IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

IMAGE AND VIDEO CAPTIONING USING DEEP LEARNING

[1]Shailaja Jadhav, [2]Pranalee Walunj, [3]Sampada Dodake, [4]Vaishnavi Thete

[1]Lecturer in Computer Engineering Department, Marathwada Mitra Mandal's College of Engineering, Karvenagar, Pune, Maharashtra, India

[2,3,4]Student, Computer Engineering Department

[2,3,4] Marathwada Mitra Mandal's College of Engineering, Karvenagar, Pune, Maharashtra, India

Abstract: In this world of advanced technology where everything is developing at a very fast pace, image and video processing has become extremely important for various reasons. Applications for guiding Self driving cars and for building software that guides blind people. Also, it can help in military operations and surveillance by detecting threat and help weapons and soldiers to destroy them. Video caption generator uses video encoder as well as caption decoder framework. Image and Video Captioning can be done by using Deep Learning Models. In this paper we will be using Neural Networks for the image and Video captioning. Convolutional Neural Network is used as encoder which access the image features and Recurrent Neural Network (Long Short Term Memory) is used as decoder which generates the captions for the images with the help of features of the images and vocabulary that is built. In Video Captioning, for the generator module, we take an existing video captioning concept using LSTM network. For the discriminator, we apply a novel realization specifically tuned for the problem of video captioning and taking both the video and sentences features as input..

Index Terms **- Convolutional Neural Network, Recurrent Neural Network, Deep Learning, Long-Short Term Memory.**

I. INTRODUCTION

Machine Learning is a vast topic and also part of Artificial intelligence. Deep Learning can be described as part of Machine learning, capable of learning from unstructured and unlabeled data. In recent times Deep learning has extremely changed the world of Computer Vision. By using the features of deep learning features and representations, machine can give comparable or better performance than human beings in object recognition, image classification and video segmentation, but still there are developments needed in segments like image and video captioning. Caption generation could be a challenging AI issue, wherever a textual clarification for a given image must be made. It needs both computer vision strategies to know the image content, and a language model from the language process (NLP) field to remodel the image understanding into words within the right order. VIDEO captioning is the problem of producing a textual description for a given video content The multidisciplinary nature of this problem opens vast new possibilities for inter-relating with video collections and there has been an increase in research effort on this subject observable over the past years. This interdisciplinary nature, however, also poses crucial research challenges at the intersection between the fields of natural language processing and computer vision. With the progress in technology that the efficiency of image caption generation is also growing. Captioning of image can be used in many Machine Learning tasks for Recommendation Systems. There are many models being proposed for image captioning like object detection model, visual attention. In Deep Learning there are different models like VGG model, Inception Model ,ResNet-LSTM model, traditional CNN-RNN Model. In this analysis, we are going to elaborate the model we have followed for captioning the images and videos i.e CNN, RNN and LSTM.

II. Related work

Sr.No	Author	Title	Description
1.	Prof. Sandeep Samleti, Ashish Mishra,Alok Jhajhria,Shivam Kumar Rai,Gaurav Malik	Real Time Video Captioning Using Deep Learning	In this research paper we have discussed the two models, first one is Hierarchical model and second one is Multi stream hierarchical Boundary Model. The Hierarchical model is combined with steered captioning.
2.	Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati	Image Caption Generating Deep Learning Model	In this paper we will be using Neural Networks for the image captioning. Convolution Neural Network (ResNet) is used as encoder which access the image features and Recurrent Neural Network (Long Short Term Memory) is used as decoder which generates the captions for the images with the help of image features and vocabulary that is built.
3.	S. Das, L. Jain and A. Das	Deep Learning for Military Image Captioning	This paper presents the proof-of-concept demonstration for caption generation. The generative model is based on a deep recurrent architecture combined with the pre-trained image-to-vector model Inception V3 via a Convolutional Neural Network (CNN) and the word-to-vectors model word2vec via a skip-gram model.
4.	A. Hani, N. Tagougui and M. Kherallah	Image Caption Generation Using A Deep Architecture	In this paper, we presented a model that generates natural language description of an image. We used a combination of convolutional neural networks to extract features and then used recurrent neural networks to generate text from these features.
5.	Soheyla Amirian , Khaled Rasheed, Thiab R. Taha, Hamid R. Arabnia	Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap	This article is a concise review of both image captioning and video captioning methodologies based on deep learning. This study treats both image and video captioning by emphasizing the algorithmic overlap between the two.

[6] In this research paper we have discussed the two models, first one is hierarchical model and second one is multi stream hierarchical boundary model. The hierarchical model is combined with steered captioning.

[7] In this paper we will be using neural networks for the image captioning. Convolution neural network (resnet) is used as encoder which access the image features and recurrent neural network (long short term memory) is used as decoder which generates the captions for the images with the help of image features and vocabulary that is built.

[3] This paper presents the proof-of-concept demonstration for caption generation. The generative model is based on a deep recurrent architecture combined with the pre-trained image-to-vector model inception v3 via a convolutional neural network (cnn) and the word-to-vectors model word2vec via a skip-gram model.

[8] In this paper, we presented a model that generates natural language description of an image. We used a combination of convolutional neural networks to extract features and then used recurrent neural networks to generate text from these features.

[5] This article is a concise review of both image captioning and video captioning methodologies based on deep learning. This study treats both image and video captioning by emphasizing the algorithmic overlap between the two.

III. Proposed System

Photo Feature Extractor: With the assistance of 16-layer VGG (CNN) model, we've got pre-trained the Flicker8k dataset. This pre-processes the photos with the VGG model (without the output layer) and can use the extracted features expected by this model as input.
Sequence Processor: It is a word embedding layer used for handling the text input, followed by a long short-term memory (LSTM) that is the Recurrent neural network layer. This model is trained to predict every word of the sentence once the image is generated.
Decoder: A fixed-length vector is the output of the feature extractor and sequence processor. These are aligned with one another and processed by a dense layer to create a final prediction. In the end, an image caption is generated.
The model image function Extractor expects features of the input image to be a vector of 4096 elements. These are processed via a dense layer to form a photographic illustration of 256 elements. Input Sequences with a predefined length are needed by the model of Sequence Processor to be fed into an embedding layer using a mask to avoid padded values.

The input models generate a vector of 256 parts as LSTM works with units of 256 memory units.to reduce overfitting within the training set each input models use 50% dropout regularization. The Decoder model uses an extra operation to merge the vectors from each input models. this is then fed into a Dense 256 layer of neurons and a final output Dense layer that permits a softmax prediction for the next word within the sequence over the complete output vocabulary.

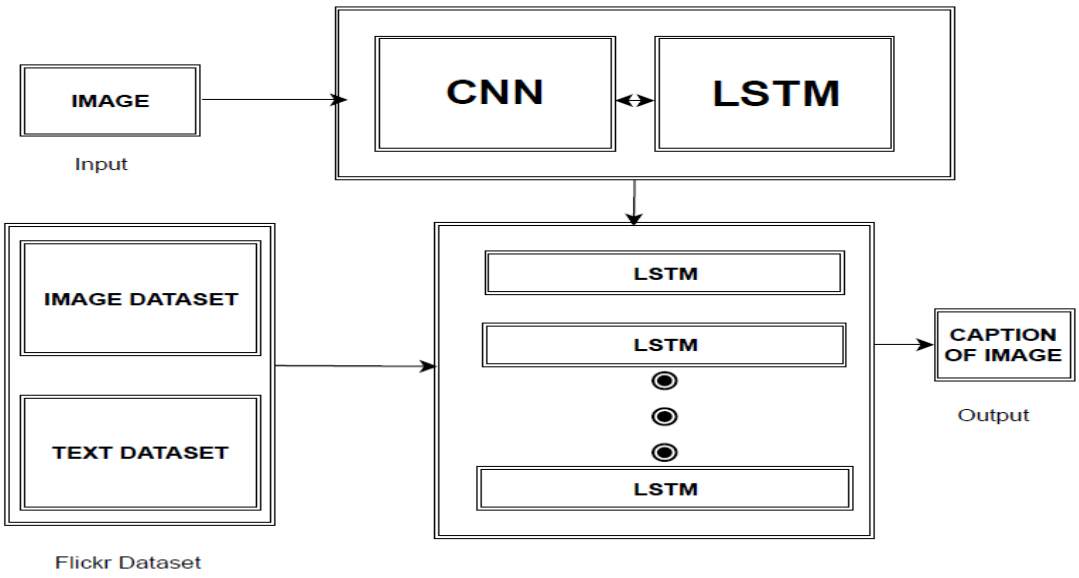

Fig.

IV. Analysis of Algorithms

1.1 CNN(Convolutional Neural Network)

In deep learning, a convolutional neural network (CNN) is a a part of deep neural networks, most ordinarily applied to research visual image. currently once we consider a neural network we expect about matrix multiplications however that's not the case with ConvNet. It makes use of a special technique known as Convolution. In arithmetic, convolution could be a mathematical operation on two functions that produces a third function that expresses how the shape of one is changed by the other.

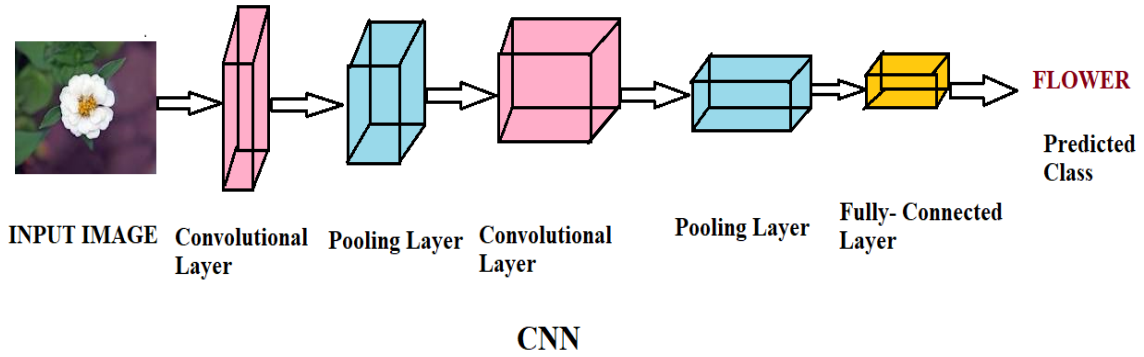

Fig. 1 CNN(Convolutional Neural Network)

1.2 RNN(Recurrent Neural Network)

RNNs were the quality suggestion for operating with sequential data before the advent of attention models. Specific parameters for every element of the sequence may be required by a deep feed forward model. It may even be unable to generalize to variable-length sequences. Recurrent Neural Networks use a similar weights for every element of the sequence, decreasing the amount of parameters and permitting the model to generalize to sequences of variable lengths. RNNs generalize to structured data apart from sequential data, like geographical or graphical knowledge, because of its style.

Recurrent neural networks, like many other deep learning techniques, are comparatively old. They were first developed within the 1980s, however we tend to didn't appreciate their full potential till late. the advent of long short-term memory (LSTM) within the 1990s, combined with a rise in computational power and also the large amounts of information that we tend to currently need to deal with, has very pushed RNNs to the forefront.

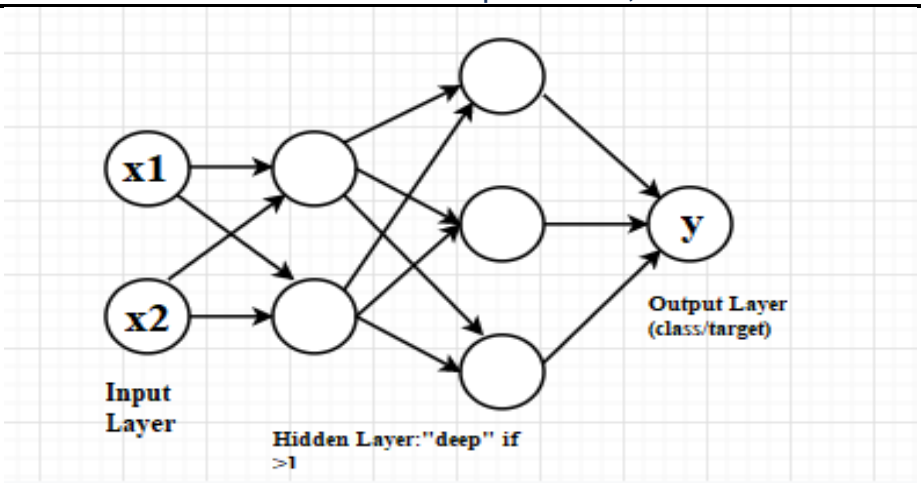

Fig. 2 RNN(Recurrent Neural Network)

