International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT

IJIEMR Transactions, online available on 16th Mar 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 03

10.48047/IJIEMR/V12/ISSUE 03/16

Title **AN AUTOMATIC IMAGE CAPTION GENERATION APPRAOCH USING LSTM AND CNN**

Volume 12, ISSUE 03, Pages: 122-128

Paper Authors
K. SAI CHARAN LAHIRI, M. ANITHA LAKSHMI, P. PREM KUMAR,SK. ALTAF, M. YASWANTH KUMAR, SLVVD SARMA

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

AN AUTOMATIC IMAGE CAPTION GENERATION APPRAOCH USING LSTM AND CNN

[1]K. SAI CHARAN LAHIRI, [2]M. ANITHA LAKSHMI, [3]P. PREM KUMAR,[4]SK. ALTAF, [5]M. YASWANTH KUMAR, [6]SLVVD SARMA

[1,2,3,4,5] B. Tech Students, Dept of CSE, Kallam Haranadhareddy Institute of Technology, Chowdavaram, Guntur, A.P, India.

[6]Assistant Professor, Dept of CSE, Kallam Haranadhareddy Institute of Technology Chowdavaram, Guntur, A.P, India.

ABSTRACT: Automatic image caption generation is one of the frequent goals of computer vision. Image description generation models must solve a larger number of complex problems to have this task successfully solved. The objects in the image must be detected and recognized, after which a logical and syntactically correct textual description is generated. For that reason, description generation is a complex problem. It is an extremely important challenge for machine learning algorithms because it represents an impersonation of a complicated human ability to encapsulate huge amounts of highlighted visual pieces of information in descriptive language. As the deep learning techniques are growing, huge datasets and computer power are helpful to build models that can generate captions for an image. Hence in this work, an automatic image caption generation approach using LSTM and CNN is presented. In this project, deep learning techniques like CNN (Convolutional Neural Network) and LSTM (Long Short Term Memory) are used to identify the caption of the image. Image caption generator is a process which involves natural language processing and computer vision concepts to recognize the context of an image and present it in English. In this project, some of the core concepts of image captioning and its common approaches are followed. Keras library, numpy and jupyter notebooks are used for making of this project. This project also discusses about flickr_dataset and CNN used for image classification.

KEYWORDS: Automatic Image Caption Generation, Deep Learning (DL), Convolutional Neural Network) and Long Short Term Memory (LSTM).

I. INTRODUCTION

As we are living in the 21st century, image caption is one of the most needed tools these days. With the rise of users in internet the number of images and videos This data is usually unstructured and raw data which does not carry much information. Whether to extract and use the right data from the internet or to organize the files captioning plays a huge role. This application has a built-in function for producing captions for a specific image. Image captioning means automatic generation of a caption for an image [1].

As a recently emerged research area, it is attracting more and more attention. Nowadays, an image caption generator has become the need of the hour, be it for social media enthusiasts or visually impaired people. It can be used as a plug in in currently trending social media platforms to recommend suitable captions for people to attach to their post or can be used by visually impaired people to understand the image content on the web. Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications [4].

To achieve the goal of image captioning, semantic information of images needs to be captured and expressed in natural languages. Connecting both research communities of computer vision and natural language processing, image captioning is a quite challenging task. Various approaches have been described to solve this problem. The number of digital

images increases rapidly; hence, categorizing these images and retrieving the relevant web images are a difficult process. For people to use numerous images effectively on the web, technologies must be able to explain image contents and must be capable of searching for data that users need. Moreover, images must be described with natural sentences based not only on the names of objects contained in an image, but also on their mutual relations.

Photo captions aim to describe objects, actions, and details found in an image using natural language. Most image caption research focuses on single-sentence captions, but the descriptive capabilities of this form are limited; one sentence can only describe in detail a small part of an image [2]. This task of automatically generating captions and describing the image is significantly harder than image classification and object recognition. The description of an image must involve not only the objects in the image, but also relation between the objects with their attributes and activities shown in images. Most of the work done in visual recognition previously has concentrated to label images with already fixed classes or categories leading to the large progress in this field. Eventually, vocabularies of visual concepts which are closed, makes a suitable and simple model for assumption.

