International Journal of Research in Engineering, Science and Management
Volume-3, Issue-6, June-2020
www.ijresm.com | ISSN (Online): 2581-5792

Two Step Semantic Phrase Construction from Images

Prashant Kumar Bhardwaj[1*], Ujjwal Singh[2], Tapan Yadav[3], Ritesh Srivastava[4]
[1,2,3]*Student, Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India*
[4]*Professor, Department of Computer Science and Engineering, Galgotias College of Engineering and Technology, Greater Noida, India*
*Corresponding author: prashantbhardwaj282@gmail.com

***Abstract*: With the growing technology, we all have experienced there is an extensive growth in the field of artificial intelligence and natural language processing. But what most attracted the attention of researchers is Image's Feature Collection, which itself generate the captions of an image. Its application is wide and extensive as the traffic in social networking is increasing day by day. In this paper we have used CNN and LSTM which is special kind of RNN used to store data for a longer period of time firstly the input image is processed by a CNN and the output of that is given as an input to the RNN which generates textual descriptions, that's what this paper summarizes. The dataset which we've used is flicker_8k containing the set of 8k images to train the model well enough. The different libraries are also involved in this project making like Keras, Numpy, etc. Finally, this paper underlines the advantages and shortcomings of this architecture as well.**

***Keywords*: Wide and expensive, Processes by a CNN, Advantages and shortcoming.**

1. Introduction

Image Captioning is a very important field of work in deep learning technology. Image Captioning can help to machines to understand the features embedded in the images and can do the analysis on that data.

Image Captioning by images works similarly that how we humans understands the images, we first need to know the individual objects in the images and then relate them according to their position in the images and generate the phrase for the images. Similarly, the computer needs to train so that it can understand the images correctly according to their features.

Image Captioning process contains two parts mainly. The first is to extract features from the image, and the second is to combine the features and create a phrase in the natural language. In this project we will discuss how we can implement these two steps using Deep Learning techniques. We will use CNN for feature extraction and LSTM for caption generation in Natural Language.

Now-a-days there is a lot of Image data is on the internet so we can extract very much information from that data but we cannot do it manually so we need a computer program that understands the image and gather the information.

Image Captioning is used in various fields, one of them is Social media apps where millions of images are posted in a day, and companies are extracting the information from the user's images and finds the user interest according to the images that user post. This image processing is done by Computers only.

So in this project, we will look at how we can create the captions for the images with the help of different Deep Learning techniques, we will also discuss the past researches and their concepts which are already done in this field.

2. Literature Review

With the advancement in the field of Artificial Intelligence, the techniques to resolve daily basis problems are effectively done by Machine Learning whether it would be with the help of IoT or with CNN, RNN, etc. The major portion of the problems related to growing technology can be solved by these advanced trained machines or methods. Day by day as the social sites are increasing and so as a user the companies invest so much money on this, this led to research in the field of Image Captioning Techniques. Earlier many techniques have been made for generating image caption images with the help of Supervised Learning, Reinforcement learning, etc. but some of them are discarded due to the inefficiency or due to system fault and so on. Many of the researchers had done excellent work in this area but the problems to develop caption for blurred images or the images containing several items still remain untouched.

The datasets which are prominently used to build the system are flicker8k, flicker30k, MSCOCO, etc. These datasets differ on the number of images they contain, we know to build effective and powerful systems we have to train the system with a high amount of data.

Several researchers who had worked on this field are listed below,

Ahmet Aker and Robert Gaizauskas. (2010) [1] proposed a method of automatic captioning of an image using relational patterns. They generate via collecting and summarising all the documents which contain images location. The places are traced to that location to provide an effective caption to the user. They had experimented with n-gram models which provides

automatically generated summaries. But the weakness of this model is it captures very basic knowledge about short term sequences and cannot model large dependencies.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. (2016) [2] proposed a method where they had used Semantic proposition for Image Caption Evaluation. They had also proposed SPICE in their project which is used as a principled metric for the evaluation of automatic image caption. They had demonstrated that SPICE performance can be decomposed to answer the questions like 'which caption generator best understands colors' and 'caption generators count'.

Dzmitry Bahdanau, Kyunghyun Cho, and YoshuaBengio. (2015) [3] they had used a neural machine translation for generating the captions. They used translation and aims to build a single neural network that can be turned out into a maximum translation performance. This translation approach contains RNN-encoder and decoder which can generate caption of any length and here models are learned to align and translate. The dataset which they used is WMT' 14. They trained two types of models, firstly as we have discussed the RNN Encoder-Decoder approach, and secondly is a proposed model that is RNN search.

