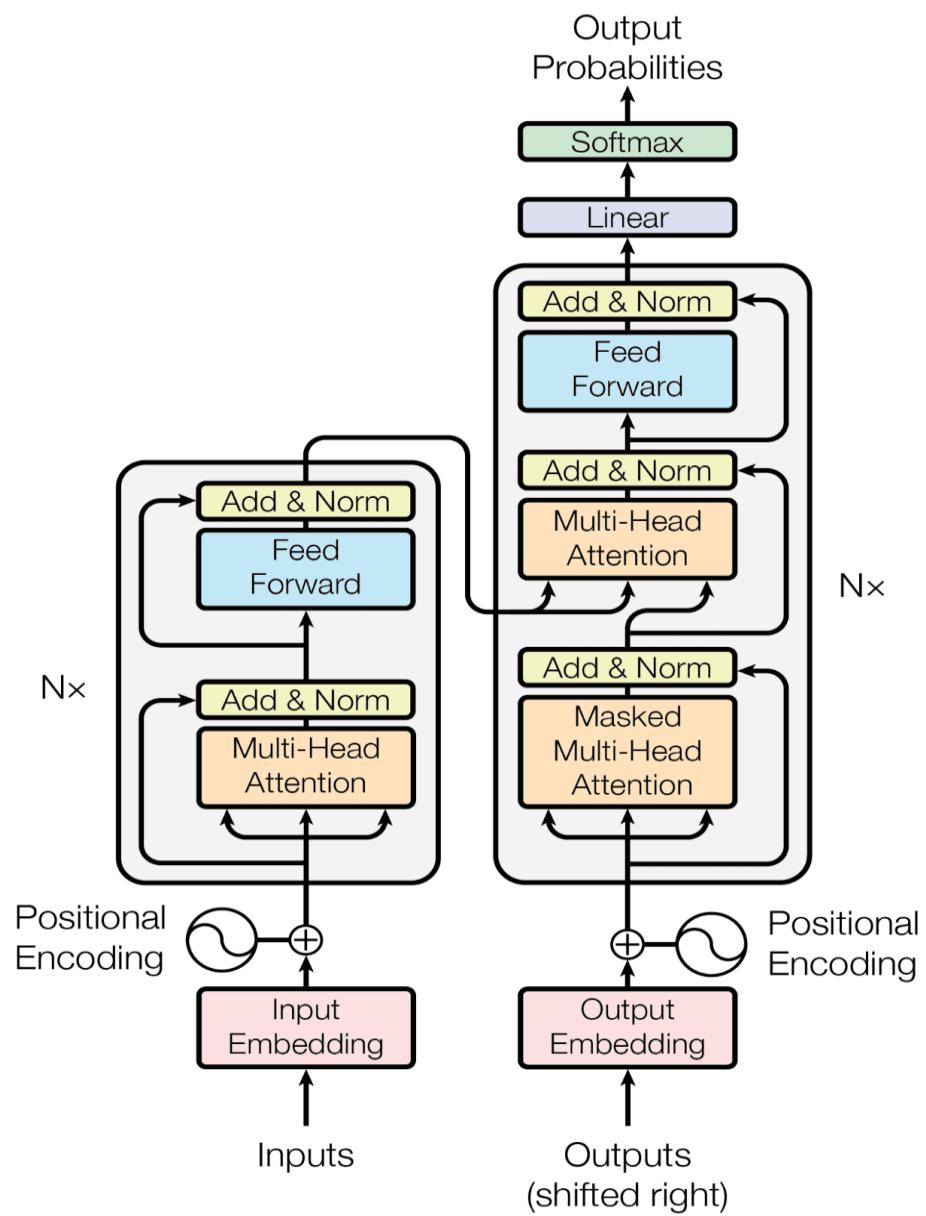


Fig. 2.4 Architecture of Transformer

The Transformer architecture follows an encoder-decoder structure but does not rely on recurrence and convolutions in order to generate an output. In a nutshell, the task of the encoder, on the left half of the Transformer architecture, is to map an input sequence to a sequence of continuous representations, which is then fed into a decoder. The decoder, on the right half of the architecture, receives the output of the encoder together with the decoder output at the previous time step to generate an output sequence[5].

