IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 13, May 2023

Image Caption Generator with Speech Using CNN and LSTM

Akshara Madhusoodanan[1] and Shyma Kareem[2]

Student, Department of Computer Applications[1]

Assistant Professor, Department of Computer Applications[2]

Musaliar College of Engineering & Technology, Pathanamthitta, Kerala

Abstract: *Image processing remains one of the most cutting-edge technologies employed by Google, the medical industry, and other sectors of the economy. Due to its free and open source tool, which every developer can afford, this technology has recently drawn many programmers and developers. Since it is currently used as a primary technique of gathering information from images, processing those images for various purposes, and performing various operations on those images, image processing also aids in learning a lot of information from a single image. The problem of creating voice-based image captions uses the idea of NLP (natural language processing) to comprehend the description of an image. The optimal method for this project is the combination of CNN and LSTM; the primary goal of the suggested research is to find the ideal caption for an image. The description will be translated into text after being obtained, and the text will then be given voice. For persons who are blind and cannot understand visuals, image descriptions can be obtained as a speech output using TTS engine.*

Keywords: CNN, LSTM, Deep learning, TTS engine

I. INTRODUCTION

It is difficult to express the image in plain English words in the Deep Learning space. There are many uses for picture captioning. For those who are visually challenged, it can have a significant influence. Additionally, it might be helpful in the area of automating the task of an individual interpreting the image. It will be extensively employed in fields where text is often used and where it is possible to deduce or produce text from a picture. Additionally, it can be useful for frame-by-frame examination of videos. Social media platforms might also draw conclusions straight from the photographs. There has been a significant amount of research While it is simple for humans to describe sights or situations, it can be difficult for machines to do the same. While humans can easily explain sights or circumstances, machines may find it challenging to do so. Image captioning is the process of providing in-depth details and explanations for certain photographs. The many types of elements that make up an image include objects, the surrounding environment, and interactions between the subjects or circumstances shown in the picture. Similar to how language serves as a vehicle for elucidating and communicating important details from the circumstances in the photographs. The captions created from the images contain the information related to it and give a brief overview of diverse worldwide situations. The development of image description technology could one day enable blind people to "see" the outside world. The system uses the flicker 8K dataset to retrieve the partial vectors of the words and the vector information of the images, merges them to create the perfect captions from the image, and then outputs the captions as speech to help the blind and visually impaired receive information from the outside world.

II. LITERATURE REVIEW

Sumathi, T., Hemalatha, M,"A combined hierarchical model for automatic image annotation and retrieval[1]."t used SVM and JEC to extract the depth feature for an image by using the Gaussian effect in order to comprehend a user-provided picture better. They used JEC for the image feature extraction method in [1]. A feature vector with various dimensions will be created for image annotation. It only entails processing the image using several models. The image was rotated in flat, axis, and position-wise ways to map the key properties that the picture's annotations

DOI: 10.48175/568

comprise after the features and JEC were separated from each other using the SVM model. It helps to understand or be able to identify the object.

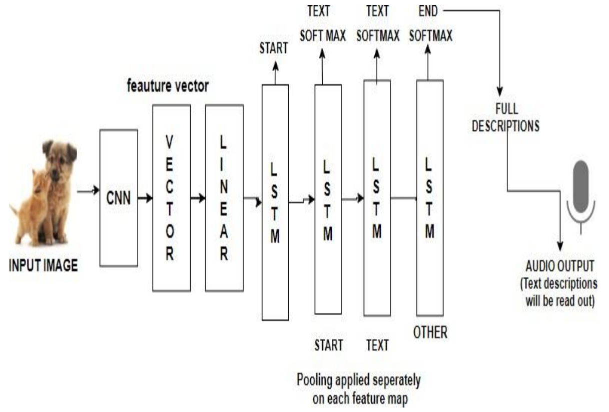
Dong-Jin Kim, Donggeun Yoo, Bonggeun Sim, In So Kweon, "Senetence Learning Deep convolutional neural Network for Image Caption Generation " [2], The caption generating technique makes use of the CNN and RNN algorithms as well as an attention model to foretell the suitable phrases. Given that LSTM is more effective than RNN in our model, we choose to employ it. They were able to recognise the odd things by using CNN. After receiving information from the picture, the first stage of long short term memory will correctly anticipate the single word at the beginning of the sentence. To get around this, they used the Guide's long short memory. The instruction is applied in this way throughout the whole procedure, making use of the term that was previously taught and remaining

