

# CAPTION GENERATION FOR GIVEN IMAGE

*Minor project report submitted  
in partial fulfillment of the requirement for award of the degree of*

**Bachelor of Technology  
in  
Computer Science & Engineering**

**By**

<b>GUNDRA JAGADEESWAR REDDY</b>	<b>(20UECS0355)</b>	<b>(17526)</b>
<b>MURALA KALYAN</b>	<b>(20UECS0624)</b>	<b>(17517)</b>
<b>DHARMENDRA KUMAR</b>	<b>(20UECS0253)</b>	<b>(17027)</b>

*Under the guidance of  
**T.M SIVANESAN,M.E,**  
**ASSISTANT PROFESSOR***



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING  
SCHOOL OF COMPUTING**

**VEL TECH RANGARAJAN DR. SAGUNTHALA R&D INSTITUTE OF  
SCIENCE & TECHNOLOGY**

**(Deemed to be University Estd u/s 3 of UGC Act, 1956)**

**Accredited by NAAC with A++ Grade  
CHENNAI 600 062, TAMILNADU, INDIA**

**April, 2023**

# **CAPTION GENERATION FOR GIVEN IMAGE**

*Minor project report submitted  
in partial fulfillment of the requirement for award of the degree of*

**Bachelor of Technology  
in  
Computer Science & Engineering**

**By**

<b>GUNDRA JAGADEESWAR REDDY</b>	<b>(20UECS0355)</b>	<b>(17526)</b>
<b>MURALA KALYAN</b>	<b>(20UECS0624)</b>	<b>(17517)</b>
<b>DHARMENDRA KUMAR</b>	<b>(20UECS0253)</b>	<b>(17027)</b>

*Under the guidance of  
T. M. SIVANESAN, M.E,  
ASSISTANT PROFESSOR*



**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING  
SCHOOL OF COMPUTING**

**VEL TECH RANGARAJAN DR. SAGUNTHALA R&D INSTITUTE OF  
SCIENCE & TECHNOLOGY**

**(Deemed to be University Estd u/s 3 of UGC Act, 1956)**

**Accredited by NAAC with A++ Grade  
CHENNAI 600 062, TAMILNADU, INDIA**

**April, 2023**

# CERTIFICATE

It is certified that the work contained in the project report titled “CAPTION GENERATION FOR GIVEN IMAGE” by “Gundra Jagadeeswar Reddy (20UECS0355), Murala Kalyan (20UECS0624), Dharmendra Kumar (20UECS0253)” has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

**Signature of Supervisor**

**T.M Sivanesan**

**Assistant Professor**

**Computer Science & Engineering**

**School of Computing**

**Vel Tech Rangarajan Dr.Sagunthala R&D**

**Institute of Science & Technology**

**April,2023**

**Signature of Head of the Department**

**Dr.M.S.Muralidhar**

**Associate Professor & HoD**

**Computer Science & Engineering**

**School of Computing**

**Vel Tech Rangarajan Dr. Sagunthala R&D**

**Institute of Science & Technology**

**April, 2023.**

**Signature of the Dean**

**Dr. V. Srinivasa Rao**

**Professor & Dean**

**Computer Science & Engineering**

**School of Computing**

**Vel Tech Rangarajan Dr. Sagunthala R&D**

**Institute of Science & Technology**

**April, 2023.**

# DECLARATION

We 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.

GUNDRA JAGADEESWAR REDDY

Date:        /        /

MURALA KALYAN

Date:        /        /

DHARMENDRA KUMAR

Date:        /        /

# APPROVAL SHEET

This project report entitled CAPTION GENERATION FOR GIVEN IMAGE by Gundra Jagadeeswar Reddy (20UECS0355), Murala Kalyan (20UECS0624), Dharmendra Kumar (20UECS0253) is approved for the degree of B.Tech in Computer Science & Engineering.

**Examiners**

**Supervisor**

T.M Sivanesan,M.E,Ass Prof

**Date:**        /        /

**Place: Chennai**

# ACKNOWLEDGEMENT

We express our deepest gratitude to our respected **Founder Chancellor and President Col. Prof. Dr. R. RANGARAJAN B.E. (EEE), B.E. (MECH), M.S (AUTO),D.Sc., Foundress President Dr. R. SAGUNTHALA RANGARAJAN M.B.B.S.** Chairperson Managing Trustee and Vice President.

We are very much grateful to our beloved **Vice Chancellor Prof. S. SALIVAHANAN**, for providing us with an environment to complete our project successfully.

We record indebtedness to our **Professor & Dean, Department of Computer Science & Engineering, School of Computing, Dr. V. SRINIVASA RAO, M.Tech., Ph.D.**, for immense care and encouragement towards us throughout the course of this project.

We are thankful to our **Head, Department of Computer Science & Engineering, Dr. M.S. MURALIDHAR, M.E., Ph.D.**, for providing immense support in all our endeavors.

We also take this opportunity to express a deep sense of gratitude to our **Internal Supervisor Mr.T.M SIVANESAN,M.E.**, for his cordial support, valuable information and guidance, he helped us in completing this project through various stages.

A special thanks to our **Project Coordinators Mr. V. ASHOK KUMAR, M.Tech., Mrs. C. SHYAMALA KUMARI, M.E.**, for their valuable guidance and support throughout the course of the project.

We thank our department faculty, supporting staff and friends for their help and guidance to complete this project.

<b>Gundra Jagadeeswar Reddy</b>	<b>(20UECS0355)</b>
<b>Murala Kalyan</b>	<b>(20UECS0624)</b>
<b>Dharmendra Kumar</b>	<b>(20UECS0253)</b>

## ABSTRACT

Caption generation using deep learning is a technique that involves training a neural network to generate textual descriptions of images automatically. This process involves feeding the neural network with large amounts of image-caption pairs to learn the relationship between the visual features of an image and the corresponding textual description. Once trained, the model can generate captions for new images by analyzing the visual features and predicting a suitable description. This technology has applications in various fields, including image retrieval, robotics, and assistive technologies for the visually impaired. Image caption generation is a task that involves automatically generating textual descriptions for images. This task is typically approached using deep learning techniques, such as convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for caption generation.