Automatic caption generation for an image is one of the challenging problems in artificial intelligence. Image captioning models not only solve computer vision challenges of object recognition but also capture and express their relationships in natural language. This task is more complicated as compared to well-studied image classification and object recognition tasks, which have been the main focus in the computer vision community [5].

Recent advancements in language modeling and object recognition have made image captioning an essential research area in computer vision and natural language processing. Caption generation of an image has a great impact by helping visually impaired people to better understand the contents on the web [3]. Automatic caption generation is a tough undertaking that can aid visually challenged persons in understanding the content of web images. It may also have a significant impact on search engines and robots. This problem is substantially more difficult than image categorization or object recognition, both of which have been extensively researched.

Recently, deep learning methods have achieved state-of the-art results on examples of this problem. It has been demonstrated that deep learning models are able to achieve optimum results in the field of caption generation problems. Hence in this work, an automatic image caption generation approach using CNN and LSTM is presented in this work. The rest of the work s organized as follows: The section II demonstrates literature survey. The section III presents an automatic image caption generation approach using LSTM and CNN. The section IV evaluates the result analysis of presented approach. Finally the work is concluded in section V.

II. LITERATURE SURVEY

Xiangqing Shen, Bing Liu, Yong Zhou & Jiaqi Zhao et. al., [7] describes Remote sensing image caption generation via transformer and reinforcement learning. A new model using the Transformer to decode the image features to target sentences is presented. For making the Transformer more adaptive to the remote sensing image captioning task, we additionally employ dropout layers, residual connections, and adaptive feature fusion in the Transformer. Reinforcement

Learning is then applied to enhance the quality of the generated sentences. We demonstrate the validity of our proposed model on three remote sensing image captioning datasets. This model obtains all seven higher scores on the Sydney Dataset and Remote Sensing Image Caption Dataset (RSICD), four higher scores on UCM dataset, which indicates that the proposed methods perform better than the previous state of the art models in remote sensing image caption generation.

Songtao Ding, Shiru Qu, Yuling Xi, Arun Kumar Sangaiah, Shaohua Wan et. al., [8] describes Image caption generation with high-level image features. A novel image captioning model based on high-level image features is presented. They combine low-level information, such as image quality, with high-level features, such as motion classification and face recognition to detect attention regions of an image. We demonstrate that our attention model produces good performance in experiments on MSCOCO, Flickr 30K, PASCL and SBU datasets. This approach gives good performance on benchmark datasets.

Xinlei Chen, C. Lawrence Zitnick et. al., [9] describes learning a Recurrent Visual Representation for Image Caption Generation. A novel recurrent visual memory is presented that automatically learns to remember long-term visual concepts to aid in both sentence generation and visual feature reconstruction. Authors evaluated this approach on several tasks. These include sentence generation, sentence retrieval and image retrieval. State-of-the-art results are shown for the task of generating novel image descriptions. When compared to human generated captions, our automatically generated captions are preferred by humans over 19.8% of the time. Results are better than or comparable to state-of-the-art results on the image and sentence retrieval tasks for methods using similar visual features.

Philip Kinghorn, Li Zhang, Ling Shao et. al., [10] presents A region-based image caption generator with refined descriptions. A novel region-based deep learning architecture for image description generation is presented. It employs a regional object detector, recurrent neural network (RNN)-based attribute prediction, and an encoder–decoder language generator embedded with two RNNs to produce refined and detailed descriptions of a given image. Most importantly, the proposed system focuses on a local based approach to further improve upon existing holistic methods, which relates specifically to image regions of people and objects in an image. Evaluated with the IAPR TC-12 dataset, the proposed system shows impressive performance and outperforms state-of-the-art methods using various evaluation metrics

