A

Mini Project Report on

## Caption Craft

Submitted in partial fulfillment of the requirements for the degree of

BACHELOR OF ENGINEERING

IN

### Computer Science & Engineering

### Artificial Intelligence & Machine Learning

by

Vaibhav Bura (22106067)

Vivek Dalvi (22106108)

Yash Desai (22106005)

Pratik Dhas (22106063)

Under the guidance of

## Prof. Poonam Tiware



### Department of Computer Science & Engineering

### (Artificial Intelligence & Machine Learning)

**A. P. Shah Institute of Technology**

**G. B. Road, Kasarvadavali, Thane (W)-400615**

**University Of Mumbai**

**2024-2025**

## A. P. SHAH INSTITUTE OF TECHNOLOGY

## CERTIFICATE

This is to certify that the project entitled “**Title of the Project”** is a bonafide work of Vaibhav Bura(22106067), Vivek Dalvi(22106108), Yash Desai(22106005), Pratik Dhas(22106063) submitted to the University of Mumbai in partial fulfillment of the requirement for the award of **Bachelor of Engineering** in **Computer Science & Engineering (Artificial Intelligence & Machine Learning).**

|  |  |
| --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
| Prof. Poonam Tiware | Dr. Jaya Gupta |
| Mini Project Guide | Head of Department |

## 

## A. P. SHAH INSTITUTE OF TECHNOLOGY

## Project Report Approval

This Mini project report entitled “**Caption Craft*”*** by **Vaibhav Bura, Vivek Dalvi, Yash Desai and Pratik Dhas**is approved for the degree of ***Bachelor of Engineering*** in ***Computer Science &Engineering***, (AIML) ***2024-25***.

##### External Examiner:

##### Internal Examiner:

Place: APSIT, Thane

Date:

**Declaration**

##### We declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I 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.

|  |  |  |  |
| --- | --- | --- | --- |
| Vaibhav Bura | Vivek Dalvi | Yash Desai | Pratik Dhas |
| (22106067) | (22106108) | (22106005) | (22106063) |

#### ABSTRACT

The rapid advancements in computer vision and natural language processing have paved the way for innovative applications in artificial intelligence. This mini project focuses on developing Caption Craft (an Image Caption Generator), a system designed to automatically produce descriptive captions for images. The primary objective is to bridge the gap between visual content and textual interpretation, making images more accessible and informative.

The project utilizes a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to achieve this goal. Specifically, a pre-trained CNN, such as InceptionV3 or ResNet, is employed to extract high-level features from images. These features are then fed into an RNN, typically a Long Short-Term Memory (LSTM) network, which generates coherent and contextually relevant captions based on the visual content.

A key component of the project involves training the model on a large dataset of images paired with corresponding captions, such as the Flicker 8k dataset. The training process involves fine-tuning the model to accurately capture the relationships between image features and descriptive text.

The outcomes of the project demonstrate the potential of combining computer vision and language models to enhance image understanding and accessibility. The Image Caption Generator can be applied in various domains, including content management, assistive technologies for visually impaired individuals, and automated media tagging. Future work could involve expanding the model's capabilities to handle more diverse and complex images, as well as integrating it into real-time applications.

**Keywords**: Caption Craft, CNN, Flicker 8k dataset, contextually relevant captions

**Index**

|  |  |  |  |
| --- | --- | --- | --- |
| Index | | | Page no. |
| Chapter-1 | | |  |
|  | Introduction | |  |
|  |  |  |  |
| Chapter-2 | | |  |
|  | Literature Survey | |  |
|  | 2.1 | History |  |
|  | 2.1 | Review |  |
|  |  |  |  |
| Chapter-3 | | |  |
|  | Problem Statement | |  |
|  |  |  |  |
| Chapter-4 | | |  |
|  | Experimental Setup | |  |
|  | 4.1 | Hardware setup |  |
|  | 4.2 | Software Setup |  |
|  |  |  |  |
| Chapter-5 | | |  |
|  | Proposed system and Implementation | |  |
|  | 5.1 | Block Diagram of proposed system |  |
|  | 5.2 | Description of Block diagram |  |
|  | 5.3 | Implementation |  |
|  |  |  |  |
| Chapter-6 | | |  |
|  | Conclusion | |  |
|  |  |  |  |
| References | | |  |
|  |  |  |  |

# CHAPTER 1 INTRODUCTION

### INTRODUCTION

In the era of digital information, the ability to understand and describe visual content through automated systems is increasingly valuable. Caption Craft (an Image Caption Generator) is a sophisticated tool that addresses this need by creating descriptive textual captions for images. This technology holds promise across various fields, including content creation, accessibility for visually impaired individuals, and enhancing user interactions in social media platforms.

The core of an image captioning system lies in its ability to translate visual information into coherent language. This process involves two key components: feature extraction and language generation. Convolutional Neural Networks (CNNs) play a pivotal role in the first component—feature extraction. CNNs are a class of deep neural networks designed to process and analyze visual data. They are highly effective in identifying and extracting hierarchical patterns and features from images, such as edges, textures, and object shapes.

