Image Caption Generator

Ashwini Pandit, Rohini Mane, Kirti Mane
Under The Guidance of Prof. Vikram Deokate
DEPARTMENT OF COMPUTER ENGINEERING, AI Ameen college of Engineering, Bhima Koregaon, Pune

Abstract
An Image Caption Generator is a sophisticated AI system that combines computer vision and natural language processing to automatically create descriptive textual captions for images. This technology utilizes deep learning, particularly Convolutional Neural Networks (CNNs), to analyze and extract meaningful visual features from the input image. These features capture details about the objects, scenes, and elements within the image. Subsequently, a natural language processing model, often built on Recurrent Neural Networks (RNNs) or Transformers, processes these visual features and generates coherent, contextually relevant captions. Post-processing steps may be applied to enhance the quality of the generated text. The primary aim of Image Caption Generators is to facilitate image understanding, improve accessibility, and enhance content search ability by providing human-readable descriptions for visual content. This technology is instrumental in various fields, including content tagging, accessibility tools for the visually impaired, and enhancing user experiences in multimedia content management systems, ultimately bridging the gap between visual and textual information for a more comprehensive and humanlike interpretation of images

CHAPTER1 INTRODUCTION

1.1 OVERVIEW

An Image Caption Generator harnessing the power of Convolutional Neural Networks (CNNs) represents a remarkable fusion of computer vision and natural language processing, propelling our capacity to decipher and describe visual content. CNNs, a cornerstone of this technology, have demonstrated extraordinary prowess in image analysis by enabling machines to perceive intricate details within images, recognize objects, and discern complex visual contexts. The synergy with advanced language models such as Recurrent Neural Networks (RNNs) or Transformers facilitates the generation of textual captions that are not only contextually accurate but also linguistically coherent. This dynamic pairing of CNNs and language models forms the backbone of Image Caption Generators, equipping them with the ability to provide vivid and contextually relevant descriptions for a wide range of images. Such systems have far-reaching implications, from assisting individuals with visual impairments to enhancing image retrieval and content management in digital repositories. As technology continues to evolve, the marriage of CNN algorithms and language generation is at the forefront of improving our interaction with the everexpanding visual content landscape.

1.1.1 Motivation

The motivation behind an Image Caption Generator using CNN algorithms stems from the need to efficiently understand and describe the ever-increasing volume of visual content on the internet. These generators enhance accessibility for visually impaired individuals, aid in content organization and retrieval, and improve humancomputer interaction by providing accurate, contextually relevant textual descriptions for images. This technology responds to the evolving landscape of digital content and strives for more inclusive, intuitive, and user-friendly interactions with visual data.

2

1.1.2 Problem Defination

The problem addressed by an Image Caption Generator using CNN algorithms is the challenge of automatically generating descriptive and contextually relevant textual captions for images. Given the vast and diverse array of visual content available, this technology aims to bridge the gap between the visual and textual domains, facilitating image understanding, accessibility for the visually impaired, efficient content management, and more intuitive human-computer interaction. By leveraging CNNs, these systems extract valuable visual features and combine them with advanced language models to create accurate and coherent image captions, addressing the need for a more comprehensive and accessible interpretation of visual data.

1.1.3 Objective

1.To automatically generate accurate and contextually relevant textual descriptions for a wide range of images.

2.To provide a more inclusive experience for visually impaired individuals by offering detailed image captions.

3.To assist in organizing and retrieving images in multimedia databases and digital repositories.

4.To create more intuitive and natural ways for users to interact with machines, enhancing human-computer communication in the context of visual data.

5.To facilitate a comprehensive interpretation of visual content, bridging the gap between the visual and textual domains.

CHAPTER2

LITERATURESURVEY

2.1 STUDY OF RESEARCH PAPER

1.Paper Name:Image Caption Generator

Author:Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, Vrinda Mathur Abstract :In the modern era, image captioning has become one of the most widely required tools. Moreover, there are inbuilt applications that generate and provide a caption for a certain image, all these things are done with the help of deep neural network models. The process of generating a description of an image is called image captioning. It requires recognizing the important objects, their attributes, and the relationships among the objects in an image. It generates syntactically and semantically correct sentences.In this paper, we present a deep learning model to describe images and generate captions using computer vision and machine translation. This paper aims to detect different objects found in an image, recognize the relationships between those objects and generate captions. The dataset used is Flickr8k and the programming language used was Python3, and an ML technique called Transfer Learning will be implemented with the help of the Xception model, to demonstrate the proposed experiment. This paper will also elaborate on the functions and structure of the various Neural networks involved. Generating image captions is an important aspect of Computer Vision and Natural language processing. Image caption generators can find applications in Image segmentation as used by Facebook and Google Photos, and even more so, its use can be extended to video frames.

They will easily automate the job of a person who has to interpret images. Not to mention it has immense scope in helping visually impaired people.

