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.2022.1006110 |

Image Caption Generation Using Hybrid CNN-LSTM Model

P. Dinesh Kumar [1], Harish.V[2], Y. Basheer Ahmed[3], M. Mohamed Omar Jafran[4], A.J. Adaikalaraj[5]

Assistant Professor, Dept. of C.S.E, Saranathan College of Engineering, Trichy, India[1]

U.G Students, Dept. of C.S.E, Saranathan College of Engineering, Trichy, India [2,3,4,5]

ABSTRACT: Image captioning has become one of the most widely required tools. The process of generating a description of an image is called image captioning. It requires recognizing the important objects, their attributes, and the relationships among the objects in an image. It generates syntactically and semantically correct sentences. This paper aims to detect different objects found in an image, recognize the relationships between those objects and generate captions using Transfer Learning and Supervised learning to generate semantically and syntactically correct captions for new unseen images.

KEYWORDS: TRANSFER LEARNING; SUPERVISED LEARNING; CONVOLUTIONAL NEURAL NETWORKS; RECURRENT NEURAL NETWORK; LONG SHORT-TERM MEMORY; VISUAL GEOMETRY GROUP

I. INTRODUCTION

Every day, we encounter a large number of images from various sources such as the internet, news articles, document diagrams, and advertisements. These sources contain images that viewers would have to interpret themselves. Most images do not have a description, but humans can largely understand them without their detailed captions. However, the machine needs to interpret some form of image captions if humans need automatic image captions from it. Image captioning is important for many reasons. Captions for every image on the internet can lead to faster and more descriptively accurate image searches and indexing. Ever since researchers started working on object recognition in images, it became clear that only providing the names of the objects recognized does not make such a good impression as a full human-like description. Image caption generation is a task that involves image processing and natural language processing concepts to recognize the context of an image and describe them in a natural language like English or any other language Image captioning has various applications in various fields such as biomedicine, commerce, web searching, military, etc. Social media like Instagram, Facebook, etc

II. RELATED WORK

The[5]Image captioning has recently gathered a lot of attention specifically in the natural language domain. There is a pressing need for context based natural language description of images, however, this may seem a bit far fetched but recent developments in fields like neural networks, computer vision and natural language processing has paved a way for accurately describing images i.e. representing their visually grounded meaning. We are leveraging state-of-the-art techniques like Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and appropriate datasets of images and their human perceived description to achieve the same. We demonstrate that our alignment model produces results in retrieval experiments on datasets such as Flicker.[1]Template-based approaches have fixed templates with a number of blank slots to generate captions. In these approaches, different objects, attributes, actions are detected first and then the blank spaces in the templates are filled. For example, Farhadi et al. use a triplet of scene elements to fill the template slots for generating image captions. Li et al. extract the phrases related to detected objects, attributes and their relationships for this purpose. A Conditional Random Field (CRF) is adopted by Kulkarni et al. to infer the objects, attributes, and prepositions before filling in the gaps. Template-based methods can generate grammatically correct captions. However, templates are predefined and cannot generate variable-length captions. Moreover, later on, parsing based language models have been introduced in image captioning which are more powerful than fixed template-based methods. Therefore, in this paper, we do not focus on these template based methods.[3] Convolutionalnetworks were inspired by biological processes in that the connectivity pattern between neurons resembles the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field. CNNs use relatively little pre-processing compared to other image classification algorithms.The[2] Convolutional Neural Network consists of an input and an output layer,

