2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. June 4-5, 2020

Generating Image Captions based on Deep Learning and Natural language Processing

Smriti Sehgal
Assistant Prof.
Dept. of Computer Science and Engineering
ASET, Amity University, Noida
smriti1486@gmail.com

Jyoti Sharma
Dept. of Computer Science and Engineering
ASET, Amity University, Noida
jyoti03sharma98@gmail.com

Natasha Chaudhary
Dept. of Computer Science and Engineering
ASET, Amity University, Noida
natasha.chaudhary660@gmail.com

***Abstract*—This model enables an individual to input an image and output a description for the same. The research paper makes use of the functionalities of Deep Learning and NLP (Natural Language Processing). Image Caption Generation is an important task as it allows us automate the task of generating captions for any image. This functionality enables us to easily organize files without paying heed to the task of captioning. It is also important for making dynamic web pages. This paper is for people who are visually impaired or suffer from short sightedness. So, rather than looking at an image with trouble they can easily read the caption generated by this model in a larger format. It can also be used to give description of a video in real time on later implementation for a video.**

Keywords—DeepLearning; Convolutional neural network; filters;recurrent neural network; natural language processing; LSTM

I. INTRODUCTION

Image Caption Generation is based on the functionalities of CNN and RNN. Use of Keras Library has been made and the development has been done in Jupyter Notebook. The implementation for this paper has been done using Python Language. This paper is going to make use of COCO dataset. COCO (Common Objects and Contexts) data set is a specialized dataset which contains 1 lakh images with each containing 5 captions associated with each. This is one of the apt dataset for our model and allows to develop a well trained model after rigorous training.

CNN is majorly used to extract objects and other spatial information patterns from our input image whereas RNN works efficiently with any kind of sequential data fed to it. CNN is often called the encoder whereas RNN is associated with the decoder.

Natural Language Processing is also being used here to generate image captions alongside CNN. LSTM are the specialized Recurrent Neural Networks which allow for information to persist. Use of the VGG16 model which is pre – trained on the ImageNet dataset for the classification of images has been made.

II. PLATFORM USED

In 2014, Fernando Pérez announced a spin-off project from IPython called project Jupyter. It is a platform which is open source web application in nature and is used to develop and create code to implement models in an effective manner. Enables us to incorporate live code, visualize and narrative texts. The implementation of the model is being done on Jupyter Notebook.

III. WORKFLOW

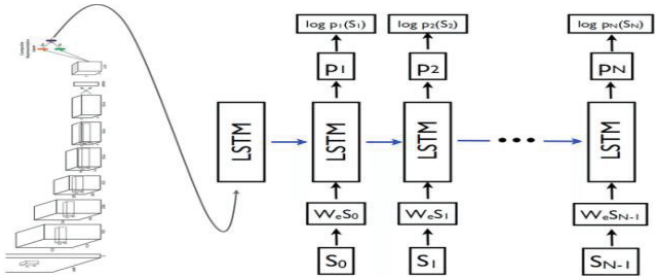


Fig. 1. Basic Workflow of the model[8]

The two deep learning algorithms used here are convolutional neural network and recurrent neural network.

Firstly, the input image is passed through the Convolutional Neural Network (CNN) to identify the objects and scenes present in the image. Also, use of transfer learning is done here for the pre-processed model. In CNN various concepts such as pooling, padding, filters etc. are used. A set of words/objects will be generated as the output from the CNN model. Next, use of Natural Language Processing (NLP) is done which helps us to communicate with the computer.

Finally Recurrent Neural Network is trained using the Flickr8k_text dataset. The objects detected are passed to the RNN after some desired processing and the RNN produces some desired meaningful caption. Refer the image to understand the workflow more clearly.

IV. DEEP LEARNING

Deep Learning can be defined as an advancement of machine Learning which is based on Artificial Neural Networks and contains more than one hidden Layer. Deep in deep Learning itself is used to denote the number of layers in the Network.