2.Paper Name:Multiple Perspective Caption Generation with Attention Mechanism Author:Hidekazu Yanagimoto,Maaki Shozu

Abstract :In caption generation, a caption generation system generates a caption, which describes the content of the image with natural language and needs to understand both an image and a text. So caption generation is an essential task in natural language processing and image processing. Many researchers recently pay attention to deep learning as a key technique to construct the caption generation system because deep learning can construct an intermediate representation, which is shared in both image processing and natural language processing. First, the system generates a feature from a given image with convolutional neural networks. Eventually, the system generates a word sequence, a caption, from the feature. It means that the system consists of two modules, the image processing module and the language model module and both of the modules are simultaneously trained with a training dataset. The deep learning based caption generation system is a blackbox system and it is difficult to collaborate with a human. So, we introduce attention mechanism in the caption generation system and we control caption generation by the attention weights. The usual caption generation systems can generate only a single caption from one image because the system can generate a caption from an image directly. The proposed system can generate some different captions form the same image because we can control the initial attention weights. Our results demonstrate how we collaborate with deep learning and the fact that the collaboration improves caption generation. Index Terms—multimodal learning, deep learning, caption generation, natural language processing, image processing

3.Paper Name:GENERATION OF VIEWED IMAGE CAPTIONS FROM HUMAN
BRAIN ACTIVITY VIA UNSUPERVISED TEXT LATENT SPACE
Author:Saya Takada , Ren Togo , Takahiro Ogawa and Miki Haseyama

Abstract:Generation of human cognitive contents based on the analysis of functional magnetic resonance imaging (fMRI) data has been actively researched. Cognitive contents such as viewed images can be estimated by analyzing the relationship between fMRI data and semantic information of viewed images. In this paper, we propose a new method generating captions for viewed images from human brain activity via a novel robust regression scheme. Unlike conventional generation methods using image feature representations, the proposed method makes use of more semantic text feature representations, which are more suitable for the caption generation. We construct a text latent space with unlabeled images not used for the training, and the fMRI data are regressed to the text latent space. Besides, we newly make use of unlabeled images not used for the training phase to improve caption generation performance. Finally, the proposed method can generate captions from the fMRI data measured while subjects are viewing images. Experimental results show that the proposed method enables accurate caption generation for viewed images.

4.Paper Name:Building A Voice Based Image Caption Generator with Deep Learning Author:Mohana priya R,Dr.Maria Anu,Divya S

Abstract:Image processing is used in various industries and it is remaining as one of the most advanced technologies used in Google, medical field etc. Recently, this technology has also attracted many programmers and developers due to its free and open source tool, which every developer can afford it. Image processing also helps in finding out lot of information from a single image since it is currently utilized as a primary method for collecting the information from image and processing it for some purpose and some operations will also be performed on the image. A voice based image caption generation is a task that involves the NLP (natural language processing) concept for understanding the description of an image. The combination of CNN and LSTM is considered as the best solution for this project; the main target of the proposed research work is to obtain the perfect caption for an image. After obtaining the description, it will be converted into text and the text into a voice. Image description is a best solution used for a visually impaired people who are unable to comprehend visuals. With the use of a voice based image caption generator, the descriptions can be obtained as a voice output, if their vision can't be resorted. In

future, image processing will emerge as a significant research topic, which will be primarily utilized to save human lives.
8

5.Paper Name:Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach
Author:N. Komal Kumar1 , D. Vigneswari2 , A. Mohan3 , K. Laxman4 , J. Yuvaraj5 Abstract:Image Caption Generator deals with generating captions for a given image. The semantic meaning in the image is captured and converted into a natural language. The capturing mechanism involves a tedious task that collaborates both image processing and computer vision. The mechanism must detect and establish relationships between objects, people, and animals. The aim of this paper is to detect, recognize and generate worthwhile captions for a given image using deep learning. Regional Object Detector (RODe) is used for the detection, recognition and generating captions. The proposed method focuses on deep learning to further improve upon the existing image caption generator system. Experiments are conducted on the Flickr 8k dataset using python language to demonstrate the proposed method.

CHAPTER3

SOFTWARE REQUIREMENT

SPECIFICATIONS

3.1 INTRODUCTION

3.1.1 Project Scope

The project scope of an Image Caption Generator using CNN algorithms encompasses the development of a software system that can automatically generate descriptive textual captions for images. This system will incorporate CNNs for image analysis, advanced language models for text generation, and may include features such as post-processing and user interfaces. The scope also covers the potential application areas, which could range from assisting visually impaired users to enhancing content management systems and improving user experiences with multimedia databases. Additionally, it involves defining the project's limitations, including the types of images it can handle and the accuracy of generated captions.

3.1.2 User Classes Characteristics

User Classes for an Image Caption Generator using CNN algorithms: 1.General Users:Everyday individuals seeking to add captions to personal photos or understand image content.
2.Content Creators:Bloggers, social media influencers, and photographers looking to enhance their visual content.
3.Researchers and Data Scientists:Professionals using the technology for academic, scientific, or research purposes.
4.Visually Impaired Users:Individuals with visual impairments relying on the generator for accessibility.
5.Content Managers and Librarians:Professionals responsible for organizing and managing image collections.
6.Developers and IT Professionals:Those integrating the generator into applications or systems.

