AN EMBRACIVE STUDY OF IMAGE CAPTION GENERATION USING PRE-TRAINED NEURAL NETWORKS

A PROJECT REPORT

Submitted by

NANDANA ANIL (TKM20MCA2024)

to

The APJ Abdul Kalam Technological University

in partial fulfilment of requirements for the award of degree of

MASTER OF COMPUTER APPLICATIONS

Thangal Kunju Musaliar College of Engineering
Kerala

DEPARTMENT OF COMPUTER APPLICATIONS

JULY 2022

DECLARATION

I hereby declare that the project report on **"An Embracive study of image caption generation using pre-trained neural network",** submitted for partial fulfillment of the requirements for the award of degree of Master of Computer Applications of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by me under supervision of **Prof. Jasmin M.R**. This submission represents my ideas in my own words and where ideas or words of others have been included, I have adequately and accurately cited and referenced the original sources. I also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. The report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Kollam Nandana Anil

12/07/22

DEPARTMENT OF COMPUTER APPLICATIONS
TKM COLLEGE OF ENGINEERING

C E R T I F I C A T E

This is to certify that, the project report entitled **"AN EMBRACIVE STUDY OF IMAGE CAPTION GENERATION USING PRE-TRAINED NEURAL NETWORKS"** submitted by **NANDANA ANIL (TKM20MCA2024)** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the degree of Master of Computer Applications, is a bonafide record of the project work carried out by her under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Internal Supervisor Head of the Department External Examiner

ACKNOWLEDGEMENT

First and foremost I thank GOD almighty and my parents for the success of this project. I owe sincere gratitude and heart full thanks to everyone who shared their precious time and knowledge for the successful completion of my project.

I am extremely grateful to **Dr. Fousia.M.Shamsudeen**, Head of the Department, for providing me with best facilities.

I would like to thank my project guide **Prof. Jasmin M.R**, Department of Computer Applications, who motivated me throughout the work of my project.

I profusely thank all other faculty members in the department and all other members of TKM College of Engineering, for their guidance and inspirations throughout my course of study.

I owe my thanks to my friends and all others who have directly or indirectly helped me in the successful completion of this project.

NANDANA ANIL

ABSTRACT

It's a difficult task to automatically generate natural language descriptions of an image's content. Though, unlike humans, it does not come readily to machines. However, implementing this capability would surely alter how machines interact with us. The recent advancement of object recognition from photos has resulted in a paradigm for captioning images based on their object relationships. Various picture caption producing models based on pre-trained neural networks are presented in this research, with an emphasis on the various CNN architecture and LSTM to examine their influence on phrase synthesis. For creating a caption from an image, a combination of neural networks is more suited. The quality of generated captions is calculated using BLEU Metrics.

CONTENTS

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Automatic image caption generation is a common activity in machine vision. Using machine vision and computational linguistics, AI technology instantaneously analyses the aspects of image. To appropriately caption visual context, deep learning is applied. Deep learning takes a considerable amount of power to solve more sophisticated and large assignments. As machine learning algorithms progress, it is critical to evaluate existing complex models. If at all possible, such procedures should be adjusted and used to address a portion of the entire problem. Despite the fact that this topic is fairly complex and extensive, many academics have worked on it over the last ten years due to significant progress in the field of machine learning, and they have offered several fairly successful solutions.

Deep learning and machine learning are two of the most advanced technologies accessible today. Artificial intelligence is currently being compared to human intelligence, with AI outperforming humans in some categories. New research in this topic is published every day. This field is quickly increasing since we now have the processing capability to fulfil this task. Deep learning is a subset of machine learning that employs multi-layered neural networks.

In this section, we will integrate picture and text processing to create a valuable Deep Learning application, known as Image Captioning. Image captioning is a process of creating written descriptions based on the items and events in an image. This method has numerous real-world applications. One such example is to save an image's captions so that they can be conveniently retrieved later. Basic knowledge of two Deep Learning approaches, including LSTM (Long Short Term Memory) and Convolutional Neural Networks (CNN) are necessary. The CNN model Xception is trained using an image net dataset. Xception is in charge of extracting picture features. These features are given into the LSTM model, which generates the image caption. Natural language processing has a subfield called text generation (NLP). It uses computational linguistics and artificial intelligence knowledge to generate natural language documents that can meet specific communicative requirements.

1.1 Problem Statement

Due to the long execution time, CNN algorithm is used for feature extraction. It fastens the execution time. CNN is great for extracting features from images and has been shown to be very effective at finding patterns that are difficult to detect with traditional methods.

Data preparation eliminates the possibility of imbalanced data. Preprocessing data makes it simpler to utilize and analyze. By deleting any inconsistencies or duplicates in the data, the accuracy of a model is improved. The co-occurrence of identical objects in a visual context is also eliminated. Preprocessing the data makes sure there aren't any erroneous or missing values brought on by bugs or human error.

The final output to be evaluated is the caption that was generated, and the quality of the captionis an important factor. Due to the less quality of generated caption, bleu metrics is used for caption evaluation. The caption quality has improved.

1.2 Objective

- The principle objective is to generate a meaningful caption for a picture.
- Preprocessing is accomplished to enhance the unbalanced dataset and put off inconsistencies.
- To extract features, a pre-trained CNN model is employed as it can obtain higher and rapid overall performance with much less labeled information.
- LSTM is used to generate photograph caption.
- Bleu metrics are used to calculate the best-generated caption.

CHAPTER 2

LITERATURE SURVEY

A literature overview is the complete test and interpretation of literature that relates to a selected situation depend. Even as one uses literature to compare research questions diagnosed, then one seeks to answer those research questions via looking for and analyzing relevant literature. Some importance of literature critiques is that new insights can be advanced via re-analyzing the consequences of the look. A literature evaluation is each a summary and explanation of the whole and present day-day united states of facts on a subject as found in academic books and magazine articles. There are forms of literature evaluations you could write at college: one which university students are requested to put in writing down as a stand-by myself venture in a course, and the opposite, this is written as a part of the introduction to, or schooling for, extended paintings, normally a thesis or studies report. One way to learn the differences between these types is to read online literature evaluations or the first chapters of theses and dissertations on your personal topic. Examines the structure of their arguments and how they address the issues.

2.1 Cause of the Literature Review

1. It gives readers easy admission to investigate a specific subject matter using deciding on excessive satisfactory articles or research which are applicable, meaningful, vital, and legitimate and combining them into a single comprehensive document.
2. It provides an excellent starting point for scholars beginning to do study in a new location by pushing them to compile, assess, and contrast actual research in that specific area.
3. It guarantees that investigators do not duplicate previously completed work.
4. It may provide hints as to where future study should go or indicate areas to be aware of.
5. It covers the most important facts.
6. Discrepancies, omissions, and ambiguities in the literature are identified.
7. It offers a fair assessment of other researchers' techniques and actions.

2.2 Related works

2.2.1 Automatic Caption era for annotated photos by way of the use of the clustering algorithm

The paintings make use of an annotation method to address the issue of pictures. The 2 methods employed in this situation are: manually annotating the photograph with the use of a human interface and storing the annotated snapshot in a repository. They're accumulated through computerized annotation utilizing a characteristic extraction and grouping method. SIFT algorithm is used for extracting functions from the picture. For characteristic extraction in this research, pre-knowledgeable CNN models were used. The final mechanism is clustering-based annotation studying. The conventional method for developing picture descriptions makes use of a level framework that includes floor reputation and content desire. The number one degree examines the image's content material, whilst the second degree chooses the manner to set up the diagnosed pics. Usually, every step is created with the resource of a hand. It makes an area of expertise of captioned images placed in online courses and teaches clients the way to annotate and categorize snapshots using numerous fashions. Throughout trying out, given an image with the embedded record and asked to write down a caption for that precise picture. The method learns about the snapshots, the captions that may be used to represent them, and the related documents that describe snapshots and the content material based totally on snapshots throughout the training section. The method is to write captions for the internet photos, in assessment to the majority who strive to develop photograph descriptions. It takes a long time as the SIFT set of rules is used for function extraction.

2.2.2 SD-RSIC: Summary-driven deep remote sensing image captioning

Experimental results on the current remote sensing dataset show that the proposed approach is effective compared to the modern approach. For remote sensing photos, a deep neural network is deployed to generate correct captions. Interpretation of remote sensing data is difficult. It is worth noting that the proposed method's performance can be enhanced further by implementing the attention strategy of selecting the most informative area in the remote sensing photograph. Any tactics that attract attention can be easily incorporated into the proposed approach. These measurements do not compare the real value of the created subtitle to the true value of the subtitle.

