International Journal of Future Generation Communication and Networking
Vol.14, No. 1, (2021), pp. 3901 – 3908

Generating Image Captions Using Deep Learning

Prof. Archana Shinde[1], Prajwal Jadhav[2], Aditya Mankar[3], Deepak[4], Daimi Zaid[5]

Computer Department, Sinhgad Academy of Engineering, Kondhwa

[1]asshinde.sae@sinhgad.edu

[2]prajwaljadhav300@gmail.com

[3]*adityamankar09@gmail.com*

[4]deepakdpkgmlcm@gmail.com

[5]*zaidaimi8055@gmail.com*

Abstract

In Artificial Intelligence (AI), image content is automatically generated including computer vision and NLP (Natural Language Processing). This model is used for producing natural sentences that ultimately describe the image. This model contains Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). CNN is used for feature image extraction and RNN are used for sentence production. The model is trained for such a way that when an input image is provided for a model, we create captions that almost define the image.

Keywords— *Neural Network, Image, Caption, Description, Long Short-Term memory(LSTM), Deep Learning*

I. INTRODUCTION

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. This task of automatically generating captions and describing the image is signifi-cantly harder than image classification and object recognition.

Image captioning is a popular research of Artificial Intelligence (AI) that deals with image understanding and language description for that image. 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 concept which are closed, makes a suitable and simple model for assumption.

However, the machines need to interpret image captions in some form if humans need automatic image captions from it. Image understanding needs to detect and rigidize objects. It also needs to understand scene type or location, object properties and their interactions. Every day, we see a large number of images from various sources like the internet, articles, documents, diagrams, and advertisements. Most of these images do not have a description, but the human can easily understand them without their detailed captions.

II. RELATED WORK

The paper, A short Review in Image Caption generation with Deep Learning, by Soheyla Amirian, Khaled Rasheed and Thiab R. Arabnia, have shown the methodologies that utilize Deep Learning are

ISSN: 2233-7853 IJFGCN

shown which offer great potential for applications that automatically attempt to generate captions or descriptions about images.

In the paper, Comparison of Image Captioning Methods, by Jeel Sukhadiya, Harsh Pandya and Vedant Singh, we have seen that humans can give insight descriptions of the images or the scenes presented to them. Thus, image captioning is a task of generating a basic description of all the objects, their relationship with the environment around them present in the images, to effectively describe the image.

Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin and Hamid Laga, have shown the general concept of Image Caption Generation. They have also discussed the datasets and evaluation metrics which are used in deep learning-based image captioning. Deep learning-based techniques are capable of handling the complex and challenging tasks of image captioning.

In the paper, A systematic literature Review on Image Captioning from Department of Information Technology, Vilnius Gediminas Technical University, they have shown the recent progress in artificial intelligence (AI) has greatly improved the performance of models. Due to the increasing amount of information on this topic, it is very difficult to keep on track with the newest researches and results achieved in the image captioning field. Moreover, it is still not clear if MS COCO and Flickr30k datasets are enough for model evaluation and if they serve sufficiently well when having in mind the diverse environment.

III. CONVOLUTIONAL NEURAL NETWORK

Convolution Neural Networks or ConvNets are very popular in computer vision applications thanks to their ability to detect and identify various sorts of objects in images. In layman's terms, CNN is essentially a totally connected neural network alongside convolution operations at the beginning. These convolution operations are often used to detect defining patterns in images. It is almost like the neurons within the lobe of the human brain.

The architecture of ConvNets is made using 3 layers which are then stacked to make a full ConvNet architecture. Following are the three layers: Convolution layer, Pooling layer, Full Connection.

A. Convolution Layer

The convolution layer is that the core is a part of a ConvNet and performs all the computationally heavy tasks. A kernel or a filter of a selected pattern is traversed through the whole image to detect a selected sort of feature. The output of this traversal leads to a two-dimensional array called Feature maps. Each value during this feature map is a ReLU function to get rid of non-linearity.

B. Pooling Layer

This layer is liable for reducing the size of knowledge because it reduces the computations and time required for processing. There are two sorts of pooling: average and max. As the name suggests, Max pooling returns the max value and Average pooling returns the typical values of the portion of the image covered by the kernel.

C. Full Connections

The two-dimensional output array received from the previous step is converted into a column vector through the flattening process. This vector is passed to a multi-layered neural network which through a series of epochs learns to classify the pictures using the Softmax function.

ISSN: 2233-7853 IJFGCN

III. RECURRENT NEURAL NETWORK

Recurrent Neural Network (RNN) is a special type of Neural Network where the output from the previous step is given as an input to the current step. In traditional neural networks techniques, all the inputs and outputs are independent of each other, but in cases like when it is required to predict the next word of a sentence, the previous words are required and hence there is a need to remember the previous words.

Recurrent Neural Networks were created which solved this issue with the help of a Hidden Layer in the network. The main and most important feature of Recurrent Neural Network is the Hidden state layer, which remembers some information about a sequence.

Recurrent Neural Networks have a memory which remembers information about what has been calculated. It uses the same parameters for each input as it performs the same task on all the inputs or hidden layers to produce the output. This reduces the complexity, unlike other neural networks.

IV. LONG SHORT-TERM MEMORY(LSTM)

Long Short Term Memory networks are an extension of recurrent neural networks (RNNs) mainly introduced to handle situations where recurrent neural networks fail. Talking about recurrent neural networks, it is a network that works on the present input by taking into consideration the previous output (feedback) and storing in its memory for a short period of time (short-term memory). Out of its various applications, the most popular ones are in the fields of speech processing, non-Markovian control, and music composition.

The basic difference between the architectures of recurrent neural networks and Long Short Term Memory is that the hidden layer of Long Short Term Memory is a gated unit or gated cell. It consists of four layers that interact with each other in a way to create the output of that cell along with the cell state. These two things are then given onto the next hidden layer. Unlike recurrent neural networks which have gotten the sole single neural net layer of tanh, LSTMs comprises three logistic sigmoid gates and one tanh layer.

Gates are introduced so as to limit the knowledge that's skilled the cell. They determine which part of the knowledge is going to be needed by subsequent cells and which part is to be discarded. The output is typically within the range of 0-1 where ‘0’ means ‘reject all’ and ‘1’ means ‘include all’. Long Short Term Memory models got to be trained with a training dataset before its employment in real-world applications. Some of the foremost demanding applications are discussed below:

Language modelling or text generation, that involves the computation of words when a sequence of words is fed as input. Language models are often operated at the character level, n-gram level, sentence level or maybe paragraph level. Image processing, that involves performing analysis of an image and concluding its result into a sentence.

V. SENTENCE GENERATION

ISSN: 2233-7853 IJFGCN

The output of Long Short Term Memory is the probability of each word in the vocabulary. Beam search is used to generate sentences from images. Beam search is a heuristic search algorithm that explores a graph by expanding the most promising node in the neural network. In addition to beam search, we also use k-best search to generate sentences from images in our dataset. It is similar to the time synchronous Viterbi search. The method iteratively selects the k best sentences from all the candidate sentences up to time t, and keeps only the resulting best k of them.

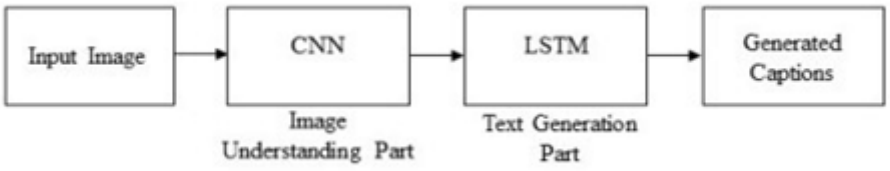

Fig. 1 A block diagram of simple Encoder-Decoder architecture-based image captioning