7.Government and Public Sector:Government agencies applying the technology for accessibility and security purposes.
Characteristics of these users vary based on their specific needs, technical expertise, and application scenarios, ranging from personal use to professional and educational

contexts.

3.1.3 Assumptions and Dependencies

Assumption :Convolutional Neural Network is one of the main categories to do image classification and image recognition in neural networks. Scene labeling, objects detections, and face recognition, etc., are some of the areas where convolutional neural networks are widely used.CNN takes an image as input, which is classified and process under a certain category such as dog, cat, lion, tiger, etc. The computer sees an image as an array of pixels and depends on the resolution of the image. Based on image resolution, it will see as h * w * d, where h= height w= width and d= dimension. For example, An RGB image is 6 * 6 * 3 array of the matrix, and the grayscale image is 4 * 4 * 1 array of the matrix.In CNN, each input image will pass through a sequence of convolution layers along with pooling, fully connected layers, filters (Also known as kernels). After that, we will apply the Softmax function to classify an object with probabilistic values 0 and 1.
Dependencies: Python:

Python is an interpreted, high-level and general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects. Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), objectoriented and functional programming. Python is often described as a ”batteries included” language due to its comprehensive standard library. Python was created in the late 1980s, and first released in 1991, by Guido van Rossum as a successor to the ABC programming language. Python 2.0, released in 2000, introduced new features, such as list comprehensions, and a garbage collection system with reference counting, and was discontinued with version 2.7 in 2020.Python 3.0, released in 2008, was a major revision of the language that is not completely backward-compatible and much Python 2 code does not run unmodified on Python

3. With Python 2's endof-life (and pip having dropped support in 2021), only Python 3.6.x and later are supported, with older versions still supporting e.g. Windows 7 (and old installers not restricted to 64-bit Windows). Python interpreters are supported for mainstream operating systems and available for a few more (and in the past supported many more). A global community of programmers develops and maintains CPython, a free and open-source reference implementation. A non-profit organization, the Python Software Foundation, manages and directs resources for Python and CPython development. As of January 2021, Python ranks third in TIOBE's index of most popular programming languages, behind C and Java, having previously gained second place and their award for the most popularity gain for 2020.

3.2 EXTERNAL INTERFACE REQUIREMENT

3.2.1 User Interface

- Application Based On Image Caption Generator.

3.2.2 Hardware Interfaces:

- Hardware : intel i5

- Speed : 2.80 GHz

- RAM : 8GB

- HardDisk : 500 GB

- Key Board: Standard Windows Keyboard

3.2.3 Software Interfaces

- Operating System: Windows 10

- IDE: Spyder

- Programming Language : Python

3.3 NON FUNCTIONAL REQUIREMENT

3.3.1 Performance Requirements

- The performance of the functions and every module must be well. The overall performance of the software will enable the users to work decently. Performance of encryption of data should be fast. Performance of the providing virtual environment should be fast Safety Requirement

- The application is designed in modules where errors can be detected and steadily. This makes it easier to install and update new functionality if required.

3.3.2 Safety Requirement

- The application is designed in modules where errors can be detected and fixed easily. This makes it easier to install and update new functionality if required.

3.3.3 Software Quality Attributes

- Our software has many quality attribute that are given below:-

- Adaptability: This software is adaptable by all users.

- Availability: This software is freely available to all users. The availability of the software is easy for everyone.

- Maintainability: After the deployment of the project if any error occurs then it can be easily maintained by the software developer.

- Reliability: The performance of the software is better which will increase the reliabilityof the Software.

- User Friendliness: Since, the software is a GUI application; the output generated is much user friendly in its behavior.

- Integrity: Integrity refers to the extent to which access to software or data by unauthorized persons can be controlled.

- Security: Users are authenticated using many security phases so reliable security is provided.

- Testability: The software will be tested considering all the aspects.

3.4 SYSTEM REQUIREMENTS

3.4.1 Database Requirements

SQLite is one of the most popular and easy-to-use relational database systems. It possesses many features over other relational databases. Many big MNCs such as Adobe, use SQLite as the application file format for their Photoshop Lightroom product.SQLite is an embedded, server-less relational database management system. It is an in-memory open-source library with zero configuration and does not require any installation. Also, it is very convenient as it's less than 500kb in size, which is significantly lesser than other database management systems.

3.4.2 Software Requirements

Anaconda Navigator: Anaconda is an open-source distribution of the Python and R programming languages for data science that aims to simplify package management and deployment. Package versions in Anaconda are managed by the package management system, conda, which analyzes the current environment before executing an installation to avoid disrupting other frameworks and packages.The Anaconda distribution comes with over 250 packages automatically installed. Over 7500 additional open-source packages can be installed from PyPI as well as the conda package and virtual environment manager. It also includes a GUI (graphical user interface), Anaconda Navigator, as a graphical alternative to the command line interface. Anaconda Navigator is included in the Anaconda distribution, and allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages, install them in an environment, run the packages and update them.