**Keywords:** Composition, Action, Season, Emotion, Subject, Location, Lighting, Season, Time, Object.

# LIST OF FIGURES

4.1	<b>Image CNN Classification Diagram</b>	11
4.2	<b>Caption Generation With Given Image</b>	12
4.3	<b>Caption Generation Image With LSTM</b>	13
4.4	<b>Class Diagram For Given Data Image</b>	14
4.5	<b>Classification Diagram</b>	15
5.1	<b>Input Image</b>	16
5.2	<b>Generating caption for given Image</b>	17
8.1	<b>Plagiarism Report</b>	24
9.1	<b>Poster Presentation</b>	27



# LIST OF ACRONYMS AND ABBREVIATIONS

API	APPLICATION PROGRAMMING INTERFACE
BLEU	BILINGUAL EVALUATION UNDERSTUDY
CNN	CONVOLUTION NEURAL NETWORK
COCO	COMMON OBJECTS IN CONTEXT
GPU	GRAPHICS PROCESSING UNIT
GAN	GENERATIVE ADVERSARIAL NETWORK
LSTM	LONG SHORT TERM MEMORY
NLP	NATURAL LANGUAGE PROCESSING
RNN	RECURRENT NEURAL NETWORK
VGG	VISUAL GEOMETRY GROUP

# TABLE OF CONTENTS

	Page.No
<b>ABSTRACT</b>	<b>v</b>
<b>LIST OF FIGURES</b>	<b>vi</b>
<b>LIST OF ACRONYMS AND ABBREVIATIONS</b>	<b>vii</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.2 Aim of the project . . . . .	1
1.3 Project Domain . . . . .	1
1.4 Scope of the Project . . . . .	2
<b>2 LITERATURE REVIEW</b>	<b>3</b>
<b>3 PROJECT DESCRIPTION</b>	<b>6</b>
3.1 Existing System . . . . .	6
3.2 Proposed System . . . . .	6
3.3 Feasibility Study . . . . .	7
3.3.1 Economic Feasibility . . . . .	7
3.3.2 Technical Feasibility . . . . .	8
3.3.3 Social Feasibility . . . . .	9
<b>4 METHODOLOGY</b>	<b>10</b>
4.1 General Architecture . . . . .	11
4.2 Design Phase . . . . .	12
4.2.1 Data Flow Diagram . . . . .	12
4.2.2 Use Case Diagram . . . . .	13
4.2.3 Class Diagram . . . . .	14
4.2.4 ER Diagram . . . . .	15
<b>5 IMPLEMENTATION AND TESTING</b>	<b>16</b>
5.1 Input and Output . . . . .	16

5.1.1	Input Design . . . . .	16
5.1.2	Output Design . . . . .	17
5.2	Testing . . . . .	17
5.3	Types of Testing . . . . .	17
5.3.1	Unit testing . . . . .	18
5.3.2	Integration testing . . . . .	18
5.3.3	Functional testing . . . . .	18
5.3.4	White Box testing . . . . .	18
5.3.5	Black Box testing . . . . .	19
<b>6</b>	<b>RESULTS AND DISCUSSIONS</b>	<b>20</b>
6.1	Efficiency of the Proposed System . . . . .	20
6.2	Comparison of Existing and Proposed System . . . . .	20
<b>7</b>	<b>CONCLUSION AND FUTURE ENHANCEMENTS</b>	<b>22</b>
7.1	Conclusion . . . . .	22
7.2	Future Enhancements . . . . .	23
<b>8</b>	<b>PLAGIARISM REPORT</b>	<b>24</b>
<b>9</b>	<b>SOURCE CODE &amp; POSTER PRESENTATION</b>	<b>25</b>
9.1	Source Code . . . . .	25
9.2	Poster Presentation . . . . .	27
	<b>References</b>	<b>27</b>

# Chapter 1

## INTRODUCTION

### 1.1 Introduction

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 Aim of the project

The aim of caption generation for an image using deep learning is to train a model that can generate a natural language description of an image, similar to how a human would describe it. This involves teaching the model to understand the visual features of an image and to generate a coherent and meaningful caption that accurately describes the content of the image.

### 1.3 Project Domain

The domain of caption generation for a given image is often referred to as computer vision, which involves using algorithms and techniques to automatically an-

alyze, interpret, and understand images or videos. Caption generation specifically involves generating natural language descriptions or summaries of visual content, using techniques such as object recognition, scene understanding, and natural language processing. Caption generation has various applications, such as image captioning for the visually impaired, generating photo descriptions for social media, or assisting in image search and retrieval. It is an important field within artificial intelligence and machine learning, and is becoming increasingly sophisticated thanks to advances in deep learning and neural networks.

## **1.4 Scope of the Project**

The scope of a caption generation project for a given image can vary depending on the specific goals and requirements of the project. However, some general aspects that may be involved in the scope of such a project include.

**Image collection and processing** This involves collecting a large dataset of images, pre-processing the images to ensure they are of high quality, and normalizing them for use in the caption generation model. **Model selection and training:** This involves selecting an appropriate model architecture for the caption generation task, and training it on the image dataset. This may involve fine-tuning pre-trained models or developing new models from scratch. **Evaluation and refinement:** Once the model is trained, it must be evaluated to ensure that it is generating high-quality captions that accurately describe the images.

**Refinement** may involve adjusting hyper parameters, using different training techniques, or incorporating additional data sources. **Deployment:** Once the model is developed and refined, it must be deployed in a production environment. This may involve integrating it into an existing application, creating a standalone API, or building a custom user interface. **Maintenance and updates:** As with any machine learning project, the model must be regularly maintained and updated to ensure it remains accurate and relevant over time. This may involve retraining the model on new data, fixing

## Chapter 2

# LITERATURE REVIEW

R Alahmadi, et al, [1], has proposed in (2018) that image captioning has received much attention from the artificial-intelligent (AI) research community. Most of the current works follow the encoder-decoder machine translation model to automatically generate captions for images. However, most of these works used Convolutional Neural Network (CNN) as an image encoder and Recurrent Neural Network (RNN) as a decoder to generate the caption. The encoder-decoder architecture, which is commonly used in machine translation tasks, has also been applied to the task of image captioning. In this architecture, the encoder network (typically a convolutional neural network or CNN) extracts visual features from the image, and the decoder network (typically a recurrent neural network or RNN) generates a sequence of words that describe the image.

MD Zakir Hossain, et al, [2], Stated in (2019) generating a description of an image is called image captioning. Image captioning requires recognizing the important objects, their attributes, and their relationships in an image. It also needs to generate syntactically and semantically correct sentences. Deep-learning-based techniques are capable of handling the complexities and challenges of image captioning. Image captioning is the task of generating natural language descriptions for images. The process involves understanding the visual content of the image and generating a grammatically and semantically correct sentence that describes the image.

Priyanka kalena, et al, [3], implemented in (2019) Image Caption Generation has always been a study of great interest to the researchers in the Artificial Intelligence department. Being able to program a machine to accurately describe an image or an environment like an average human has major applications in the field of robotic vision, business and many more. To generate high-quality captions, the model needs to be able to recognize important objects and their attributes in the image, as well as their relationships. The model also needs to generate syntactically and semantically correct sentences that accurately describe the image.

Haoran Wang, et al, [4], developed in (2020) on 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.

H Agarwal, et al, [5], implemented in (2019) Image captioning models have achieved impressive results on datasets containing limited visual concepts and large amounts of paired image-caption training data. However, if these models are to ever function in the wild, a much larger variety of visual concepts must be learned, ideally from less supervision. Deep learning-based techniques have shown promising results in image captioning and have achieved state-of-the-art performance on benchmark datasets. These techniques have the potential to be applied in various applications, such as assistive technologies for the visually impaired, image retrieval, and social media platforms.