2.2 Review of Existing System

There have been many existing systems for image caption generation, with the majority being based on deep learning techniques. One of the early systems was proposed by Vinyals et al. (2015) called "Show and Tell" that used a CNN-RNN architecture to generate captions. The system achieved impressive results on the COCO dataset, producing captions that were accurate and contextually relevant to the input image.

Another notable system is "NeuralTalk" proposed by Karpathy et al. (2015), which used a CNN-LSTM architecture to generate image captions. The system was able to generate captions that were not only accurate but also demonstrated creativity and humor.

More recently, some systems have utilized attention mechanisms to focus on specific regions of the input image. One such system is "Show, Attend and Tell" proposed by Xu et al. (2015), which used a soft attention mechanism to weight the importance of different regions of the image. The attention mechanism improved the overall quality of the generated captions, making them more informative and descriptive.

Overall, these systems and many others have made significant progress in the field of image caption generation, achieving impressive results on various datasets. However, there is still room for improvement in terms of generating more diverse and semantically meaningful captions that accurately capture the underlying visual content of the input image.

2.3 Limitation of Existing System/Research Gap

While the existing systems for image caption generation have achieved impressive results, there are still some limitations and research gaps that need to be addressed. One of the limitations is the lack of diversity in the generated captions. Many of the existing systems tend to produce captions that are generic and lack specificity to the input image. This can be partly attributed to the training datasets used, which may not cover a wide range of visual concepts and semantic relationships.

Another limitation is the lack of contextual understanding in the generated captions. While the systems are able to describe the visual content of the image, they may not fully capture the contextual information that is essential for generating meaningful and informative captions. There is also a research gap in developing systems that can generate captions for a wide range of visual concepts, including abstract and complex concepts.

Many of the existing systems have been evaluated on datasets that mostly contain concrete and easily recognizable objects. Finally, there is a need for developing systems that can generate captions that are more aligned with human preferences in terms of language style, creativity, and humor. While some systems have demonstrated creativity and humor in their generated captions, there is still room for improvement in terms of generating captions that are more human-like and engaging.

Chapter 3

Proposed System

3.1 Analysis/Framework

This chapter includes a detailed analysis of the project which is further broken down into functional and non-functional requirements. It also consists of the proposed system along with a diagrammatic representation of the system.

3.1.1 Functional Requirements

1. **jpg as input:** The image file should be only in the format of .jpg.
2. **Size of input:** The size of image should not be more than 2MB..

3.1.2 Non-Functional Requirements

1. **User Friendly:** User Interface and experience is simple and each function is evident to the user.
2. **Correctness:** Each form is validated so that invalid details are not added to the database. Also, each function is tested to verify that it is working well.
3. **Maintainability:** Each function is written in its own file to obtain modularity. This makes it easier for developers to maintain the codebase.
4. **Portability:** System can run and it can be deployed on any OS or cloud service.

3.2 Design Details

In this section, the GUI and flowchart can be combined to provide a comprehensive overview of the image captioning project. The GUI provides an interface for users to interact with the project, while the flowchart offers a visual representation of the image captioning process. Together, these elements can help users and readers understand the technical details of the project in a more accessible way.

3.2.1 Graphical User Interface

A graphical user interface (GUI) is a digital interface in which a user interacts with graphical components such as icons, buttons, and menus. In a GUI, the visuals displayed in the user interface convey information relevant to the user, as well as actions that they can take [6].

Figure 3.1 is an out view of the Graphic User Interface of the project. The GUI is quite simple in which the user will have to upload an image. The image will be displayed along with a list of predicted captions for the image.

