International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

IMAGE CAPTION GENERATOR USING CNN & LSTM

Dr. Sakshi Paithane
Department of Electronics and Telecommunication,
Jspm's Rajarshri shahu college of engineering (Tathawade)
Pune, India
sapaithane_entc@jspmrscoe.edu.in

Prathmesh Naik
Department of Electronics and Telecommunication,
Jspm's Rajarshri shahu college of engineering (Tathawade)
Pune, India
prathmeshn661@gmail.com

Utkarsha Patil
Department of Electronics and Telecommunication,
Jspm's Rajarshri shahu college of engineering (Tathawade)
Pune, India
utkarshapatil014@gmail.com

Kunal Fulzele
Department of Electronics and Telecommunication,
Jspm's Rajarshri shahu college of engineering (Tathawade)
Pune, India
kunalf18@gmail.com

ABSTRACT

In the last few years, the problem of generating descriptive sentences automatically for images have gained arising interest.Image captioning is a task where each image must be understood properly and are able generate suitable caption with proper grammatical structure. Automatically creating the description or caption of an image using any natural language sentences is a very challenging task. It requires both methods from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order.Our project aims to implement an Image caption generator that responds to the user to get the captions for a provided image.

The ultimate purpose of Image caption generator is to make users experience better by generating automated captions. The combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) such as long short-term memory (LSTM) has been shown to be effective in addressing this challenge.

In this study, we introduce an image caption generator that utilizes both CNN and LSTM architectures. The CNN extracts visual features from the input images, which are then fed to the LSTM for generating descriptive sentences

INTRODUCTION

Image caption generator is a task that involves computer vision and natural language processing concepts to recognize the context of an image and describe them in a natural language like English. The objective of our project is to learn the concepts of a CNN and LSTM model and build a working model of Image caption generator by implementing CNN with LSTM. In this Python project, we will be implementing the caption generator using CNN (Convolutional Neural Networks) and LSTM (Long short term memory). The image features will be extracted from CNN model trained on the dateset and then we feed the features into the LSTM model which will be responsible for generating the image captions. Nowadays , Machine learning is a trend in Artificial Intelligence . Recently , we apply AI in building a powerful performance and highly intelligent machines . Machine learning has a subset called deep learning , it provides high accuracy with its results so its performance is high too through its output . In our research paper ,deep learning is used in apps of image description . Image description provides the process of describing the content from an image . The idea is based on the detection of objects and what actions in the input image . Bottom-up and top-down approaches are two main approaches of image description . Bottom-up approaches generate contents in an input image , and then combine them into a caption . Top-down approaches generate a semantic representation of an input image that is then decoded into a caption using various architectures like recurrent neural networks .□ Image description could have many benefits , for instance by helping visually impaired people better understand the content of images on the web Image captioning can be regarded as an end-to-end Sequence to Sequence problem, as it converts images, which is regarded as a sequence of pixels to a sequence of words. For this purpose, we need to process both the language or statements and the images. For the Language part, we use recurrent Neural Networks and for the Image part, we use Convolutional Neural Networks to obtain the feature vectors respectively.

Image captioning is one of the most innovative model in the field of neural networks and artificial intelligence.Caption generation is becoming a growing business in the world, and many data annotation firms are earning billions from this.We are going to build one such annotation tool which is capable of generating very relevant captions for the image with the help of datasets. Image Captioning is the process of generating a textual description for given images. It has been a very important and fundamental task in the Deep Learning domain. Image captioning has a huge amount of application. NVIDIA is

using image captioning technologies to create an application to help people who have low or no eyesight. Basic knowledge of two techniques of Deep learning including LSTM(a type of Recurrent Neural Network) and Convolutional Neural Networks(CNN) is required for the same.
Image caption generator is a process of recognizing the context of an image and annotating it with relevant captions using deep learning, and computer vision. To build an image caption generator model we have to merge CNN with LSTM (Long short term memory).

We can drive that:
Image Caption Generator Model (CNN-RNN model) = CNN + LSTM.

CNN- To extract features from the image. A pretrained model called Xception is used for this.
LSTM- To generate a description from the extracted information of the image.

Image caption generator is a process of recognizing the context of an image and annotating it with relevant captions using deep learning, and computer vision. It includes the labeling of an image with English keywords with the help of datasets provided during model training. Dataset is used to train the CNN model called Xception. Xception is responsible for image feature extraction. These extracted features will be fed to the LSTM model which in turn generates the image caption.

AIM

To Generate Image Caption Generator using Deep learning algorithms, CNN and LSTM.