It is a technique in which computers learn to perform tasks that can just be performed by humans. Deep Learning algorithms are capable of giving results that can perform better than humans. There are many algorithms such as convolutional neural network, recurrent Neural Network etc that are applied in different applications like speech recognition, video recognition, NLP etc

In deep Learning very large sets of labelled data are required for training models containing many hidden layers. Therefore, large data sets and high computing power are two most important requirements of deep learning.

978-1-7281-7016-9/20/$31.00 ©2020 IEEE

V. NEURAL NETWORKS

Human beings are blessed with the ability to think and memorize. Artificial Intelligence tries to mimic this behaviour. And this is what forms the basis for Neural Networks.They can be considered as a series of algorithms which are trying to mimic the functionalities of a human brain. The human brain is based on networks of neurons which act as messengers of signals to the brain. Similarly, in neural networks there are several layers. They try to find out the underlying relationships in the input data. It contains different layers. Each node in a neural network is called a perceptron. A perceptron is a single neuron model that was a precursor to larger neural networks. It feeds the signal produced by a multiple linear regression into an activation function that may be nonlinear.

VI. CONVOLUTIONAL NEURAL NETWORK

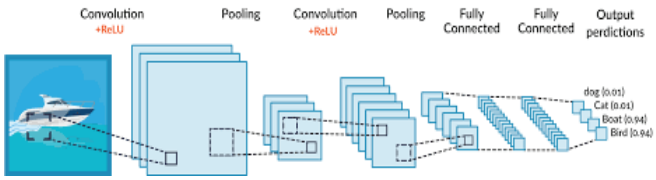


Fig. 2. Convolutional Neural Network[8]

Convolutional neural network also known as CNN or ConvNet is a kind of artificial neural network which is used by most people for the analysis of images. It can be also used for problems of classification and data analysis. CNN can be thought of as an Artificial Neural Network that uses classification for detecting patterns and tries to find a meaning from them.

More formally, CNN is an algorithm of deep learning which allow user to input an image, and the algorithm can assign learnable biases and weights to various objects or aspects that are present in the image and hence be able to differentiate one object from the other.

A. Important concepts in Convolutional neural network

1) Filters

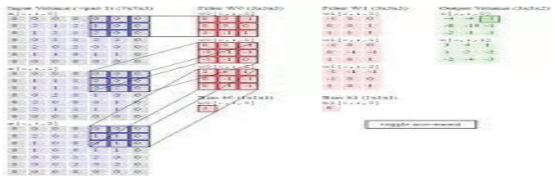


Fig. 3. Filters[1]

In Convolutional Neural Network every layer that is present contains filters. Filters are responsible for finding specific patterns or features in the data as input is given. In each layer the number of filters that the layer should have are given.

At the start of the network filters are pretty simple and detect patterns like edges, circles, etc. but as we move to further layers these filters become more advanced and are able to identify complete figures like mouse, cat etc. Filters can be imagined to be a matrix of specified number of rows and columns. Any random number can be used to initialize the matrix blocks.

The input given is larger than the filter. Dot product between a patch of the input image that is of the same size of the filter and the filter itself is applied. This gives a single value If we multiply the filter with array of input images we get a single value but if the same operation is repeated multiple times we get a 2-D array of filtered values and this is called as **feature map**.

2) Padding

Padding is used to protect the length and the breadth of the input image when going from one layer to the next. Deeper network can be designed with the help of padding. The performance is said to be improved here as the information is kept at the borders. Padding is of two types, valid padding and same padding. In Valid Padding the feature that is convolved through the layers is smaller than the input image. In same padding the feature that is convolved through the layers is either of same size or bigger size than the input image.

3) Activation function

A node that is kept in the middle or at the finish point of a neural network is called the activation function. An activation function helps us to identify whether to fire the neuron or not. It is a non-linear function that is applied on the input signal which is then transformed and sent to the further layers of neurons which then treat it as input. We have used RELU and softmax activation function.