In the context of image captioning, CNNs are employed to transform an image into a set of high-level features that capture its visual content. For instance, a pre-trained CNN like InceptionV3 or ResNet can be used to generate a feature vector that summarizes the visual elements of an image. These feature vectors serve as input for the subsequent language generation phase.

The language generation phase is typically handled by Recurrent Neural Networks (RNNs), often utilizing Long Short-Term Memory (LSTM) units or Transformer architectures. These models are adept at generating sequences of words based on the input feature vectors. By learning from vast datasets of images and their corresponding captions, the system learns to generate descriptive and contextually relevant captions that mirror human-like language.

The integration of CNNs for feature extraction and RNNs for language generation enables the development of a Caption Craft that can automatically produce meaningful and accurate descriptions of images. This capability is not only transformative for content management and search engine optimization but also crucial for advancing accessibility and enhancing user engagement in various applications.

In summary, the Caption Craft using CNNs represents a significant advancement in bridging the gap between visual and textual information. By leveraging the power of deep learning techniques, it offers a promising solution for understanding and describing the rich content embedded in images.

# CHAPTER 2 LITERATURE SURVEY

#### LITERATURE SURVEY

###### 2.1-HISTORY

The history of image caption generation research spans several decades and reflects the evolution of both computer vision and natural language processing technologies. Early efforts in this field focused on simple, rule-based approaches where descriptions were manually crafted based on predefined templates or keywords extracted from images. These methods were limited in their ability to generate diverse and contextually accurate captions due to their reliance on fixed rules and shallow feature extraction.

The advent of machine learning marked a significant shift in image captioning research. In the early 2000s, researchers began exploring statistical models and more sophisticated image analysis techniques. The introduction of object detection algorithms allowed systems to identify and label objects within images, providing a foundation for more meaningful captions. However, these models still struggled with integrating visual features into coherent and contextually rich text.

The real breakthrough came with the rise of deep learning in the 2010s. Convolutional Neural Networks (CNNs), particularly those like AlexNet and VGG, demonstrated significant improvements in image classification and feature extraction, paving the way for more advanced image captioning systems. Researchers started employing CNNs to extract high-level features from images, which were then combined with Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, to generate natural language descriptions. The encoder-decoder architecture, where CNNs served as the encoder and LSTMs as the decoder, became a widely adopted approach. This paradigm shift enabled the generation of more fluent and coherent captions by effectively mapping visual features to textual descriptions.

Further advancements in this era included the integration of attention mechanisms, which allowed models to focus on specific regions of an image while generating captions. The Bahdanau attention model and its variants significantly enhanced the capability of captioning systems by improving their ability to handle spatial and temporal dependencies.

The introduction of Transformer models, such as those used in BERT and GPT, further advanced the field. These models excel in capturing complex dependencies and context within text, leading to more accurate and context-aware image captions. The integration of Transformers with CNNs and other vision models, such as Vision Transformers (ViTs), has led to state-of-the-art results in image captioning.

Recent research has also explored multimodal approaches that combine vision and language models with external knowledge sources and large-scale pre-trained models. These advancements aim to address challenges like contextual understanding and fine-grained descriptions, pushing the boundaries of what automated captioning systems can achieve.

Overall, the evolution of image captioning technology reflects broader trends in AI research, transitioning from simple, rule-based methods to sophisticated, deep learning-driven systems that leverage advanced neural architectures and multimodal integration.

#### 2.2-LITERATURE REVIEW

###### Vision Semantics Image Captioner

###### This paper introduces the completed project development of a cutting-edge Vision Semantics Image Captioner., a comprehensive platform aimed at generating contextually rich descriptions for images. Focused on leveraging advancements in vision semantics, our system utilises state-of-the-art techniques to provide detailed and meaningful captions for a wide array of images. The primary objective of this research is to enhance the understanding and interpretation of images through automated captioning, contributing to the fields of computer vision and natural language processing. The image captioner goes beyond basic descriptions by incorporating semantic nuances, ensuring that the generated captions capture the essence and context of the visual content. Our platform addresses the need for a reliable and efficient image captioning system by offering a user-friendly interface and incorporating robust algorithms. The implementation plan involved thorough literature surveys, enabling us to integrate the latest advancements in vision semantics into the image captioning process. In addition to the technical aspects, the paper discusses the practical applications of the Vision Semantics Image Captioner, emphasizing its potential in various domains such as accessibility, content indexing, and assistive technologies. The research also includes guidelines for optimising and customizing the system for specific use cases. By presenting a comprehensive solution for image captioning, this research contributes to the advancement of automated visual understanding. The platform's capabilities are demonstrated through a series of evaluations, showcasing its effectiveness in generating accurate and meaningful image captions. We believe that the Vision Semantics Image Captioner has the potential to revolutionize the way images are interpreted, providing valuable insights across diverse applications.

