[1]Harshal J

[2]Nithish TU

[3]Chandana

[4]Girish Y P

[5]Dr. Sivagamasundari G

Image Captioning Using Dl and Matlab

Abstract*:* - This project seeks to create a system that produces captions for photos using MATLAB and Deep Learning methodologies, particularly Convolutional Neural Networks (CNN). The produced captions are then transformed into audio via a text-to-speech converter. The system also recognizes the primary subject in the picture and then plays an appropriate audio file (e.g., a barking sound for a dog) after the narration of the caption. This novel method improves the accessibility and engagement of visual material via the use of audio input, resulting in a more immersive experience. picture captioning refers to the process of generating descriptions for the content shown in a picture. Image captioning is used to give descriptions that contextualize the images. The examination of vast amounts of unlabeled photographs, the identification of obscure patterns for machine learning applications in autonomous vehicles, and the development of software that assists the visually impaired are only a few instances of the many domains where image captioning proves to be really advantageous. Deep learning models are applicable for picture captioning. Advancements in deep learning and natural language processing have facilitated the generation of descriptions for supplied images. This article will use neural networks for image captioning.

Keywords*:* examination, Advancements, advantageous

1. Introduction

The capacity to automatically produce meaningful captions for photographs has become a crucial tool for a variety of applications in this age of digital media. Some of these applications include accessibility for the visually impaired, content management, and social media. Through the use of Deep Learning methods, in particular CNNs, the objective of this research is to bridge the gap between the representation of visual material and the representation of aural content. For the purpose of giving an auditory explanation of the picture, the incorporation of text-to-speech conversion is a further enhancement that further increases the functionality of the system. An extra feature that allows the system to play a relevant audio track that corresponds to the key player in the picture adds a depth that is both interactive and interesting to the system.

The process of captioning images used to be a challenging one, and the captions that are generated for a picture are not always particularly helpful even when they are made. Thanks to the development of text processing methods such as Natural Language Processing and Neural Networks, many jobs that were challenging and difficult to complete using Machine Learning have become easy to complete with the assistance of Deep Learning and Neural Networks. This is because of the progress that has been made in both of these areas. These are very useful in a wide variety of applications of artificial intelligence, such as image identification, image classification, image captioning, and a great deal of other applications. In its most basic form, the act of producing descriptions of what is happening in the input picture is included in the process of captioning images. From a fundamental standpoint, this model receives input in the form of photographs and produces output in the form of a caption. Alongside the development of new technologies comes an increase in the effectiveness of the process of making picture captions.Image captioning is very helpful for a wide range of applications, including the ever-increasingly popular

[1] (Pes University)
email: harshalj4623@gmail.com
(Pes University)
email: nithishu15@gmail.com
(Pes University)
email: Chandanav971@gmail.com
(Pes University)
email: girishyp610@gmail.com
Project Guide:
(Associate Professor, Pes University)
email: sivagamasundari@pes.edu

self-driving automobiles. Several applications of machine learning, such as recommendation systems, may benefit from the usage of image captioning. Object recognition models, deep learning-based captioning pictures, and visual attention-based captioning images are just some of the many approaches that have been developed for the purpose of captioning photographs since their introduction.

Problem Statement

Understanding a picture's visual information and producing a logical, contextually appropriate description is a difficult undertaking known as automatic image captioning. Although current solutions concentrate on creating subtitles, there aren't many that provide audio feedback that corresponds to the main character in the picture. By creating a system that not only creates subtitles but also plays a matching sound for the main actor in the picture, our project seeks to close this gap.

Objectives

- To develop a system that generates captions for images using CNN and MATLAB.

- To convert the generated caption text into an audio format using a text-to-speech converter.

- To identify

Hardware and Software Requirements

Hardware:

- A computer with sufficient processing power and memory to run MATLAB and Deep Learning

algorithms.

- Audio output device (speakers or headphones).

Software:

- MATLAB with the Deep Learning Toolbox.