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

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

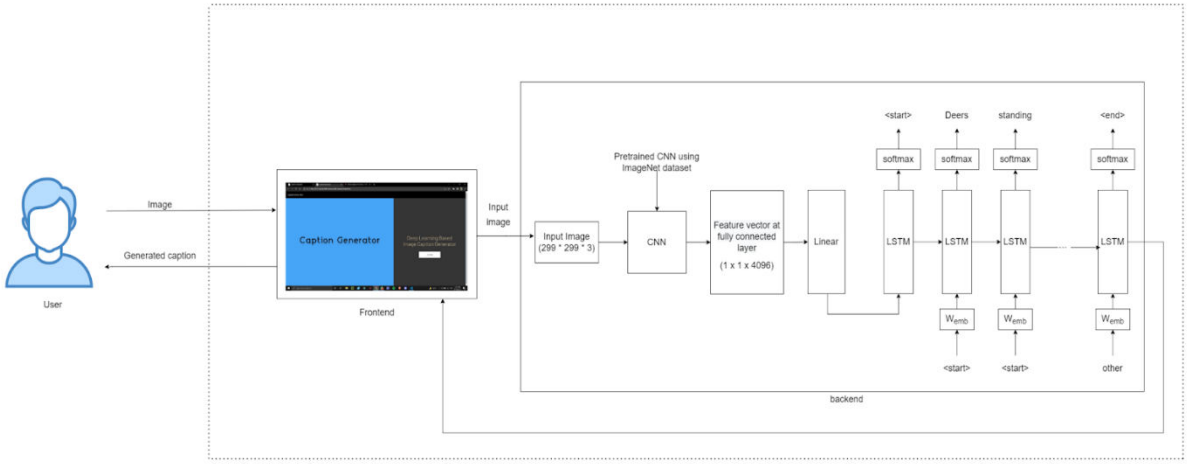
| DOI: 10.15680/IJIRCCE.2022.1006110 |

as well as multiple hidden layers. The hidden layers of a CNN typically consist of convolutional layers, RELU layer i.e. activation function, pooling layers, fully connected layers, and normalization layers.[4] LSTM (long-short-term memory) was developed from RNN, with the intention to work with sequential data. It is now considered the most popular method for image captioning due to its effectiveness in memorizing long-term dependencies through a memory cell.Usually, the vocabulary size might vary from 10,000 to 40,000 words, while their model relies on 258 words. The decrease is quite sharp—reduced by 39 times if compared to 10,000, but the results are high, with some space for improvements.

III. PROPOSED ALGORITHM

A. Design Considerations:

- Remove Noise from the Dataset
- Extract feature vector from the training images using different pre-trained CNN models
- Tokenize the captions for training
- Build,train and evaluate the LSTM model
- Integrate the trained model with an web app



B.Description of the Proposed Algorithm:

Aim of the proposed algorithm is to generate captions for new unseen images in a more efficient way. The proposed algorithm consists of the following steps.

Step 1: Removing Noise in the Dataset:

The dataset used is Flickr8k which is a publicly available dataset, the special characters and numerical characters are removed from the dataset along with single letter words like ‘a’ is removed which does not provide any contribution in generating the captions.

Step 2: Extract Features from the training images:

We had extracted the features from the training images using different CNN models which include VGG16, VGG19, ResNet50, InceptionV3, InceptionResNetV2 and saved the extracted features in a python pickle file.

Step 3: Data Generation:

Since the dataset contains 8000 images from which we are going to train our model with 7282 images, which cannot fit into limited memory. So, we had to create a data generator that will yield batches. The generator will yield the input and output sequence. Then map each word of the vocabulary with a unique index value. Keras library provides us with the tokenizer function that we will use to create tokens from our vocabulary.