2.2.3 Automatic caption generation for medical images

In this article, the authors perform a detailed evaluation of the effectiveness of automated clinical photo annotation techniques, current fashion protection, scientific image annotation benchmark datasets, and generated annotations. Significant progress has been achieved in project development to automatically generate signatures, increasing the availability of scientific images in a variety of ways utilizing the most recent Pix processor, which is quite fast, accurate, and has improved computing power. It has become an essential approach to ending clinical photography, improving healthcare and making more impact with less. Similar to the technique used, this paper uses a generative mode based on deep neural networks. The opportunity approach is primarily search-based, selecting compelling captions for new images from a database of photo-caption pairs. The automatic annotation time of medical snapshots is the main topic of this study. The GRU is used for translating sentences. We will introduce our past efforts, statistics, evaluation methods, etc. Captions are still not prevalent in the clinical arena due to various aspects such as small pieces of information, text content types with very long captions containing many different facts, and professional nature. A metric is used to calculate the beauty of a name. Do not properly compare measurements with imbalanced records. It is difficult to narrow down the abnormal part.

2.2.4 Automatic generation of subtitles for news images

This method makes use of the internet's large photo library, as well as the fact that many of them are categorized and placed next to documents with comparable themes. The model trained on the labelled dataset is based on the assumption that images and textual information are generated by a common collection of latent variables. According to research, it is possible to create captions that are tailored to the content of an image and the article with which it is associated, while still allowing for creative expression in the description. For the visually challenged video retrieval technologies were used in this paper. However, the retrieval technique will be inaccurate. Captions are generated using a recurrent neural network.

2.2.5 Stimulus- and concept-driven analysis to generate captions

This work uses attention theory in the psychology of subtitle generation. There are two methods: stimulus-driven model with object recognition to recognize elements of a particular class and finds them using a bounding box, and Visual Question Answering (VQA), which uses a wide range of A-concept-driven embedded solutions. Distribute the images and project them into a shared semantic space. Annotations have made significant improvements in machine learning and cognitive computing recently. For high-level visual tasks, however, the language

was unable to deliver the expected outcomes. Providing good subtitles for complex multi-target scenarios is difficult. The recommended technique works perfectly on the MSCOCO test server.

2.2.6 Computer system for generating long video subtitles

In this work, for large video retrieval, a long video caption creation algorithm is introduced. Prior to caption synthesis, they employed STIPs to identify and remove extraneous frames, segment the video using a nonlinear combination of distinct visual signals, and finally choose the key video frames. A video description is built using the most important video frames created with the LSTM variation model in conjunction leveraging the attention mechanism. People's videos are frequently associated with significant events in their lives. However, as the big data era approaches, the amount of time required to retrieve and monitor data can be daunting. New methodologies for the application of extended video segmentation are given in this study, which can drastically shorten retrieval time. The improved spatiotemporal interest points (STIPs) identification technique detects the length of a long video's motion. The filtered long video is then super frame segmented to obtain an interesting piece of the long video. The STIP previously acquired on the video clips during the key frame selection process is used to form the region of interest, and the feature recognition of these regions of interest is used to screen out video key frames. Finally, attention vectors are added to the standard LSTM to generate video captions. The approach is assessed using the bleu and rouge on the Video Set dataset.

2.2.7 Deep learning based image caption generator

The author created a model for capping images in Hindi in this work. This is the first time subtitles in Hindi have been generated. Manually translated MSCOCO records from English to Hindi are used to construct the records. In addition, many sorts of attention-based designs have been developed for Hindi subtitles. These attention mechanisms are unique in Hindi; the proposed model's outcomes are compared to multiple BLEU score baselines, and the outcomes reveal that the suggested model surpasses the others. This method can usually generate semantically meaningful and grammatically correct captions by extracting knowledge from image and caption pairings. In the proposed methodology, the visual system attempts to describe the scene by recording a two-dimensional array image. It is an easy and rewarding task to use natural language to automatically explain the content of an image. It has the potential to have a huge influence. It can, for example, help visually impaired persons grasp the content of photos on the Internet.

2.2.8 From Show to Tell: Investigating Deep Learning-Based Captions

The goal of this work is to offer a thorough overview of the caption technique, including visual coding and text production, as well as training methods, data sources, and statistics. In this context, we analyse various relevant snipping approaches quantitatively in order to find the most significant improvements in design and staff development. It also outlines numerous problem variants and associated open challenges. To accommodate the demands and description styles of diverse users, several domain-specific ideas and job modifications were also tested. When reporting implicit and contextual information, go non-visual. While the latter is usually acknowledged as the caption task's purpose, this definition focuses on distinct features and includes descriptions at various levels of depth. One presents a basic overview of recently developed approaches, models, and task variants in order to chronicle how subtitling has been done thus far and to inspire new ideas.

The drawbacks found in the above-related works are filled up in this project named "An Embracive Study of Image Caption Generation using Pre-trained Neural Networks". The techniques used in the project are clearly described in the next chapter, Methodology. It specifies the whole working flow of the project.

CHAPTER 3

METHODOLOGY

3.1 Proposed System

It turned into as soon as the notion that computer systems couldn't manage visual representation, but breakthroughs in deep getting to know and massive amounts of facts get admission may be leveraged to expand models that may generate descriptions for pictures. The suggested system is a caption generator for images that provides useful captions. When an input image is given, a caption suitable for the image is generated. The features of an image are extracted using pre-trained neural networks. LSTM is used to produce captions. The quality of generated caption is evaluated using bleu metrics.

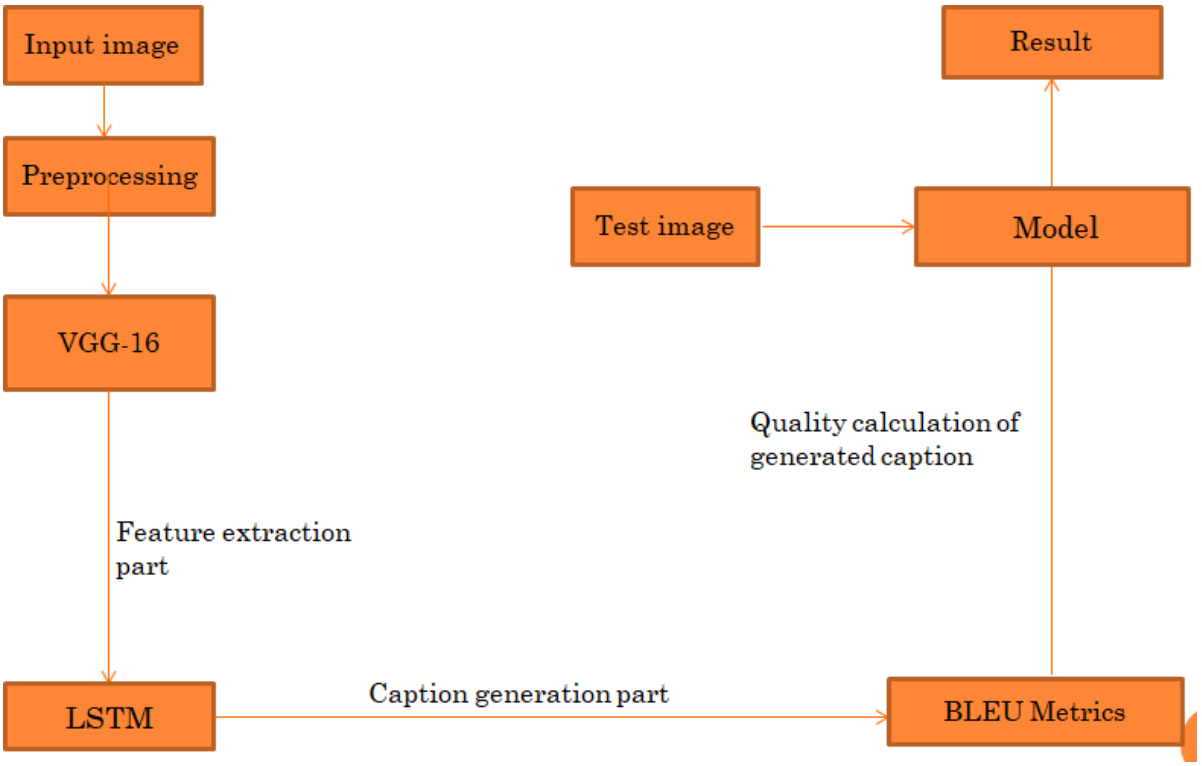

Fig.1: System Architecture

3.2 Used Dataset

The Flickr8k dataset was used for the project. There are two directories in the dataset.