- Pre-trained CNN model (e.g., VGG16, ResNet) for image feature extraction.

- Text-to-Speech (TTS) converter library or API.

- Audio files corresponding to the major actors (e.g., barking sound for a dog).

2. Literature Survey

The literature survey is the key phase in the software development process. Prior to tool development, it is essential to assess the temporal factor, economic considerations, and organizational capacity. Upon satisfying these criteria, the subsequent stage is to ascertain the appropriate operating system and programming language for the development of the tool. Once the programmers begin the development of the tool, they need substantial external help. This help may be obtained from senior programmers, books, or online. Prior to constructing the system, the aforementioned concerns are considered in the development of the proposed system.
The primary segment of the project development sector assesses and thoroughly evaluates all necessary requirements for project advancement. A literature study is the most critical component of the software development process for any project. Prior to the development of tools and their corresponding design, it is essential to assess the time factor, resource requirements, personnel, financial considerations, and organizational capacity.
To enhance and customize the user experience of its products, photographs use picture categorization. Intraclass variation, occlusion, deformation, size variation, viewpoint variation, and illumination are common challenges in computer vision shown by the image classification problem. Techniques effective for image classification are anticipated to be as effective for other significant computer vision tasks such as detection, localization, and segmentation.
Image captioning is a wonderful demonstration of this. Given a picture, the image captioning task is to construct a sentence description of the image. The picture captioning challenge parallels the image categorization problem in that it demands more depth and encompasses a broader range of options. Image classification functions as a

black box system in contemporary picture captioning systems; hence, enhanced image classification results in improved captions. The picture captioning issue is compelling since it integrates two major domains of artificial intelligence: computer vision and natural language processing. An image captioning system demonstrates its comprehension of both picture semantics and plain language. Image Captioning Through Deep Learning Techniques This study by Dr. P. Srinivasa Rao, Thipireddy Pavankumar, Raghu Mukkera, Gopu Hruthik Kiran, and Velisala Hariprasad investigates the use of CNNs, RNNs, and LSTM models for the generation of picture captions by the decomposition of images and words into their constituent components. The LSTM approach exhibited superior efficiency, with 80% accuracy as measured by BLEU metrics on the Visual Genome role-caption database.

Image Captioning Through Deep Learning Mr. N. Raghu, Sai Srikar, Aaftaab, Ruthvik Sai - This study tackles the issue of picture captioning via the integration of computer vision, natural language processing, and machine learning methodologies. A model using convolutional neural networks for feature extraction and recurrent neural networks with an attention mechanism for caption generation was assessed on the Flicker8k database, producing promising and competitive results.

Generating Image Captions Based On Deep Neural Networks T.Sandhya, Dr.Kondapalli Venkata Ramana - This study employs Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to construct an image caption generator capable of explaining pictures in English. The study examines essential principles and approaches in picture captioning, using technologies such as Keras, NumPy, and Jupyter Notebooks.

3. Methodology

1. Image Captioning Using CNN:

- Utilize a pre-trained CNN model in MATLAB to extract features from the input image.

- Use these features to generate a caption using a language model.

2. Text-to-Speech Conversion:

- Convert the generated caption into audio using a TTS converter.

3. Identifying the Major Actor:

- Analyze the generated caption to identify the major actor in the image.

4. Playing Corresponding Audio:

- Play a pre-recorded audio file that corresponds to the identified major actor in the image.

5. Integration and Testing:

- Integrate all components into a unified system and test the system using various images.

We propose this model in this research in order to improve both the efficiency and the accuracy of the caption creation for the picture. This is something that we are doing in order to ease the difficulty that is associated with gradient descent.