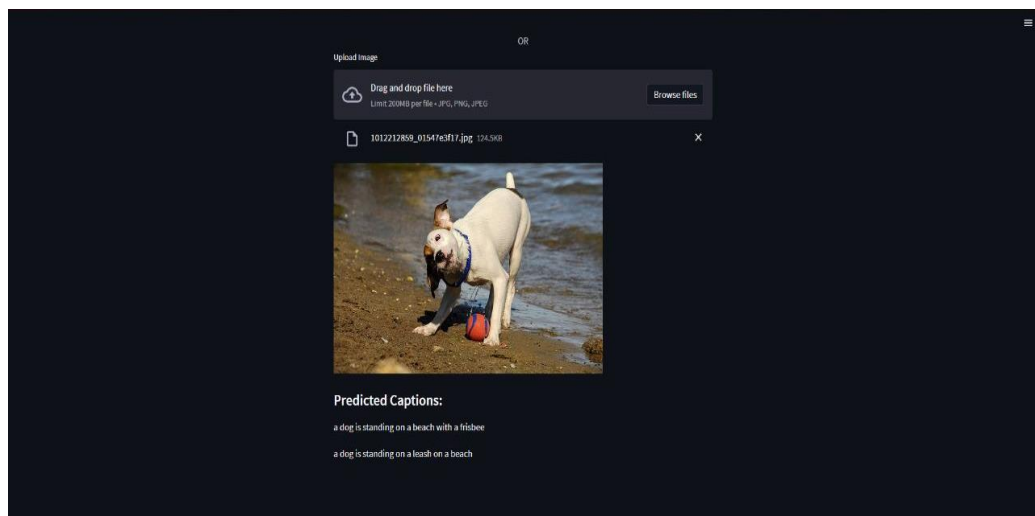


Fig 3.1 Predicted Result on Website

Figure 3.1 and 3.2 show the predicted results from the website when we upload the images. We have single as well as multiple results for the images in our dataset that is why we can see two predicted results in the above figures.

