



INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

IMAGE CAPTION GENERATOR

Using CNN-LSTM

¹Anjali Barigela, ²Gopidi Meghana Reddy, ³Pashikanti Shivani, ⁴Varakala Satheesh Kumar, ⁵Bejjam Vasundhara Devi

¹Student, ²Student, ³Student, ⁴Assistant Professor, ⁵Assistant Professor

¹Computer Science Engineering,

¹Sreenidhi Institute of Science and Technology, Ghatkesar- Hyderabad, India

Abstract: Creating a caption for an image is the aim of the project. Making a description for an image is called photograph captioning. It demands an understanding of the crucial elements, as well as their traits and interrelationships. The elements of an image We can now develop models that can forecast the future since deep learning techniques have advanced and massive datasets and computer power are now readily available. Construct captions for a photo. This is what we did in our Python-based research using CNN (Convolutional Neural Networks), a form of the neural network, and LSTM (Long Short-Term Memory), another type of neural network. combining many RNN (Recurrent Neural Network) classes to enable computer vision A photograph can be shown in the appropriate context by a computer after being recognized by that context.

Index Terms – CNN, LSTM, RNN.

I. INTRODUCTION

A more convincing description is one that is full and human-like. As long as machines do not think, talk, or act like people, natural language descriptions will remain a challenge. There are several uses for image captioning in a variety of industries, including biomedicine, business, online search, and the military, among others. Social media platforms like Facebook, Instagram, and others can automatically create captions from photographs. A caption can be provided for any pixel in a color or monochrome image using image captioning. To recognize the context of a picture and explain it in a natural language like English, an image caption generator uses computer vision and natural language processing techniques. A thorough human-like description creates a better first impression, and we will have constructed the caption generator using CNN (Convolutional Neural Networks) and LSTM in this Python-based project. As long as machines do not think, speak, or act like humans, natural language descriptions will remain a challenging problem to solve. Many industries, including the military, business, web search, and medicine, use image captioning. On social networking platforms like Facebook and Instagram, captions can be produced automatically from images. Any pixel in a color or monochrome image may produce subtitles through image captioning. In order to understand the context of a picture and describe it in a natural language like English, an image caption generator must use computer vision and natural language processing concepts. In this Python-based project, the caption generator will be implemented using CNN Neural Networks. The CNN model Xception, trained on the Flickr8k dataset, will be used to provide the image features, which will then be used to feed the features into the LSTM model, which will be in charge of generating captions for the images. A deep neural network that can process a lot of data is called a convolutional neural network. It accepts a 2D matrix as input, and images are easily made.

II. IMAGE CAPTIONING TECHNIQUES

2.1 CNN

Convolutional Neural Networks, or CNNs, are essential, specialized neural networks that can generate data with a certain shape, like a 2D lattice. CNNs are useful when working with images. To extract important details from a picture and combine the elements to characterize it, it looks at the picture from left to right and all the way through. It is capable of handling interpreted, rotated, scaled, and altered images. The Convolutional Neural System is a sophisticated learning algorithm that absorbs the information picture, assigns importance to various components and protests in the picture, and distinguishes it from other pictures.

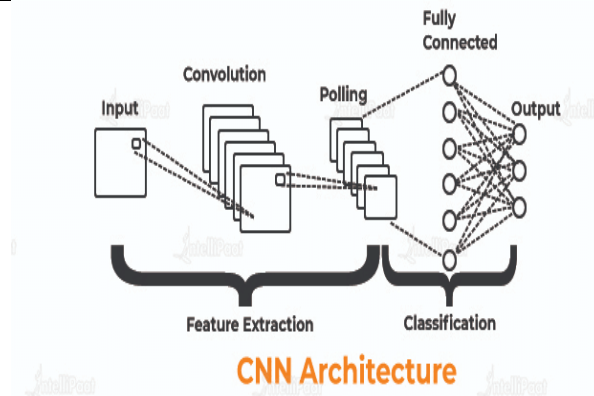


Fig 1. CNN Architecture

2.2 How does CNN function?

According to CNN, the neurons in a cell may be connected with a specific cell area before it rather than all the neurons in a completely similar way. As we have previously discussed, a fully connected neural network is advantageous for the task at hand because every input in the preceding layers is connected to every input in the following layers. This helps the neural network become less complex and use less computational power. According to a new computer's standard image with pixel numbers. When typically comparing two pictures We examine each pixel's pixel values. This method only works when we are comparing two identical photos; otherwise, the comparison will not work. CNN performs piece-by-piece image comparisons.

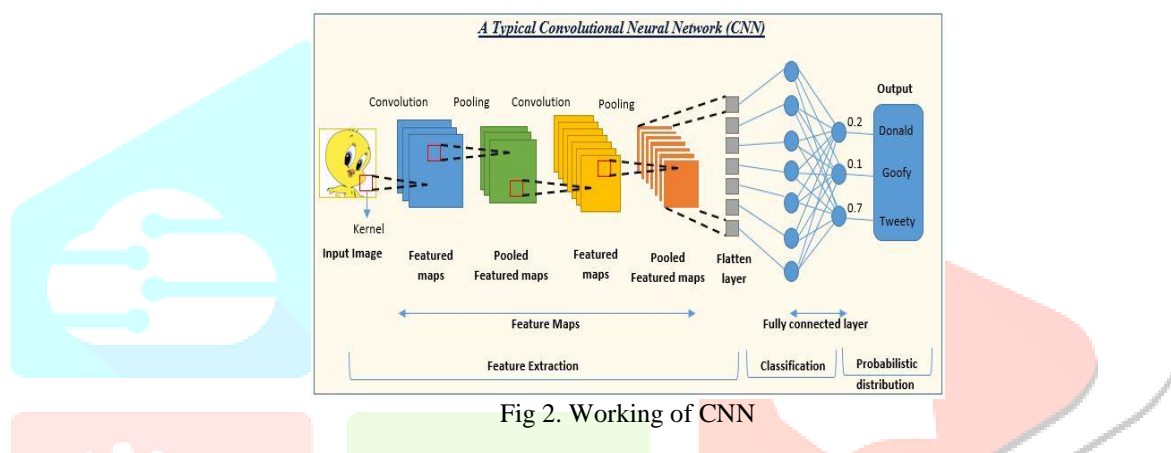


Fig 2. Working of CNN

In the CNN model, there are three different types of layers: 1. Convolutional 2. Pooling 3. Fully Connected The input image is read through the CNN in the first layer, and a feature map is built on top of that. The next layers, such as the Pooling layer, take their input from that feature map. The feature map is divided into additional, simpler pieces in the pooling layer so that the context of the image can be examined in depth. In order to find the most important details about the image, this layer makes the feature map denser. Depending on the image, the first and second layers, i.e., convolutional and pooling, are repeatedly used to obtain dense information about the image. These two layers combine to provide the extra dense feature map. And the last layer, Fully Connected, makes use of this deep feature map. In this layer, classification is done. The pixels are sorted according to similarities and differences. Classification is done to a very high degree in order to capture the core of the image and aid in object, person, and item identification. These layers aid CNN in precisely locating and identifying visual elements. The extraction of key details from a fixed-length input image is converted into outputs of a fixed size. CNN techniques are widely used in: In the field of medical sciences, image analysis is only done by CNNs. With its use, one can easily study the inner workings of the body. It has been employed in mobile phones for a variety of purposes, including determining a person's age and unlocking the device using a camera image. It is frequently used in industries to create patents or copyright for certain images that have been taken. Pharmaceutical discovery — it has been widely utilized to locate the best medication to treat a specific ailment by studying the chemical characteristics of potential drugs.

III. Source of LSTM:

Sepp Hochreiter and Jurgen Schmidhuber, two German researchers, performed the LSTM's initial search in 1997. the long short-term memory, or LSTM. The LSTM maintains a significant position in the recurrent neural network field of deep learning. The unique feature of LSTM is that in addition to storing the input data, it can also generate predictions for the following datasets using its own. This LSTM network keeps the stored data for a specific amount of time and then forecasts or assigns the data future values based on that information. This is the primary reason that LSTM is utilized in this situation more so than conventional RNN.

3.1 Working of LSTM:

LSTMs are essentially a subset of RNNs, which have a greater ability to store data values than RNNs. Today, LSTMs are widely used across all industries. Below is a picture of the LSTM's most basic diagram. There are three main gates in it: the forget gate, the input gate, and the output gate. These gates are equipped with the ability to store data and produce the desired output. The three gates are a constant when the LSTM network is discussed. The LSTM's simplest architecture is depicted in the diagram below:

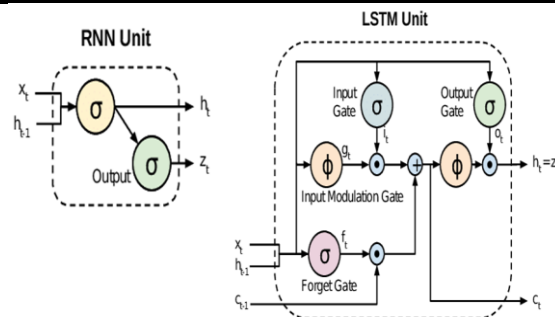


Fig 3. Working of LSTM

3.2 Long Short-Term Memory Network Applications: -

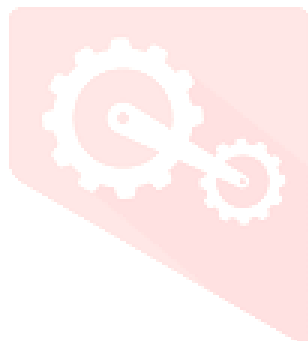
LSTMs are extensively and primarily utilised for a wide range of deep learning tasks, the majority of which involve forecasting data based on historical data. The two outstanding examples are stock market and text prediction.

Text Prediction - The LSTM is extensively utilised in text prediction. Because of its grasp of long-term memory, LSTM is able to predict the words that will come after them in sentences. This is the outcome of the LSTM network's self-generated prediction of the subsequent words. The LSTM first saves the data, including the words' sounds, styles, and usage in specific contexts, etc., and then predicts the subsequent words based on that information. The input data that has been saved is then used again in the future.

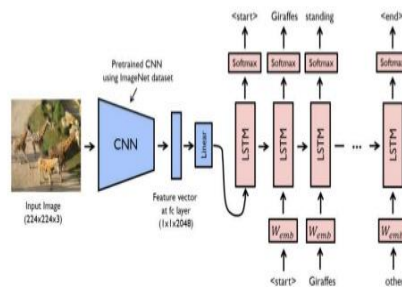
A Chatbot, which is commonly used by eCommerce websites and mobile applications, serves as the best example of text prediction. **Stock market Prediction** - LSTM also works to anticipate future variations and trends in the stock market by storing the data or trends that describe how the market operates at a specific moment in time. Predicting stock market fluctuations is a difficult endeavour because they are very difficult to predict and forecast. To provide consumers with accurate values, the LSTM model must be trained properly. For such, a significant amount of data must be kept on hand for days at a time.

IV. IMAGE CAPTION GENERATOR

Picture caption generators, which recognize the context of a picture and annotate it with applicable captions, are made using deep learning and computer vision. One of its components is the labelling of a picture with English keywords using datasets from model training. The Imagenet dataset is used to train the CNN model Xception. Xception is in charge of extracting picture features. The LSTM model will receive these extracted features and output the image caption.



Model



4.1 Project file organisation:

The following files make up the data set that we downloaded for our research:

This file, titled [Flickr8k_Datasets](#), contains all the images for which we must initially train our model. It has 8091 pictures in it.

[Flickr8k_texts](#) are a folder that houses text documents and prepared captions for the images.

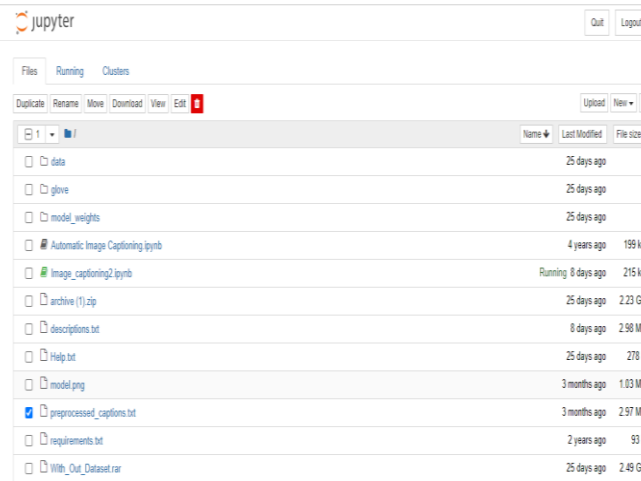
The following files are set up so that this system can be used by us to test the CNN-LSTM model's functionality.

[Models](#) - This folder will hold all of the initially trained models. The model would be trained once using this approach.

This file, called [Description.txt](#), contains the names of the images and the captions that go with them.

[Models.png](#) is a diagrammatic illustration of the CNN-LSTM model's extension.

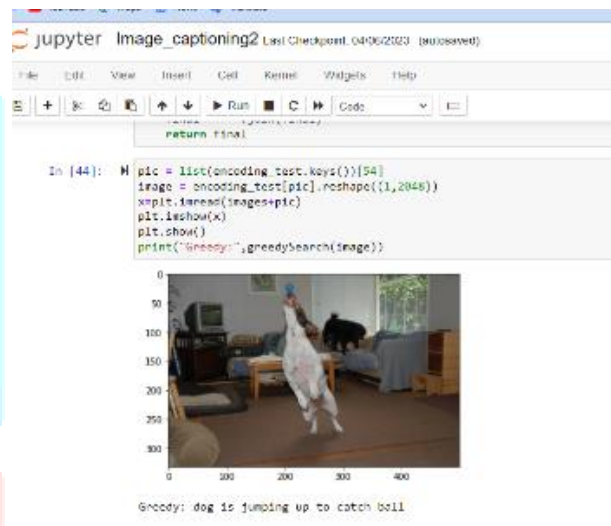
[Pre-processed captions.txt](#): This text file contains five captions for every image.



Name	Last Modified	File size
data	25 days ago	
glove	25 days ago	
model_weights	25 days ago	
Automatic Image Captioning.ipynb	4 years ago	195 kB
Image_captioning2.ipynb	Running 8 days ago	215 kB
archive (1).zip	25 days ago	2.23 GB
descriptions.txt	8 days ago	2.90 MB
Help.txt	25 days ago	270 B
model.png	3 months ago	1.03 MB
preprocessed_captions.txt	3 months ago	2.97 MB
requirements.txt	2 years ago	93 B
With_Out_Dataset.rar	25 days ago	2.49 GB

Fig 4. Project Files

V. OUTPUT SCREENS:



```

return final

In [44]: In: pic = list(encoding_test.keys())[54]
image = encoding_test[pic].reshape((1,2048))
xplt.imshow(images-pic)
plt.imshow(x)
plt.show()
print("greedy:", greedySearch(image))

0
50
100
150
200
250
300
350
400
0
100
200
300
400

Greedy: dog is jumping up to catch ball

```

Fig 5. Output

We take one image from the dataset and need the caption that describes it.

VI. CONCLUSION:

Our model which is used to create the Deep Neural Networks will generate captions in a simple understandable language. We will present the way to understand the image using the Convolutional Neural Network (CNN) and Long Short-term Memory (LSTM) neural network. It allows us to leverage the useful aspects of powerful models in tasks that have never been used before. Depending on the datasets, our model performance increases. As the caption generated can be of diverse nature. We can make our caption that are able to generate as diverse as we can depending on the fine details or we can make our model that gives attention to the certain parts and be able to generate different captions depending on attention to those parts. We can improve our model's accuracy by increasing the number of epochs and using better pretrained.

REFERENCES:

- [1] Alon Lavie and Abhaya Agarwal. Meteor, m-bleu, and m-ter: Machine translation output evaluation metrics with strong correlation to human rankings. In the Proceedings of the Third Workshop on Statistical Machine Translation. 115–118, Association for Computational Linguistics.
- [2] Robert Gaizauskas and Ahmet Aker. 2010. Utilising dependency relationship patterns to provide image descriptions. published in the Association for Computational Linguistics' 48th Annual Meeting Proceedings. 1250–1258, Association for Computational Linguistics.
- [3] Stephen Gould, Mark Johnson, Basura Fernando, and Peter Anderson. 2016. Spice: Semantic evaluation of image captions. at the European Computer Vision Conference. 382–398 in Springer.
- [4] The study was written by Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Picture captioning and visual quality assessment with a focus on both the bottom-up and top-down. arXiv preprint arXiv:1707.07998 (2017).
- [5] Alexander G. Schwing, Aditya Deshpande, and Jyoti Aneja. 2018. Convolutional captioning for images.