Shuang Bai and Shan An. (2018) [4] both Chinese researchers proposed a neuro computing model to generate the image caption. So what they did is they collected semantic information of images and then they try to express in natural languages. They used several neural network-based methods which are further subdivided into categories based on the framework which they are using. The usage of Retrieval based, Template-based, Deep neural network-based and based on multimodal learning are some of the key features of their paper.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. (2016) [5] proposed a method that can generate an unambiguous description of a specific object or region in an image. So they avoid such objects in the image which causes ambiguity in the scene. They had used the MSCOCO dataset which is known profoundly for his collection of large images. Their research also released a toolbox for visualization and evaluation.

3. Image Captioning Techniques

A. Convolutional Neural Networks

In our project we will use CNN a deep learning technique for the image processing. Convolutional Neural Network technique is used for image processing i.e. for extracting the information from the image.

CNN is the matrix based approach which converts the image in a 2D matrix and scan the matrices form top-left corner to the bottom-right corner and extract the features included in that image.

An image can contain multiple objects and features so to differentiate these objects CNN uses the leasable weights and biases to the objects with which it differentiates from each other.

CNN can also have worked on transformed images so it is reliable for any kind of images. The below diagram identifies the basic flow of the convolutional network.

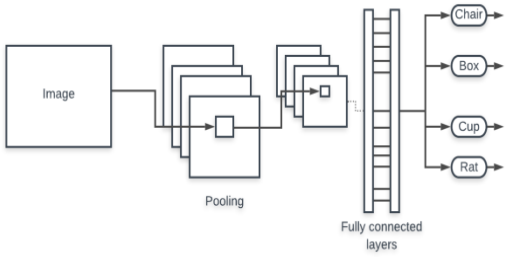

Fig. 1. Working of Deep CNN

CNN works in three layers that are convolutional, pooling, and fully connected.

Working of the first layer is all about getting the input image matrix and traverse it from top to bottom, in this first level CNN creates a featured map of features included in that image and classify the image's features.

The second part is known as Pooling, in this stage, the dimensions of the features reduce such that most important information of features remains. Since feature's dimensions are reduced in this step, so this is also called as Down sampling.

These two stages are repeated multiple times until the right degree of Granularity is achieved.

The last stage is a neural-network which works on the important features and trains the Model according to the input and output of the image.

Nowadays CNN is used in many different fields such as in Computer vision for image recognition, Natural Language Processing for sentence generation, etc.

B. Origin of LSTM

LSTM which is the abbreviation of a very popular concept Long Short Term Memory is introduced by a German named Hochreiter and Schmidhuber in 1997 in their research paper.

LSTM is a type of Recurrent Neural Network (RNN) capable of learning sequences over long period of time, which is needed to develop longer and detailed sentences for captioning, but RNN having issues to retain information over longer periods of time and cannot generate long phrases.

C. The Vanishing Gradient in RNN

Vanishing gradient is a major problem which occurs in RNN based Models, because of which the performance of RNN based models is also much low. RNN was able to remember the things for a shorter time which restricts its long term capacities, due to which RNN based models can remember only a few sequences at a point of time. This problem restricts the use of RNN based models because of this issue RNN now uses only for small sequence and short time based problems.

Due to the Vanishing Gradient model can't remember the old things or in technical language, it cannot store the older sequences in the memory also it cannot store too many

timestamps for a series of inputs also it forgets the information on using backpropagation so we cannot use it for complex models.

Long Short Term Memory acronym of LSTM is the solution of vanishing gradient problem, LSTM is constructed mainly to eliminate the vanishing gradient and make the models remember the sequences for a longer time in contrast to models which are RNN based. The main logic which is behind the LSTM is that it constantly maintains a non-variable error in the model which allows the model to learn continuously over various time stamps and also it helps in backpropagation and enables you to propagate back in time.

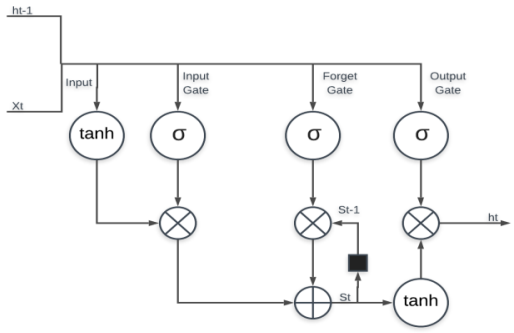

Fig. 2. LSTM architecture

LSTM concept uses the gated cells which helps it to store the features or information which makes it superior to RNN. The cells are also on their own can make the decisions on the information and can also work on these decisions by changing the state (open and close) of gates. With the help of gates, pieces of information can also be manipulated by the network, with the help of gates network can read and write the stored information in the cells.

Long time access of the information makes the LSTM models much better than old RNN based Models.

The above section says how to prepare a subsection. Just copy and paste the subsection, whenever you need it. The numbers will be automatically changes when you add new subsection. Once you paste it, change the subsection heading as per your requirement.