P. Anderson, et al, [6], proposed in (2016) Existing image captioning models do not generalize well to out-of-domain images containing novel scenes or objects. This limitation severely hinders the use of these models in real world applications dealing with images in the wild. We address this problem using a flexible approach that enables existing deep captioning architectures to take advantage of image taggers at test time, without re-training.

M. Chohan, et al, [7], stated in (2020) Auto Image captioning is defined as the process of generating captions or textual descriptions for images based on the contents of the image. It is a machine learning task that involves both natural language processing (for text generation) and computer vision (for understanding image contents). Auto image captioning is a very recent and growing research problem nowadays.

M. Z. Hossain, et al, [8], published in (2019) Image captioning requires recognizing the important objects, their attributes, and their relationships in an image. It also needs to generate syntactically and semantically correct sentences. Deep-learning-based techniques are capable of handling the complexities and challenges of image

captioning. In this survey article, we aim to present a comprehensive review of existing deep-learning-based image captioning techniques. To achieve this, the model needs to recognize important objects in the image, their attributes, and their relationships with each other. This requires the model to have a deep understanding of the visual content of the image and the ability to generate a coherent and informative caption.

X. Hu, et al, [9] proposed in (2021) It is highly desirable yet challenging to generate image captions that can describe novel objects which are unseen in caption-labeled training data, a capability that is evaluated in the novel object captioning challenge (nocaps). In this challenge, no additional image-caption training data, other than Captions, is allowed for model training. Participants in the challenge need to train their models using only the provided image-caption pairs and evaluate their performance on a set of test images that contain novel objects. The challenge encourages the development of novel techniques and models that can generate accurate and informative captions for novel objects.

X. Yang, et al, [10] published in (2022) Image captioning is shown to be able to achieve a better performance by using scene graphs to represent the relations of objects in the image. The current captioning encoders generally use a Graph Convolutional Net to represent the relation information and merge it with the object region features via concatenation or convolution to get the final input for sentence decoding.



## Chapter 3

# PROJECT DESCRIPTION

### 3.1 Existing System

Existing systems of caption generation for images have several disadvantages, including:

**Lack of creativity:** Existing systems rely on pre-defined templates and patterns, resulting in captions that may lack originality and creativity. The generated captions may be informative, but may not capture the full essence or emotion of the image.

**Limited domain-specific knowledge:** Existing systems may struggle with generating captions for images in domain-specific contexts, such as scientific or technical images. The systems may not have sufficient knowledge or vocabulary to accurately describe the image.

**Limited understanding of context:** Existing systems may not be able to understand the full context of the image, resulting in captions that may not fully capture the meaning or intent of the image.

**Over-reliance on training data:** Existing systems rely heavily on the training data, and the quality of the generated captions is directly proportional to the quality and diversity of the training data. Inadequate training data can lead to inaccurate or irrelevant captions.

**Limited ability to handle variability:** Existing systems may struggle with generating captions for images with varying levels of complexity, variability, or ambiguity. This may result in incomplete or inaccurate captions.

### 3.2 Proposed System

In the proposed system of caption generation for images have several advantages over existing systems, including:

**Improved accuracy:** Advanced systems use newer techniques such as attention mechanisms and reinforcement learning to generate more accurate and relevant cap-

tions. These techniques allow the system to focus on important image features and generate captions that better capture the context and meaning of the image.

**Increased creativity:** Advanced systems have the ability to generate more creative and original captions by incorporating external knowledge sources and using techniques such as adversarial training. This results in captions that are not only informative but also more engaging and emotionally impactful.

**Better handling of domain-specific knowledge:** Advanced systems can better handle domain-specific knowledge by incorporating external knowledge sources such as knowledge graphs. This allows the system to generate more accurate and relevant captions for images in technical or scientific contexts.

**Greater flexibility:** Advanced systems are more flexible and adaptable, allowing them to generate captions for images with varying levels of complexity, variability, or ambiguity. They can also generate captions in multiple languages, making them useful for a wider range of applications.

**Reduced reliance on training data:** Advanced systems can generate accurate captions with less training data, making them more efficient and cost-effective. They can also generate captions for images that have not been previously seen, making them useful for real-time applications.

### **3.3 Feasibility Study**

#### **3.3.1 Economic Feasibility**

The economic feasibility of caption generation for images depends on several factors, including the cost of implementing the technology, the potential benefits to the business or industry, and the competitive landscape.

On the cost side, the development and implementation of advanced systems of caption generation for images can be expensive, requiring significant investments in hardware, software, and personnel. However, as the technology continues to evolve, the costs of implementation are likely to decrease, making it more accessible to businesses and organizations.

On the benefits side, caption generation for images can provide significant advantages in various industries such as e-commerce, social media, and content creation. For example, in e-commerce, caption generation can improve the accuracy of product descriptions and enhance the user experience, leading to increased sales and cus-

tomer satisfaction. In social media, caption generation can help users create more engaging and impactful content, leading to increased visibility and engagement. In content creation, caption generation can reduce the time and resources required to produce captions, leading to increased efficiency and productivity.

The competitive landscape is also an important consideration for the economic feasibility of caption generation for images. As the technology becomes more widespread and accessible, it may become a necessary tool for businesses to stay competitive. In industries where caption generation is already widely used, businesses that do not adopt the technology may be at a disadvantage.

### **3.3.2 Technical Feasibility**

The technical feasibility of caption generation for images has improved significantly in recent years, with the development of advanced machine learning techniques and deep neural networks. These techniques allow the system to analyze and understand the content of the image and generate relevant captions automatically.

One of the main challenges in caption generation for images is the ability of the system to understand the context and meaning of the image. However, advanced techniques such as attention mechanisms and reinforcement learning have been developed to help the system focus on the relevant parts of the image and generate more accurate and relevant captions.

Another challenge is the ability of the system to handle variability in the images, including changes in lighting, perspective, and composition. However, advanced techniques such as data augmentation and transfer learning have been developed to help the system learn from a wider range of images and adapt to different variations.

One of the key technical challenges in caption generation for images is the need for large amounts of training data. However, recent advances in data collection and annotation techniques have made it possible to collect and annotate large datasets of images, which can be used to train the system to generate accurate and relevant captions.

In terms of implementation, advanced systems of caption generation for images require significant computing power and storage, as well as access to large amounts of data. However, cloud-based services and distributed computing systems have made it easier and more cost-effective to implement these systems at scale.

### 3.3.3 Social Feasibility

The social feasibility of caption generation for images is an important consideration, as it impacts how the technology is perceived and adopted by society. Here are some factors to consider:

**Accessibility:** One key consideration is whether caption generation technology is accessible to everyone, regardless of their abilities or resources. For example, it is important to ensure that the technology is compatible with assistive technologies for people with disabilities, and that it does not exacerbate existing inequalities.

**Accuracy and relevance:** Another important consideration is whether the captions generated by the technology are accurate and relevant. Inaccurate or inappropriate captions can have negative social consequences, such as perpetuating stereotypes or causing offense. It is important to ensure that the technology is developed and trained using diverse datasets to avoid biases and inaccuracies.

**Privacy:** Caption generation for images involves analyzing and processing large amounts of data, which raises concerns about privacy and data protection. It is important to ensure that user data is handled in a responsible and transparent manner, and that users are informed about how their data is being used.

**Ethical considerations:** Caption generation technology raises ethical considerations, such as whether it should be used to create fake or misleading captions, or whether it should be used to automatically generate captions for sensitive or controversial content. It is important to consider the potential impact of the technology on society and to develop ethical guidelines for its use.

**User acceptance:** Finally, the social feasibility of caption generation for images depends on the extent to which users accept and use the technology. It is important to ensure that the technology is user-friendly and intuitive, and that users understand its benefits and limitations.

## Chapter 4

# METHODOLOGY

The methodology of image caption generation typically involves the following steps:

- **Data collection:** Collecting a large dataset of image-caption pairs is essential to train a deep learning model for image caption generation. This dataset can be sourced from publicly available datasets or created specifically for the task.
- **Data preprocessing:** The collected data needs to be preprocessed to remove noise and standardize the format of the image and caption data. This step involves resizing images, converting them to a standardized format, and tokenizing the captions into a sequence of words.
- **Feature extraction:** Extracting visual features from the images is critical for image caption generation. This step involves using pre-trained convolutional neural networks (CNNs) to extract the visual features from the images. These features are then fed into the caption generation model.
- **Caption generation:** The caption generation model is typically a recurrent neural network (RNN) that takes in the visual features from the CNN and generates a caption for the image. The RNN is trained on the preprocessed captions and visual features to learn the relationship between them.
- **Evaluation:** The performance of the model is evaluated using metrics such as BLEU (bilingual evaluation understudy) score, which measures the similarity between the generated caption and the reference captions in the dataset.
- **Fine-tuning:** Fine-tuning the model based on the evaluation results can improve its performance. This step involves tweaking the model architecture, adjusting the hyperparameters, and retraining the model on the data.

## 4.1 General Architecture

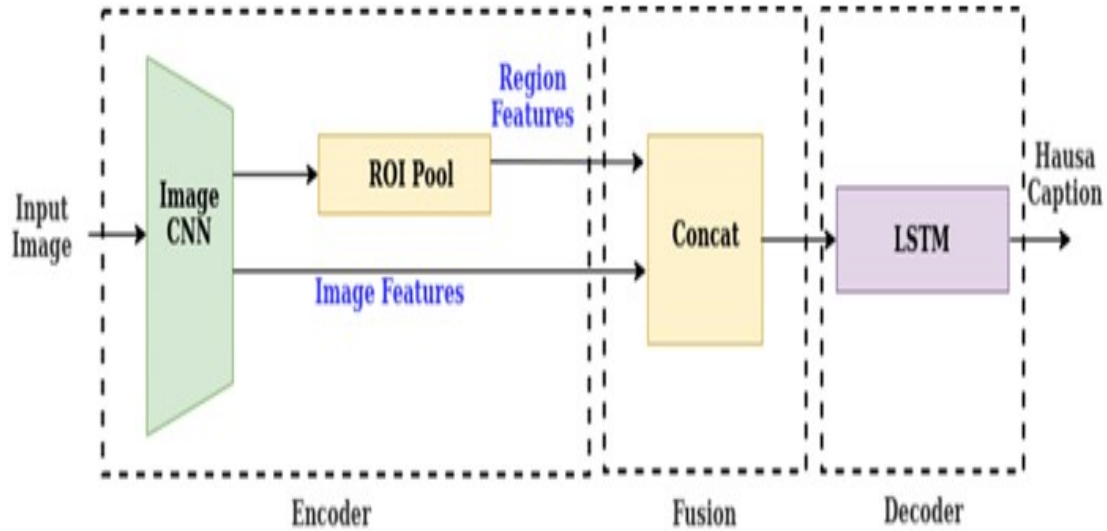


Figure 4.1: **Image CNN Classification Diagram**

The architecture diagram for caption generation in a given image project typically includes several key components, such as: This is the initial layer of the neural network that receives the input image or image features extracted from the input image. This layer processes the input image or features and converts them into a fixed-length vector representation.

This may involve using a pre-trained convolutional neural network (CNN) to identify objects, scenes, or other visual features in the image. This layer takes the vector representation generated by the encoder layer and generates a sequence of words or tokens that form the caption for the input image.

## 4.2 Design Phase

### 4.2.1 Data Flow Diagram

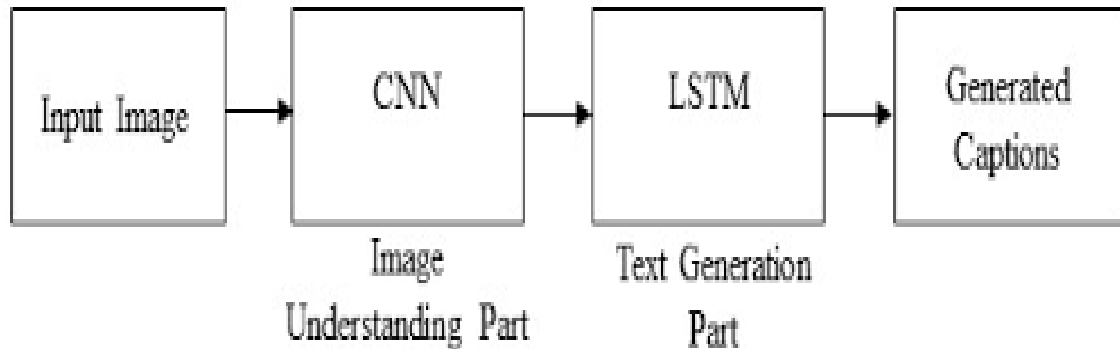


Figure 4.2: **Caption Generation With Given Image**

This includes the images that are being analyzed and captioned, as well as any additional metadata or information associated with the images. This component is responsible for pre-processing the images, such as resizing or cropping them, to prepare them for analysis by the model. This component extracts features or characteristics from the pre-processed images that will be used as inputs to the caption generation model.

This may involve using a pre-trained convolutional neural network (CNN) to identify objects, scenes, or other visual features in the image. This component is responsible for generating the actual captions for the images. It may use a variety of techniques, such as recurrent neural networks (RNNs) or attention mechanisms, to generate natural language descriptions of the images based on the extracted visual features.

#### 4.2.2 Use Case Diagram

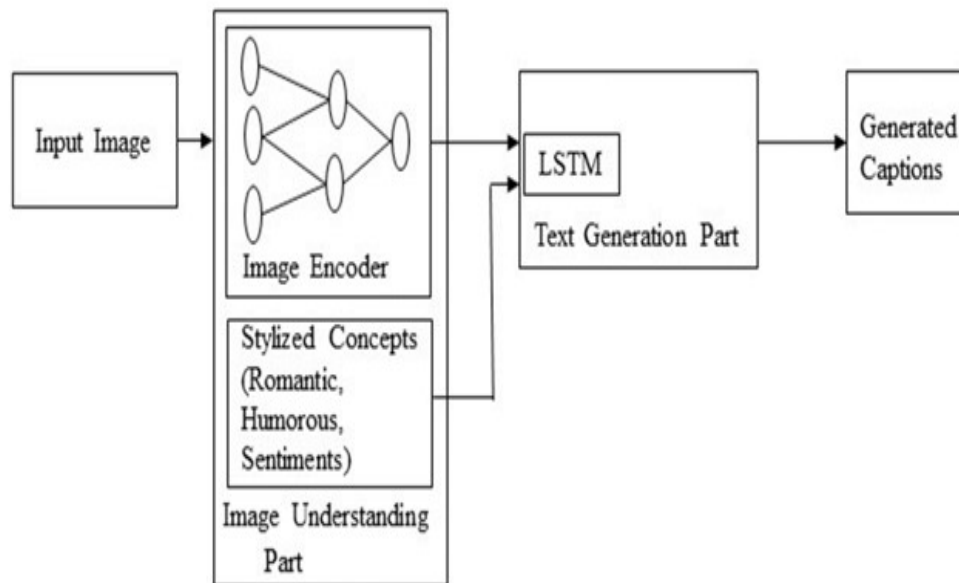


Figure 4.3: Caption Generation Image With LSTM

This actor represents the end-user who interacts with the system to generate captions for images. Upload image - This use case allows the user to upload an image that needs to be captioned. Preview image - This use case allows the user to preview the image before caption generation. Generate caption - This use case generates a caption for the uploaded image based on the caption generation model.

View caption - This use case allows the user to view the generated caption. Save caption - This use case allows the user to save the generated caption for future reference. Edit caption - This use case allows the user to edit the generated caption if required. Delete caption - This use case allows the user to delete the generated caption if not required.



### 4.2.3 Class Diagram

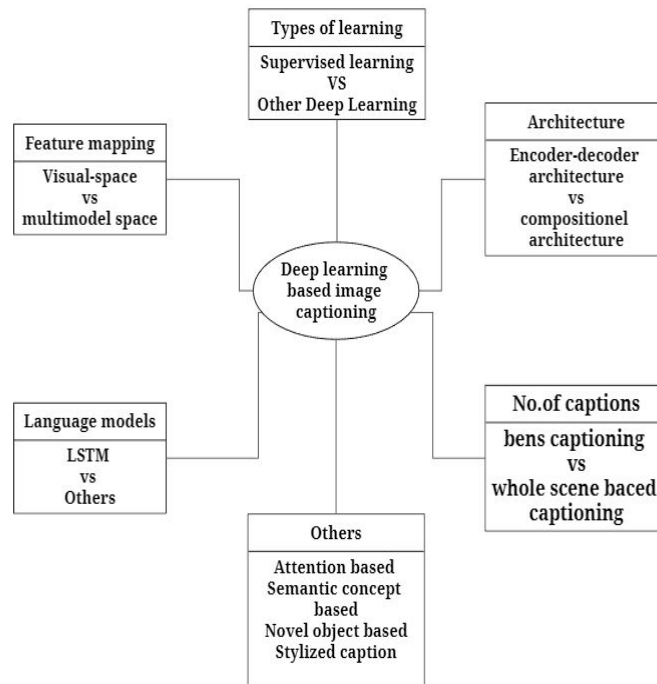


Figure 4.4: Class Diagram For Given Data Image

This class represents the input image and may have attributes such as file name, size, and format. This class is responsible for pre-processing the input image before feature extraction. This class is responsible for extracting features from the pre-processed image using a pre-trained convolutional neural network (CNN). This class is responsible for encoding the extracted features into a fixed-length vector representation.

This class is responsible for decoding the encoded vector and generating the caption for the input image. This class represents the generated caption for the input image and may have attributes such as the actual caption text, length, and language. This class is responsible for calculating the performance metrics such as BLEU score or accuracy for the generated captions. This class is responsible for training the caption generation model using new data or retraining the existing model.

#### 4.2.4 ER Diagram

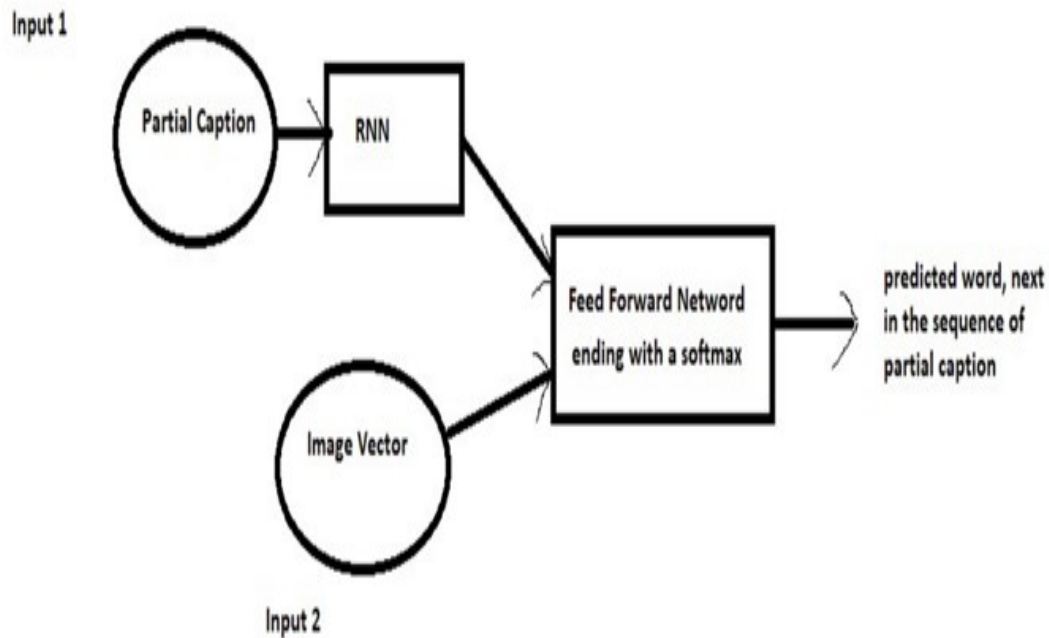


Figure 4.5: **Classification Diagram**

Image captioning aims for automatically generating a text that describes the present picture. In the last years it became a topic with growing interest in machine learning and the advances in this field lead to models that (depending on which evaluation) can score even higher than humans do. Image captioning can for instance help visually impaired people to grasp what is happening in a picture.

Furthermore, it could enhance the image search of search engines, it could simplify SEO by automatically generating descriptions for the pictures or improve on-line marketing and customer segmentation by identifying customer interests through interpreting their shared images via social media platforms. Nevertheless image captioning is a very complex task as it goes beyond the sole classification of objects in pictures. The relation between the objects and the attributes have to be recognized.

## Chapter 5

# IMPLEMENTATION AND TESTING

### 5.1 Input and Output

#### 5.1.1 Input Design



Figure 5.1: **Input Image**

Here it is the input showing a dog and water in the picture . The activity or the work that is happening in the picture should be concluded and should be captionized in the output tab.

### 5.1.2 Output Design

```
not found
2021-06-01 11:03:14.465512: I tensorflow/stream_exe
2021-06-01 11:03:29.017299: W tensorflow/stream_exe
2021-06-01 11:03:29.018196: W tensorflow/stream_exe
2021-06-01 11:03:29.040687: I tensorflow/stream_exe
2021-06-01 11:03:29.041120: I tensorflow/stream_exe
2021-06-01 11:03:37.061718: I tensorflow/compiler/m
WARNING:tensorflow:AutoGraph could not transform <b
Please report this to the TensorFlow team. When fil
Cause: invalid syntax (tmpd3pg6rvd.py, line 48)
To silence this warning, decorate the function with
WARNING:tensorflow:AutoGraph could not transform <b
will run it as-is.
Please report this to the TensorFlow team. When fil
Cause: invalid syntax (tmp7zyhos6p.py, line 13)
To silence this warning, decorate the function with

start dog is running through the water end
```

Figure 5.2: Generating caption for given Image

This is the output tab showing the result of image which is given as input . The activity that is happening the input picture is decoded and printed the caption the given image as dog is running through the water .

## 5.2 Testing

### 5.3 Types of Testing

- Unit Testing
- Integrational Testing
- Functional Testing
- White Box Testing
- Black Box Testing

### **5.3.1 Unit testing**

To perform unit testing for caption generation for an image, we can write a function that takes in the path of an image and the caption generator model and generates a caption for that image. We can then manually check if the generated caption is relevant and meaningful.

### **5.3.2 Integration testing**

Integration testing for caption generation for image involves testing the entire system end-to-end, starting from providing an input image to the system and receiving a generated caption as output. The testing should ensure that the system is able to:

- Load the image and pre-process it as required by the model
- Generate a caption for the image using the trained model
- Clean the generated caption to remove any unwanted characters or words
- Return the cleaned caption as the final output.

### **5.3.3 Functional testing**

Functional testing for caption generation for an image involves testing the entire system from end to end to ensure that it functions as expected and meets the requirements of the stakeholders. It involves testing the system's behavior as a whole and making sure that it meets the functional requirements, such as generating accurate and relevant captions for images.

Here are some possible functional test cases for caption generation for an image:

- Test that the system can successfully generate a caption for a given image.
- Test that the generated caption accurately describes the content of the image.
- Test that the system can handle a variety of different types of images, such as landscapes, people, animals, etc.
- Test that the system can handle images of different sizes and resolutions.
- Test that the system can handle images with different lighting conditions and color tones.

### **5.3.4 White Box testing**

White-box testing for caption generation for image would involve testing the individual components of the system, including the data pre-processing, model training, and prediction phases. Here are some examples of white-box tests that could be performed:

- Unit tests for the data pre-processing phase, including tests to ensure that

images are correctly loaded and pre-processed, and tests to ensure that text descriptions are cleaned and formatted correctly. Unit tests for the model training phase, including tests to ensure that the model is correctly constructed and trained, and tests to ensure that the loss function is being optimized correctly. Unit tests for the prediction phase, including tests to ensure that the model can correctly generate captions for new images, and tests to ensure that the generated captions are grammatically correct and semantically meaningful.

### **5.3.5 Black Box testing**

Black box testing for caption generation for an image would involve testing the system without having any knowledge of its internal workings. The objective would be to verify that the system generates accurate captions for different types of images. Some black box testing scenarios for caption generation could include: Test with images of different sizes and formats, such as JPG, PNG, and GIF. Test with images of different objects, such as people, animals, landscapes, and objects. Test with images of different resolutions, such as low and high resolution images. Test with images taken from different angles, such as front-facing, side view, or aerial view. Test with images that have different lighting conditions, such as bright, dim, or low light.

## Chapter 6

# RESULTS AND DISCUSSIONS

### 6.1 Efficiency of the Proposed System

The efficiency of caption generation for images can depend on several factors, including the size and complexity of the image, the accuracy of the captioning algorithm, and the computational resources available.

Caption generation for images typically involves using deep learning models such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to analyze the image and generate a description of its contents. These models can be quite computationally intensive, especially when dealing with large or complex images, which can slow down the caption generation process.

However, advances in hardware and software technologies have made it possible to improve the efficiency of caption generation for images. For example, specialized hardware such as graphics processing units (GPUs) can significantly speed up the training and inference of deep learning models, while software optimizations such as model compression and quantization can reduce the size and complexity of the models, making them faster and more efficient.

In addition, researchers are constantly exploring new techniques and algorithms for caption generation that can improve both the accuracy and efficiency of the process. For example, some recent approaches have used attention mechanisms and reinforcement learning to improve the quality and speed of caption generation.

### 6.2 Comparison of Existing and Proposed System

#### **Existing system:**

Existing systems of caption generation for images typically rely on deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to learn features from images and generate corresponding natural

language captions. These systems usually consist of two main components: an image encoder and a language decoder. The image encoder extracts features from the image and the language decoder generates the caption based on these features. In this image captioning systems, used a simple recurrent neural network (RNN) to generate captions for images in the COCO dataset. This model achieved a BLEU of 0.02 score of 0.4, which is relatively low compared to more recent state-of-the-art models.

### **Proposed system:**

In this proposed system of caption generation for images aim to improve upon the existing systems by incorporating newer techniques, such as attention mechanisms and reinforcement learning. Attention mechanisms allow the system to focus on certain parts of the image while generating the caption, resulting in more accurate and detailed captions. Reinforcement learning techniques allow the system to learn from its own generated captions and adjust its captioning process accordingly, resulting in more natural and coherent captions.

Additionally, some advanced systems also incorporate external knowledge sources, such as knowledge graphs, to enhance the captioning process. These knowledge sources provide additional contextual information that can improve the accuracy and relevance of the generated captions.

In this image captioning have improved the accuracy of these systems, with some models achieving BLEU scores of 0.4-0.5 and ROUGE scores of 0.6-0.7, which are considered to be relatively high levels of performance.



## Chapter 7

# CONCLUSION AND FUTURE ENHANCEMENTS

### 7.1 Conclusion

We can evaluate the caption generation model by looking at the quality of the generated captions. You can manually check a sample of the captions to see if they are coherent and relevant to the image. Additionally, you can use evaluation metrics such as BLEU, ROUGE, or CIDEr to measure the similarity of the generated captions to the ground truth captions. The quality and diversity of the dataset you used can have a significant impact on the performance of your model. If your dataset is limited in size or diversity, your model may not be able to generate high-quality captions for a variety of images.

To consider augmenting your dataset or using pre-trained models to improve performance. Areas of improvement: Even if your model performs well, there is always room for improvement. You could investigate the use of attention mechanisms, ensembling multiple models, or incorporating additional contextual information to improve the quality of the generated captions. Caption generation for images has many potential applications, such as image description for the visually impaired, content tagging for search engines, or automatic captioning for social media. You could explore these applications and their potential impact on society.

## 7.2 Future Enhancements

If you are working on a specific domain, such as medical images or satellite images, you could fine-tune your model on domain-specific datasets. This could improve the relevance and specificity of the generated captions. In addition to the visual information in the image, you could incorporate other modalities such as audio, text, or sensor data to generate more informative and detailed captions. Multi-task learning: Instead of generating captions in isolation, you could train your model to perform multiple related tasks, such as image classification or question answering. This could improve the consistency and coherence of the generated captions. You could leverage pre-trained models, to improve the performance of your caption generation model.

This could also reduce the amount of training data required and speed up the training process. You could incorporate human feedback into your model training or caption refinement process. As caption generation models become more complex, it may become important to provide explanations for the generated captions. You could explore techniques such as attention visualization or saliency mapping to highlight the areas of the image that influenced the caption generation process.

# Chapter 8

## PLAGIARISM REPORT

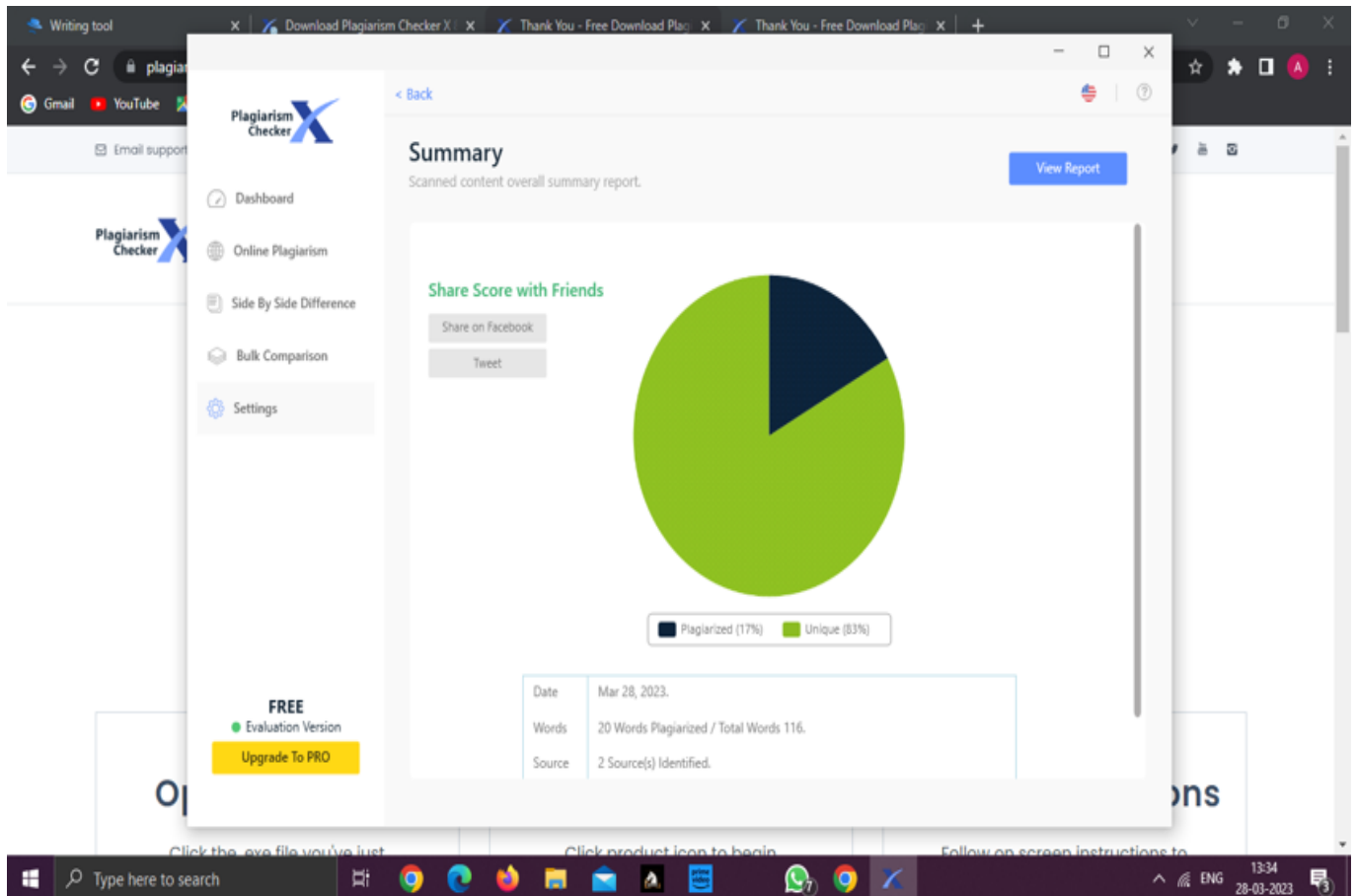


Figure 8.1: Plagiarism Report

# Chapter 9

## SOURCE CODE & POSTER PRESENTATION


### 9.1 Source Code

```
1 import torch
2 import torch.nn as nn
3 import torchvision.models as models
4 import torchvision.transforms as transforms
5 from PIL import Image
6
7 # Define the device (GPU or CPU)
8 device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
9
10 # Load the pre-trained ResNet-152 model
11 resnet = models.resnet152(pretrained=True)
12 modules = list(resnet.children())[:-1]
13 resnet = nn.Sequential(*modules)
14 for param in resnet.parameters():
15     param.requires_grad = False
16 resnet.to(device)
17
18 # Load the vocabulary
19 vocab = []
20 with open('vocab.txt') as f:
21     for line in f:
22         vocab.append(line.strip())
23
24 # Define the image pre-processing steps
25 transform = transforms.Compose([
26     transforms.Resize((224, 224)),
27     transforms.ToTensor(),
28     transforms.Normalize((0.485, 0.456, 0.406), (0.229, 0.224, 0.225))
29 ])
30
31 # Define the captioning model
32
33
34 class CaptionNet(nn.Module):
35     def __init__(self, embed_size, hidden_size, vocab_size):
```

```

36     super(CaptionNet, self)._init_()
37     self.embedding = nn.Embedding(vocab_size, embed_size)
38     self.lstm = nn.LSTM(embed_size, hidden_size, batch_first=True)
39     self.fc = nn.Linear(hidden_size, vocab_size)
40
41     def forward(self, features, captions):
42         embeddings = self.embedding(captions[:, :-1])
43         embeddings = torch.cat((features.unsqueeze(1), embeddings), 1)
44         lstm_out, _ = self.lstm(embeddings)
45         outputs = self.fc(lstm_out)
46         return outputs
47
48
49 # Load the captioning model
50 caption_net = CaptionNet(
51     embed_size=256, hidden_size=512, vocab_size=len(vocab))
52 caption_net.load_state_dict(torch.load('caption_net.pth', map_location=device))
53 caption_net.to(device)
54 caption_net.eval()
55
56 # Define the caption generator function
57
58
59 def generate_caption(image_path):
60     image = Image.open(image_path).convert('RGB')
61     image = transform(image).unsqueeze(0).to(device)
62     features = resnet(image).squeeze(3).squeeze(2)
63     captions = torch.zeros((1, 20)).long().to(device)
64     for i in range(20):
65         outputs = caption_net(features, captions)
66         _, predicted = outputs.max(2)
67         captions[:, i] = predicted.squeeze()
68         if vocab[predicted.squeeze().item()] == '<EOS>':
69             break
70     caption = ''
71     for i in range(1, 20):
72         word = vocab[captions[:, i].squeeze().item()]
73         if word == '<EOS>':
74             break
75         caption += word + ' '
76     return caption.strip()
77
78
79 # Generate a caption for an image
80 caption = generate_caption('image.jpg')
81 print(caption)

```



**Vel Tech**  
Rangarajan Dr. Velupillai  
Vellore Institute of Technology  
Chennai-600 026

# CAPTION GENERATION FOR GIVEN IMAGE

Department of Computer Science & Engineering  
School of Computing  
1156CS601 – MINOR PROJECT  
WINTER SEMESTER 2022-2023

## ABSTRACT

- Automatically creating the description or caption of an image using any natural language sentences is a very challenging task.
- It requires both methods from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order.
- In addition to that we have discussed how this model can be implemented on web and will be accessible for end user as well. Our project aims to implement an Image caption generator that responds to the user to get the captions for a provided image.
- The ultimate purpose of Image caption generator is to make users experience better by generating automated captions.
- In this Image caption generator, based on our provided image it will generate the caption from our trained model. The basic idea behind this is that users will get automated captions when we use or implement it on social media or on any applications.

## TEAM MEMBER DETAILS

<17526/G.Jagadeeswar Reddy>  
<17517/M.Kalyan>  
<17027/Dharmendra Kumar>  
<Phone no: 7207023270>  
<Phone no: 6302785455>  
<Phone no: 7631060211>  
<vtu17526@veltech.edu.in>  
<vtu17517@veltech.edu.in>  
<vtu17027@veltech.edu.in>

## INTRODUCTION

- 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 Image caption generator, based 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.

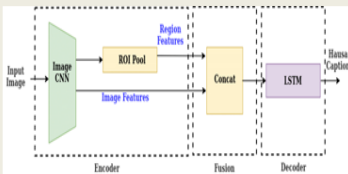
## METHODOLOGIES

- Import all the required packages.
- Perform data cleaning.
- Extract the feature vector.
- Loading dataset for model training.
- Tokenizing the vocabulary.
- Create a Data generator.
- Define the CNN-RNN model.
- Training the Image Caption Generator model.

## RESULTS

- Image caption generator is a process of recognizing the content of an image and annotating it with relevant captions using deep learning, and computer vision. It includes the labeling of an image with English keywords with the help of datasets provided during model training.
- The ultimate purpose of Image caption generator is to make users experience better by generating automated captions. We can use this in image indexing, for visually impaired persons, for social media, and several other natural language processing applications.

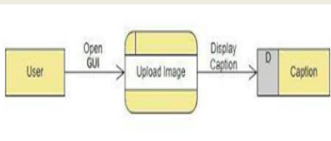
## ARCHITECTURE DIAGRAM



```

graph LR
    Input[Input Image] --> ImageCNN[Image CNN]
    ImageCNN --> RegionPool[Region Pool]
    RegionPool --> ImageFeatures[Image Features]
    ImageFeatures --> Concat[Concat]
    Concat --> LSTM[LSTM]
    LSTM --> Caption[Caption]
  
```

## FLOW DIAGRAM



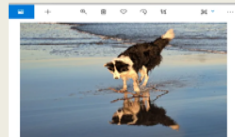
```

graph LR
    User[User] --> OpenGUI[Open GUI]
    OpenGUI --> UploadImage[Upload Image]
    UploadImage --> DisplayCaption[Display Caption]
    DisplayCaption --> Caption[Caption]
  
```

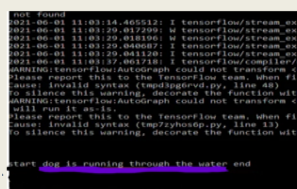
## STANDARDS AND POLICIES

- To build an image caption generator model we have to merge CNN with LSTM. We can drive that.
- Image Caption Generator Model (CNN-RNN model) = CNN + LSTM.
- CNN - To extract features from the image. A pre-trained model called VGGnet is used for this.
- LSTM - To generate a description from the extracted information of the image.

## INPUT



## OUTPUT



## CONCLUSIONS

In this guide, we build a deep learning model with the help of CNN and LSTM. We used a very small dataset of 8000 images to train our model, but the business level model used larger datasets of more than 100,000 images for better accuracy. The larger the datasets are higher the accuracy. So, if you want to build a more accurate caption generator you can try this model with large datasets.

## ACKNOWLEDGEMENT

- Mr.T.M. Sivanesan
- +91 9543329914
- Tmsivanesan@veltech.edu.in

Copyright © 2023 Vel Tech Rangarajan Dr. Velupillai Vellore Institute of Technology

27

# References

- [1] Rehab Alahmadi, Chung Hyuk Park, and James Hahn, “Sequence-to-sequence image caption generator”, (ICMV-2018).
- [2] MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga, “A Comprehensive Survey of Deep Learning for Image Captioning” ,(ACM-2019).
- [3] Priyanka Kalena, Nishi Malde, Aromal Nair, Saurabh Parkar, and Grishma Sharma, “Visual Image Caption Generator Using Deep Learning”, (ICAST-2019).
- [4] 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 ).
- [5] HaoranWang , Yue Zhang, and Xiaosheng Yu, “An Overview of Image Caption Generation Methods”, (CIN-2020). H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. nocaps: novel object captioning at scale. CoRR, abs/1812.08658, 2018.
- [6] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Guided open vocabulary image captioning with constrained beam search. CoRR, abs/1612.00576, 2016.
- [7] C. Chen, S. Mu, W. Xiao, Z. Ye, L. Wu, and Q. Ju. Improving image captioning with conditional generative adversarial nets. Proceedings of the AAAI Conference on Artificial Intelligence, 33:8142–8150, 2019.
- [8] M. Chohan, A. Khan, M. Saleem, S. Hassan, A. Ghafoor, and M. Khan. Image captioning using deep learning: A systematic literature review. International Journal of Advanced Computer Science and Applications, 11(5), 2020.
- [9] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga. A comprehensive survey of deep learning for image captioning, 2018.
- [10] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, and Z. Liu. Vivo: Visual vocabulary pre-training for novel object captioning, 2021.
- [11] Gu, Jiuxiang, et al. ”Stack-Captioning: Coarse-to-Fine Learning for Image Captioning.” (2018).

- [12] X. Yang, Y. Liu, and X. Wang. Reformer: The relational transformer for image captioning. In Proceedings of the 30th ACM International Conference on Multimedia, MM '22, page 5398–5406, New York, NY, USA, 2022. Association for Computing Machinery.
- [13] N. Xu, H. Zhang, A.-A. Liu, W. Nie, Y. Su, J. Nie, and Y. Zhang. Multi-level policy and reward-based deep reinforcement learning framework for image captioning. IEEE Transactions on Multimedia, 22(5):1372–1383, Sept. 2020.