- Flickr8k images: It contains 8092 jpeg images.
- Flickr8k text: A collection of files containing image captions from various sources.

3.3 Preprocessing

3.3.1 Image Preprocessing

Image preprocessing refers to processes done on images at their most basic degree of abstraction. If entropy is an information metric, these actions diminish rather than increase picture information content. Preprocessing is a technique for improving picture data by removing unwanted aberrations or enhancing key visual qualities that are important for further evaluation and processing tasks.

BGR to RGB conversion is the picture preprocessing technique used. Blue (255, 0), Green (0, 255, 0), and Red are abbreviated as BGR (0, 0, 255). When an image is opened in pattern recognition with cv2, the default colour image is BGR. Color-changing algorithms based on cv2 are also included. The shade of each pixel in RGB image format is generated by combining the red, green, and blue concentrations featured in each hue plane at the pixel's location. RGB images are recorded in graphics file formats as 24-bit images, with the red, green, and blue elements each being 8 bits. BGR image is read in imread format. The cvtColor(9) method transforms a BGR image to an RGB image and vice versa.

The purpose of converting an RGB image to a BGR image is to:

- The pixel orderings in many image processing frameworks differ
- Color complexity
- Learning image processing becomes easier
- Easier visualization

3.3.2 Caption Preprocessing

In the case of caption preprocessing, the raw text cannot be taken straight away and fit in Deep learning or machine learning models. To begin, tidy up the information by dividing it up into words and addressing punctuation and case sensitivity issues. Although computers cannot grasp English words, they express them with numbers and assign each word in the lexicon a unique index value. Each word should be encrypted into a remedied vector and expressed as a number. Then will the system be able to interpret the text and generate image captions.

The library used for data cleaning is Matplotlib. Matplotlib is a Python tool that allows you to generate static, animated, and interactive data visualizations. Matplotlib makes simple things simple and complex things possible. To increase the vocabulary size, the text should be cleansed. To do so, the data must first be loaded. Then, develop an image description dictionary. Punctuation is deleted, all text is converted to lowercase, and words containing digits are removed. All of the distinct words are separated. All of the descriptions are combined to form a vocabulary. A descriptions.txt file is produced to keep track of all the captions.

3.4 CNN

In various deep learning methodologies, data is utilised to train neural network algorithms to perform various machine learning tasks, such as classifying distinct object classes. Convolutional neural networks (CNNs) are deep learning algorithms that excel at picture analysis. Deep learning has several applications, such as image processing and natural language processing. It's also employed in medicine, media and entertainment, and self-driving cars. CNN is a highly effective image processing technique. These are the finest strategies for automated image processing right now. Many businesses employ these algorithms to recognize objects in photos. The image contains RGB combination data. Matplotlib is used to read a picture from a file and store it in memory. The image is invisible to the computer, which simply sees an array of numbers. Color photographs are stored in a 3D array. The image's height and width are the first two dimensions (in pixels). The final dimension corresponds to the red, green, and blue hues found in each layer. CNN is divided into three categories. Convolutional neural networks are used to perceive images and videos. CNNs are typically used for image analysis applications like detection, identification, and classification.

Convolutional neural networks comprise three layers:

- Convolutional Layer: In a convolutional neural network, each input neuron is routed to a subsequent hidden layer. Only a few neurons in CNN's input layer are linked to neurons in the hidden layer.
- Pooling Layer: The pooling layer reduces the dimensionality of the feature map. There will be multiple activation and pooling levels within CNN's hidden layer.
- Fully-Connected layer: The pooling layer reduces the dimensionality of the feature map. There will be several triggering and pooling layers within the CNN's hidden layer.

3.4.1 Pre-trained VGG-16 Architecture

VGG-16 is a deep convolutional neural network with 16 layers. It is possible to load a pre-trained version of the network that has been trained on over a million images from the ImageNet database. The network has been pre-trained to learn over 1000 different object classes, including keyboards, mice, pens, and other animals. As a result, the network has learned comprehensive feature representations for several photos. The network's image input size is 224 by 224. On ImageNet dataset, VGG16 was identified as the best performing model out of all configurations. Let's have a look at the real architecture of the arrangement.

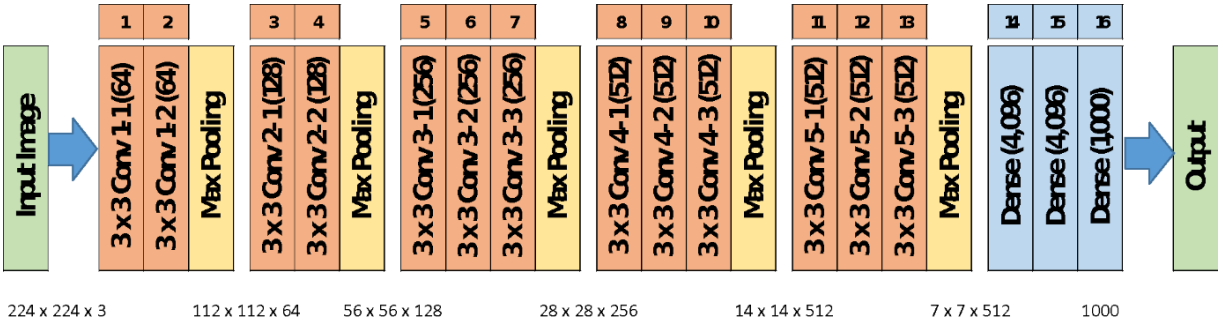

Fig.2: VGG-16 Architecture

Any network configuration regards the input as a 224 x 224 image with three channels (R, G, and B). The only preprocessing that is done is to standardise the RGB values of each pixel. This is accomplished by removing the pixel's average value. As shown in Fig. 1, convolutional layer stacks are followed by three totally linked layers separated by a smoothing layer. The first two layers each include 4,096 neurons, while the final fully connected layer acts as the output layer, with 1,000 neurons representing the ImageNet dataset's 1,000 possible classifications. The image is passed through the first stack of two convolution layers with 3 3 receptive sizes using post-rectified linear activation functions. Each of these two layers has 64 filters. Convolution stride and padding are both set to one pixel. The spatial resolution is preserved in this setting, and the scale of the resulting activation map is equal to the parameters of the input image. After that, the activation maps are spatially pooled across a 2 x 2-pixel window with a stride of 2 pixels. As a result, the size of activation is cut in half. The first stack has 112 x 64 activations at the bottom. The activations are then routed through a second stack, which is similar to the first but with 128 filters instead of 64. As a result, the final size after the second stack is 56 x 56 x 128.

3.5 LSTM (Long Short Term Memory)

The long short-term memory (LSTM) is a deep learning and AI artificial neural network. LSTM has feedback connections, as opposed to standard feedforward neural networks. This type of recurrent neural network can examine large data sequences as well as single input points (such as photos) (such as speech or video).

The preceding step's output is sent into the current RNN stage. LSTM was developed by Hochreiter and Schmidhuber. It addressed the issue of RNN long-term reliance, which occurs when an RNN is unable to predict words stored in long-term memory but can make more accurate predictions based on current input. As the gap length increases, RNN becomes inefficient. LSTM can save information for an extended period of time by default. It is used to predict and classify time series data. Captions are created with it.