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio et. al., [11] describes Neural Image Caption Generation with Visual Attention. An attention based model is described that automatically learns to describe the content of images. Authors described how this model is trained in a deterministic manner using standard back-propagation techniques and stochastically by maximizing a variational lower bound. They also show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. They validate the use of attention with state-of-theart performance on three benchmark datasets: Flickr9k, Flickr30k and MS COCO.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

III. AUTOMATIC IMAGE CAPTION GENERATION APPROACH

An automatic image caption generation approach using LSTM and CNN is presented in this section. The main objective of this project is to develop a web based interface for users to get the description of the image and to make a classification system in order to differentiate images as per their description. It can also make the task easier which is complicated as they have to maintain and explore enormous amounts of data. The fig. 1 shows the system architecture of automatic caption generation approach using CNN and LSTM.

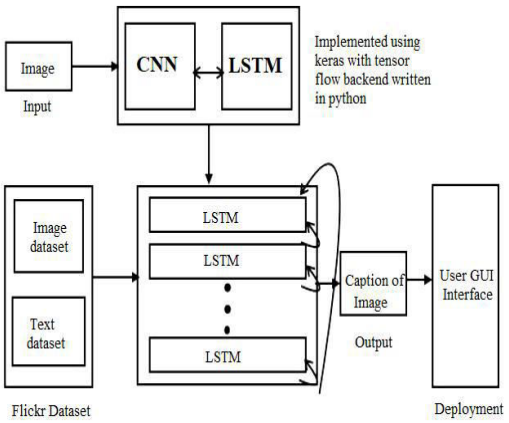


Fig. 1: System Architecture of Automatic Caption Generation Approach using CNN and LSTM

Here, a dataset called flickr8k which is collected from kaggle. Flickr8k dataset is a public benchmark dataset for image to sentence description. This dataset consists of 8000 images with five captions for each image. These images are extracted from diverse groups in Flickr website. Each caption provides a clear description of entities and events present in the image. The dataset depicts a variety of events and scenarios and doesn't include images 37 containing well known people and places which makes the dataset more generic. The dataset has 6000 images in training dataset, 1000 images in development dataset and 1000 images in test dataset. Features of the dataset making it suitable for this project are: Multiple captions mapped for a single image makes the model generic and avoids overfitting of the model. Diverse category of training images can make the image captioning model to work for multiple categories of images and hence can make the model more robust. Dataset is collected from various source of internet.

Data preprocessing is done in this step includes Data cleaning, Data reduction, Image data preparation. For instance, punctuations, digits, single length words are removed from the text dataset. Two deep learning models have been selected i.e, CNN and LSTM. Firstly, CNN takes image as input and extract features such as background, objects in the image.

CNN stands for Convolutional Neural Networks. It is a deep learning algorithm which takes image as an input.CNN scans images from left to right and top to bottom to pull out important features from the image and combines the features to classify images. Pre-processing required in convolutional neural networks is much lower as compared to other classification algorithms.

The architecture of a Conv Net is analogous to that of the connectivity pattern of Neurons in the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons respond to stimuli only in a restricted region of the visual field known as the Receptive Field. A collection of such fields overlaps to cover the entire visual area. The CNNs are inspired by visual system of human brain. The idea behind the CNNs thus is to make the computers capable of viewing the world as humans view it. This way CNNs

can be used in the fields of image recognition and analysis, image classification, and natural language processing.

CNN is a type of deep neural networks which contain the convolutional, max pooling, and activation layers. The convolutional layer, considered as a main layer of a CNN, performs the operation called "convolution" that gives CNN its name. Kernels in the convolutional layer are applied to the layer inputs. All the outputs of the convolutional layers are convolved as a feature map. In this study, the Rectified Linear Unit (ReLU) has been used in the activation function with a convolutional layer which is helpful to increase the non-linearity in input image, as the images are fundamentally nonlinear in nature.

The pooling layer is an important building block of CNN. Pooling can be the max, average, and sum in the CNN model. In this study, max pooling has been used because others may not identify the sharp features easily as compared to max pooling. The dropout layer has also been used, which drops the neurons during the training chosen at random to reduce the overfitting problem. CNN is used to extract features from the image. A pre-trained model called Xception is used for this.CNN can be used in the fields of image recognition, image classification, and natural language processing.