Varsha Kesavan, Vaidehi Muley,Megha kolhekar, "Deep Learning based Image Caption Generation" [3] A transfer learning approach is used to create automated image captioning for user-provided photos. The VCG16 is being used in this instance for the encoding process. The input is then encoded using a recurrent neural network to provide a constant-dimensional vector with the appropriate description. In order to select the most effective method, they compared the accuracy they obtained using a number of techniques, including VCG16, RESNET, and the inception model.

Ren C. Luo, Yu-Ting, Hsu, Yu-Cheng, Wen, Huan-Jun, Ye , "Visual Image Caption Generation for Service Robotics and Industrial Application"[4] It detects everything in front of the cameras using image detection, much like the face and object identification in self-driving cars, and predicts the appropriate phrase for the object, such as a box, pen, or bottle. It adds the newly formed unknown words to the values of the prior dataset and combines this prediction with the earlier pre-trained model. This operation will take a long time, and irrelevant descriptions will be produced.

III. ARCHITECTURE

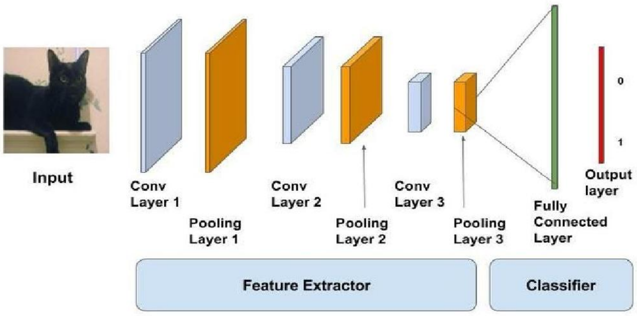
Convolutional neural networks employ user-provided input images to detect items in the picture. They then extract key features from the images, store the feature vector values, and forecast the features using pooling functions. When the process is finished, it moves on to the long short term memory layer for the prediction of a sequence based on the previous one. Here, the softmax function is used to predict the output accurately and for overcoming the over fitting problem. When working with neural networks, the majority of nodes have outputs that are related to the previous one, which Figures and Tables



IV. ALGORITHM

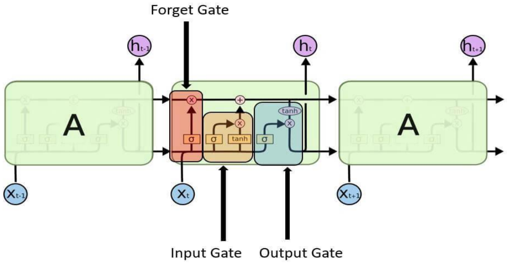
4.1 Convolutional Neural Network

Convolutional neural networks, often known as CNNs, are a type of deep neural network that are primarily used for classifying and analyzing visual pictures. They are also utilized in other fields, such as voice and image identification and natural language processing. Convolutional, pooling, and fully linked layers make up its three layers. The primary benefit of employing a convolutional neural network is its ability to recognize the objects and faces in a picture. If a photo is indexed with cats and dogs, CNN will be able to tell what the two animals' main characteristics are and how they relate to one another, as well as their distinctive behavioural patterns. Because it can forecast the image with the maximum degree of precision, it is significantly more efficient than the other algorithms.