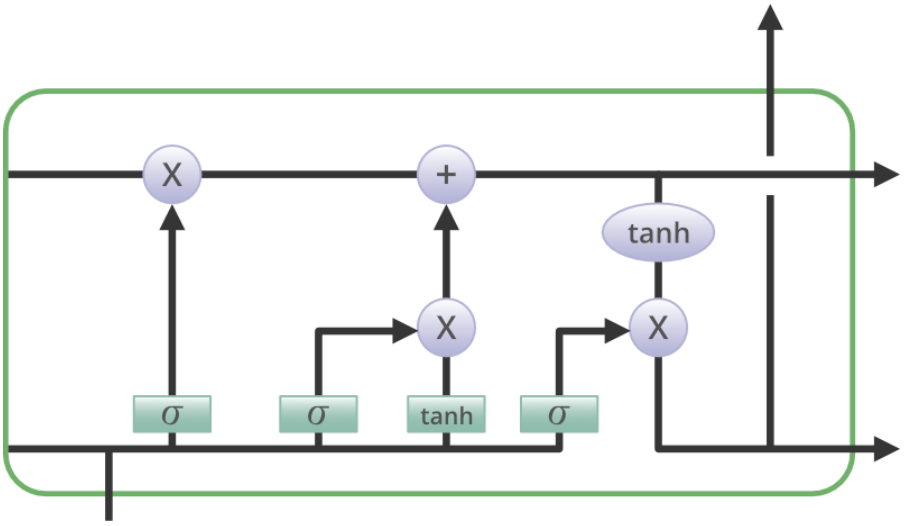

Fig.3: Structure of LSTM

As illustrated in Fig. 2, The LSTM is composed of four neural networks and many memory units known as cells. The cells store information, but the gates change it. The LSTM is composed of four neural networks and many memory units known as cells. The cells store information, but the gates change it. The LSTM is composed of four neural networks and many memory units known as cells. The cells store information, but the gates change it. The LSTM is composed of four neural networks and many memory units known as cells. The cells store information, but the gates change it. Cell state is a word used to describe long-term memory. The looping arrows show that the cell is recursive. This permits past interval information to be kept within the LSTM cell. Forget gate, which is placed beneath the cell state, modifies it, and the input modulation gate adjusts it.

There are three ways in—

1. Forget gate: The forget gate purges the cell of unneeded information. Prior to biasing, two inputs are sent into the gate and multiplied by weight matrices, x t (at-the-time input) and h t-1 (prior cell output). As shown in Fig. 3, the result is processed via an activation function, which produces a binary output. If the output for a particular cell state is zero, the data is lost; if it is one, the data is preserved for later use.

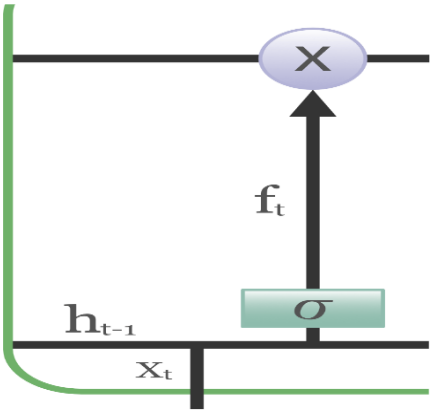

Fig.4: Forget gate

2. Input gate: The input gate is in charge of adding important information to the cell state. To begin, the sigmoid function is used to regulate the information, and the values to be remembered are filtered using inputs h t1 and x t in a way similar to the forget gate. The vector is then built using the tan h function, which returns a value between -1 and +1 and contains all of the potential h t-1 and x t values. Finally, the vector and controlled values are merged to yield useable information.

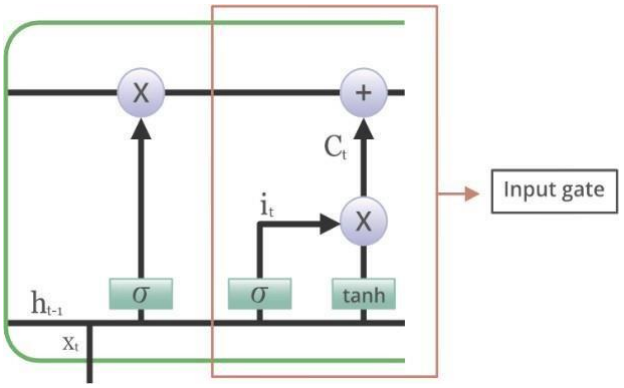

Fig.5: Input gate

3. Output gate: The output gate gathers critical information from the current cell state and outputs it. To begin, the cell uses the tanh function to produce a vector. The data is then filtered by the values to be remembered via the h t-1 and x t inputs, and it is governed by the sigmoid function. Finally, the vector and controlled values are multiplied and provided as output and input to the next cell, as shown in Fig. 5.

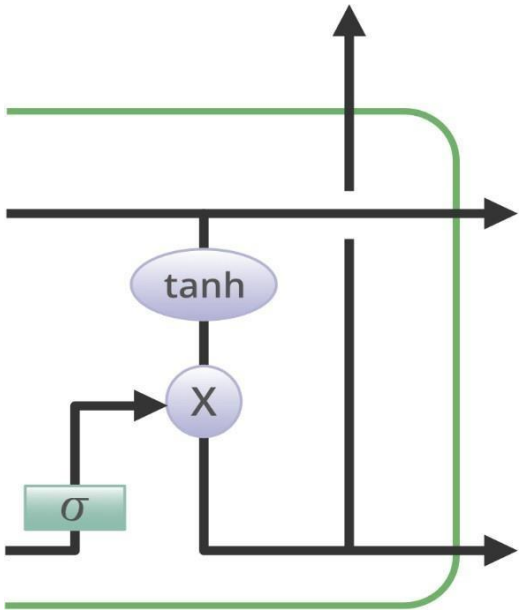

Fig.6: Output gate

3.6 BLEU METRICS

BLEU (bilingual evaluation understudy) evaluates machine-translated text between two natural languages. According to the primary concept underlying BLEU, the quality of a machine's output is supposed to be related to that of a human: "the better a machine translation is, the closer it is to a competent human translation." BLEU was one of the first measures to show a strong relationship with human quality assessments, and it is still one of the most commonly used automated and low-cost metrics. Individual chunks of translated text, typically sentences, are rated by comparing their translations to a set of high-quality reference translations. These ratings are then summed throughout entire corpus to establish the overall translation quality.

An exact match will score 1.0 and an exact match will score 0.0. BLEU and BLEU-derived metrics are most commonly used for machine translation. A score was developed to assess the predictions of automated machine translation systems. It's not perfect, but it has five compelling benefits:

- It can be calculated quickly and inexpensively.
- Easy to understand.
- Language independent.
- Human ratings have a high association.
- Widely adopted.

BLEU has been widely reported to correlate well with human judgment and remains a benchmark for assessing new metrics. However, some criticisms have been expressed. BLEU can, in principle, evaluate translations in any language, but knows that it cannot process a language without word boundaries in its current form. While BLEU has significant benefits, it is claimed that there is no guarantee that an increase in BLEU scores will indicate an improvement in translation quality. BLEU compares generations' n-grams to references' n-grams.

3.7 TRAINING THE MODEL

The dataset uses 6000 images for training purpose. First, load the CNN model's extracted features. A dictionary containing the captions for each image in the image list is created. Then the vocabulary is tokenized. This matches the tokenizer given the descriptive text for the loaded photo. Next, give each word in the dictionary a unique index value. The Keras package includes routines for generating tokens from vocabulary and generating tokens. The data will be created. Provide the model with training inputs and outputs to turn this into supervised learning activity. The model is trained with 6000 images, each image contains feature vector of 4096 lengths, and the label for the image is also displayed numerically. This massive amount of data produced by 6000 images cannot be held in memory, so a stack-generating generator method is used. The generator provides input and output sequences. In this model, the training process is performed when the input image is taken from the dataset. The image is trained to extract features. Once the image features are extracted, the LSTM is used to generate corresponding image captions.

3.8 TESTING THE MODEL

The model has undergone training. Designers apply the same tokenizer PKL file to obtain the words from the index values because the prognostication contains the maximum length of index values. During testing, the process of generating subtitles is mainly carried out. When an image is loaded from a dataset, an appropriate caption is created for that image. The data set uses 1000 images for testing. These images are tested on each testing process. After testing, the best model is obtained on a single input.