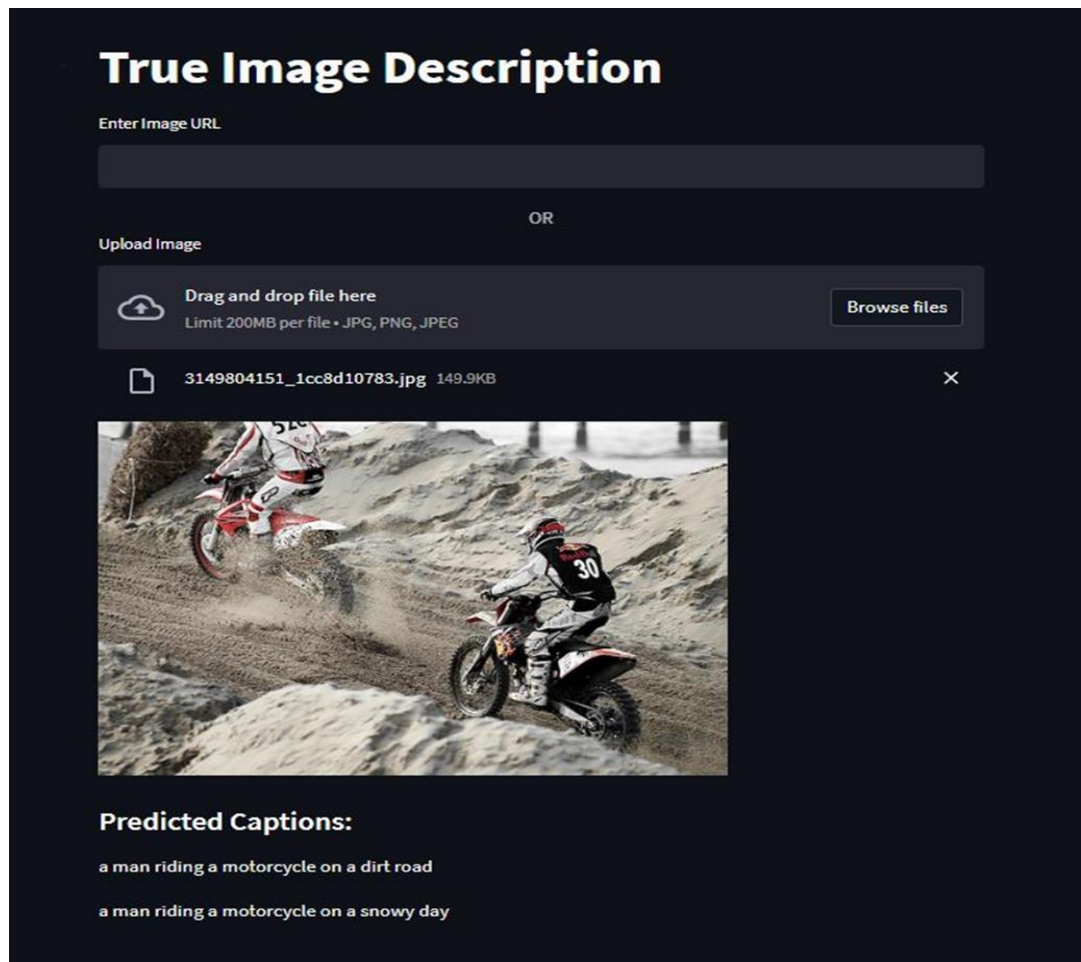


Fig 3.2 Another Predicted result on website aster R-CNN Architecture

3.2.2 Flowchart

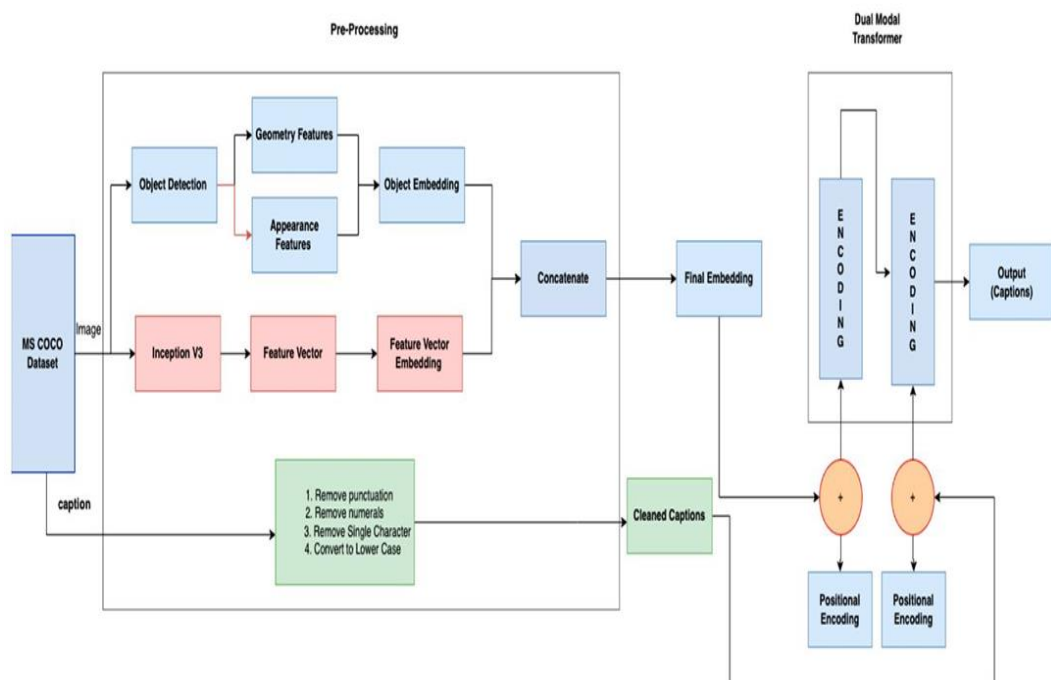


Fig. 3.3 Flowchart of True Image Description

The input image is first fed through the Inceptionv3 convolutional neural network, which extracts features from the image. These features are then passed through a feature extraction layer, resulting in a vector representation of the image. The vector representation of the image is then passed through the visual embedding layer of the dual-mode transformer. This layer maps the image features to a visual embedding space, where they are represented as a sequence of visual tokens.