4.2 Long Short Term Memory

A kind of RNN called long short term memory is applied to sequence prediction issues. Long short-term memory (LSTM) has the efficient performance when comparing to RNN, and it can be sustainably obtain the information with the long period of time. The non-relevant information will be deleted by utilizing LSTM. It may be able to make predictions about the information based on upcoming or past data. The biggest issue with LSTM is that it will take longer to drain the data as the dataset size increases. In order to extract information from the image, CNN will be employed, and LSTM will provide captions for the input image.



V. METHODOLOGY

- Dataset Collection and Data Cleaning
- Feature Extraction
- Loading dataset for Training the model
- Tokenization
- Data Generator
- Text to speech conversation

a) It is simple to map the description with the input photos when utilizing the flicker dataset, which comprises images and descriptions that are in the form of dictionaries with keys and values. Data cleaningmust be completed for eachtext dataset. In order to achieve it, special symbols like an asterisk, a semicolon, a colon, and double quotes must be cleared.

b) The CNN model extracts features from the original pictures, which are then compressed into smaller feature vectors that are compatible with each other. It also goes by the name Encoder. Using the VGG16 model to extract features from a picture is a The neural network condenses a large number of feature extractions from the original input into a smaller feature vector that is compatible with recurrent neural networks. It is the main justification behind calling CNN as "encoder."

c) When we train our model, we normally evaluate the length of the training photos, the descriptions, and the features as well. The created model is specifically dependent on the epoch values that we have provided during the training process, hence the best description can only be gained through training. As a consequence, an accurate description of an input may be predicted.

d) Data must be tokenized, or compressed into smaller, readable pieces when language processing is utilized, in order to create unique content.

e) Data is delivered to the CNN layer, where operations like pooling are carried out, when we employ the data generator. In order to fit the initial input with the second produced word utilizing dense, the data is then transmitted via the LSTM model, which utilizes the output of the CNN model. The proper description may be predicted by comparing each pixel in an image over time.

f) In addition, the system provides a mechanism to turn this produced caption into spoken speech utilizing a TTS engine, which will also benefit the blind and crippled. At the conclusion of all this, the image's captions will be provided as a final output, and the output caption will also be simultaneously translated into voice as another output and can also download the generated speech.

VI. TESTING THE MODEL

First process is to uploading the image it may be from the dataset which we have gathered or else it may be user own image. After that step it enter into various module then it will print the related description for an user given input, once the captions is created then it will play the audio of an caption generated

VII. CONCLUSION AND FUTURE SCOPE

Using a CNN-LSTM model, a voice-based picture caption generator has been created. The suggested model is trained to evaluate human input in addition to the dataset, allowing it to predict descriptions from external photos. This is one of the project's primary takeaways.

The future enhancements would include describing the captions based on multiple goals. Can generate caption in variety of languages as a part of future improvements. Training and testing the model with larger datasets and on different architectures. This technology can help in various grounds like in the medical field which can help the doctors to get the information produced by the x-rays or the MRI scans. May make it easier for those who are blind to comprehend their surroundings and the environment around them by employing video captions. Also useful in the field of human computer interaction, traffic and surveillance and many more.

REFERENCES

[1] Sumathi, T., and Hemalatha, M. presented "A combined hierarchical model for automatic image annotation and retrieval" at the 2011 International Conference on Advanced Computing (ICAC).

[2] "Senetence Learning Deep Convolutional Neural Network for Image Caption Generation" was presented at the 13th International Conference on Ubiquitous Robots and Ambient Intelligence by Dong-Jin Kim, Donggeun Yoo, Bonggeun Sim, and In So Kweon.

[3] Varsha Kesavan, Vaidehi Muley, and Megha Kolhekar presented "Deep Learning based Image Caption Generation" at the 2019 Global Conference for Advancement in Technology (GCAT).

[4] "Visual Image Caption Generation for Industrial and Service Robotics Applications" by Yu-Ting Hsu, Yu-Cheng Wen, Huan-Jun Ye, and Ren C. Luo was published in IEEE-2019.

[5] M.T. Yu and M.M. Sein presented their "Automatic image captioning system using integration of N cut and color-based segmentation method"

DOI: 10.48175/568