OBJECTIVES

[1] To perform extensive literature survey and component selection.
[2] To import required python libraries.
[3] To analyze the features of the loaded image.
[4] To extract features from the given image .
[5] To train the model by loading the data set.
[6] To test and predict the final output.
[7] To analyze the proposed system systematically

LITERATURE SURVEY

Abstract Automatically creating the description or caption of an image using any natural language sentences is a very challenging task. It requires both methods from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. In addition to that we have discussed how this model can be implemented on web and will be accessible for end user as well. Our project aims to implement an Image caption generator that responds to the user to get the captions for a provided image. The ultimate purpose of Image caption generator is to make users experience better by generating automated captions. We can use this in image indexing, for visually impaired persons, for social media, and several other natural language processing applications. Deep learning methods have demonstrated state- of-the-art results on caption generation problems. What is most impressive about these methods is a single end-to-end model can be defined to predict a caption, given a photo, instead of requiring sophisticated data preparation or a pipeline of specifically designed models. In this an Image caption generator, Basis on our provided image It will generate the caption from our trained model. The basic idea behind this is that users will get automated captions when we use or implement it on social media or on any applications.Deep Learning algorithms are designed in such a way that they mimic the function of the human cerebral cortex. These algorithms are representations of deep neural networks i.e. neural networks with many hidden layers. Convolutional neural networks are deep learning algorithms that can train large datasets with millions of parameters, in form of 2D images as input and convolve it with filters to produce the desired outputs.CNN models are built to evaluate its performance on image recognition and detection datasets. The algorithm is implemented on MNIST and CIFAR-10 data-set and its performance are evaluated.The accuracy of models on MNIST is 99.6 %, CIFAR-10 is using real-time data augmentation and dropout on CPU unit.Several variants of the Long Short-Term Memory (LSTM) architecture for recurrent neural net-works have been proposed since its inception in1995. In recent years, these networks have be-come the state-of-theart models for a variety of machine learning problems. This has led to a renewed interest in understanding the role and utility of various computational components of typical LSTM variants.. Deep learning methods have demonstrated state-of-the-art results on caption generation problems. What is most impressive about these methods is a single end-to-end model can be defined to predict a caption, given a photo, instead of requiring sophisticated data preparation or a pipeline of specifically designed models. In this an Image caption generator, Basis on our provided image It will generate the caption from our trained model. The basic idea behind this is that users will get automated captions when we use or implement it on social media or on any applications.

FINDINGS FROM LITERATURE SURVEY

Generating textual descriptions for images is a challenging task in artificial intelligence that requires both computer vision and natural language processing methods. Recently, deep learning techniques have achieved state-of-the-art results for this problem, with recurrent neural networks (RNNs) being particularly powerful for sequential data modeling. This has been explained in detail by Andrej Karapathy in his blog post "The Unreasonable Effectiveness of Recurrent Neural Networks." In this systematic literature review (SLR), we analyze different deep learning models that are commonly used for image captioning. We searched for relevant articles in three academic databases and selected 61 primary studies after applying inclusion and exclusion criteria. Through data extraction and analysis, we discovered various models and techniques used for image captioning. Convolutional neural networks (CNNs) are

frequently used for image content extraction, while RNNs and long short-term memory (LSTM) models are used for language generation. Our analysis shows that LSTM outperforms RNN in this context. We also found that different studies have used various mechanisms for scene understanding, including encoder-decoder and attention mechanisms. Overall, this literature review provides an in-depth analysis of the current state-of-the-art deep learning models used for image captioning. It highlights the strengths and weaknesses of different models and techniques and provides insights into the current research trends in this field.

METHODOLOGY

The following is a step-by-step guide for building an image caption generator using Python and deep learning techniques:
[Step 1]: Install and import the necessary libraries, such as TensorFlow, Keras, pillow, Numpy, tqdm, and jupyterlab, in Jupyter Notebook.
[Step 2]: Load the descriptions in CSV format and map them to their corresponding images using a dictionary.
[Step 3]: Clean the text data by removing noise in the form of special characters, hashtags, punctuation, and numbers.
[Step 4]: Generate the vocabulary, which is a set of unique words present in the text corpus.
[Step 5]: Load the training images and map them to their corresponding descriptions using the image name as the key and a list of descriptions as the value. Add unique words at the beginning and end of each sentence to identify the start and end of the sentence.
[Step 6]: Analyze the features of the text by converting the image into an encoding that the machine can understand.
[Step 7]: Define the model using the CNN model, which involves processing the sequence from the text and extracting the feature vector from the image.
[Step 8]: Train the model using a function named model.fit_generator() to fit the batches to the model. Save the model to the models folder.
[Step 9]: Test the model accuracy by inputting test image data and predicting an output, i.e., a caption, for the image.

Following these steps can help you build an effective image caption generator that can generate descriptive and accurate captions for images.

