IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 6, June 2022

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.165

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006129 |

Image Captioning and Summary Generation Using CNN

Shruthi D V, Ashwin Athreya H S, Hemanth Kumar Gongadimath, Nikhil S Y, Yashas R

Asst. Professor, Dept. of Information Science and Engineering, Malnad College of Engineering, Hassan,

Karnataka, India

Students, Dept. of Information Science and Engineering, Malnad College of Engineering, Hassan, Karnataka, India

ABSTRACT: Attention based captioning methods are used for image captioning that automatically creates natural language description for images that has lot of research. However, the focus of attention-based image capturing is to creating visual information at the regions that has interest for creating sentence and forget to focus on regions of interest in image that has relational reasoning. Attention based image captioning does not consider the previous regions which can be used as a guide for attention selection. When compared to existing methods, the methods we are using advanced methods that has relation aware for the visual representation for image capturing. We are using methods such as CNN, RNN and NLP for the experimental results that can be used for various outperform results.

I. INTRODUCTION

Human beings have the ability to make differentiate between the actions but the computers don't have that ability. For computers to make differentiate the images and to analyze is hard but not like it is impossible. In recent years deep learning advancement has made the impact for the analyzing of creating the captioning of image for the computer to store and save.

During the recent time deep learning is used as a tool for image captioning for its improvement. Usually, encoder and decoder systems are implemented, CNN will be used as encoder for getting information of the image and RNN will be used as a decoder, that converts the image into natural language description. The CNN will collect the data and send to the decoder. While collecting the data the CNN may compress the data into sequence which will result in data loss. The focus on a single part of the image while creating a single world is used by CNN-RNN image captioning. Using this the visual attention will reduce the impact of irrelevant words in a single word.Currently retrieval and templates-based methods are used in image captioning, where retrieval method will extract captioned images to query images by leveraging distance metric and bound together all the captions to create new caption. Similarly in template method, it creates collection of sentence templates according to the situation in the image and try to fill the remaining template by actions.

II. RELATIVE WORK

The partake of human is usually way more necessary to annotate images. Deep learning can help in the dynamic annotation of the images. Earlier, object detection & image classification tasks were used to identify objects within the image. Putting forth an encoder-decoder mechanism for an attention-based picture captioning model. This above-mentioned model sheds the lights on certain parts of the visuals of the images rather than emphasizing on the whole image from the dataset. The purpose of picture caption extraction is to provide a detailed and comprehensive description of the image's content. When extracting the sentence, it evaluates not just the objects in the image, but also their relationships to create sequences.

A comprehensive work of CNN and RNN for image captioning includes:

1) A CNN and RNN based combined network generate captions.
2) Another CNN-RNN based network checks for the captions and sends the feedback to the first input layer of the network to produce high quality captions.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006129 |

III. PROPOSED SYSTEM

Our proposed model make use of 'Convolution Neural Network' or 'CNN' to encode the pictures of the images provided as an input from a respective dataset. Loosely translated, the pictures here consist of the nuances of the images fed from the dataset as an input. There upon, the above pictures are used to build a context vector by the attention layer. Finally, the 'Recurrent Neural Network', fundamentally called 'RNN' helps in decoding the information produced by the attention model into a comprehensible, sequential statement.

A convolutional neural network is a specific type of neural network with one or more layers. It processes the data which consists of a grid like structure. One of the prime advantages of using CNN is that, one need not to perform extra pre-processing on the data input. The main difference between a regular neural network and a convolutional neural network is that, CNNs make use of convolutions to manage the mathematics behind the scenes. Convolutions are used instead of matrices in at least any of the CNNs. CNNs employs by applying the filters to the pass-throughs at the input layer. The specialty of CNNs is that, they fine tune the input data during the training. The convolution part of CNN consists of artificial neurons called nodes. These nodes functions by calculating the weighted sum of inputs at the input layer and helps in returning the activation map. Each node in the network is described by its weight values. When one passes an input in the input layer (say images), it takes the values of the pixel and picks some of the visualized features. Some of the main applications of CNNs are as follows:

1) Helps in recognizing different handwritten patterns.
2) Helps in recognizing the images with minimum pre-processing.
3) Mainly used in postal services to read the zip codes on an envelope.
4) Also used in banks to rad out the digits on check.
5) CNNs finds its application in Compute vision.