1.3 LSTM(Long Short Term Memory)

Long Short Term Memory (LSTM) networks are a kind of recurrent Neural Network that may learn order dependence. The output of the previous step is used as input within the current step in RNN. Hochreiter & Schmidhuber created the LSTM. It addressed the issue of RNN long-term dependency, within which the RNN is unable to predict words stored in long-term memory however will create more correct predictions based on current data. RNN doesn't offer an efficient performance because the gap length rises. The LSTM could keep information for a long time by default. it is used for time-series processing, prediction, and classification. LSTM has feedback connections, not like conventional feed-forward neural networks. It can handle not solely single knowledge points (like photos) however conjointly complete knowledge streams (such as speech or video). LSTM will be used for tasks like nonsegmental, joined handwriting recognition, or speech recognition.

• CNN:

The number of filters is equal to dd (in that case, the conv layer does not change the depth dimensionality). So, in that case, the time complexity indeed amounts to O(knd2)O(knd2) because we're repeating the O(knd)O(knd) routine described in the question for each of the dd filters.

• RNN:

RNN has a Complexity of O(n.d2).

• LSTM:

LSTM is local in space and time, which means that the input length doesn't affect the storage requirements of the network and for every time step, the time complexity per weight is O(1).

V. CONCLUSION

In this paper we have proposed Image captioning deep learning model. We have used CNN model to get captions for each of the given image. For the purpose of training the model Flicker8k dataset has been used. RESNET is the design of convolution layer. The RESNET architecture is used for extracting the image features and then this image features are given as input to Long Short Term Memory units and captions are generated with the assistance of vocabulary generated throughout the training process. This model works with efficiency once. The model is processed using Graphic Processing Unit. With the implementation of the algorithm, it is realized that an end-to-end neural network system can automatically view an image and generates a responsible description in natural language. Image captioning based on the Convolution Neural network encodes a picture into a representation followed by a recurrent neural network that generates corresponding text. This Image Captioning deep learning model is very much useful for analyzing the massive amounts of unstructured and unlabeled data to find the patterns in those images for guiding the Self driving cars, for building the software system to guide blind people.

REFERENCES

[1] Liu, Shuang Bai, Liang Hu, Yanli Wang, Haoran. (2018). Image Captioning Based on Deep Neural Networks. MATECWeb of Conferences. 232. 01052. 10.1051/matecconf/201823201052.

[2] A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture", 2019 International Arab Conference on Information Technology (ACIT), 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998.

[3] S. Das, L. Jain and A. Das, "Deep Learning for Military Image Captioning " , 2018 21st International Conference on Information Fusion (FUSION), 2018, pp. 2165-2171, doi:10.23919/ICIF.2018.8455321

[4] G. Geetha,T.Kirthigadevi,G GODWIN Ponsam,T.Karthik,M.Safa," Image Captioning Using Deep Convolutional Neural Networks(CNNs)" Published under licence by IOP Publishing Ltd in Journal of Physics :Conference Series ,Volume 1712, International Conference On Computational Physics in Emerging Technologies(ICCPET) 2020 August 2020,Manglore India in 2015.

[5] Soheyla Amirian , Khaled Rasheed, Thiab R. Taha, Hamid R. Arabnia "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap", December 2020

[6] Prof. Sandeep Samleti, Ashish Mishra,Alok Jhajhria,Shivam Kumar Rai,Gaurav Malik "Real Time Video Captioning Using Deep Learning" December 2021

[7] Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati "Image Caption Generating Deep Learning Model", September 2021

[8] A. Hani, N. Tagougui and M. Kherallah, "Image Caption Generation Using A Deep Architecture," 2019 International Arab Conference on Information Technology (ACIT), 2019, pp. 246-251, doi: 10.1109/ACIT47987.2019.8990998.