The input caption is tokenized into a sequence of words, and each word is represented as an embedding vector using an embedding layer. The embedding vectors for the words in the caption and the visual tokens from the image are then passed through the dual-mode transformer. This transformer has two separate attention mechanisms - one for the visual tokens and one for the textual tokens - that allow it to jointly model the interaction between the image and the caption.

The transformer outputs a sequence of tokens, which are then converted into a sequence of words using a decoding layer. This sequence of words forms the generated caption for the input image. The generated caption is evaluated using a metric such as BLEU score to measure its similarity to the ground truth caption for the image.

The model is trained using a combination of backpropagation and gradient descent to optimize its parameters to minimize the loss between the generated captions and the ground truth captions.

3.3 Methodology

There are many simple libraries for development, the most popular being Pytorch and Tensorflow along with datasets like Flickr and MS COCO that simulate growth in different deep learning tasks and their applications. During the planning process, a dual-mode transformer is used to generate the correct image caption for the MS COCO dataset. The embeddings used are obtained by combining embeddings from object search, the Inception V3 model, and the clean text from the MS COCO dataset. CNNs and geometry features are used to create the embeds. This is done to obtain an enhanced feature vector based on two modalities viz. image-based embedding and text-based embedding. The use of CNNs has proven suitable to extract the features of a multimodal transformer used in the image captioning, so Inception V3 is used in the proposed approach[7].

Chapter 4

Implementation Details

4.1 Experimental Setup

An experimental setup is a critical component of any research project, as it allows for the systematic testing and evaluation of hypotheses. In the context of a report, the experimental setup refers to the specific procedures and conditions used to carry out the research.

4.1.1 Dataset Details

The dataset used for image captioning is the MS COCO dataset, which contains 330k images and five captions corresponding to each image along with 1.5 object instances. This dataset has been considered one of the most promising datasets in the domain of image captioning as it contains non-iconic images, which makes it different and stands apart from other datasets. The term non-iconic here signifies multiple objects overlapping in the image whereas iconic images are those which constitute a single object. This advantage of the MS COCO dataset becomes very useful while incorporating the labelling task for the images. Figure 4.1 shows the image examples present in the dataset.



Fig 4.1 Dataset Sample

4.2 Software and Hardware Setup

1. Hardware Requirements:

Windows from XP to all further versions, MAC, Linux operating systems. In windows system requirements are usually any core processor may work but specifically, Intel(R) Core (TM) with i5-4300U CPU @ 1.90GHz 2.50 GHz processor, and operating systems can be 32 bit or 64 bit but 64 bit would be more efficient and faster.

2. Software Requirements

- OS-windows 98 or Linux: An operating system (OS) is system software that manages computer hardware, software resources, and provides common services for computer programs.
- Python 3 - Python is an interpreted, high-level, general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object- oriented approach aim to help programmers write clear, logical code for small and large-scale project [8].
- NumPy - NumPy is a Python package which stands for 'Numerical Python'. It is the core library for scientific computing, which contains a powerful n-dimensional array object, provide tools for integrating C, C++ etc. It is also useful in linear algebra, random number capability etc [9].
- Pandas - A library for data manipulation and analysis, provides a powerful and flexible toolset for working with data.

- Matplotlib - A library for creating static, animated, and interactive visualizations in Python.
- TensorFlow - TensorFlow is an end-to-end open source platform for machine learning. TensorFlow is developed by Google and has integrated the most common units in deep learning frameworks. It supports many up-to-date networks such as CNN with different settings. TensorFlow is designed for remarkable flexibility, portability and high efficiency of quipped hardware.[10]
- matplotlib.pyplot - It is a sublibrary of the matplotlib data visualization library for Python. It provides a simple and convenient interface for creating a wide range of static visualizations such as line plots, scatter plots, bar plots, histograms, and more.
- Re – It is a built-in module in Python that provides support for regular expressions (regex). Regular expressions are a powerful tool for pattern matching and text manipulation.