3.4.3 Hardware Requirements

RAM : 8 GB

As we are using Machine Learning Algorithm and Various High Level Libraries Laptop RAM minimum required is 8 GB.
Hard Disk : 500 GB

Processor : Intel i5 Processor

IDE : Spyder

3.4.4 Analysis Models: SDLC Model to be applied

The software development cycle is a combination of different phases such as designing, implementing and deploying the project. These different phases of the software development model are described in this section. The SDLC model for the project development can be understood using the following figure The chosen SDLC model is the waterfall model which is easy to follow and fits bests for the implementation of this project.

Requirements Analysis: At this stage, the business requirements, definitions of use cases are studied and respective documentations are generated.

Design: In this stage, the designs of the data models will be defined and different data preparation and analysis will be carried out.

Implementation: The actual development of the model will be carried out in this stage. Based on the data model designs and requirements from previous stages, appropriate algorithms, mathematical models and design patterns will be used to develop the agent's back-end and front-end components.

Testing: The developed model based on the previous stages will be tested in this stage. Various validation tests will be carried out over the trained model.

Deployment: After the model is validated for its accuracy scores its ready to be deployed or used in simulated scenarios.

Maintenance: During the use of the developed solution various inputs/scenarios will been countered by the model which might affect the models overall accuracy. Or with passing time the model might not fit the new business requirements. Thus, the model must be maintained often to keep its desired state of operation.

3.4.5 System Implementation Plan

The System Implementation plan table, shows the overall schedule of tasks compilation and time duration required for each task.

Sr. No.	Name/Title	Start Date	End Date
1	Preliminary Survey		
2	Introduction and Problem Statement		
3	Literature Survey		
4	Project Statement		
5	Software Requirement And Specification		
6	System Design		
7	Partial Report Submission		
8	Architecture Design		
9	Implementation		
10	Deployement		
11	Testing		
12	Paper Publish		
13	Report Submission		

CHAPTER 4 SYSTEM ANALYSIS

4.1 SYSTEM ARCHITECTURE

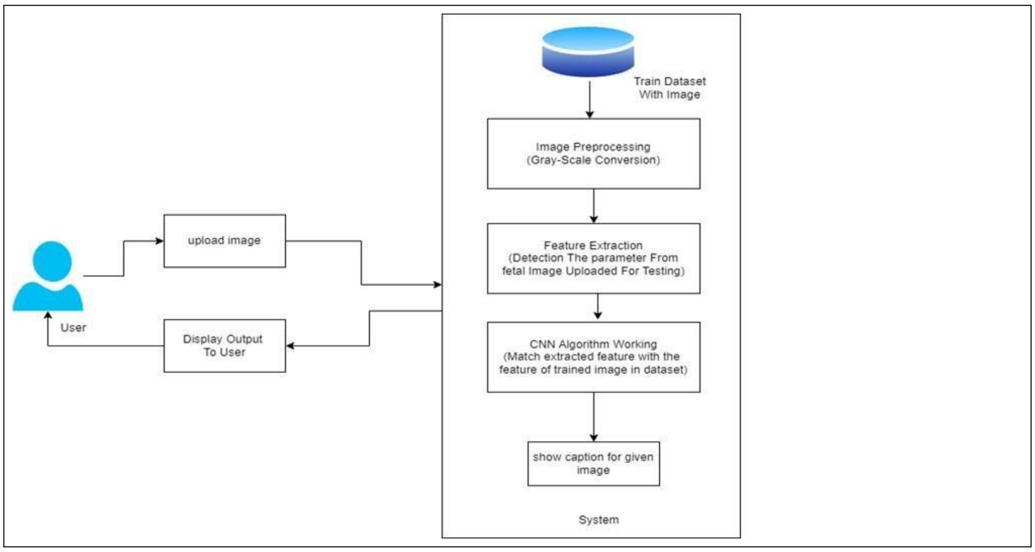

Figure 4.1: System Architecture

4.1.1 Data Flow Diagram

In Data Flow Diagram,we Show that flow of data in our system in DFD0 we show that base DFD in which rectangle present input as well as output and circle show our system,In DFD1 we show actual input and actual output of system input of our system is text or image and output is rumor detected like wise in DFD 2 we present operation of user as well as admin.