3.8 SOFTWARE REQUIREMENTS AND SPECIFICATION

The software used for the project:

- Python
- Google Colaboratory
- Django

1. Python

Guido Rossum created Python, an object-oriented programming language, in 1989. It is great for sophisticated application fast prototyping. It may be adapted to C or C ++ and supports a variety of operating system functions and libraries. Several companies, such as NASA, Google, YouTube, and Bit Torrent, use the Python programming language. Python programming is widely used in advanced areas like AI, natural language production, neural networks, and other computer sciences. Python is a complex artificial language developed by Guido van Rossum in the late 1980s and now controlled by the Python Software Foundation. It derives from his ABC language, which he helped create early in his career. Python is a complex programming language that may be used to create games, graphical user interfaces (GUIs), and online applications. It is a very advanced language. Reading and writing Python scripts are similar to reading and writing standard English statements. As a result, they are not written in a computer language and must be processed by Python code before being executed by a system. Python is a simple language. This implies that when the program runs, the interpreter evaluates the code and converts it into bytecode that is machine-readable. Python is an object-oriented programming language that teaches users how to manage and manipulate data structures or objects in order to build and execute programs. Python has it all. Languages die and become obsolete when they fail to meet expectations and are replaced by more capable languages. Python is a programming language that has proven and continues to be useful not just for industry and business, but also for programmers and individual users. It is a vibrant, thriving, and extremely useful programming language that is highly regarded as a significant programming language.

2. Google Colaboratory

Google Colab was created to provide free access to GPUs and TPUs to anyone who needs them for creating machine learning or deep learning models. Google Colab is a more advanced version of the Jupyter notebook. Jupyter Notebook is a browser extension or integrated environment for development. Notebooks are used instead of files. Notebook documents can include executable code, as well as text, photos, diagrams, tables, charts, equations, and a variety of other graphic data. To put it simply, a notebook document is a method of creating a human-readable executable document. Cells are building blocks in notebooks. A notebook is built up entirely of cells. An Execute button to the left of the code cell can be used to execute the contents of the cell. The output is presented below the cell after the cell has been performed.

Google Colab includes several fascinating features that current IDEs do not. The Following are some of the most exciting aspects:

- Through interactive tutorials, one can learn about machine learning and neural networks.
- Create and run Python 3 scripts without the need for a local setup.
- Data can be imported from a third-party source, such as Kaggle.
- Notebooks should be saved to Google Drive.
- Google drive notebooks can be imported.

3. Django

Django is a Python-based web framework that adheres to the MTV (Model-Template-Views) design philosophy. Django's main goal is to make it simple to create complex database-driven websites. The framework demonstrates component reuse and "plugability," code reduction, weak coupling, rapid development, and non-repetition. Python is used throughout the application, including configuration, files, and data models. Django additionally includes a management interface for options creation, reading, updating, and deletion. Contemplation generates these dynamically and assembles them using a management model.

Notwithstanding its terminology [7], such as referring to variable objects that generate HTTP responses as "views." The core Django framework is an MVC design [8]. It is a type of object-relational mapper (ORM) that functions as a bridge between a data model (provided as a Python class) and a relational database (the "model"), a system that processes HTTP request using webpage authoring framework (the "view"), and grep ("controller").

The fundamental framework additionally includes the following:

- A tiny, self-contained web server for both testing and development
- A tool that can convert between HTML forms and database-stored values for input, print version, and verification.
- A template system based on inheritance through entity coding
- A cache system that allows for the use of multiple caching techniques
- Middleware classes that can perform custom functions at various levels of request processing are supported.
- An internal dispatcher system that permits event communication between application components to one another through the use of predefined signals
- A Django model instance encoding mechanism that is capable of creating and reading XML and/or JSON representations
- A system for enhancing the template engine’s functionality

CHAPTER 4

RESULT AND DISCUSSION

Many deep learning-based algorithms for generating automatic captions have been proposed. To generate captions, an approach based on encoder/decoder architecture employs a simple CNN and a text generator. Methods based on attention-based image annotation focus on distinct prominent areas of the image and outperform methods based on coder/decoder architecture. To generate semantically rich captions, image captioning algorithms based on semantic ideas selectively focus on different aspects of an image. The visuals are trained during training. Training entails feature extraction as well as label development. The Blue Metric is used to calculate the quality of the created subtitles.

4.1 Performance and Evaluation

How to evaluate the results of creative style transfer is a difficult problem because people's perspectives on the matter vary greatly and are difficult to quantify. In this project, the quality of the generated caption is calculated using Bleu Metrics. Through bleu metrics, bleu score is obtained as shown in Fig.6. Quality is defined as the correlation between machine and human performance. "The better computer translation is, the closer it gets to professional human translation." -This is his BLEU fundamental premise. BLEU was one of the initial indicators, and it is now being used to demonstrate a strong link with human quality judgment, low-cost indicators.

```
227/227 [==============================] - 610s 3s/step - loss: 5.2408
227/227 [==============================] - 617s 3s/step - loss: 4.0278
227/227 [==============================] - 613s 3s/step - loss: 3.5940
227/227 [==============================] - 680s 3s/step - loss: 3.3255
227/227 [==============================] - 612s 3s/step - loss: 3.1263
227/227 [==============================] - 603s 3s/step - loss: 2.9778
227/227 [==============================] - 630s 3s/step - loss: 2.8627
227/227 [==============================] - 676s 3s/step - loss: 2.7662
227/227 [==============================] - 1457s 6s/step - loss: 2.6831
227/227 [==============================] - 601s 3s/step - loss: 2.6094
227/227 [==============================] - 598s 3s/step - loss: 2.5468
227/227 [==============================] - 579s 3s/step - loss: 2.4929
227/227 [==============================] - 619s 3s/step - loss: 2.4458
227/227 [==============================] - 610s 3s/step - loss: 2.3987
227/227 [==============================] - 639s 3s/step - loss: 2.3600
227/227 [==============================] - 619s 3s/step - loss: 2.3199
227/227 [==============================] - 609s 3s/step - loss: 2.2851
227/227 [==============================] - 604s 3s/step - loss: 2.2548
227/227 [==============================] - 613s 3s/step - loss: 2.2251
227/227 [==============================] - 610s 3s/step - loss: 2.1952
BLEU-1: 0.538977
```

Fig.7 Bleu score obtained

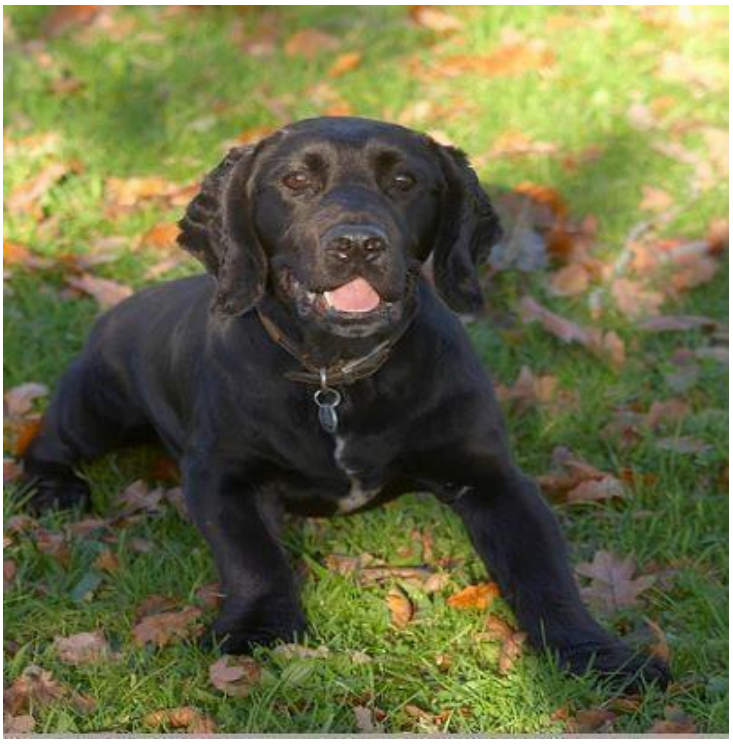

Fig.8 Input image

```
----------------------Actual----------------------
startseq black dog endseq
startseq black dog lays in the grass and looks towards the camera endseq
startseq black dog lays on grassy lawn covered in leaves endseq
startseq black puppy wearing black collar sits on the grass and leaves endseq
startseq black dog lays on grass endseq
--------------------Predicted--------------------
startseq black dog lays on the grass endseq
[02/Jul/2022 15:29:36] "POST / HTTP/1.1" 200 3213
[02/Jul/2022 15:29:36] "GET /media/2059616165_b7c99c1009.jpg HTTP/1.1" 200 151247
```

Fig.9 Output of image – Generated caption

CHAPTER 5
CONCLUSION

In this project, one presented a picture caption generator based on pre-trained neural networks. Captions are naturally difficult to create since they combine machine vision and natural language processing difficulties. The pre-trained convolutional neural network used is VGG-16. To sum up, the test's findings are clear: For many picture identification tasks, a pre-trained model is advantageous for a variety of reasons. The first justification is that utilizing a pre- trained model saves time and effort by requiring less training and model architecture development. To extract features from images, VGG-16 is employed, while LSTM is used to generate image annotations. A mapping of both image and caption is shown in this project. Figure 8 depicts the input image. The output, as seen in Figure 8, is a caption for the input image. The blue metric is used to determine the quality of the generated subtitles, a bleu score of 0.53897 is obtained for the model as shown in Fig.6. A bleu score of 0.6 or 0.7 is the best that can be achieved.