A. Collection Of DataSet

Multiple datasets are available for training a deep learning model to generate captions for images, including ImageNet, COCO, FLICKR 8K, and FLICKR 30K. The Flickr8K dataset is used for our model training. The Image Caption Generating Deep Learning Model may be effectively trained with the Flickr8k Dataset. The Flickr8K dataset comprises 8019 pictures, about 80 percent of which are allocated for training a deep learning model, while the other 20 percent are used for model development and evaluation. The text data collection contains five captions for each picture, each elucidating an activity occurring inside the image.

B. Image Prepocessing

In order to provide the photos as input to the ResNet, we must first preprocess them after importing the data sets. We must scale each image to the same size, which is 224X224X3, because we cannot feed various sized photos

through the Con- volution layer like ResNet. Additionally, we are transforming the photos to RGB using the cv2 library's built-in capabilities.

C. Text Prepocessing

After the captions have been uploaded to the Flickr data set, they need to be initialized in a manner that prevents any confusion or problems when the deep learning model builds its vocabulary from the captions. Please remove any numerical information from the captions. After that, we need to fill in the missing captions and blank spaces in the provided data set. It is necessary to change all capital letters in the captions to lowercase so that there is no room for confusion while building the vocabulary and training the model.During the model's train and test stages, you must tell the neural network where each caption starts and ends. Each caption has a connection between "startofseq" and "endofseq" at the beginning and end of it. This model's output is word-by-word captions since it takes picture characteristics and previously created words as inputs.

D. Model Training and Fitting

To train the model, we must import the prepared photographic and textual data to use it for model fitting. Subsequently, by generating input and output sequences in batches and training the model using the data.We will use fit() to train the model on all photographs from the Flickr8K dataset's training set. The first stage involves loading the prepared photographic and textual data for model fitting. The process involves loading photos and their corresponding captions into a data structure, such as a dictionary. Once the data is imported, it is separated into training and validation sets. The model training utilizes a Convolutional Neural Network (RESNET) and a Transformer (LSTM).RESNET extracts features from images, while LSTM generates captions based on those features. During training, the model receives a single picture together with its accompanying captions, and it is taught to provide the captions for each image. This technique is repeated until the model can provide a correct caption for each picture in the training dataset.Upon completion of model training, it may be evaluated using the validation set, confirming its capability to provide accurate captions for novel pictures. Employing a convolutional neural network A CNN functions by assigning weights to an image according to its distinct objects and processing it via many convolutional layers. Convolutions are a mathematical procedure used to extract information from a picture. A tiny matrix is used to scan an image during convolution, when a mathematical operation is performed on the matrix to identify characteristics within the picture. The mathematical process of convolution may integrate two forms of information. Convolution is often used to elucidate the data and construct a feature map from the input data.The size of this filter may be, for instance, 3x3. This filter is also referred to as a feature detector or kernel. Convolution is executed by element-wise matrix multiplication by the kernel while traversing the input picture. Each responsive region, or domain where convolution transpires, will be recorded in the feature map together with its outcomes. The filter must be repositioned one more before finalizing the feature map.

The feature map is then sent by an activation function, which serves to activate the network's neurons. This facilitates the model's acquisition of more picture characteristics. Activation functions like ReLU, Sigmoid, and Tanh are used to guarantee that the output conforms to the intended structure.Subsequent to the activation function, the feature map is sent via a pooling layer. Pooling is used to lower the feature map size and lessen the complexity of network. There are two kinds of pooling layers: max pooling and average pooling. The maximum value of each responsive field is obtained by max pooling, and the average value is derived using average pooling.

4. Basics of CNN Architecture

Convolutional Neural Networks (CNNs) are deep learning models that extract features from images using convolutional layers, followed by pooling and fully connected layers for tasks like image classification. They excel in capturing spatial hierarchies and patterns, making them ideal for analyzing visual data.

There are two main parts to a CNN architecture

• A convolution tool that separates and identifies the various features of the image for analysis in a process called as Feature Extraction.

• The network of feature extraction consists of many pairs of convolutional or pooling layers.

• A fully connected layer that utilizes the output from the convolution process and predicts the class of the image based on the features extracted in previous stages.