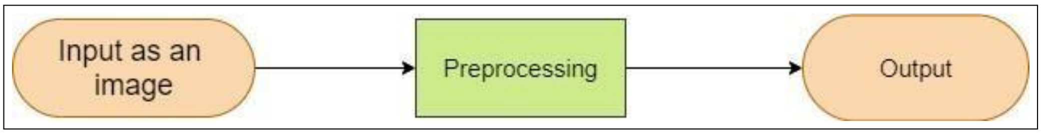

Figure 4.2: Data Flow diagram

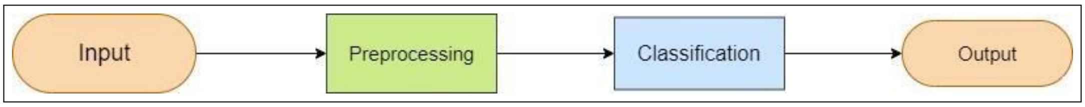

Figure 4.3: Data Flow diagram

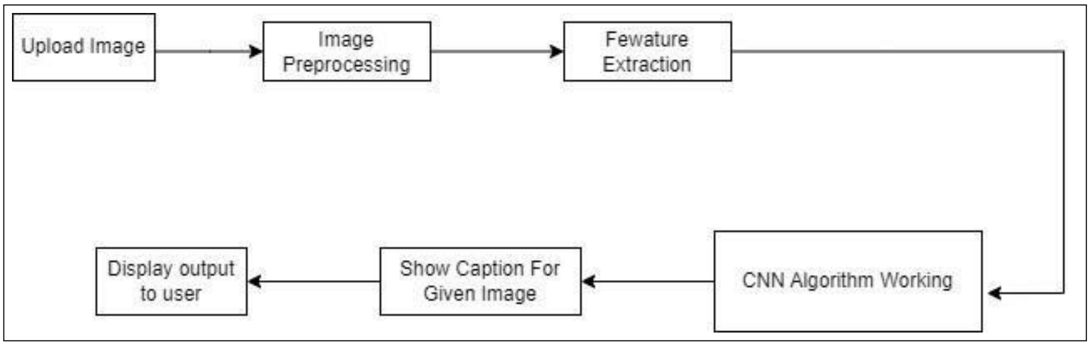

Figure 4.4: Data Flow diagram

4.2 UML DIAGRAMS

Unified Modeling Language is a standard language for writing software blueprints.The UML may be used to visualize,specify,construct and document the artifacts of a softwareintensive system.UML is process independent,although optimally it should be used in process that is use case driven,architecturecentric,iterative,and incremental.The Number of UML Diagram is available.

Use case Diagram.

Component Diagram.

Activity Diagram.

Sequence Diagram.