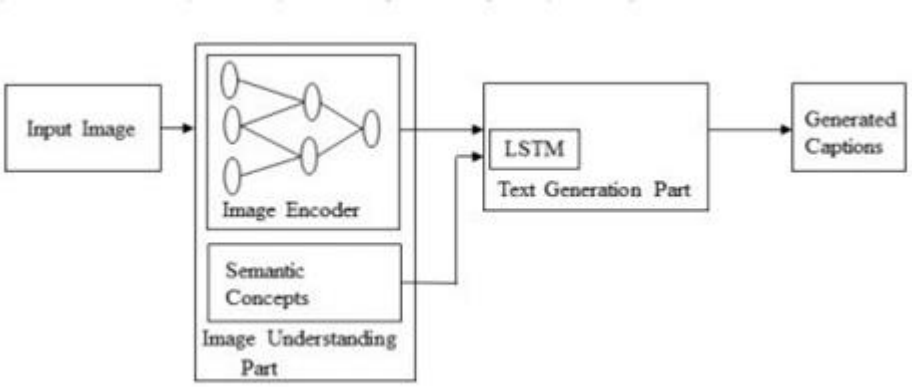

Fig. 2 A block diagram of simple Encoder-Decoder architecture-based image captioning

Attention-based methods have shown great performance and good efficiency in image captioning as well as other computer vision-based tasks. However, attention maps generated by these attention-based methods are only machine dependent tasks. They do not consider supervision from human attention which creates the necessity to think about the gaze information whether it can improve the performance greatly of these attention-based methods in image captioning.

VI. IMPLEMENTATION

The first step will be to import all the necessary libraries such as NumPy, pandas and keras along with the dataset. In the next step which is Data cleaning where we preprocess the main text file

containing all the image captions. Here we implement functions which perform operations like reducing irregularities and removing punctuations.

In the next step we will implement Transfer learning by using a pretrained CNN model which is called the Xception model. This model will extract the features from the images in our dataset. This Xception model takes the 299x299x3 image as an input. For this model we will set the pooling value to avg and include_top set to false. This step will generate a pickle file containing the features which will be stored in the features.p file.

For the next step we will perform tokenization as computers don't understand English words. We use tokenization to map each word of the vocabulary to a unique index. It is saved in a pickle object called a tokenizer.pkl file. We will also create a data generator which is used to progressively train our model as our machine cannot hold such a large dataset in memory. It will create an input and output sequence.

In the next step we will define our model using the keras library. It consists of three major parts which are discussed below. Feature extractor which takes the features extracted from the image of size 2048 will be reduced to 256 nodes using a dense layer. Sequence processor which handles the textual input using the embedding layer followed by the LSTM layer. Decoder layer merges the output from above two layers and it will be processed by the dense layer to make the final prediction.

We will train our model on 6000 images in batches using the model.fit_generator() method. The created model is also saved in the models folder in our directory. After training the model a separate file is created which will load the model and generate predictions from our images.