5.1 Future Enhancement

- Supervised learning requires a lot of data with labels for training. As a result, supervised and reinforcement learning will become increasingly prevalent in captions in the future.
- The caption of real time videos can be generated.
- Other architecture can be implemented in addition to VGG-16.

REFERENCES

[1] A. S. Reddy, N. Monolisa, M. Nathiya, and D. Anjugam, "Automatic caption generation for annotated images by using clustering algorithm," in Proceedings of the International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), pp. 1–5, IEEE, India, March 2015.

[2] Gencer Sumbul, Sonali Nayak, "SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning", in Proceedings of the journal on Geoscience and Remote Sensing, IEEE, Vol. 59, No. 8, August 2021.

[3] I. Allaouzi, M. Ben Ahmed, B. Benamrou, and M. Ouardouz, "Automatic caption generation for medical images," in Proceedings of the 3rd International Conference on Smart City Applications (SCA'18), 3rd International Conference on Smart 10 Mathematical Problems in Engineering City Applications (SCA'), pp. 1–6, Tetouan, Morocco, October 2018.

[4] Y. Feng and M. Lapata, "Automatic caption generation for news images," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, pp. 797–812, 2012.

[5] S. Ding, S. Qu, Y. Xi, and S. Wan, "Stimulus-driven and concept-driven analysis for image caption generation," vol. 398, pp. 520–530.

[6] S. Ding, S. Qu, Y. Xi, and S. Wan, "A long video caption generation algorithm for big video data retrieval," Future Generation Computer Systems, vol. 93, pp. 583–595, 2019.

[7] Pranay Mathur, Aman Gill, Aayush Yadav, "Deep learning based image caption generator", International Conference on Computational Intelligence in Data Science(ICCIDS), 2017.

[8] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara, "From Show to Tell: A Survey on Deep Learning-based Image Captioning", : : DOI 10.1109/TPAMI.2022.3148210, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[9] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep Captioning with Multimodal Recurrent Neural Networks (mRNN)," in ICLR, 2015.

[10] S. Bai and S. An, "A survey on automatic image caption generation," Neurocomputing, vol. 311, pp. 291–304, 2018.

APPENDICES

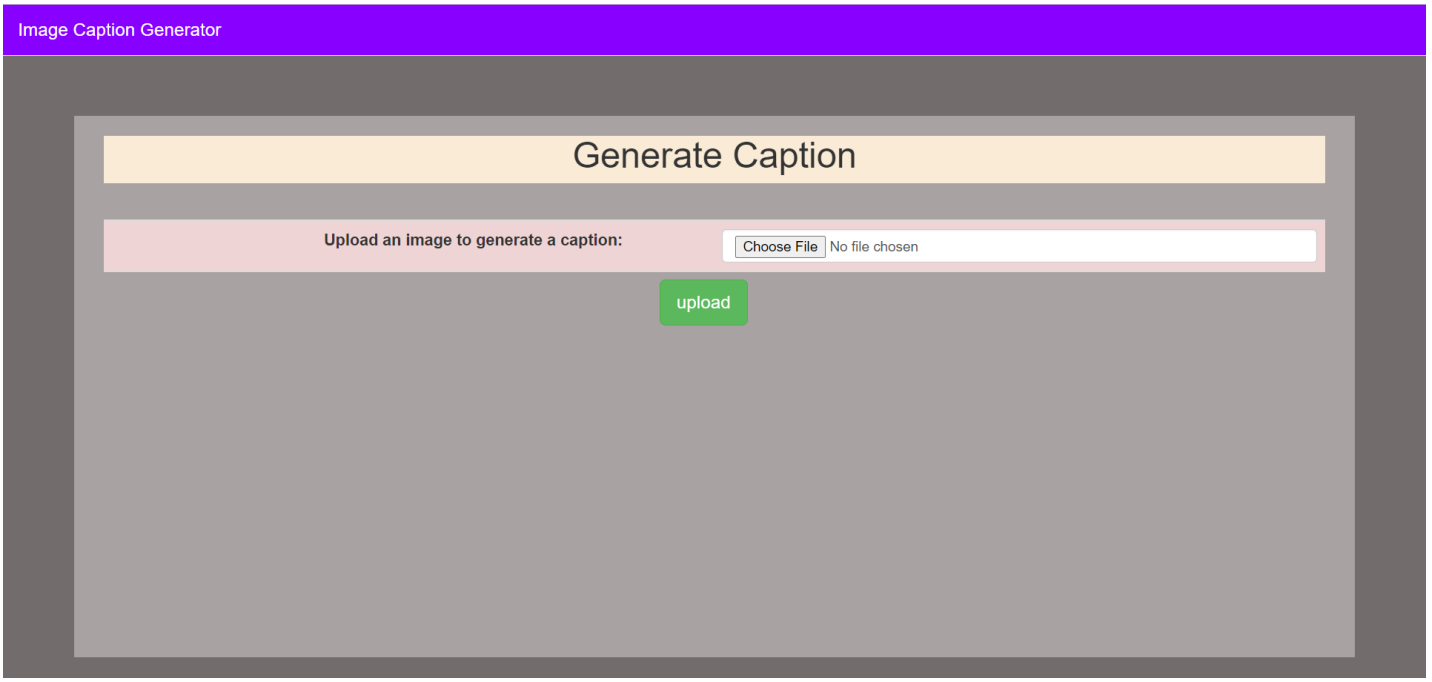
Image Caption Generator
Generate Caption
Upload an image to generate a caption:
Choose File
No file chosen
upload

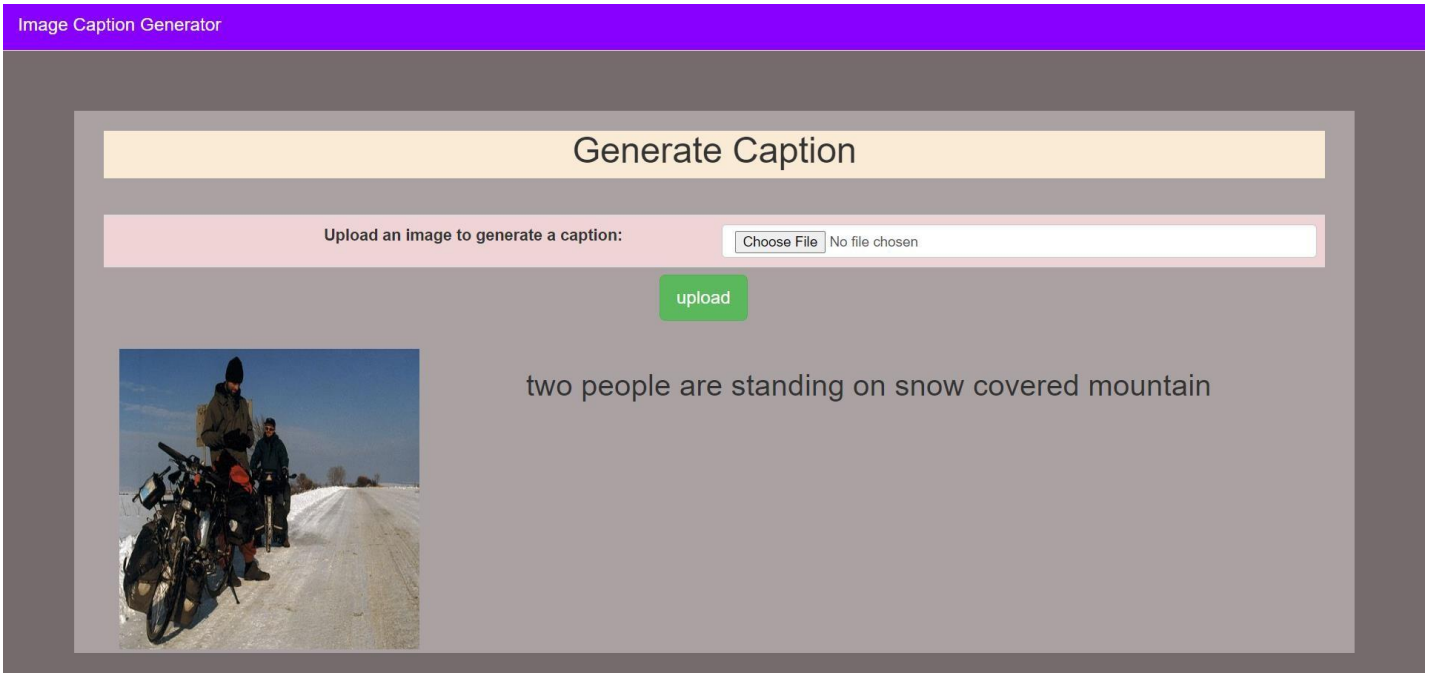
Image Caption Generator
Generate Caption
Upload an image to generate a caption:
Choose File
No file chosen
upload
two people are standing on snow covered mountain