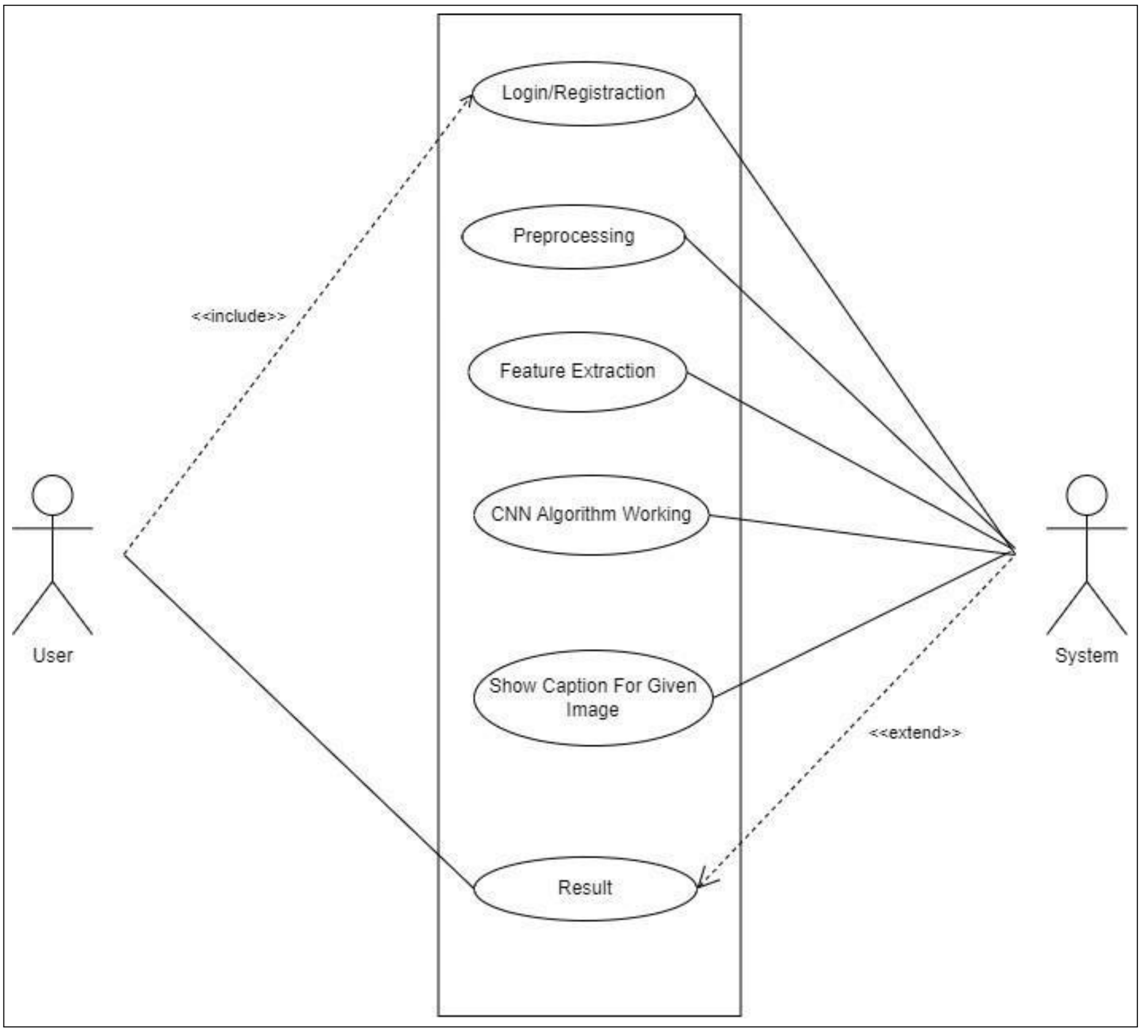

Figure 4.5: Use case Diagram

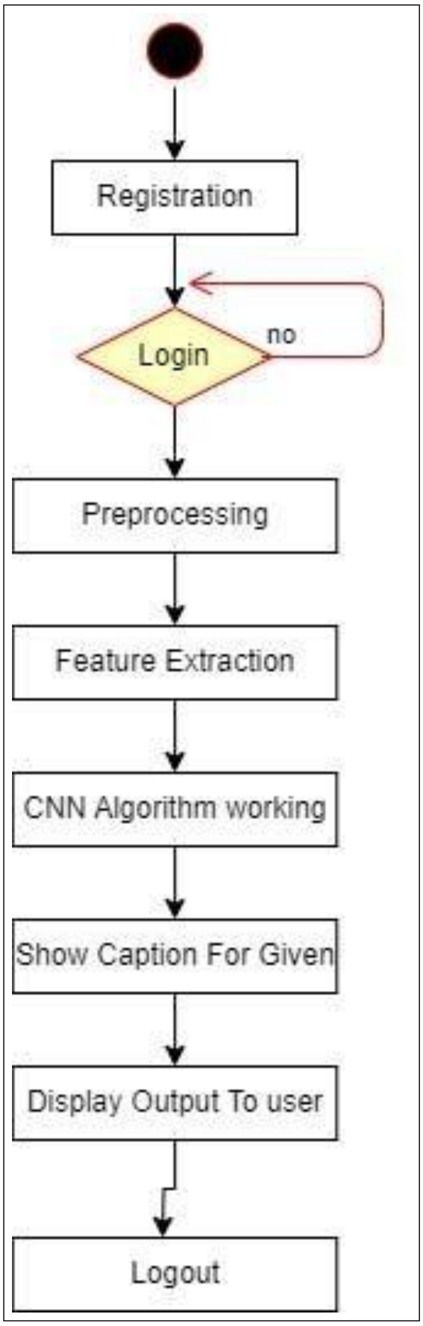

Figure 4.6: Activity Diagram

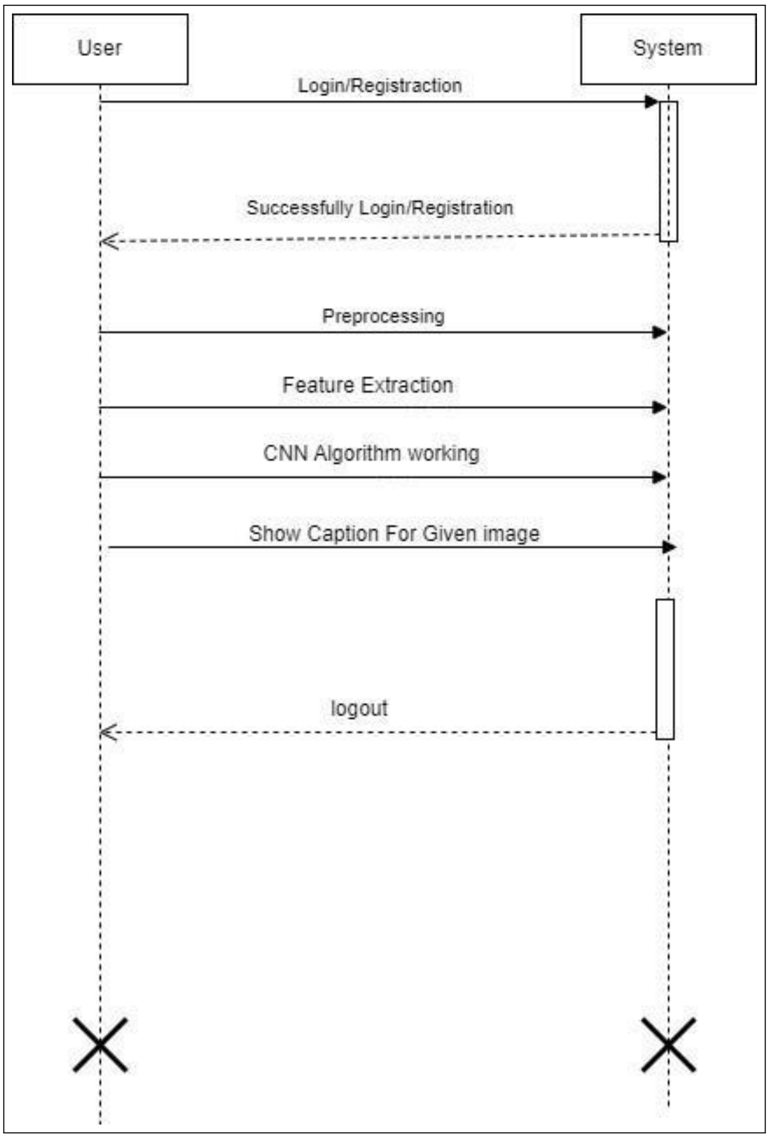

Figure 4.7: Sequence Diagram

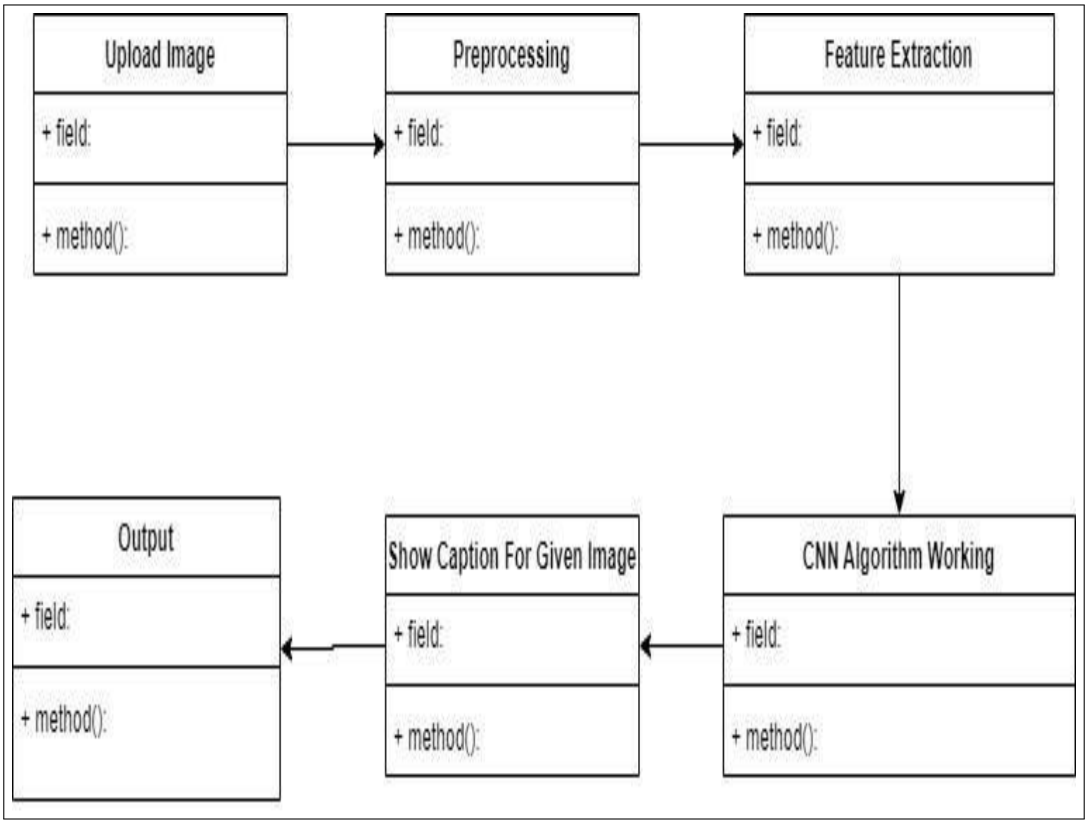

Figure 4.8: Class Diagram

CHAPTER 5

OTHER SPECIFICATION

5.1 ADVANTAGES

- Automated Image Description: Generates accurate and contextually relevant descriptions for images, saving time and effort.

- Enhanced Accessibility: Improves content accessibility for visually impaired users by providing detailed image captions.

- Efficient Content Management: Assists in organizing and retrieving images in multimedia databases, streamlining content management.

- Improved User Experience: Enhances user interaction by making content more understandable and relatable.

- Consistency and Scalability: Provides consistent and scalable image descriptions, useful for handling large volumes of images.

5.2 LIMITATIONS

- Accuracy and Ambiguity: Caption accuracy can vary, and the generator may struggle with interpreting ambiguous or complex images.

- Limited Understanding: The system may not grasp nuanced or abstract concepts in images.