D. Long Short-Term Memory Architecture

LSTM has a chain like structure which binds up the information for more time so that much older information can be processed to generate the new information.

Gates used are the basic building blocks of LSTM, a brief summary of these gates is given below with the structural diagrams.

The forget gate is used as a broom which sweeps out the not necessary things like this forget gate is used to delete the not necessary information from the memory. Forget gate is only used to increase the reliability of the model.

The input gate acts like an entry gate which adds new data to the gated cells.

The output gate is used to select and sends important pieces of information to the user other model as output.

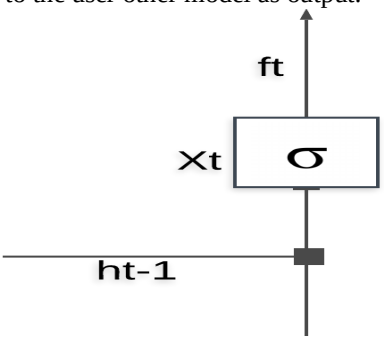

Fig. 3. Forget Gate

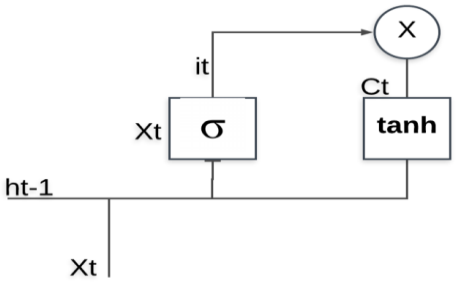

Fig. 4. Input Gate

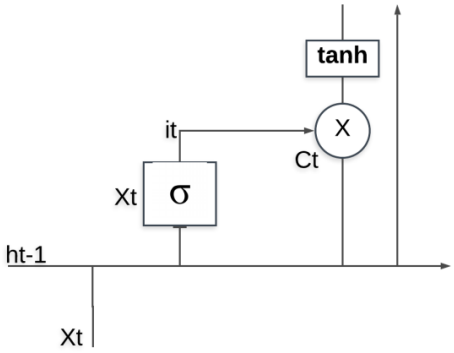

Fig. 5. Output Gate

E. Real-time uses of LSTM

Due to their long period memory retention LSTM are used in various of deep learning models, there are plenty of applications where LSTM are nowadays in use like:

The newer and non-general one of its applications is Semantic Parsing, Semantic parsing is used to transform the general spoken language words into a computer processable logical form in which computers can make new words and sentences.

In this process, LSTM are used to extract useful data to find the meaning of a language's utterance.

4. Structure of Our Model

As the name of this project suggest our caption generator model we will use two concept CNN and LSTM. Also we will make use of Python language to code our model.

CNN is used for extracting the important features from the image, CNN uses the matrix approach to define the image features.

LSTM is used in place RNN will work on the information extracted by the CNN and use Natura Language Processing concept to generate the captions according to the output given by CNN.

A. Dataset and File Structure

Dataset is also an important part of the model, it gives our training and testing data for our project we will need a very large number of images. There are various databases provided free on the internet like Flicker8k, Mosoco, Flicker30k, etc., these are very large data sets.

We will use Flicker8k data set for our project, this dataset contains 8091 images and their information in total of two files.

One of the two files is dataset file which contains images with their names and another file only contains the text in which the captions for each image are stored. For every image in the Dataset file, there are 5 captions given in the Text file, mapping of images to their captions is done with the help of image names. Figure 6 describes how the images are stored in the dataset file.

Fig. 6. Flickr dataset

5. Conclusion

In this survey paper, we have reviewed different deep learning image captioning methods. We discussed the implementation of image captioning via CNN and LSTM with the help of proper diagram. We have used Flickr_8k dataset which includes nearly 8000 images, and the corresponding captions are also stored in the text file. Caption generator with deep learning concept enabled are very optimized and used by various peoples in the past years. An ideal image caption generator is one that can generate the caption of all images which are given to it. These models are improving day by day and continuous research is going on these concepts which making them better.

References

[1] Ahmet Aker and Robert Gaizauskas. 2010. “Generating image descriptions using dependency relational patterns,” in Proceedings of the 48th annual meeting of the association for computational linguistics. Association for Computational Linguistics.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. “Spice: Semantic propositional image caption evaluation,” in European Conference on Computer Vision. Springer.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and YoshuaBengio. 2015. “Neural machine translation by jointly learning to align and translate,” in International Conference on Learning Representations (ICLR).

[4] Shuang Bai and Shan An. 2018. “A Survey on Automatic Image Caption Generation,” Neurocomputing. ACM Computing Surveys.

[5] Junhua Mao, Jonathan Huang, Alexander Toshev, OanaCamburu, et al. 2016. “Deep compositional captioning: Describing novel object categories without paired training data,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition