INTERNATIONAL JOURNAL ENGINEERING RESEARCH

International Journal for Innovative Engineering and Management Research

PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL

www.ijiemr.org

IJIEMR Transactions, online available on 21st Feb 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 03

10.48047/IJIEMR/V12/ISSUE 02/88

Title **DEEP LEARNING BASED AUTOMATED IMAGE CAPTION GENERATOR**

Volume 12, ISSUE 02, Pages: 573-579

Paper Authors

B.Kawshik, G.Sai Mahesh, M.Sai kumar,
Jonnadula Narasimharao

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

DEEP LEARNING BASED AUTOMATED IMAGE CAPTION GENERATOR

B.Kawshik[1], G.Sai Mahesh[2], M.Sai kumar[3],
Jonnadula Narasimharao[4]
[1,2,3] B.Tech Student, Department of Computer Science and Engineering,
CMR Technical Campus,Medchal, Hyderabad,Telangana, India,
[1]kbhyroju@gmail.com, [2]gajula797@gmail.com, [3]saikumar10684@gmail.com,
[4]Associate Professor, Department of Computer Science and Engineering,
CMR Technical Campus, Medchal, Hyderabad, Telangana, India,
jonnadula.narasimharao@gmail.com

ABSTRACT: In the past few years, the problem of generating descriptive sentences automatically for images has garnered a rising interest in natural language processing and computer vision research. An image caption is something that describes an image in the form of text. It is widely used in programs where one needs information from any image in automatic text format. Image captioning is a fundamental task which requires semantic understanding of images and the ability of generating description sentences with proper and correct structure. With the exponential development in the field of artificial intelligence in recent years, many researchers have focused their attention towards the topic of image caption generation. With advanced deep learning techniques, accessibility of big datasets and computer power one can build an efficient model to generate captions. Hence, in this work Deep Learning based Automated image Caption Generator is presented. The model is trained in such a way that if input image is given to model it generates captions which nearly describes the image. In this approach, two deep learning algorithms like LSTM (Long Short Term Memory) and CNN (Convolutional Neural Networks) are used. Feature extraction is done first and then captions are generated. The flickr_8k dataset is used for training the model. The dataset which we are using contains 8000 images and each image is mapped with five different captions.

KEYWORDS: Image Caption Generator, Convolutional Neural Network (CNN), Long Short Term Network (LSTM).

I. INTRODUCTION

Image captioning is the process of generating a textual description of an image that aims to describe the salient parts of the given image.

Image caption includes the multi-level use of image information. From the target in the image, the relationship between the targets, to the description of the image, and the construction of the scene graph, all belong to the category of image description research. Each task in image caption has great research value and great practical application value [4].

Caption generation is an interesting artificial intelligence problem where a descriptive sentence is generated for a given image. It involves the dual techniques from computer vision to understand the content of the image and a language model from the field of natural language processing to turn the understanding of the image into words in the right order. 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.

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 [2].Gesture recognition becomes a popular analytics tool for extracting the characteristics of user movement and enables numerous practical applications in the biometrics field. Despite recent advances in this technique, complex user interaction and the limited amount of data pose serious challenges to existing methods [1].

With the growing advancement in in-depth reading techniques, the availability of large databases, and the ability to integrate, models are often developed to produce image captions. Image captioning is a process that involves image processing and natural language processing concepts to identify the context of an image and interpret it in a natural language such as English or any other language. Image caption generator incorporates the concept of reconfiguring an image and interpreting it in the native language such as English [3].

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 [5].

Image caption methods can be divided into template-based methods, retrieval-based methods and deep learning-based methods. The template-based method first obtains some visual concepts for the image, and then generates a sentence through sentence templates, syntactic rules, or combined methods. Retrieval-based methods usually need to save a large database, and then obtain a sentence or a group of sentences through image retrieval, and then obtain a complete image description. The image description based on deep learning mainly uses the structure of codec to complete the image caption task. Further, the effect of image description can be improved through the attention machine or other methods of enhancing the deep learning model.

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, deep learning based automated image caption generator is presented. The rest of the work s organized as follows: The section II demonstrates literature survey. The section III presents deep learning based automated image caption generator.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., [6] 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. 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.

Nayan Mehta, Suraj Pai,Sanjay Singh et. al., [7] describes Automated 3D sign language caption generation for video. This paper aims to present a useful technology that can be used to leverage online resources and make them accessible to the hearing-impaired community in their primary mode of communication. First, the video gets converted to text via subtitles and speech processing methods. The generated text is understood through NLP algorithms and then mapped to avatar captions which are then rendered to form a cohesive video alongside the original content. We validated our results through a 6-month period and a consequent 2-month study, where we recorded a 37% and 70% increase in performance of students taught using Sign captioned videos against student taught with English captioned videos.