- Overfitting: The model might perform well on specific datasets but struggle with generalization.

- Computational Demands: CNN algorithms can be computationally intensive, impacting real-time performance on some platforms.

- Vocabulary and Language: Limited to the vocabulary and language structures present in its training data, potentially resulting in unnatural-sounding cap- tions.

28

5.3 APPLICATIONS

1.Content Tagging

2.Accessibility Tools

3.Social Media and Blogging

4.E-commerce

5.Multimedia Content Management

6.Education

7.Security and Surveillance

CHAPTER 6

ALGORITHM

Convolutional Neural Network

Convolution layer: A convolution layer is a fundamental component of the CNN architecture that performs feature extraction, which typically consists of a combination of linear and nonlinear operations, i.e., convolution operation and activation function.

Nonlinear activation function: The outputs of a linear operation such as convolution are then passed through a nonlinear activation function.The most common nonlinear activation function used presently is the rectified linear unit (ReLU).

Pooling layer: A pooling layer provides a typical down sampling operation which reduces the in-plane dimensionality of the feature maps in order to introduce a translation invariance to small shifts and distortions, and decrease the number of subsequent learnable parameters.

Fully connected layer: The output feature maps of the final convolution or pooling layer is typically flattened, i.e., transformed into a onedimensional (1D) array of numbers (or vector), and connected to one or more fully connected layers, also known as dense layers, in which every input is connected to every output by a learnable weight. Once the features extracted by the convolution layers and down sampled by the pooling layers are created they are mapped by a subset of fully connected layers to the final outputs ofthe network, such as the probabilities for each class in classification tasks. The final fully connected layer typically has the same number of output nodes as the number of classes.

Last layer activation function: The activation function applied to the last fully connected layer is usually different from the others. An activation function applied to the multiclass classification task is a softmax function which normalizes output real values from the last fully connected layer to target class probabilities, where each value ranges between 0 and 1 and all values sum to 1

CHAPTER 7

7.1 CONCLUSION

In conclusion, the Image Caption Generator using CNN algorithms is a transformative application of artificial intelligence that bridges the visual-textual divide. By automatically generating accurate and contextually relevant captions for images, it enhances accessibility, content management, and user experiences across various domains. While it offers substantial advantages, it’s not without limitations, including accuracy challenges and potential biases. However, its versatility and potential for customization make it a powerful tool with wideranging applications, from improving content tagging and accessibility to advancing research and education, and enriching human-computer interaction.

7.2 FUTURE SCOPE

The future scope of Image Caption Generator technology is exceptionally promising, with several exciting developments on the horizon. Advancements in artificial intelligence and deep learning techniques are expected to significantly enhance the accuracy and precision of image captions, making them even more contextually relevant and reducing ambiguities. Multilingual capabilities are likely to become more widespread, enabling these generators to serve a global audience. Furthermore, the incorporation of emotion and sentiment analysis may enable the generation of captions that reflect the emotional content within images, opening up new avenues for creative and empathetic applications. Customization is another area with potential for growth, offering users more flexibility and user-friendly interfaces to tailor captions to their specific preferences. Additionally, as processing power continues to increase, we can anticipate real-time image captioning, making this technology even more responsive and versatile across numerous domains and industries.

CHAPTER 8

REFERENCES

- 1.HaoranWang , Yue Zhang, and Xiaosheng Yu, “An Overview of Image Caption Generation Methods”, (CIN-2020)

- 2. B.Krishnakumar, K.Kousalya, S.Gokul, R.Karthikeyan, and D.Kaviyarasu, “IMAGE CAPTION GENERATOR USING DEEP LEARNING”, (interna- tional Journal of Advanced Science and Technology- 2020)

- 3. MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga,

“A Comprehensive Survey of Deep Learning for Image Captioning” ,(ACM-2019)

- 4. Rehab Alahmadi, Chung Hyuk Park, and James Hahn, “Sequence-tosequence image caption generator”, (ICMV-2018)

- 5. Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan, “Show and Tell: A Neural Image Caption Generator”,(CVPR 1, 2- 2015)

- 6. Priyanka Kalena, Nishi Malde, Aromal Nair, Saurabh Parkar, and Grishma

Sharma, “Visual Image Caption Generator Using Deep Learning”, (ICAST2019)

- 7. Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra, and Nand Kumar Bansode,“Camera2Caption: A Real-Time Image Caption Generator”, International Conference on Computational Intelligence in Data Science(ICCIDS) - 2017

- 8. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, et al., ”Show, attend and tell: Neural image caption generation with visual attention”, Proceedings of the International Conference on Machine Learning (ICML), 2015.

- 9. J. Redmon, S. Divvala, Girshick and A. Farhadi, ”You only look once: Unified real-time object detection”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

- 10. D. Bahdanau, K. Cho, and Y. Bengio. “Neural machine translation by jointly learning to align and translate.arXiv:1409.0473”, 2014.