4) Stride

The number of pixels of input matrix over which filter is moved is termed as stride. If stride is one, the pixel is moved one time. Similarly, if stride is two or three times the pixel is moved one or two times. Thus, stride decides how much our filter slides over the input.

B. Layers in Convolutional Neural Network

Input image in convolutional neural network goes through different layers. Each layer is responsible for its own functions that are assigned to them. All layers are discussed as follows -

1) Input Layer

The very first layer present in the convolutional neural network is the input layer. The input layer contains artificial input neurons. This layer is responsible for bringing the very initial data into the system. The input layer consists of a characteristic that the other layers do not have. Input layers consists of neurons that are passive in nature. Being the first layer it does not take information from previous layers. An input layer contains neurons which do not contain weighed inputs or we can say where the weights are calculated in a way different than other layers as input is coming for the very first time. In convolutional neural network the input layer can take any image of 3 dimensions. The image can be coloured or black and white.

2) Convolutional Layer

The convolutional layer is the most important layer as most of the computation takes place in this layer. It contains many hidden layers. The purpose of this Convolutional operation is to extract features as well as minute details from the input image. Convolutional neural network is named after this operation. There can be more than one convolutional layer in CNN. This layer is responsible for detecting patterns and objects from the input image. It contains specified number of filters which are applied to the

input image. Each object that is given in the input image is identified using filters in complex mathematical steps to reach an individual value in the output map.

In Convolutional Layer the filter is slid throughout the whole image and a dot product of the part of the image that is overlapped by the filter and the filter itself is performed to give a scaler product which is stored in a matrix. This output matrix helps us to detect patterns in an image. This output is then given as an input to the further convolutional layers for the detection of more complex patterns.

3) Pooling Layer

One of the most commonly known limitation of feature map is that they capture the positions of all the features of the input very precisely. So, if there are small variations in the input image we get a different feature map. These variations can be possible due to various different reasons such as rotating, cropping, shifting etc.

Pooling layers are responsible for breaking the features present into patches of feature map. Hence, we have an image of lower resolution but containing the major elements that are important ignoring all the minute useless details. This technique can be referred as down sampling. Pooling layer is used between two convolutional layers. The two methods of pooling are -

- Average pooling – The average of all the blocks of the matrix that are overlapped by the filter are given by average pooling. Average pooling helps in reducing the dimensions.
- Max pooling – The highest value of the input figure that is overlapped by the filter is given by Max Pooling. Another function of max pooling is removing the noisy activations and hence supressing the noise in the image along with the reduction of dimensions. Performance of max Pooling is better than the performance of average pooling.

4) Fully Connected Layer

This is mainly present at the ending of the convolutional network and is also known as the feed forward layer or FC layer. Fully connected layer is a dense layer that is each node in this layer is connected to each and every node that is present in the previous layer. Fully connected Layer is responsible for the full classification of the features obtained by us from the previous layers such as the pooling or convolutional layer.

The input that is given to the fully connected layer is the output that is received from the from the final convolutional layer or the final pooling layer. The output is flattened. Flattening means converting a matrix into a vector before it enters a the fully connected layer. After the image passes the fully connected layer, the softmax activation function is used and not the ReLU activation function to obtain the probability of input being present in some definite class.

5) Output Layer

The outputs received are obtained from the fully connected layer are calculated on the basis of highest probability obtained by the FC Layer. The outputs are objects that are present in our input image such as dog, cat, mouse etc.

C. Work Flow of CNN

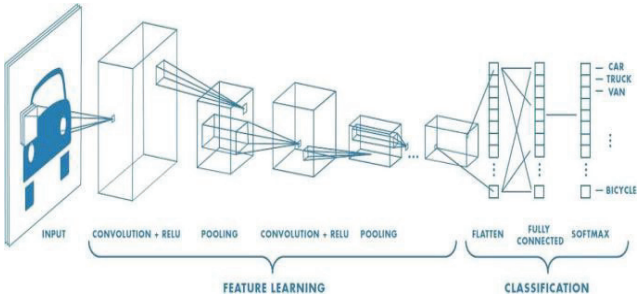


Fig. 4. Workflow of CNN[1]

All concepts required in convolutional neural network have been explained above. Firstly, an input is given to our network through the input layer, the input should be a 3-dimensional image. It could be either coloured or black and white.