PROPOSED BLOCK DIAGRAM

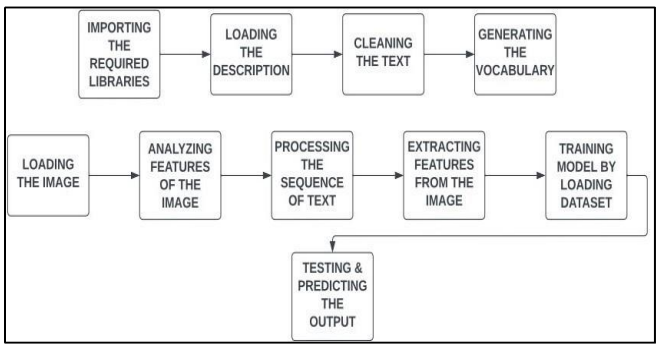

Fig 1. System Methodology

SOFTWARE SPECIFICATION

IDE USED

PyCharm :
PyCharm is a dedicated Python Integrated Development Environment (IDE) that offers a wide range of essential tools for Python engineers and is securely integrated to create an environment that fosters effective Python, web, and data science development.

Python :
Python is an object-oriented, high-level programming language with integrated dynamic for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options. Python is relatively simple, so it's easy to learn since it requires a unique syntax that focuses on readability. Developers can read and translate Python code much easier than other languages. In turn, this reduces the cost of program maintenance and development because it allows teams to work collaboratively without significant language and experience barriers. Additionally, Python supports the use of modules and packages, which means that programs can be designed in a modular style and code can be reused across a variety of projects.

CNN :
It is a Deep Learning algorithm which takes in an input image and assigns importance (learnable weights and biases) to various aspects/objects in the image, which helps it differentiate one image from the other. One of the most popular applications of this architecture is image classification.The neural network consists of several convolutional layers mixed with nonlinear and pooling layers. When the image is passed through one convolution layer, the output of the first layer becomes the input for the second layer. This process continues for all subsequent layers. Convolutional neural networks are distinguished from other neural networks by their superior performance with image, speech, or audio signal inputs. They have three main types of layers, which are: □ Convolutional layer □ Pooling layer □ Fully-connected (FC) layer The convolutional layer is the first layer of a convolutional network. While convolutional layers can be followed by additional convolutional layers or pooling layers, the fully-connected layer is the final layer. With each layer, the CNN increases in its complexity, identifying greater portions of the image.

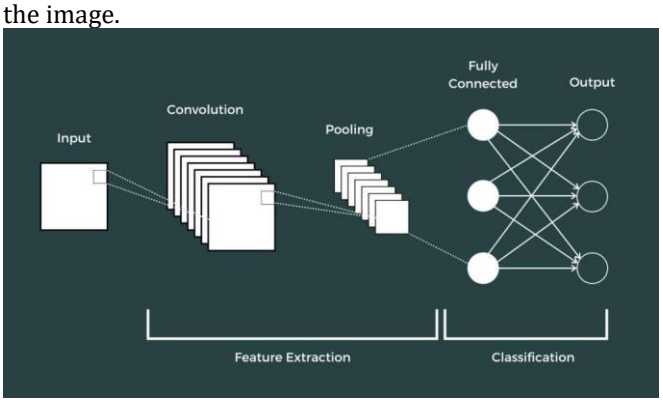

Fig 2 Convolutional neural network layers

LSTM :
These networks are a type of Recurrent Neural Network (RNN) capable of learning order dependence in sequence prediction problems. This is most used in complex problems like Machine Translation, Speech Recognition, and many more.The reason behind developing LSTM was, when we go deeper into a neural network if the gradients are very small or zero, then little to no training can take place, leading to poor predictive performance and this problem was encountered when training traditional RNNs. LSTM networks are well-suited for classifying, processing, and making predictions based on time series data since there can be lags of unknown duration between important events in a time series.LSTM is way more effective and better compared to the traditional RNN as it overcomes the short term memory limitations of the RNN. LSTM can carry out relevant information throughout the processing of inputs and discards non-relevant information with a forget gate. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell. LSTM networks are well-suited to classifying, processing and making predictions based on time series data,since there can be lags of unknown duration between important events in a time series.

CNN LSTM Architecture
The CNN-LSTM architecture involves using CNN layers for feature extraction on input data combined with LSTMs to support sequence prediction. This model is specifically designed for sequence prediction problems with spatial inputs, like images or videos. They are widely used in Activity Recognition, Image Description, Video Description and many more. CNN-LSTMs are generally used when their inputs have spatial structure, such as the 2D structure or pixels in an image or the 1D structure of words in a sentence, paragraph, or document and also have a temporal structure in their input such as the order of images in a video or words in text, or require the generation of output with temporal structure such as words in a textual description.