Chetan Amritkar, Vaishali Jabade et. al., [8] describes Image Caption Generation using Deep Learning Technique.This model is used to generate natural sentences which eventually describe the image. This model consists of Convolutional Neural Network(CNN) as well as Recurrent Neural Network(RNN). The CNN is used for feature extraction from image and RNN is used for sentence generation. The model is trained in such a way that if input image is given to model it generates captions which nearly describes the image. The accuracy of model and smoothness or command of language model learns from image descriptions are tested on different datasets. These experiments show that model is frequently giving accurate descriptions for an input image.

Philip Kinghorn, Li Zhang, Ling Shao et. al., [9] 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 presneted 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., [10] 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.

III. DEEP LEARNING BASED AUTOMATED IMAGE CAPTION GENERATOR

In this section, Deep learning based automated image caption generator is presented. The block diagram of presented approach is shown in Fig. 1.

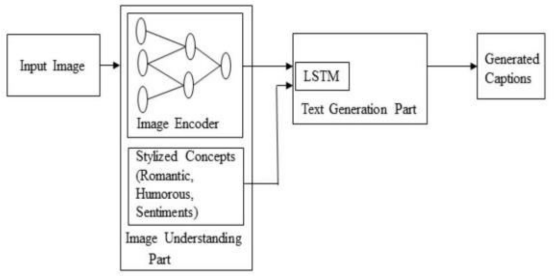


Fig. 1:Architecture of Deep learning based automated image caption generator

Image captioning isa more complicated but meaningful task in the age of artificial intelligence. In this an Image caption generator, basis on the provided oruploaded image file it will generate the caption from a trained model which is trainedusing algorithms and on a large dataset. The main idea behind this is that users will get automated captions when they use or implement it on social media or on any applications.

This project is totally based upon to generate the relevant natural languagecaption to the given input image, instead of just describing a single target object themodel detects multiple target objects for generating grammatically correct caption. Wehave used flickr_8k data set .The given input image is pre-processed.The flickr_8k dataset is used for training the model. The dataset which weare using contains 8000 images and each image is mapped with five differentcaptions.These applications in image captioning have important theoretical andpractical research value. Image captioning is a more complicated but meaningful task inthe age of artificial intelligence.

Given a new image, an image captioning algorithmshould output a description about this image at a semantic level. In this an Imagecaption generator, basis on our provided or uploaded image file It will generate thecaption from a trained model which is trained using algorithms and on a large dataset.The main idea behind this is that users will get automated captions when we use orimplement it on social media or on any applications.

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. Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. Feature extraction helps to reduce the amount of redundant data from the data set. In the end, the reduction of the data helps to build the model with less machine effort and also increases the speed of learning and generalization steps in deep learning process.

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.In this approach, CNN (Convolutional Neural Networks)and LSTM are used to generate captions for this Python-based application (Long Short TermMemory). The photograph features will be taken from VGG16, a CNN version trainedon the image net dataset, and then fed into the LSTM model, which will be responsible for creating the photograph captions. Convolutional neural networks are deep neural networks that have been customized to process data in the form of a second matrixPhotographs are easy. It is able to handle the photos that have been translated, circled,scaled and modifications in angle.

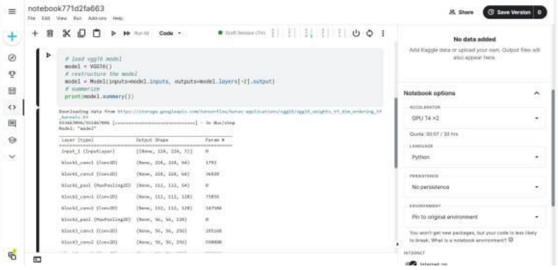
Long ShortTerm Memory (LSTM) is a form of RNN(Recurrent Neural Network) that excels at sequence prediction. Based on the prior textual content, one can estimate what the next phrase will be.

LSTM may performrelevant statistics throughout the input processing and, using an overlook gate, it caneliminate irrelevant data. So that it will be easy for the user to recognize the image inless time.

Convolutionneural network (CNN) is used to extract features from the provided input image .Theinformation from the CNN is used by the LSTM for generating the relevant caption forthe given image. The proposed model is to generate the relevant natural languagecaption to the given input image, instead of just describing a single target object themodel detects multiple target objects for generating grammatically correct caption.