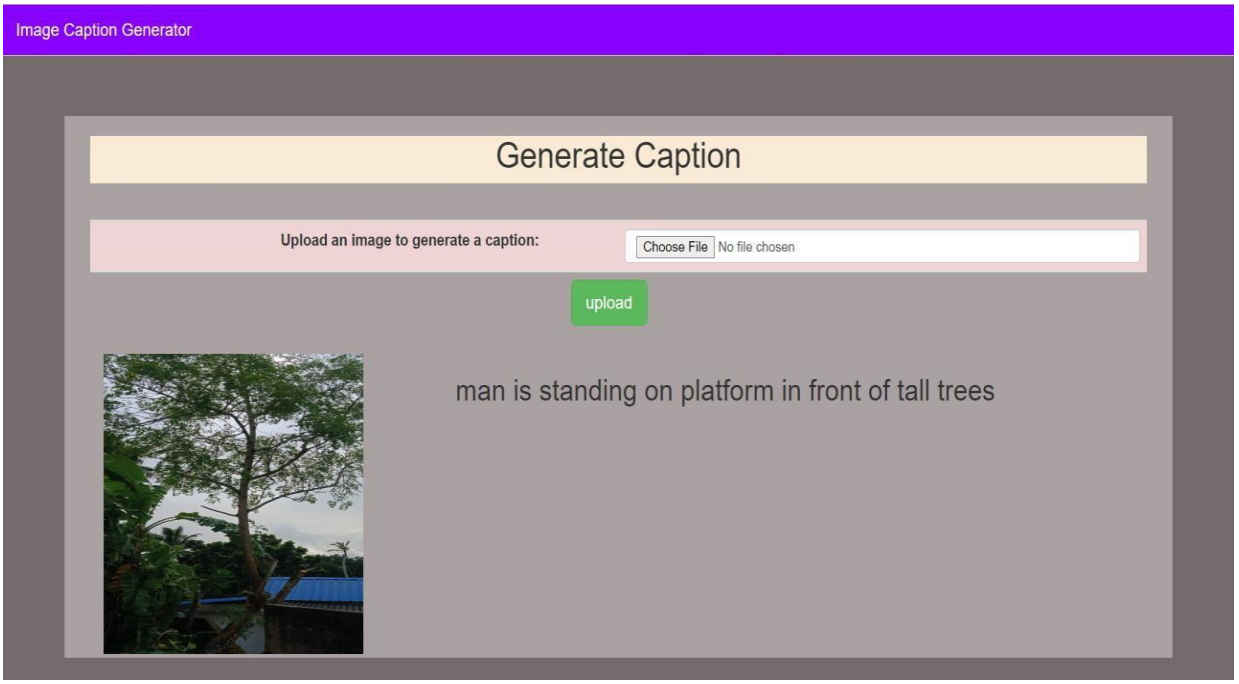
Image Caption Generator
Generate Caption
Upload an image to generate a caption:
Choose File
No file chosen
upload
man is standing on platform in front of tall trees

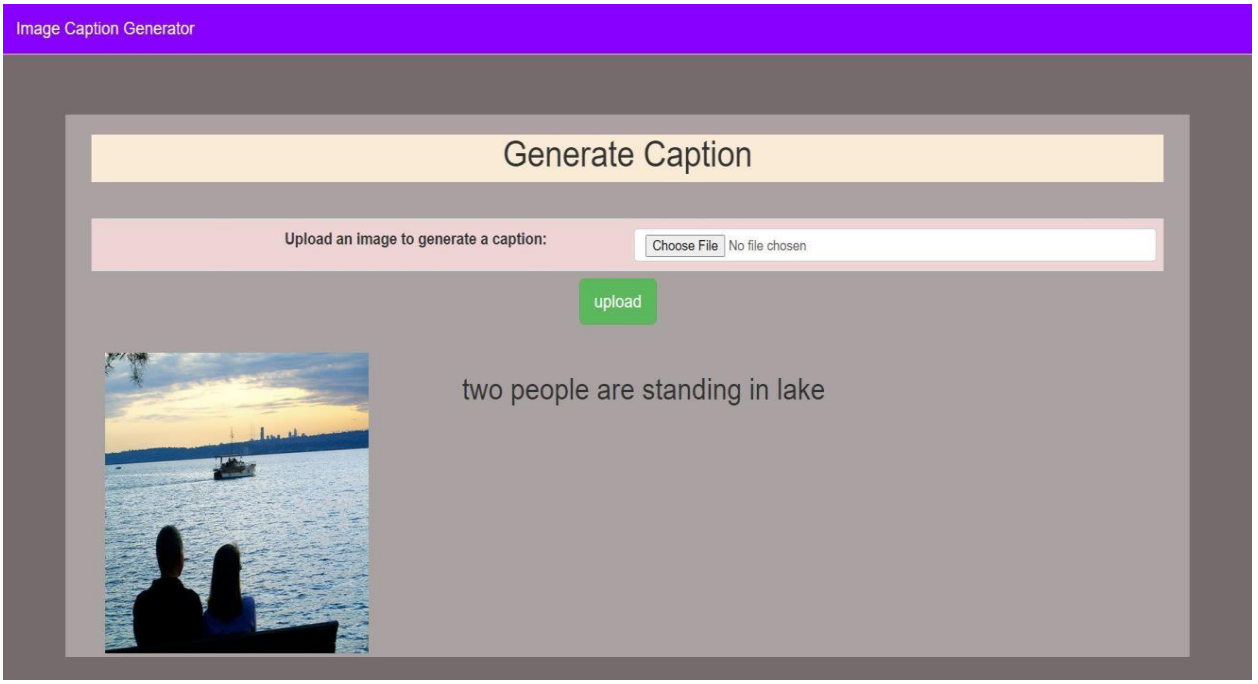
Image Caption Generator
Generate Caption
Upload an image to generate a caption:
Choose File
No file chosen
upload
two people are standing in lake