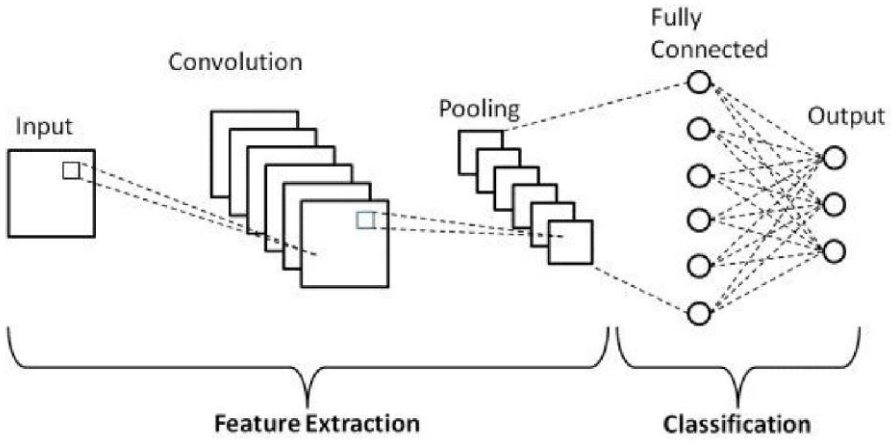

Figure 4.1

There are three types of CNN architecture which are the convolutional layers, pooling layers, and fully-connected (FC) layers. When these layers are stacked, a CNN architecture will be formed. In addition to these three layers, there are two more important parameters which are the dropout layer and the activation function which are defined below.

1. Convolutional Layer

This layer is the first layer that is used to extract the various features from the input images. In this layer, the mathematical operation of convolution is performed between the input image and a filter of a particular size MxM. By sliding the filter over the input image, the dot product is taken between the filter and the parts of the input image with respect to the size of the filter (MxM).

The output is termed as the Feature map which gives us information about the image such as the corners and edges. Later, this feature map is fed to other layers to learn several other features of the In Max Pooling, the largest element is taken from feature map. Average Pooling calculates the average of the elements in a predefined sized Image section. The total sum of the elements in the predefined section is computed in Sum Pooling. The Pooling Layer usually serves as a bridge between the Convolutional Layer and the FC Layer.

This CNN model generalises the features extracted by the convolution layer, and helps the networks to recognise the features independently. With the help of this, the computations are also reduced in a network.

Input image.

The convolution layer in CNN passes the result to the next layer once applying the convolution operation in the input. Convolutional layers in CNN benefit a lot as they ensure the spatial relationship between the pixels is intact.

2. Pooling Layer

In most cases, a Convolutional Layer is followed by a Pooling Layer. The primary aim of this layer is to decrease the size of the convolved feature map to reduce the computational costs. This is performed by decreasing the connections between layers and independently operates on each feature map. Depending upon method used, there are several types of Pooling operations. It basically summarises the features generated by a convolution layer.

3. Fully Connected Layer

The Fully Connected (FC) layer consists of the weights and biases along with the neurons and is used to connect the neurons between two different layers. These layers are usually placed before the output layer and form the last few layers of a CNN Architecture. In this, the input image from the previous layers are flattened and fed to the FC layer. The flattened vector then undergoes few more FC layers where the mathematical functions operations

usually take place. In this stage, the classification process begins to take place. The reason two layers are connected is that two fully connected layers will perform better than a single connected layer. These layers in CNN reduce the human supervision.

4. Dropout

Usually, when all the features are connected to the FC layer, it can cause overfitting in the training dataset. Overfitting occurs when a particular model works so well on the training data causing a negative impact in the model's performance when used on a new data. To overcome this problem, a dropout layer is utilised wherein a few neurons are dropped from the neural network during training process resulting in reduced size of the model. On passing a dropout of 0.3, 30% of the nodes are dropped out randomly from the neural network.