IV. RESULT ANALYSIS

Deep learning based automated image caption generator is implemented in this section. The result analysis of presented approach is evaluated in this section.This approach includes different phases, feature extraction, image extraction, data pre-processing and image caption generation. These process implemented results are shown as follows: The Fig. 2 shows the exacting features.

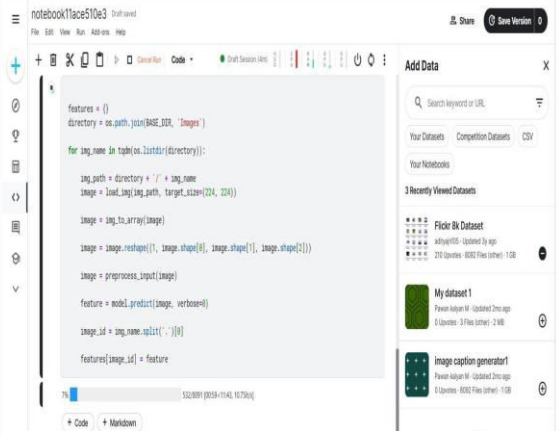


```
# load vgg16 model
model = VGG16()
# restructure the model
model = Model(inputs=model.inputs, outputs=model.layers[-2].output)
# summarize
print(model.summary())
```

```
Downloading data from https://storage.googleapis.com/tensorflow/keras-applications/vgg16/vgg16_weights_tf_dim_ordering_tf_kernels.h5
553467096/553467096 [==============================] - 3s 0us/step
Model: "model"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 224, 224, 3)]     0

 block1_conv1 (Conv2D)       (None, 224, 224, 64)      1792

 block1_conv2 (Conv2D)       (None, 224, 224, 64)      36928

 block1_pool (MaxPooling2D)  (None, 112, 112, 64)      0

 block2_conv1 (Conv2D)       (None, 112, 112, 128)     73856

 block2_conv2 (Conv2D)       (None, 112, 112, 128)     147584

 block2_pool (MaxPooling2D)  (None, 56, 56, 128)       0

 block3_conv1 (Conv2D)       (None, 56, 56, 256)       295168

 block3_conv2 (Conv2D)       (None, 56, 56, 256)       590080
```

Fig. 2:Features Extraction Process

The Fig. 3 shows the images extraction phase.

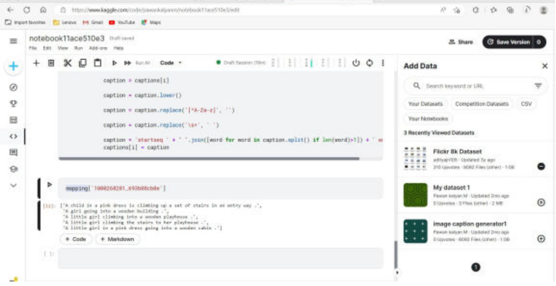


```
features = {}
directory = os.path.join(BASE_DIR, 'Images')

for img_name in tqdm(os.listdir(directory)):

    img_path = directory + '/' + img_name
    image = load_img(img_path, target_size=(224, 224))

    image = img_to_array(image)

    image = image.reshape((1, image.shape[0], image.shape[1], image.shape[2]))

    image = preprocess_input(image)

    feature = model.predict(image, verbose=0)

    image_id = img_name.split('.')[0]

    features[image_id] = feature
```

```
7%  532/8091 [00:59<11:43, 10.75it/s]
```

Fig. 3: Image Extraction Process

The Fig. 4 shows the data pre-processing phase implementation.



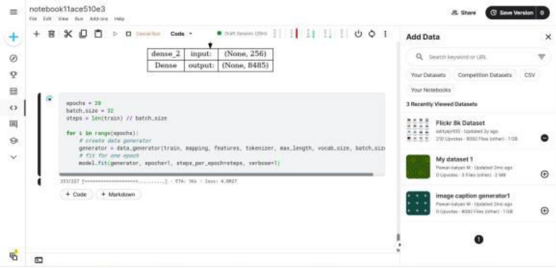
```
            caption = captions[i]

            caption = caption.lower()

            caption = caption.replace('[^A-Za-z]', '')

            caption = caption.replace('\s+', ' ')

            caption = 'startseq ' + " ".join([word for word in caption.split() if len(word)>1]) + ' e
            captions[i] = caption
```

```
mapping['1000268201_693b08cb0e']
```

```
[11]: ['A child in a pink dress is climbing up a set of stairs in an entry way .',
       'A girl going into a wooden building .',
       'A little girl climbing into a wooden playhouse .',
       'A little girl climbing the stairs to her playhouse .',
       'A little girl in a pink dress going into a wooden cabin .']
```

Fig. 4: Data-Pre-processing Process results

The Fig. 5 shows the training phase of presented approach.