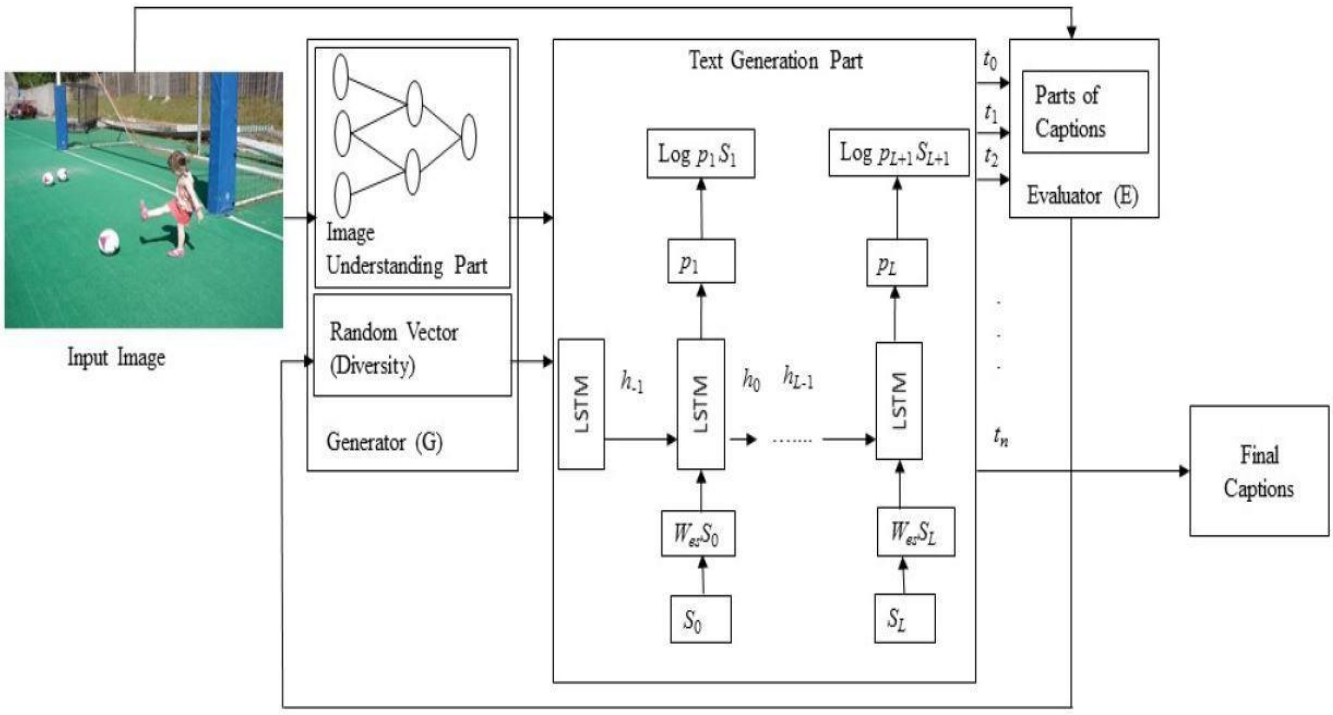

Fig. 3 A block diagram of other deep learning-based captioning

VII. RESULT

A. Datasets

The datasets we used to evaluate the performance of our methods are the MS COCO, Flickr8K, and Flickr30K. They are the most popular datasets for evaluating the generated descriptions.

By using the Flickr8k dataset for training models and running tests on the 1000 test images available in the dataset results in BLEU = 0.53356. For Flickr30k dataset, running test on the same number of test images available in the dataset results in BLEU = 0.61433 and for MSCOCO dataset running test on images results in BLEU = 0.67257.

Datasets	Vocab Size	Max Length	Total Words	Top-10 Words with Higher Occurrences
MS COCO	9486	49	6,421,733	a, on, of, the, in, with, and, is, man, to
Flickr8K	2629	37	422,800	a, in, the, on, is, and, dog, with, man, of
Flickr30K	7648	78	1,892,755	a, in, the, on, and, man, is, of, with, woman

Table 1. Comparisons of reference captions on datasets MS COCO, Flickr8K, and Flickr30K.

B. Generated Results

Fig 4.Two girls are sitting on the edge of the grass.

Fig. 5 Black dog is sitting on the floor.

VIII. CONCLUSION

This work presents a model, which is a neural network that can automatically view an image and generate appropriate captions in a natural language like English. The model is trained to produce the sentence or description from a given image. The descriptions or captions obtained from the model are categorized into:

The categories in results are due to the neighborhood of some particular words, i.e., for words like the car it's neighborhood words like a vehicle, van, cab, etc. are also generated which might be incorrect. After so many experiments, it is conclusive that the use of larger datasets increases the performance of the model. The larger dataset will increase accuracy as well as reduce losses. Also, it will be interesting how unsupervised data for both images as well as text can be used for improving the image caption generation approaches.

REFERENCES

[1] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. IEEE, 2015.J. Breckling, Ed.,
[2] Gerber, Ralf, and N-H. Nagel. "Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences." Image Processing, 1996. Proceedings., International Conference on. Vol. 2. IEEE, 1996.
[3] Yao, Benjamin Z., et al. "I2t: Image parsing to text description." Proceedings of the IEEE 98.8 (2010): 1485-1508..
[4] Farhadi, Ali, et al. "Every picture tells a story: Generating sentences from images." European conference on computer vision. Springer, Berlin, Heidelberg, 2010.
[5] Zhibin Guan 1, Kang Liu 1, Yan Ma , Xu Qian and Tongkai Ji, Sequential Dual Attention: Coarse-to-Fine-Grained Hierarchical Generation for Image Captioning, 2018
[6] Chetan Amritkar, Vaishali Jabade (Department of EnTC Vishwakarma Institute of Technology), Image Caption Generation using Deep Learning Technique, 2018

ISSN: 2233-7853 IJFGCN