###### Remote Sensing Image Captioning using CNN and LSTM

###### Image captioning is a rapidly expanding field in computer vision research and it involves generating a comprehensive description for an input aerial image. The community has been paying more attention to it lately because it offers additional semantic information about images. To properly convey the connections between the objects and the features included in remote sensing pictures with their captions, this work proposes a system for captioning those images. The system is designed to generate and exploit textual descriptions. For captioning images, an encoder-decoder structure is employed. Convolutional neural networks (CNNs) are first used to encode the image’s visual properties. After that, the encoded features are sent to a language model, or decoder. For word by word descriptions of the image content, recurrent neural networks (RNNs), often referred to as long short term memory (LSTM), are commonly employed as language models. We combined the Bahdanau attention model with LSTM to enable learning to be focused on a specific region of the image to enhance performance. Four different pre-trained CNNs were evaluated for model performance in this study. Experimental results from UAVIC dataset are presented.

###### Image Captioning using Deep Learning

The process of generating a textual description for images is known as image captioning. Now a days it is one of the recent and growing research problem. Day by day various solutions are being introduced for solving the problem. Even though, many solutions are already available, a lot of attention is still required for getting better and precise results. So, we came up with the idea of developing an image captioning model using different combinations of Convolutional Neural Network architecture along with Long Short Term Memory in order to get better results. We have used three combination of CNN and LSTM for developing the model. The proposed model is trained with three Convolutional Neural Network architecture such as Inception-v3, Xception, ResNet50 for feature extraction from the image and Long ShortTerm Memory for generating the relevant captions. Among the three combinations of CNN and LSTM, the best combination is selected based on the accuracy of the model. The model is trained using the Flicker8k dataset.

# CHAPTER 3

# Problem Statement

#### Problem Statement

Despite significant advancements in computer vision and natural language processing, generating accurate, contextually rich, and coherent textual descriptions for images remains a challenging problem. The core challenge lies in bridging the gap between visual perception and natural language understanding. The primary objective is to develop a system capable of analyzing complex visual content from diverse images and translating that visual information into precise and contextually appropriate captions. This involves several sub-problems: effectively extracting and interpreting visual features from images, understanding the relationships and interactions between objects within the image, and generating grammatically correct and contextually relevant text that accurately represents the visual content.

This problem statement highlights the complexity of translating visual information into meaningful textual descriptions and underscores the need for advanced methodologies to improve captioning accuracy and contextual relevance.

# CHAPTER 4

# Experimental Setup

#### Experimental Setup

#### 4.1 Hardware Setup

#### In this description students have to write Configuration of computer system to run the software.

#### 4.2 Software Setup

#### All software tools along with packages are need to be included here.

# CHAPTER 5

# Proposed System & Implementation

#### Proposed system & Implementation

#### 5.1 Block diagram of proposed system

#### Students have to draw block diagram of proposed system. Flow chart can be added at this point.

#### 5.2 Description of block diagram

#### Explanation of block diagram.

#### 5.3 Implementation

#### Implementation of proposed system must be included here. Students can explain implementation using screen shots of output.

#### 5.4 Advantages/ Application/ result table can be included in this subsection.

# CHAPTER 6

# Conclusion

* + 1. Conclusion

Students have to include conclusion here. Future scope can be include in next subpoint.

# References

**Research paper**

[1] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 13, no. 7, pp. 688-703, July 2003, doi: 10.1109/TCSVT.2003.815168.

[2] Wang, Yao, Jörn Ostermann, and Ya-Qin Zhang. Video processing and communications. Vol. 1. Upper Saddle River, NJ: Prentice hall, 2002.

[3] https://mpeg.chiariglione.org/who-we-are

[4] S. E. C. Osman, H. Jantan, M. T. Miskon, and W. A. K. W. Chek, “A comparative study of video coding standard performance via local area network,” in International Conference on Soft Computing in Data Science. Springer, 2015, pp. 189–197.

[5] Akramullah, Shahriar. Digital video concepts, methods, and metrics: quality, compression, performance, and power trade-off analysis. Springer Nature, 2014.

[6] Sarwer, Mohammed Golam. "Efficient Motion Estimation and Mode Decision Algorithms for Advanced Video Coding." (2011).

[7] G. J. Sullivan, J. Ohm, W. Han and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 12, pp. 1649-1668, Dec. 2012, doi: 10.1109/TCSVT.2012.2221191.

[8] S. Alamelu Mangai, B. Ravi Sankar, and K. Alagarsamy, "Taylor Series Prediction of Time Series Data with Error Propagated by Artificial Neural Network", International Journal of Computer Applications (0975 – 8887), vol. 89 , no.1, March 2014.

[9] Störr, Hans-Peter, Y. Xu, and J. Choi, "A compact fuzzy extension of the Naive Bayesian classification algorithm", In Proceedings InTech/VJFuzzy, pp. 172-177. 2002.

**URL**

[10] <https://docs.phpmyadmin.net/en/latest/require.html#php>

[11] <https://www.apachefriends.org/docs/hosting-xampp-on-aws.html>

[12] <https://youtu.be/at19OmH2Bg4>