The image is then passed further to the convolutional layer. In this layer, filters are applied on the input image. Number of filters are specified This process is called convolution. Then activation function is applied. RELU activation function is used here. Then our output here acts as an input to the pooling layer which helps to decrease the resolution of the image. Then again, the convolutional layer undergoes the same process. There can be more than one convolutional layer. As we go in further convolutional layers the filters get more complex patters such as eyes, faces, birds etc. Next, the matrix is flattened and the fully connected layer is entered. This layer helps to classify the objects in input image into classes. Softmax function is used here to find the probability of classes. Finally, the output is obtained which consists of the objects present in the image in the output layer.

VII. NATURAL LANGUAGE PROCESSING

Natural Language Processing abbreviated as NLP is a stream of artificial intelligence which helps to communicate with computers. Due to NLP, computers are now able to fulfil various tasks which were not possible before such as they are now able to read test, hear and understand what we are trying to say and even interpret the parts that are important. The various places where NLP can be used is to translate text from one language to another. It can also be used in applications such as Grammarly etc to correct any grammatical mistakes, Call centres to respond to various customers. Personal assistant applications such as alexa, sirietc use NLP to operate.

In our Research paper of image captioning, natural language processing is used after the algorithm CNN. Using NLP we apply algorithms due to which we can identify the natural language rules so that our language can be converted to a form that can be understood by the computers.

Thus, NLP here is used to help us communicate with the computer.

A. Techniques involved –

The two main techniques used in NLP are semantic analysis and syntactic analysis.

1) Syntax Analysis:

Identification is done whether the grammatical rules are being followed by the language or not. Some of the syntax techniques frequently used are Parsing, Stemming, Lemmatization

2) Semantic Analysis

Here, algorithms are applied in order to understand the meaning that NLP is trying to convey. Some of the techniques involved in this are Word sense Disambiguation, Named Entity Recognition and Natural Language Generation (NLG).

VIII. RECURRENT NEURAL NETWORK

Human beings are able to develop an understanding of the world based on their memory and past experiences. There is some persistence in thoughts which allow humans to use these as memories for reference later. Traditional neural networks are unable to store past experiences. This is one of the major advantages of Recurrent Neural Networks and why they are used.

RNN consists of loops in them. A loop helps in passing information from one step to the next in the model. RNN consists of several layers where each layer passes information to the next layer and allows for information to persist. The functionality of RNN is based on the following example. If we have a sentence like – 'The river contains' .....Then the model should be able to predict the word 'water' on its own.

A. Work flow of RNN

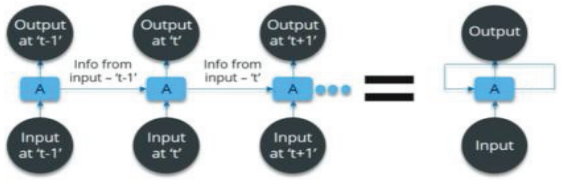


Fig. 5. Workflow of RNN [10]

In the case of RNN, the persisting information is used in conjunction with the input data for that layer and the output is predicted. Based on the above figure, information from timestamp 't-1' is passed along with the input value at timestamp 't' to predict the output at timestamp t. Back propagation through time (BTT) is the algorithm used in order to train the model. Thus, RNNs functionality is required whenever we wish to implement the model while using persistence.

B. Long Short Term Memory (LSTM)

LSTM is a type of Recurrent Neural Network which is used in cases where memory persistence for a longer period of time is required. It is often considered as the more enhanced version of RNN. It can perform almost every task that an RNN network can perform. They can understand long – term dependencies. LSTMs are designed in order to tackle the issue of long term dependency.

C. Work Flow of LSTM

LSTMs consist of numerous layers where each of these layers is in the form of chain of repeating modules. The basic idea behind LSTM is that each layer stores the output of its layer as an input for the next layer. Here there are 4 neural network layers each interacting in a unique manner.