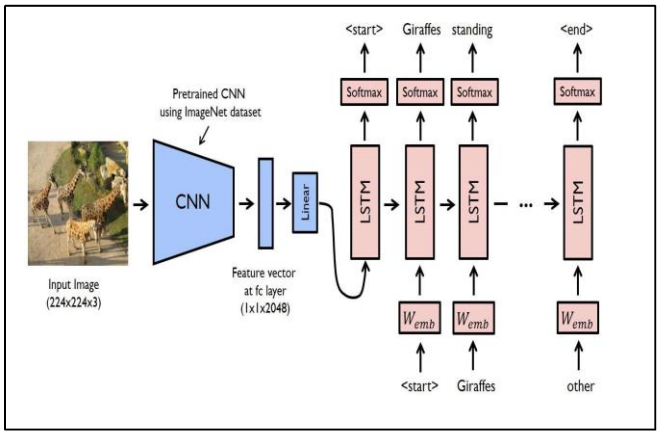

Fig 3 - CNN LSTM Architecture

EXPECTED RESULT

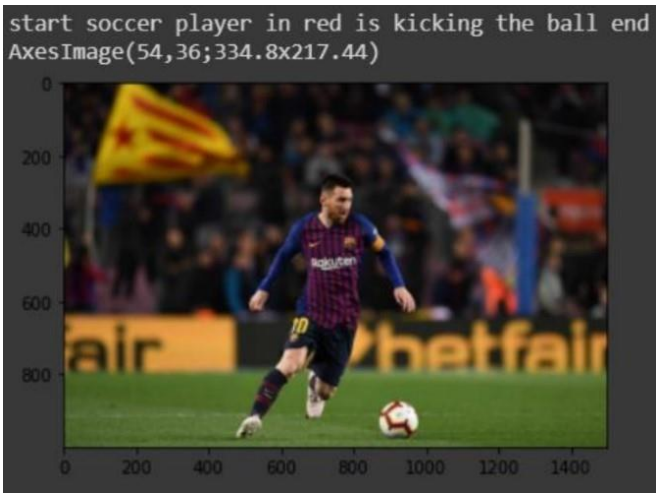

Fig 4 - Image with Relevant generated caption

CONCLUSION

In this project, we have built a deep learning model with the help of CNN and LSTM. We will be using a very small dataset of images to train our model, but the business level model used larger datasets for better accuracy. The larger the datasets are higher the accuracy. Finally, we will get the best caption for our image.

ACKNOWLEDGMENT

I would want to express my gratitude to everyone who helped me along this suggested framework. I appreciate their thoughtful leadership, invaluable developmental audit, and friendly encouragement throughout the suggested framework. You have my eternal gratitude in this regard.

FUTURE SCOPE

There are various domains that can take advantage of image captioning to automate their tasks.
1) A model can be trained in medical ultrasound or MRI images or angiographic videos to generate a complete report of a person without any consent from a doctor. Image captioning can be used to generate an automatic report by looking at those medical images of a person.
2) Image captioning can also be used in industries to automate various tasks. A model can be trained on images of a company product manufacturing environment to find out an anomaly in the environment or product automatically. It can also be used also used to detect any mishap in a company like fire or security issues.
3) Image captioning can also be used in agriculture to generate the report of crops for owners by looking at images of crops.
4) Image captioning can also be used in traffic analysis report generation by using CCTV cameras installed on streets and thus guide drivers which is the best suitable path to take and where parking is available.

REFERENCES

[1] Parth Kotak, Prem Kotakhas - Image caption generator using CNN & LSTM 2021

[2] Adil Khan,Muhammad, Saleem Mahar Image captioning using deep learning 2020

[3] Rahul Chauhan, R.C Joshi CNN for image detection and recognition 2020

[4] Rupesh Srivastava, Jan koutnik LSTM 2016

[5] Image Captioning using CNN and LSTM|IJRASET Publication- Academia.edu

[6] Image Caption Generator – IJERT

[7] Image Caption Generator using Deep Learning (analyticsvidhya.com)

[8] https://towardsdatascience.com/a-guide-to-image-captioning- e9fd5517f350

[9] ieeexplore.ieee.org/document/8276124

[10] machinelearningmastery.com/develop-a-deep-learning- caption-generation-model-in-python

[11] blog.clairvoyantsoft.com/image-caption-generator- 535b8e9a66ac

[12] www.analyticsvidhya.com/blog/2018/04/solving-an-image- captioning-task-using-deep-learning/

[13] www.clairvoyant.ai/blog/image-caption-generator