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 ||

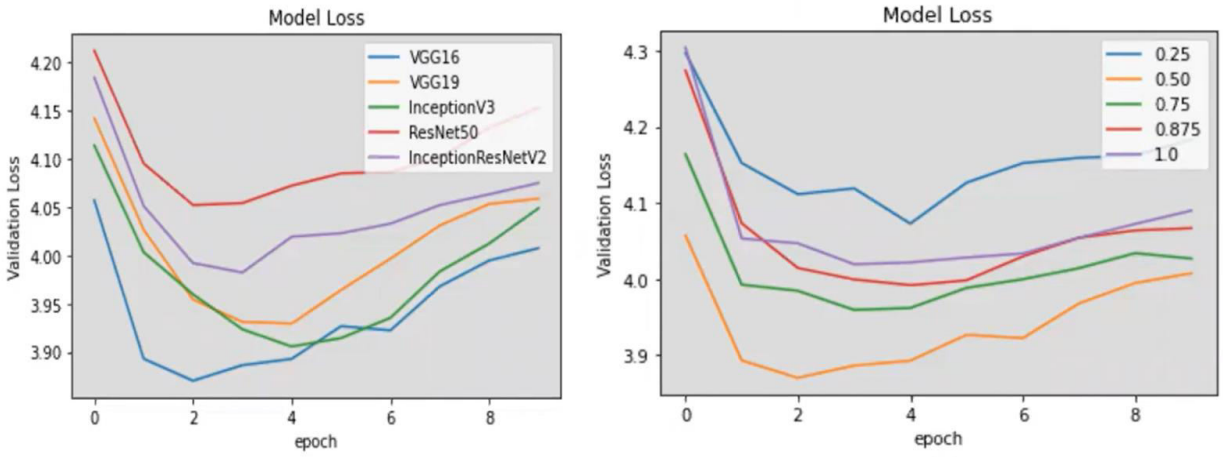
| DOI: 10.15680/IJIRCCE.2022.1006110 |

Step 4: Model Building and Selection

The proposed model consists of three major parts:

- Feature Extractor
- Sequence Processor
- Decoder

Feature Extractor – The feature extracted from the image using different pre-trained CNN models and then, with a dense layer, we will reduce the dimensions to 256 nodes.
Sequence Processor – An embedding layer will handle the textual input, followed by the LSTM layer.



The model's validation loss was low for VGG16. Hence, it was used for the feature extraction and when the dropout was set at 50% we observed a low validation loss which was used to build the LSTM model.

Step 5: Model Evaluation:

The model has been trained, The BiLingual Evaluation Understudy(BLEU) score metrics has been used to evaluate the model with the unseen testing dataset, The final results have been shown in the figure below.

Dropout = 0.5				
Model	BLEU1 Score	BLEU2 Score	BLEU3 Score	BLEU4 Score
VGG16	0.565	0.3342	0.2377	0.164
VGG19	0.5169	0.2554	0.1661	0.0932
InceptionV3	0.5283	0.284	0.199	0.1003
ResNet50	0.4533	0.1971	0.1322	0.0634
InceptionResNetV2	0.5059	0.2398	0.1634	0.0861

Model = VGG16				
Dropout	BLEU1 Score	BLEU2 Score	BLEU3 Score	BLEU4 Score
0.25	0.4232	0.1988	0.0813	0.0312
0.5	0.565	0.3342	0.2377	0.164
0.75	0.5422	0.2994	0.1855	0.1449
0.875	0.5239	0.281	0.1824	0.1253
1	0.4911	0.2672	0.1693	0.1058

| 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.2022.1006110 |

PSEUDO CODE

Step 1: Load the dataset and remove the noise in the data

Step 2: Extract the features vector of images using different CNN models

Step 3: Tokenize the vocabulary

Step 4: Build the LSTM model and train it for epochs

Step 5: Evaluate the model using BLEU score metrics

Step 6: Repeat the above steps by tuning the hyperparameters until a good BLEU score is achieved

Step 7: End

IV. CONCLUSION AND FUTURE WORK

We have implemented a CNN-LSTM model for building an Image Caption Generator. A CNN-LSTM architecture has wide-ranging applications which include use cases in Computer Vision and Natural Language Processing domains. Based on the results obtained we can see that the deep learning methodology used here bore successful results. The CNN and the LSTM worked together in proper synchronization, they were able to find the relation between objects in images. To compare the accuracy of the predicted caption, we compared them with target captions in our Flickr8k test dataset, using BLEU(Bilingual Evaluation Understudy) score from which we can observe that VGG16 gave best results with appropriate hyperparameters for the LSTM. BLEU scores are used in text translation for evaluating translated text against one or more reference translations. Over the years several other neural network technologies have been used to create hybrid image caption generators, similar to the one proposed here. Moreover, with more powerful machines we can train the model on a bigger dataset like Flickr30k or MOSCO30 dataset to further improve the accuracy of the model.

REFERENCES

1. Girish Kulkarni, VisruthPremraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35:2891–2903, June 2013.
2. Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, Tsuhan Chen, Recent advances in convolutional neural networks, Pattern Recognition, Volume 77, 2018
3. O'Shea, Keiron & Nash, Ryan. (2015). An Introduction to Convolutional Neural Networks. ArXiv e-prints.
4. Staniūtė R, Šešok D. A Systematic Literature Review on Image Captioning. Applied Sciences. 2019; 9(10):2024. https://doi.org/10.3390/app9102024
5. William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: Better text generation. arXiv preprint arXiv:1801.07736, 47, 2018

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