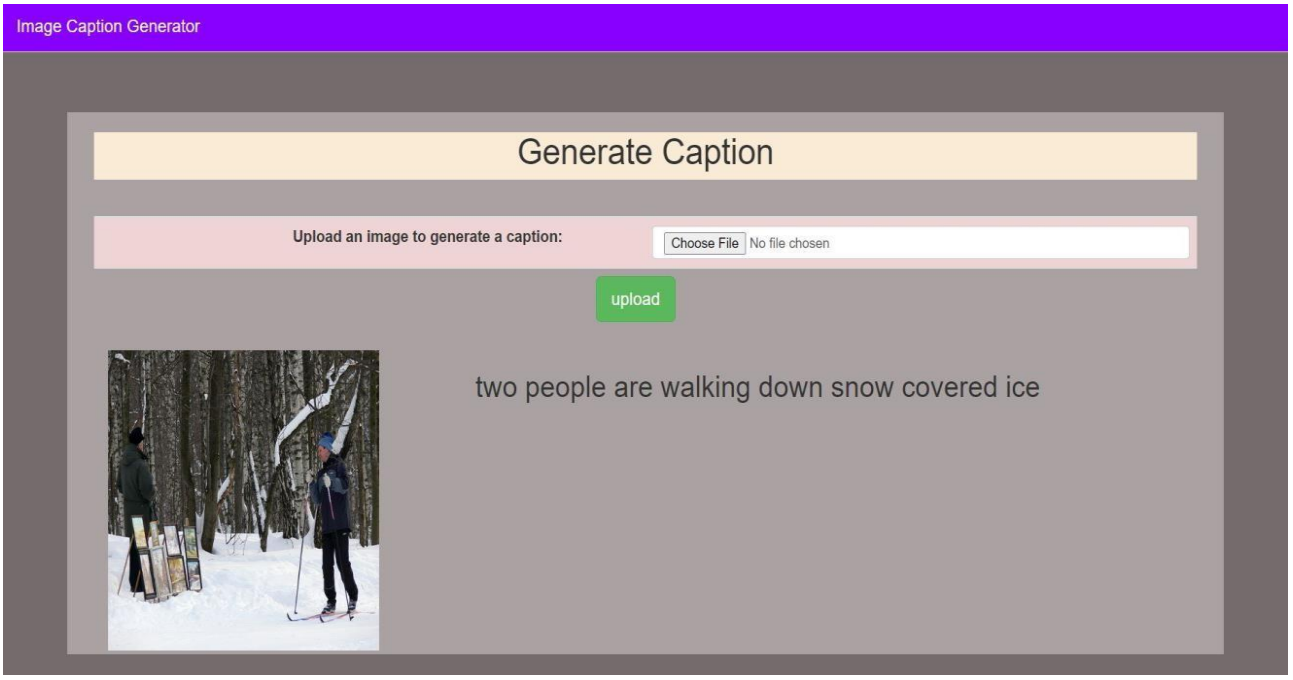
Image Caption Generator
Generate Caption
Upload an image to generate a caption:
Choose File
No file chosen
upload
two people are walking down snow covered ice

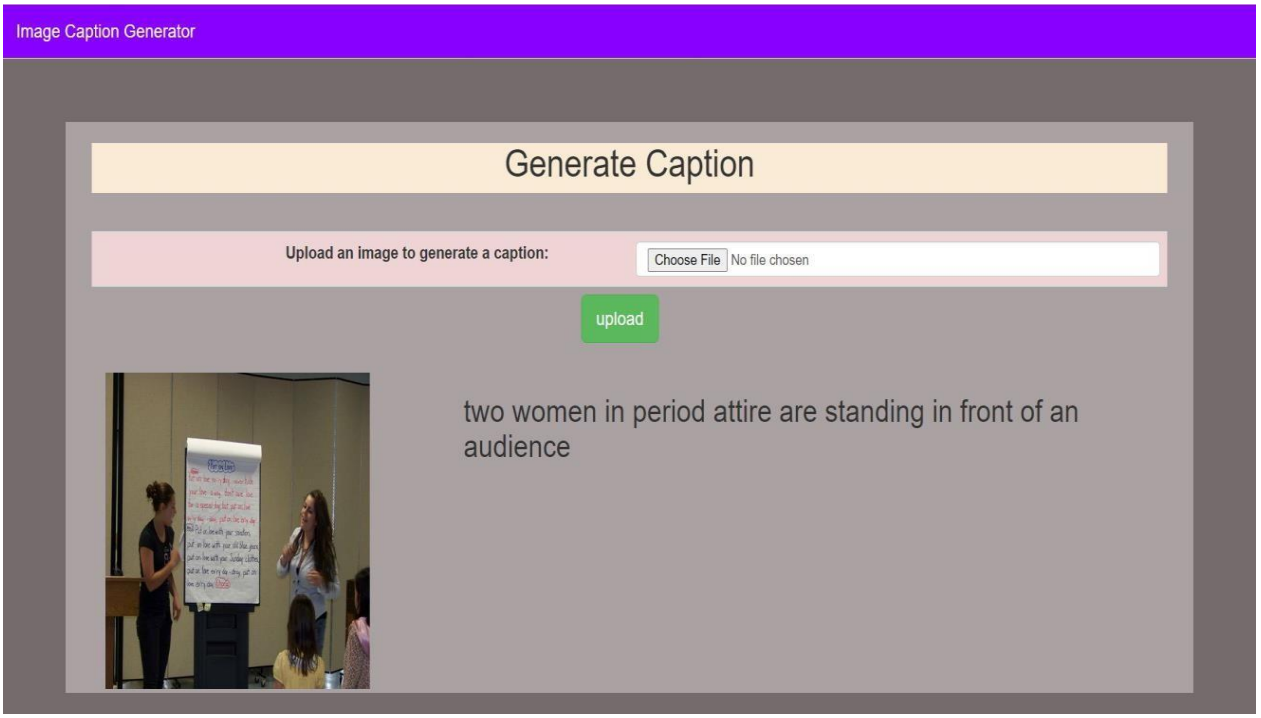
Image Caption Generator
Generate Caption
Upload an image to generate a caption:
Choose File
No file chosen
upload
two women in period attire are standing in front of an audience

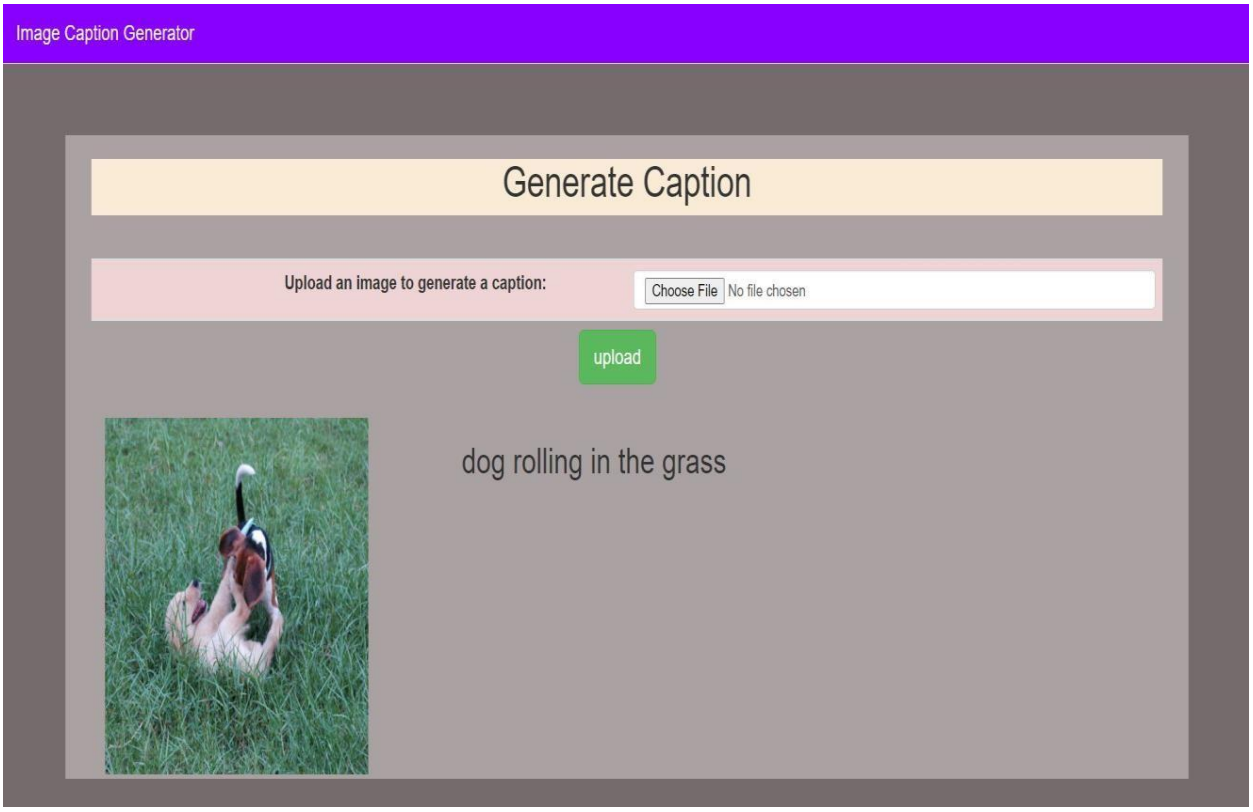
Image Caption Generator
Generate Caption
Upload an image to generate a caption:
Choose File
No file chosen
upload
dog rolling in the grass