LSTM stands for long short-term memory, it is a type of RNN (Recurrent Neural Networks. LSTM is capable of working with sequence prediction problems. It is used for word prediction purposes. In LSTM based on previous text, one can predict what the next word will be. It is the same as google search where this system will show the next word based on our previous text. LSTM can carry out relevant information throughout the process with a forget gate and discards non-relevant information. LSTM will use the information that is extracted from CNN to generate a description of the image.

The CNN LSTM architecture involves using Convolutional Neural Network (CNN) layers for feature extraction on input data combined with LSTMs to support sequence prediction. This architecture was originally referred to as a Long-term Recurrent Convolutional Network or LRCN model, although we will use the more generic name "CNN LSTM" to refer to LSTMs that use a CNN as a front end in this lesson. This architecture is used for the task of generating textual descriptions of images. Key is the use of a CNN that is pre-trained on a challenging image classification task that is re-purposed as a feature extractor for the caption-generating problem.

A pre trained model called xception is used to train LSTM which will generate captions. Features which were extracted by CNN are given to LSTM. This LSTM will generate the captions for the given image. Once all these steps were implemented, a caption will be generated for the given image. The generated captions are displayed on screen using GUI (Graphical User Interface). This whole project is implemented using keras with tensorflow backend written in python.

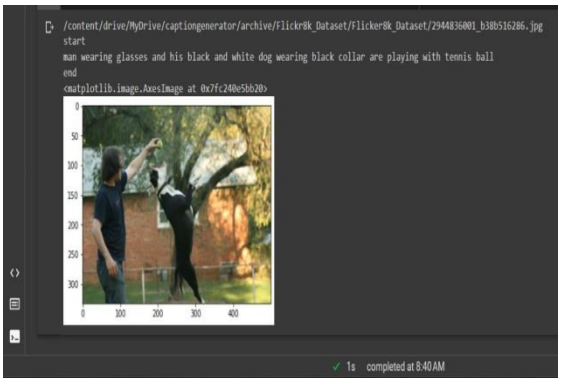
IV. RESULT ANALYSIS

In this section, an automatic image caption generation approach using LSTM and CNN is implemented using python. The result analysis of presented approach is evaluated here. In this analysis, flickr8k dataset is used. The CNN is used to extract the features and LSTM is used to generate the caption.

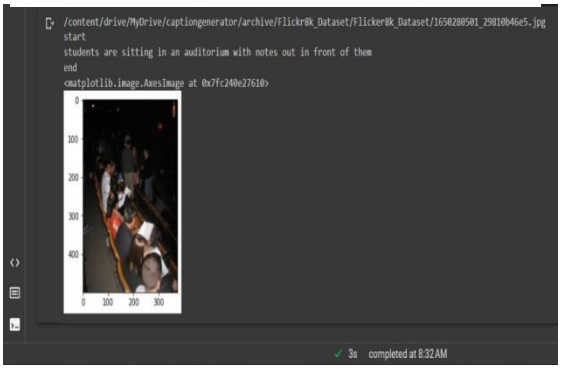
In this approach, firstly user upload an image as input this image is forwarded to CNN in which it extracts the features such as background, scene, objects in the image

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

using convolutional layer and pooling layer then these features are sent to LSTM by using fully connected layer. Now, the dataset which contains Image Dataset and text dataset is preprocessed to form an training model. This training model is used to train LSTM for which it generates captions. The user has to upload an image for which the caption has to be generated. The generated caption for the image is viewed by the user. The uploaded images and their captions are shown in following figures.