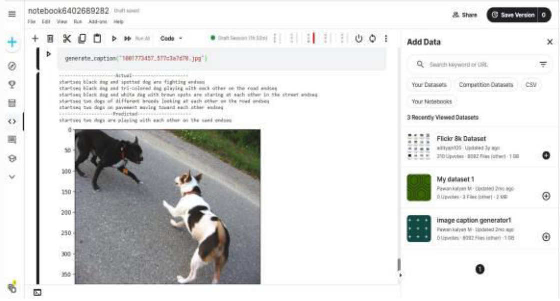
dense_2	input:	(None, 256)
Dense	output:	(None, 8485)

```
epochs = 20
batch_size = 32
steps = len(train) // batch_size

for i in range(epochs):
    # create data generator
    generator = data_generator(train, mapping, features, tokenizer, max_length, vocab_size, batch_size
    # fit for one epoch
    model.fit(generator, epochs=1, steps_per_epoch=steps, verbose=1)
```

```
153/227 [===================>..........] - ETA: 36s - loss: 4.0027
```

Fig. 5: Training Phase

The Fig. 5 shows the output of trained presented mode i.e. image caption generation.



```
generate_caption("1001773457_577c3a7d70.jpg")
```

```
---------------------Actual---------------------
startseq black dog and spotted dog are fighting endseq
startseq black dog and tri-colored dog playing with each other on the road endseq
startseq black dog and white dog with brown spots are staring at each other in the street endseq
startseq two dogs of different breeds looking at each other on the road endseq
startseq two dogs on pavement moving toward each other endseq
--------------------Predicted--------------------
startseq two dogs are playing with each other on the sand endseq
```

Fig. 5: Image Caption Generation

Hence, this approach has generated the captions for different images very effectively and accurately.

V. CONCLUSION

In this work,Deep learning based automated image caption generator is presented.Generating an appropriate and grammatically correct caption in a naturallanguage like a human is a difficult task ,this task involves feature extraction and natural language processing concepts. CNN and LSTM have worked well together in synchronization, they were able to find a connection between objects in pictures.In this analysis, The flickr_8k dataset is used for training the model which contains 8000 images and each image is mapped with five differentcaptions. The Convolutional Neural Network is used to extract features from the provided input image and Long Short Term memory algorithm is used to generate the relevant caption forthe given image. From the results,it can beconcluded that this model can be used to generate image captions for multiple images in real time.

VI. ACKNOWLEDGEMENT

We thank CMR Technical Campus for supporting this paper titled “Deep Learning based Automated Image Caption Generator”, which provided good facilities and support to accomplish our work. Sincerely thank our Chairman, Director, Deans, Head Of the Department, Department Of Computer Science and Engineering, Guide and Teaching and Non- Teaching faculty members for giving valuable suggestions and guidance in every aspect of our work

VII. REFERENCES

[1]Hao Zhou, Wei Huang, Zhuo Xiao, Shichuan Zhang, Wangzhan Li, Jinhui Hu, Tianxing Feng, Jing Wu, Pengcheng Zhu, Yanchao Mao, “Deep-Learning-Assisted Noncontact Gesture-Recognition System for Touchless Human-Machine Interfaces”, Advanced Functional Materials, 2022,doi:10.1002/adfm.202208271

[2]SaiTeja. N.R, Rashmitha Khilar, “Implementing Complexity in Automatic Image Caption Generator using Recurrent Neural Network over Long Short-Term Memory”, Journal of Pharmaceutical Negative Results, Volume13,Special Issue , 2022

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

[4] SiZhen Li, Linlin Huang, “Context-based Image Caption using Deep Learning”, 2021 IEEE 6th International Conference on Intelligent Computing and Signal Processing (ICSP 2021), DOI: 10.1109/ICSP51882.2021.9408871

[5] Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, And Pushpak Bhattacharyya, “A Hindi Image Caption Generation Framework Using Deep Learning”, Trans. Asian Low-Resour. Lang. Inf. Process. 20, 2, Article 32 (March 2021), 19 pages, doi:10.1145/3432246

[6] 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

[7] Nayan Mehta, Suraj Pai,Sanjay Singh, “Automated 3D sign language caption generation for video”,Universal Access in the Information Society volume 19, pages725–738 (2020)

[8] Chetan Amritkar, Vaishali Jabade, “Image Caption Generation using Deep Learning Technique”,2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018 IEEE, 978-1-5386-5257-2/18

[9]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

[10] 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.