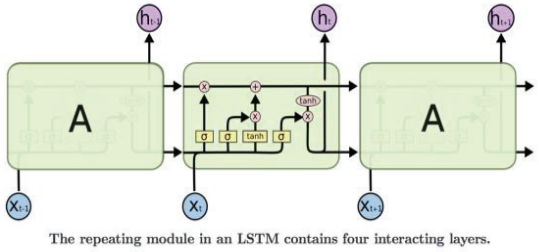
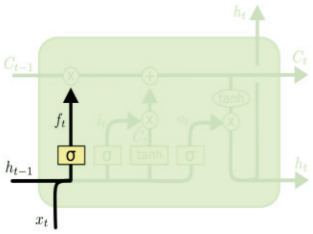


Fig. 6. Work flow of LSTM [10]

D. A Basic understanding of LSTMs

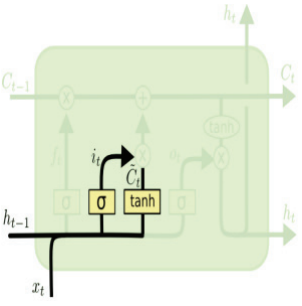
The main components of LSTM comprise of a conveyor belt like structure known as "the cell state". It is a straight horizontal line which runs across the entire structure chain and has minimal interactions. Removal of unwanted or unnecessary information from the cell state or addition new information is done using "Gates". The output that is received from the sigmoid layer is in the form of either 0 or 1. Where 1 represents "Permitting all information to pass through" and 0 represents "Dumping all information from the cell state". There are 3 gates to protect and control the cell state. The very first layer that the input data goes through in LSTM is the **"forget gate layer."** In this layer decision on which part of data do we want to discard is taken and it is done by the sigmoid layer.



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Fig. 7. Forget gate layer [10]

The second layer that the input data goes through in LSTM is the **"input gate layer".** In this layer, we decide which new information is to be added in the cell state. Here there are two components of this layer. The first is the sigmoid function and the second is the tan (h) layer.



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Fig. 8. Input gate layer [10]

IX. APPLICATIONS OF IMAGE CAPTIONING

- Helps in automating the task of labelling images.
- Can be useful for social media sites such as Facebook which can make use of these

automatically generated captions to infer data about its users solely on the basis of images.

- In web development to make websites dynamic simply add images without worrying about describing them.
- Can be of aid to visually impaired people who can simply read the text in a much larger font.
- Helps in organizing files without dwelling in the task of image captioning present in the files.
- Google photos can be helped in organizing images based on the objects present in the images.

X. CONCLUSION

Image Caption Generator is thus an effective tool which can be used for multiple purposes. It helps in automating the task of generating labels for images in an organized manner and thus aids in making lives easier. It can be used to describe images to people who are blind or have low vision and who rely on sounds and texts to describe a scene. In web development, it's a good practice to provide a description for any image that appears on the page so that an image can be read or heard as opposed to just seen.

REFERENCES

[1] Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2018. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. 0, 0, Article 0 (October 2018), 36 pages. Computing methodologies→Machine learning; Neural networks.

[2] "A gentle Introduction to deep learning Caption Generation Models", by Jason Brownlee, November 22 2017, For deep learning Natural Language Processing.

[3] O.Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge", IEEE transactions on Pattern Analysis and Machine Intelligence, 2016.

[4] Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. CoRR, abs/1506.07285, 2015

[5] Quanzeng You, HailinJin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. CoRR, abs/1603.03925, 2016.

[6] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag.

[7] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences, in: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014, pp. 2042–2050

[8] M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons:a neural based approach to answering questions about images, in: International Conference on Computer Vision, 2015

[9] X. Chen, C. Lawrence Zitnick, "Mind's eye: A recurrent visual representation for image caption generation", CVPR, 2015

[10] Understanding LSTM Networks by Colah,online blog , 2015