A Recurrent Neural Network (RNN) is an extended class of Artificial Neural Network in which the connection between different nodes gives a temporal dynamic behavior by forming a directed graph. It mainly aims in helping the model sequential data that are derived from feed forward networks. In recurrent neural network, a memory-state is added to the neurons. Here the recurrent neural network performs the operation in all available active square, hence the name 'Recurrent Neural Network'.

RNN finds its application in many industries. When it comes to financial industry, RNN helps in determining the stock prices. In Automobiles, RNN helps in avoiding an accident by anticipating the trajectory of the vehicle. Also, RNN is extensively used in image captioning, sentimental analysis and text analysis.

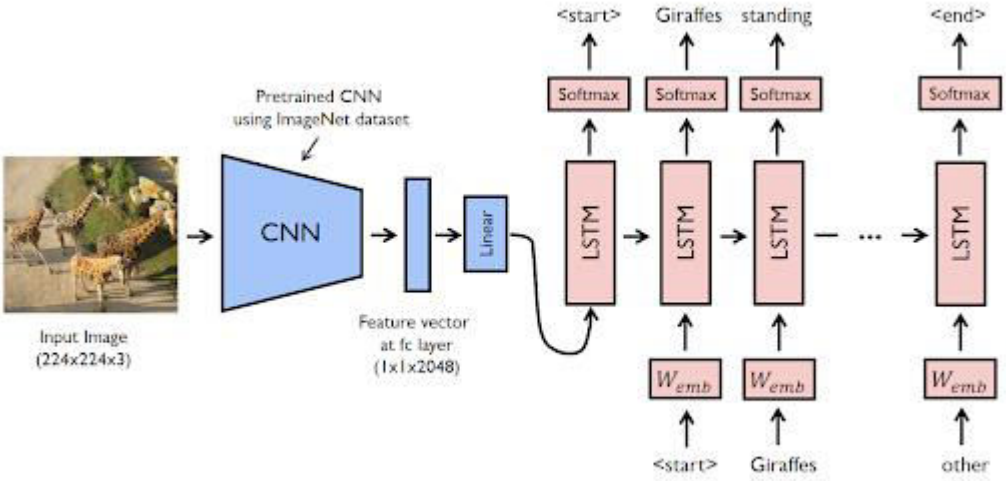


Figure 1. Model, Image Caption Generator

IV. IMPLEMENTATION

The basic idea is to have a system that is going to take an input image and produce an output in the form of sentence that describe the image that is grammatically and syntactically accurate. There are two major parts of this project. The first one is the image model where the model architecture outputs the image features as non-linear activations of pixel

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006129 |

values of the image, where these values are fed into the second part i.e., the language model that generates the summary sentence based on the output of the image model that is being obtained from the first part. To perform the processing of raw input data i.e., both images and captions into a proper format, we have written data preprocessing scripts. A Convolutional Neural Network architecture is used as an encoder to convert the encoded image features to a dimensional vector space. A Recurrent Neural Network is used as decoder to convert the encoded image features to natural language descriptions. Attention mechanism that allows to see the features of the decoder form a certain spotlighted region in the input image to enhance the overall performance.

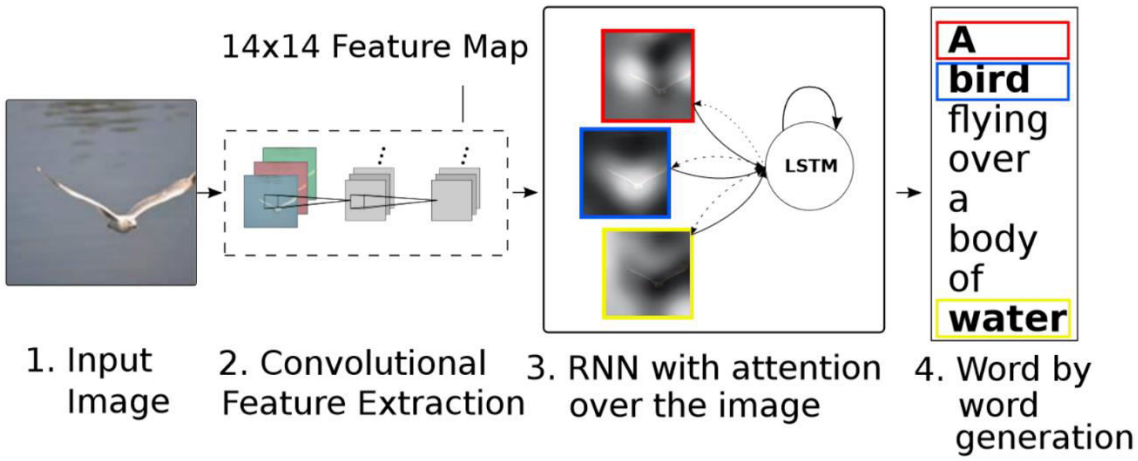


Figure 2. Show, attend and tell architecture

Dataset Sources

We have recognized in Computer Vision and research domain there are basically three most commonly used image caption training dataset i.e., COCO dataset, Flickr8k and Flickr30k. These datasets hold around 123,000, 31,000 and 8000 caption annotated images that carry around five different descriptions. Dataset used for this project is MSCOCO dataset.