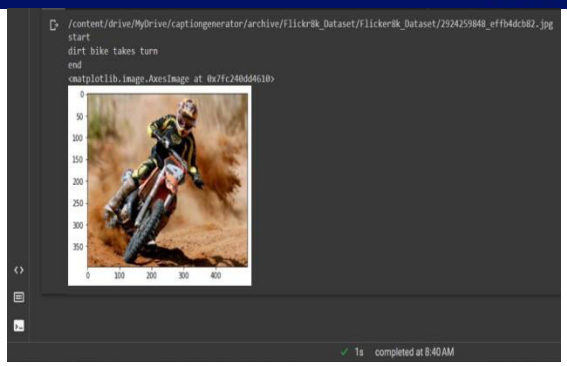


(a)



(b)

(c)



(d)

Fig. 2 (a), (b), (c) & (d): Uploaded imaged and their generated Captions

Hence this approach has generated the captions effectively and automatically for different images.

V. CONCLUSION

In this work, an automatic image caption generation approach LSTM and CNN is presented. This approach used Flickr_8k dataset which includes nearly 8000 images, and the corresponding captions are also stored in the text file. The deep learning models, Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) are employed in this analysis. The images are preprocessed and applied to CNN. CNN extracts the features like background, scene, objects in the image using convolutional layer and pooling layer. The extracted features are sent to LSTM by using fully connected layer. The LSTM generates the captions for the uploaded images. This approach has automatically generated the captions for different images. The generated captions are more accurate than state-of art approaches. The scope of image-captioning is very vast in the future as the users are increasing day by day on social media and most of them would post photos. So this project will help them to a greater extent.

VI. REFERENCES

[1] Peerzada Salman syeed, Dr.Mahmood Usman, "Image Caption Generator Using Deep Learning", Neuroquantology, October 2022, Volume 20, Issue 12, Page 2682-2691, Doi: 10.14704/Nq.2022.20.12.Nq77261

[2] Dr. P. Srinivasa Rao, Thipireddy Pavankumar, Raghu Mukkera, Gopu Hruthik Kiran, Velisala Hariprasad, "Image Caption Generation Using Deep Learning Technique", International Research Journal of Modernization in Engineering Technology and Science, Volume:04/Issue:06/June-2022, e-ISSN: 2582-5208

[3] Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha and Pushpak Bhattacharyya, "A Hindi Image Caption Generation Framework Using Deep Learning", ACM Trans. Asian Low-Resour. Lang. Inf. Process., 2021,Vol. 20, No. 2, Article 32.

[4] Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati, "Image Caption Generating Deep Learning Model", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 10 Issue 09, September-2021

[5] Moksh Grover, Rajat Rathi Chinkit, Kanishk Garg, Ravinder Beniwal, "AI Optics: Object recognition and caption generation for Blinds using Deep Learning Methodologies", 2021 International Conference on Computing, Communication, and Intelligent Systems(ICCCIS), DOI: 10.1109/ICCCIS51004.2021.9397143

[6] Omkar Nitin Shinde, Rishikesh Gawde, Anurag Paradkar, "Social Media Image Caption Generation Using Deep Learning", International Journal of Engineering Development and Research, 2020, Volume 8, Issue 4, ISSN: 2321-9939

[7] Xiangqing Shen, Bing Liu, Yong Zhou & Jiaqi Zhao, "Remote sensing image caption generation via transformer and reinforcement learning", Multimedia Tools and Applications, volume 79, pages26661–26682 (2020), doi: 10.1007/s11042-020-09294-7

[8]Songtao Ding, Shiru Qu, Yuling Xi, Ar un Kumar Sangaiah, Shaohua Wan, "Image caption generation with high-level image features", Pattern Recognition Letters, Volume 123, 15 May 2019, Pages 89-95, Elsevier, doi: 10.1016/j.patrec.2019.03.021

[9] Xinlei Chen, C. Lawrence Zitnick, "Learning a Recurrent Visual Representation for Image Caption Generation", Computer Vision and Pattern Recognition (cs.CV), arXiv:1411.5654v1 , doi:10.48550/arxiv1411.5654

[10]Philip Kinghorn, Li Zhang, Ling Shao, "A region-based image caption generator with refined descriptions", Neurocomputing, Volume 272, 10 January 2018, Pages 416-424, Elsevier, Doi:10.1016/j.neucom.2017.07.014

[11] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, "Neural Image Caption Generation with Visual Attention", Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.