Chapter 5

Results and Discussion

5.1 Performance Evaluation Parameters

BLEU (bilingual evaluation understudy) is a metric used to evaluate the quality of machine-generated translations against a reference translation. It is a widely used metric in natural language processing (NLP) and machine translation (MT) research.

BLEU-1 and BLEU-4 are two variants of the BLEU metric, which differ in the number of n-grams used for evaluation. An n-gram is a contiguous sequence of n items (words or characters) in a text. BLEU-1 and BLEU-4 are based on unigrams and 4-grams, respectively.

BLEU-1 measures the precision of unigrams (single words) in the machine-generated translation compared to the reference translation. It counts the number of unigrams in the machine-generated translation that appear in the reference translation and divides it by the total number of unigrams in the machine-generated translation.

BLEU-4, on the other hand, measures the precision of 4-grams (contiguous sequences of 4 words) in the machine-generated translation compared to the reference translation. It counts the number of 4-grams in the machine-generated translation that appear in the reference translation and divides it by the total number of 4-grams in the machine-generated translation.

BLEU-1 evaluates the precision of individual words in the machine-generated translation, while BLEU-4 evaluates the precision of longer, contiguous sequences of words. In general, higher BLEU scores indicate better translation quality, with a perfect score of 1 indicating that the machine-generated translation is identical to the reference translation [11].



true: black dog is jumping over log along beach

pred: boy is walking on dock at the beach

BLEU: 0.7071067811865476



true: black and white dog is running through the field

pred: black and white dog is running through grassy area

BLEU: 0.7259795291154771



true: child leaping from bed to bed behind the back of man

pred: child is sitting on doorway next to man in yellow shirt

BLEU: 0.7226568811456053



true: dogs fight in grass over toy

pred: two brown dogs running in grassy field

BLEU: 0.7311104457090247

Fig. 5.1 Transformer Predicted Results

Figure 5.1 displays a collection of images with their original captions and the predicted captions generated by the transformer model. Additionally, each predicted caption is accompanied by a BLEU score, which is a metric used to evaluate the quality of the generated captions by comparing them to the original captions. The higher the BLEU score, the more similar the predicted caption is to the original caption. This allows users to assess the performance of the transformer model in generating accurate and meaningful captions for images.



true: little girl covered in paint sits in front of painted rainbow with her hands in bowl
 pred: few children are running in photo chairs
 BLEU: 0.16996005791513966



true: boy smiles in front of stony wall in city
 pred: woman in skirt stands in front of woman wearing sunglasses that watch
 BLEU: 0.2790159393585827



true: collage of one person climbing cliff
 pred: man in yellow shirt climbs railing
 BLEU: 0

Fig. 5.2 LSTM Predicted Results

Figure 5.2 shows a comparison between the original captions and the captions generated by the LSTM model. Some of the generated captions are incorrect and differ significantly from the original captions, indicating that the LSTM model is not entirely accurate in caption generation.

5.2 Implementation Results

The table 5.1, displays the comparison between the image caption results generated by the Transformer model and LSTM model with respective BLEU Scores.








Image	Original	Predicted	BLEU Scores
	Black Dog is jumping over log along beach.	Boy is walking on dock at the beach.	0.71
	Black and white dog is running through the field.	Black and white dog is running through grassy area.	0.73
	Child leaping from bed to bed behind the back of man.	Child is sitting on doorway next to man in yellow shirt.	0.72
	Dogs fight in grass over toy.	Two brown dogs running in grassy field.	0.73
	Little girl covered in paint sits in front of painted rainbow with her hands in bowl.	Few children are running in photo chairs.	0.16
	Boy smiles in front of stony wall in city.	Woman in skirt stands in front of woman wearing sunglasses that watch.	0.27
	Collage of one person climbing cliff.	Man in yellow shirt climbs railing.	0