Dropout results in improving the performance of a machine learning model as it prevents overfitting by making the network simpler. It drops neurons from the neural networks during training.

5. Activation Functions

Finally, one of the most important parameters of the CNN model is the activation function. They are used to learn and approximate any kind of continuous and complex relationship between variables of the network. In simple words, it decides which information of the model should fire in the forward direction and which ones should not at the end of the network. It adds non-linearity to the network. There are several commonly used activation functions such as the ReLU, Softmax, tanH and the Sigmoid functions. Each of these functions have a specific usage. For a binary classification CNN model, sigmoid and softmax functions are preferred an for a multiclass classification, generally softmax us used. In simple terms, activation functions in a CNN model determine whether a neuron should be activated or not. It decides whether the input to the work is important or not to predict using mathematical operations.

5. Results

The result of this program is going to be a user being allowed to generate a caption for a visual image using Deep Learning, NLP, and Computer Vision. Using the concept of CNN and LSTM we have build a model and trained the model using flickr_8k data set and MS COCO for getting the appropriate captions for the given image. We have tried different methods to get the perfect outcome.

After installing the data set and loading it into the model the model will train the dataset and give the desired result as per the image. And we have used hugging face as the hosting element in which we can enter the url of the image or the image from the dataset folder for the outcome.

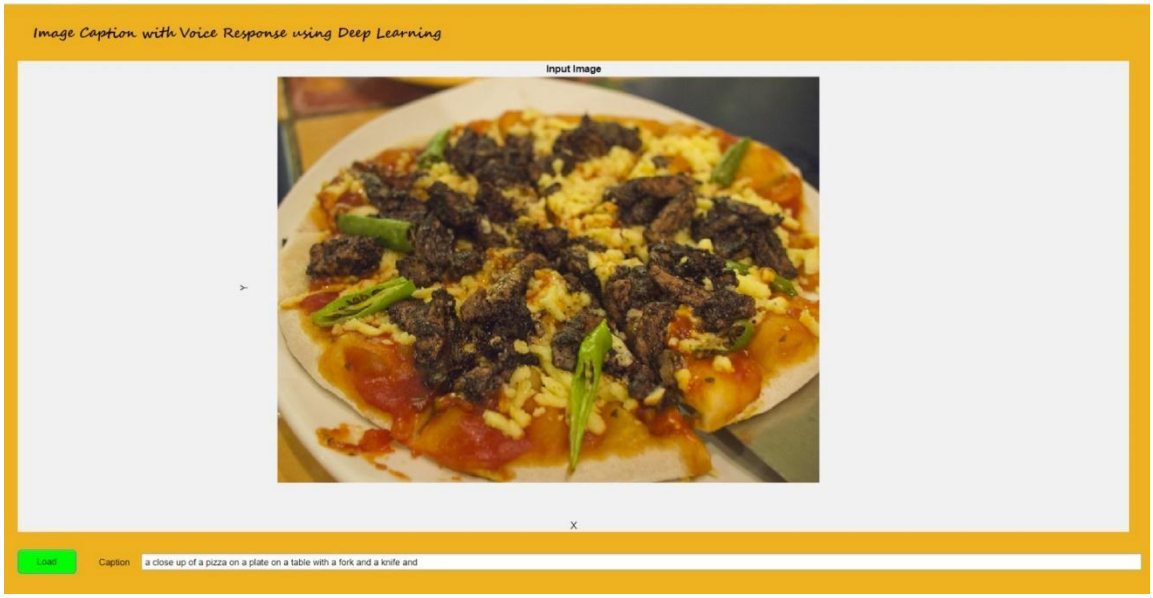

Figure 5.1

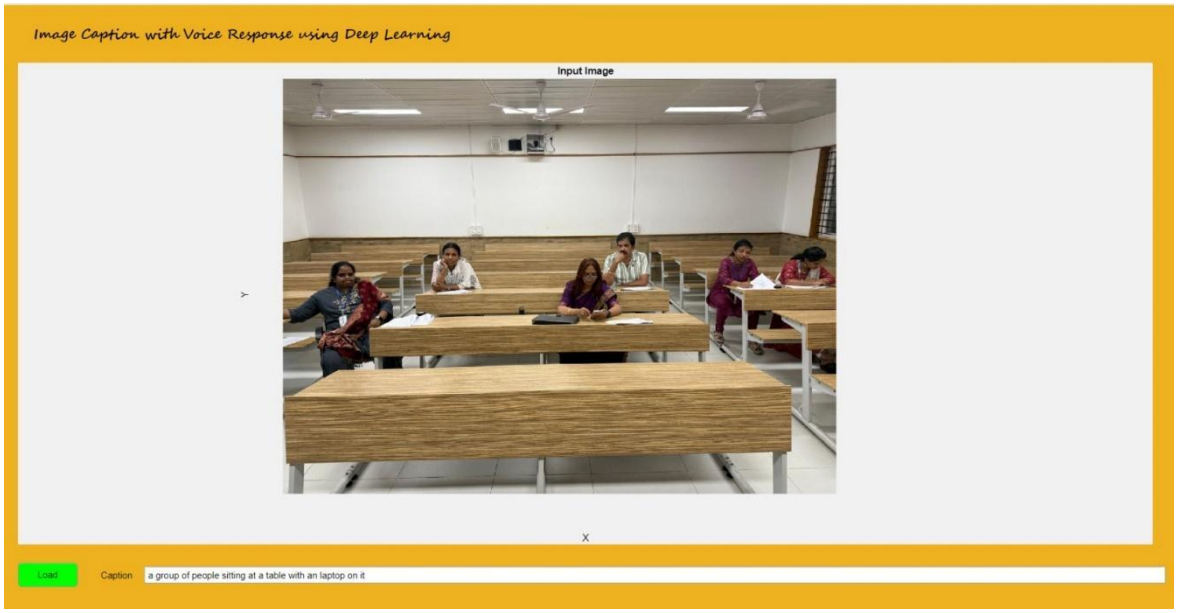

Figure 5.2

6. Conclusion