Data Preprocessing

The input datasets are composed of images and captions; thus, we have to pre-process images into correct format for CNN network and the textual captions into correct format for RNN network. Furthermore, captions are being encoded with the described dictionary and stored in a JSON file that later on can be fed into RNN model at the later stage.

Convolutional Neural Network (Encoder)

The encoder needs to extract image features of different size and encodes them into a non-linear activations of pixel values of the image that can be later fed to RNN. VGG16 and ResNet are mostly commonly recommended for image encoders. In our work CNN is used as feature encoder rather than to classify images. Further in order to support the input images of different sizes, we add another flexible layer to our CNN architecture. In our model we put out of action the gradient to reduce the computational costs and with fine tuning, we might get a better overall performance.

Soft Attention Mechanism

Next to CNN, we create the soft trainable attention mechanism which performs the action of show, attend and tell. The attention mechanism tells the network about which part of the image should be focused more on gathering the next word in the description. By adding the encoder output and historic state that is being updated in each iteration we can calculate the attention area. In the implementation of attention area, we perform transformation on both the encoder output and historic state output. The output is summed up and activation function is ReLU. The attention area returned as output is later used in decoder that involves RNN.

Recurrent Neural Network

The decoder here performs the function of generating the captions word by word using Recurrent Neural network which is able to generate the word sequentially. The input for the decoder is the encoded non-linear activations of pixel values of the image from the CNN. While receiving the encoded images and captions, we sort the encoded image and caption by encoded key length of the image in decreasing order. We process the encoded image that are having the caption length greater than or equal to the number of iterations in order to increase efficiency and reduce training time.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

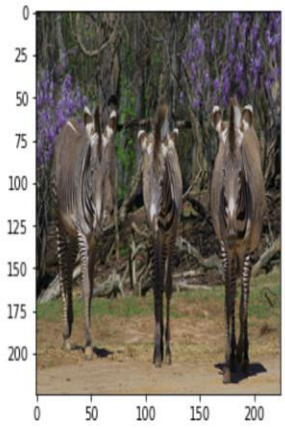
|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006129 |

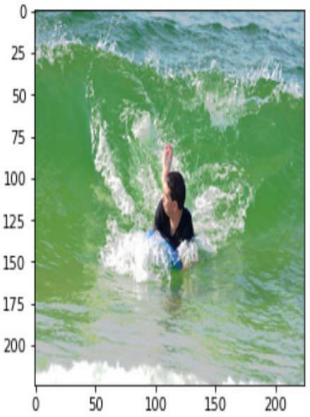
V. RESULT

Some of our results of our model are shown below

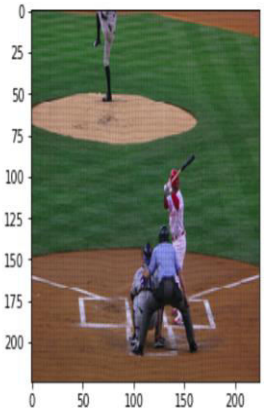
The output is : a group of zebras are standing in a field .



The output is : a man on a surfboard riding a wave .

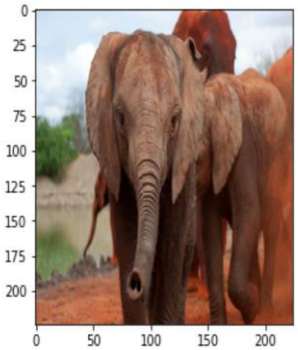


The output is : a baseball player is getting ready to throw a ball .



The output is : a man on skis is standing in the snow .

The output is : a group of elephants standing in a field .



The output is : a group of people sitting around a table with food een ddd .

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006129 |

VI. FUTURE SCOPE

Automatic image captioning and summary generation is far from mature and there lots of researchers going on in the field of more accurate image feature extraction and semantically better syntax generation. We assume that this project ignites our interest in the field of Deep Learning knowledge in Computer vision and expects to explore more in the future.

VII. CONCLUSION

An image captioning and summary generator is a model whose task is to integrate attention and is described to build a natural image from the given image. The image characteristics can be extracted using convolution neural network (CNN). CNN is a type of artificial neural network used in image recognition and processing that is specifically designed to process pixel data. These characteristics of an image are then given to the attention layer to generate a fixed length context vector. Later the recurrent neural network (RNN) is used to decode this fixed length context vector into understandable sequences. This attention mechanism also helps in the optimization of findings as well as in summarizing the appropriate captions.

REFERENCES

1. Step by Step Guide to Build Image Caption Generator using Deep Learning **https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-caption-generator-using-deep-learning/**
2. How to Develop a Deep Learning Photo Caption Generator from Scratch https://machinelearningmastery.com/develop-a-deep-learning-caption-generation-model-in-python/
3. Learn to Build Image Caption Generator with CNN & LSTM https://data-flair.training/blogs/python-based-project-image-caption-generator-cnn/
4. A Guide to Image Captioning https://towardsdatascience.com/a-guide-to-image-captioning-e9fd5517f350
5. IMAGE CAPTION GENERATOR https://www.clairvoyant.ai/blog/image-caption-generator
6. Automatic Image Captioning Using Deep Learning https://medium.com/swlh/automatic-image-captioning-using-deep-learning-5e899c127387
7. An Overview of Image Caption Generation Methods https://www.hindawi.com/journals/cin/2020/3062706/
8. Image Caption Generator using Deep Learning https://www.geeksforgeeks.org/image-caption-generator-using-deep-learning-on-flickr8k-dataset/
9. Deep Learning Based Image Caption Generator https://www.irjet.net/archives/V8/i3/IRJET-V8I392.pdf
10. Empirical Analysis of Image Caption Generation using Deep Learning https://www.researchgate.net/publication/351744861_Empirical_Analysis_of_Image_Caption_Generation_using_Deep_Learning
11. Exploring Deep Learning Image Captioning https://mobidev.biz/blog/exploring-deep-learning-image-captioning
12. Image Caption and Summary Generation https://github.com/darshandv/Image-Caption-and-Summary-Generation
13. A Deep Learning Model for Image Caption Generation https://www.ijcseonline.org/pub_paper/3-IJCSE-08233.pdf
14. Image Caption Generating Deep Learning Model
15. https://www.ijert.org/image-caption-generating-deep-learning-model
16. IMAGE CAPTION GENERATORhttp://103.47.12.35/bitstream/handle/1/1868/1613105091_ROHAN_PRASAD_FinalProjectReport%20-%20Rohan%20Gamer.pdf?sequence=1&isAllowed=y
17. A Comprehensive Survey of Deep Learning for Image Captioning https://arxiv.org/pdf/1810.04020.pdf
18. IMAGE PARAGRAPH CAPTIONING USING DEEP LEARNING AND NLP TECHNIQUES http://cse.anits.edu.in/projects/projects1920B11.pdf
19. DEEP LEARNING AND MACHINE LEARNING SOLUTIONS
20. https://www.hpe.com/in/en/compute/hpc/deep-learning.html?jumpid=ps_knpzbvc1th_aid-520061736&ef_id=Cj0KCQjwwJuVBhCAARIsAOPwGARYja6ICT1A14XgqLgCwLDlpRPoM_wy2pYuTsAUjXy1cZon2Qwd8SIaAmw7EALw_wcB:G:s&s_kwcid=AL!13472!3!541194912149!p!!g!!deep%20learning!14386686690!127123177835&
21. Image Caption Generation using Convolutional Neural Network and LSTM
22. https://www.slideshare.net/OmkarReddy7/image-caption-generation-using-convolutional-neural-network-and-lstm

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.165 |

|| Volume 10, Issue 6, June 2022 ||

| DOI: 10.15680/IJIRCCE.2022.1006129 |

23. A Survey on Various Deep Learning Models for Automatic Image Captioning https://iopscience.iop.org/article/10.1088/1742-6596/1950/1/012045/pdf
24. Far id Melgani, Genc Hoxha, Jacopo Slaghenauffi. (2020). A New CNN-RNN Framework for Remote Sensing Image Captioning. 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS).
25. Ansar Hani, Najiba Tagougui, Monji Kherallah. (2019). Image Caption Generation Using A Deep Architecture. 2019 International Arab Conference on Information Technology (ACIT) deep.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details