Table 5.1 Implemented Results

5.3 Results Discussions

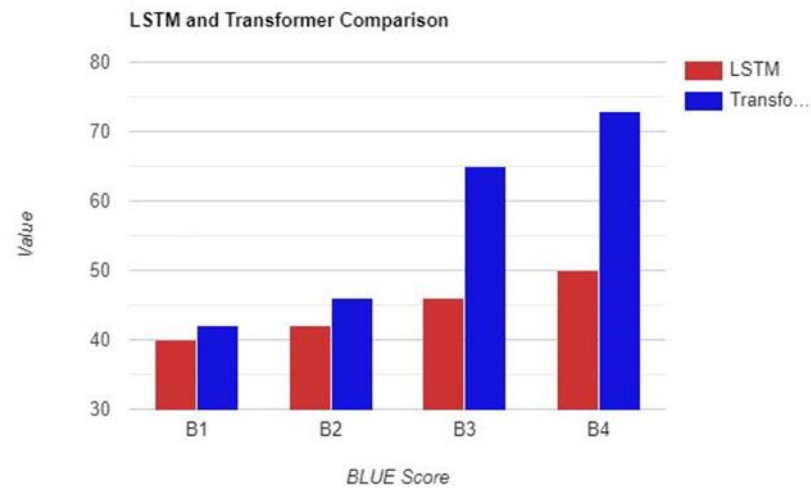


Figure 5.3 LSTM And Transformer Comparision

Figure 5.3 illustrates the difference between the BLEU score of a model with LSTM vs the BLEU score of a model with transformer. It is evident from the graph that the model with transformer is found to be better than the model with LSTM.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The outcomes of the experiments suggest that our proposed strategy is capable of achieving good results. For critical applications, more effort should be put into improving categorization performance. We will now deviate our focus to implement towards other sections of our system, namely, recommendation and recognition system. Our future efforts will be focused on increasing the system's performance and developing more relevant classifications that could be beneficial in a variety of real-world applications.

6.2 Future Work

However, there are a few areas in which there is a further scope of improvement. The methodology is not colour sensitive and not able to detect a wide range of colours accurately. Furthermore, due to the limited size of vocabulary, the generating capability of the model is also limited. Further, the model is unable to achieve high precision in the detection of the number of objects present in the image. In the future, these areas can be improved upon. Further, the model can be trained with a larger dataset and the hyper-parameters of the deep networks can be tuned to achieve better results.

References

- [1] HaoranWang , Yue Zhang, and Xiaosheng Yu, “An Overview of Image Caption Generation Methods”, (CIN-2020)
- [2] <https://github.com/Adhi-Git-hub/machine-learning>
- [3] <https://www.ibm.com/in-en/topics/machine-learning>
- [4] <https://deeptai.org/machine-learning-glossary-and-terms/convolutional-neural-network>
- [5] <https://machinelearningmastery.com/the-transformer-model/>
- [6] [https://blog.hubspot.com/website/what-is-gui#:~:text=A%20graphical%20user%20interface%20\(GUI,actions%20that%20they%20can%20take.](https://blog.hubspot.com/website/what-is-gui#:~:text=A%20graphical%20user%20interface%20(GUI,actions%20that%20they%20can%20take.)
- [7] <https://www.mdpi.com/2076-3417/12/13/6733>
- [8] <https://medium.com/ml-research-lab/python-tutorial-for-data-science-best-practice-from-easy-to-complex-problem-fa2b94e02fcd>
- [9] <https://medium.com/edureka/python-numpy-tutorial-89fb8b642c7d>
- [10] B.Krishnakumar, K.Kousalya, S.Gokul, R.Karthikeyan, and D.Kaviyarasu, “IMAGE CAPTION GENERATOR USING DEEP LEARNING”, (international Journal of Advanced Science and Technology- 2020)
- [11] <https://en.wikipedia.org/wiki/BLEU>

1. MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga, “A Comprehensive Survey of Deep Learning for Image Captioning” ,(ACM-2019)
2. Rehab Alahmadi, Chung Hyuk Park, and James Hahn, “Sequence-to sequence image caption generator”, (ICMV-2018)
3. Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan, “Show and Tell: A Neural Image Caption Generator”,(CVPR 1, 2- 2015)
4. Priyanka Kalena, Nishi Malde, Aromal Nair, Saurabh Parkar, and Grishma Sharma, “Visual Image Caption Generator Using Deep Learning”, (ICAST-2019)

5. Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra, and Nand Kumar Bansode, "Camera2Caption: A Real-Time Image.
6. Hossain, M.Z.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv. (CSUR)* 2019, 51, 1–36. [Google Scholar] [CrossRef][Green Version]
7. Chen, Y.C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; Liu, J. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 104–120. [Google Scholar]
8. Yao, T.; Pan, Y.; Li, Y.; Mei, T. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018*; pp. 684–699. [Google Scholar]
9. Yang, J.; Sun, Y.; Liang, J.; Ren, B.; Lai, S.H. Image captioning by incorporating affective concepts learned from both visual and textual components. *Neurocomputing* 2019, 328, 56–68. [Google Scholar] [CrossRef]
10. Yu, J.; Li, J.; Yu, Z.; Huang, Q. Multimodal transformer with multi-view visual representation for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* 2019, 30, 4467–4480. [Google Scholar] [CrossRef][Green Version]
11. You, Q.; Jin, H.; Wang, Z.; Fang, C.; Luo, J. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016*; pp. 4651–4659. [Google Scholar]
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* 2017, 30, 1–11. [Google Scholar]
13. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning, Lille, France, 7–9 July 2015*; pp. 2048–2057. [Google Scholar]
14. Karpathy, A.; Joulin, A.; Fei-Fei, L.F. Deep fragment embeddings for bidirectional image sentence mapping. *Adv. Neural Inf. Processing*

- Syst. 2014, 27, 1–9. [Google Scholar]
15. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled transformer for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8928–8937. [Google Scholar]
 16. Vinyals, O.; Fortunato, M.; Jaitly, N. Pointer networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 2692–2700. [Google Scholar]
 17. Wang, J.; Xu, W.; Wang, Q.; Chan, A.B. Compare and reweight: Distinctive image captioning using similar images sets. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 370–386. [Google Scholar]
 18. Yeh, R.; Xiong, J.; Hwu, W.M.; Do, M.; Schwing, A. Interpretable and globally optimal prediction for textual grounding using image concepts. *Adv. Neural Inf. Process. Syst.* 2017, 30, 1–11. [Google Scholar]
 19. Amirian, S.; Rasheed, K.; Taha, T.R.; Arabnia, H.R. A short review on image caption generation with deep learning. In Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV), Las Vegas, NV, USA, 29 July–1 August 2019; pp. 10–18. [Google Scholar]
 20. Chohan, M.; Khan, A.; Mahar, M.S.; Hassan, S.; Ghafoor, A.; Khan, M. Image Captioning using Deep Learning: A Systematic. *Image* 2020, 278–286. [Google Scholar]

Acknowledgement

We like to share our sincere gratitude to all those who help us in completion of this project. During the work we faced many challenges due to our lack of knowledge and experience but these people helped us to get over all the difficulties and in final compilation of our idea to a shaped sculpture.

We would like to thank our Head of Department (HOD), Dr. Tanuja Sarode, for approving our project. We also thank Prof. Juhi Janjua for her governance and guidance and internally reviewing our work. Special thanks to the principal, Dr. G.T. Thampi, for motivating the students and providing state-of-the-art lab facilities.

Finally, we would like to thank the Department of Computer Science of Thadomal Shahani Engineering college for providing us such an opportunity to learn from these experiences.

All of our team is thankful to the evaluation committee and the respective faculties.

We are also thankful to our peers and our parents who have inspired us to face all the challenges and overcome all the hurdles faced throughout the project.

Thank you all.

Ronak Dhingra
Sohail Gidwani
Vanshika Gurbani
Divya Panjwani

Spine

B.E (Computer Engineering)	2022-23
----------------------------	---------