This study introduces an innovative method for picture captioning by including text-to-speech conversion and supplementary aural input related to the primary subject in the image. The system utilizes MATLAB and CNNs to produce precise captions while enhancing user experience with audio, so making visual material more accessible and participatory.

Future Scope

An integrated system that creates captions for pictures by using CNNs in MATLAB, translates these captions into audio, and plays additional sounds that match to the key actor in the image is going to be developed as part of this project. Through the provision of aural feedback and interactive components, the system will improve the accessibility of visual material as well as the level of content engagement.

References

[1] R. Subash (November 2019): Automatic Image Captioning Using Convolution Neural Networks and LSTM.

[2] Seung-Ho Han, Ho-Jin Choi (2020): Domain-Specific Image Caption Generator with Semantic Ontology.

[3] Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra and Nand Kumar Bansode (2017): Camera2Caption: A Real-Time Image Caption Generator

[4] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares (June 2019): Image Captioning: Transforming Objects into words.

[5] Manish Raypurkar, Abhishek Supe, Pratik Bhumkar, Pravin Borse, Dr Shabnam Sayyad (March 2021): Deep learning-based Image Caption Generator

[6] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan (2015):Show and Tell: A Neural Image Caption Generator.

[7] K.Xu, J.Ba, K.Cho, and R.Salakhutdinov (2018): Show attend and tell: Neural image caption generator with visual attention.

[8] M.Pedersoli, T.Lucas, C.Schmid, and J.Verbeek (2017): Areas of attention for image captioning.

[9] H.R.Tavakoli,R.Shetty, B.Ali, and J.Laaksonen(2017): Paying attention to descriptions generated by image captioning models.

[10] A.Matthews, L.Xie, and X.He(2018): Sem Style-learning to generate stylized image captions using unaligned text.