International Journal of Research in Engineering, Science and Management
Volume-3, Issue-12, December-2020
https://www.ijresm.com | ISSN (Online): 2581-5792

Deep Learning Research for Eye Sight Assisting Model

Hritik Valluvar[1], Utkarsh Shukla[2*], Nirav Agarwal[3]
[1,2,3]*Department of Computer Science and Engineering, SRM Institute of Science & Technology, Lucknow, India*
*Corresponding author: 1998utkarshshukla@gmail.com

***Abstract*: According to the World Health Organization (WHO), 285 million people are visually impaired worldwide, which is nearly 3.6% of the world population. Every year, several blind people lose their lives in accidents. This paper will help in building bridges between visually impaired people and their surrounding environment. This paper proposes a model which generates descriptions of an image with the help of the robust combination of Xception model and LSTM (Long short term memory, a type of RNN). We have implemented a deep neural network recurrent architecture that automatically produces a short description of images. Our models use a CNN (Convolution Neural Network) to extract image features. We then feed these features into the NLP (Natural Language Processing) to generate a description of the image.**

***Keywords*: CNN, Google Text-to-Speech API, LSTM, NLP, Xception.**

1. Introduction

In recent years, Artificial Intelligence(AI) has turned out a lot of heads due to its impressive results. Nowadays, AI is the heart of the innovation economy and thus the base for this project is also the same. Though visually disabled people have not restricted themselves to their disability and contributed a lot to this world. But living in this world without vision, comes with several threats. In-order to reduce such threats, we have worked on a model which has the capability to generate a meaningful sentence from an image, and further will be converted to speech using Google Text-to-speech API. This might create a huge impact on the visually impaired to have a better understanding of surroundings.

This paper presents how Xception based architecture can be used for analysing the images. Many types of research have been developed for the high-level description of images. Quanzenq You [1], proposed the top-down and bottom-up approaches through a model of semantic attention. Ting Yao & Yingwei Pan [2], also proposed the 'boosting image captioning with attributes' using LSTM and RNN. They have integrated inter-attributed correlation into Multiple Instance Learning(MIL). They have used MsCOCO and GoogleNet. Krizhevsky et. al [3] presented a deep convolutional neural network having 8 layers that were used for the image classification operation. It significantly outperformed older methods with large gains. Sonu Pratap Singh Gurjar [4], has also used CNN and RNN architecture for analysing the images.

From the Study, it is observed that Convolution Neural Network(CNN) and Recurrent Neural Network(RNN) are widely used for labelling the images and then mostly LSTM and ImageNet is used to convert features into sentences. The quality dataset plays an important role in achieving better results. [5] used flickr 8k dataset whereas [1]-[4] uses Microsoft COCO model. The limitation of paper (4) is that the model doesn't identify untrained objects due to the limited size of its dictionary, which can be rectified with the help of a constructive approach as mentioned in [5]-[8]. It generates a sentence based on the firstly learning image features, which asks for the user to identify the unseen object and further it is trained on that data. This results in a 'self-learning' model.

The model proposed is built according to the need of our application. In order to make the model accurate, portable, and fast, we have used the combination of the Xception model(CNN) and LSTM(RNN). We have used the pre-trained Xception model to extract features from the input image, further LSTM uses the information generated from CNN to generate captions of the input image. The Xception model is an interpretation of the Inception model. Inception is considered to lie in between regular convolution and depth wise separable convolution. It has a similar parameter count as the Inception model. Since depth wise separable convolution shows significantly better results on larger datasets because it efficiently uses its model parameter, and it is easy-to-use. Therefore, it is well suited for our application.

Object identification and classification won't be enough for a visually impaired person to map an environment. There is a need for a more accurate description. LSTM generates meaningful sentences based on sequence predictions, i.e. based on previous text, the next word can be predicted. It can also discard non-relevant information.

2. System Model

A. System Model Architecture

The Deep Learning Model proposed in this paper primarily consists of two segments, the Xception model and LSTM. The first segment is made up of a convolutional neural network (CNN) architecture known as Xception Model. It maps cross channel correlations and spatial correlations. The CNN extracts

features from the input images and passes it on to the second segment for sentence generation.

The second segment of the model is LSTM (Long-Short Term Memory) layers. LSTM is described as a special category of Recurrent Neural Networks (RNNs). RNNs use their internal state (memory) to process sequences of inputs which means that in RNN all the inputs are related to each other. This is especially important while generating a sentence. RNNs are actively used in the field of Natural Language Processing (NLP) for sequence classification. LSTM improves upon RNN by remembering the past data in memory and by eliminating the vanishing gradient problem of RNN. The proposed model is trained on a sequence of LSTM layers by taking the input from the CNN. Apart from the two segments, there is a third segment in the architecture, and that is a text preprocessor. The job of the text-preprocessor segment is to preprocess the text. The first step of text preprocessing is to eliminate all the stopwords, numerical and special characters. This is followed by tokenization that breaks down the sentence into tokens (list of words). After this, all captions are used to map a word dictionary that contains all the unique words assigned to a numerical index. Finally, embedding is done to convert the words into a word-vector containing the corresponding numerical indexes of the words. This data along with the input received from the conder is to train the decoder LSTM component of the model.

B. Xception Model

The Xception model is considered as an interpretation or the 'extreme version' of the Inception Module. Assuming the number of independent channel space segments as parameters. A regular convolution would fall on the left side of the axis, and depth-wise separable convolution would fall on the right side of the axis. Inception module is in between regular convolution and depth-wise separable convolution. It has nearly the same number of parameter count as the Inception module. It shows significantly better performance, than the Inception model, on large datasets because it efficiently uses the model parameters.

As mentioned in [11], a convolution layer attempts to learn filters in three dimensions, with height and width as spatial dimensions, and a channel dimension. Thus a single convolution kernel has to perform two tasks simultaneously, i.e mapping cross channel correlations and spatial correlations. The Xception model starts with a 1x1 convolution, which maps cross channels correlations. It is followed by multiple 3x3 convolutions which are separately used to map the spatial correlations. In depth-wise is followed by a point-wise(1x1) convolution which projects the channel output (of depth wise convolution) onto a new channel space. There are two differences between Inception model and Depth-wise separable convolution, out of only one difference is significant. The first difference is that, order of operations in both are different, Inception performs 1x1 convolution first. The second difference is that, in Inception both operations are carried out following a ReLu non-linearity.

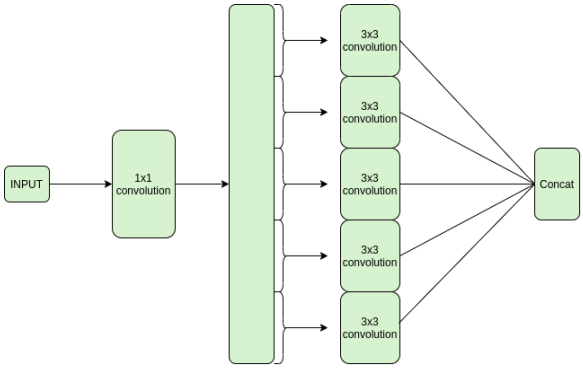

Fig. 1. Xception model

C. LSTM Approach

Long Short Term Memory Networks are a special kind of RNN which is capable of learning long-term dependencies. LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn!

All recurrent neural networks have the form of a chain of repeating modules. In standard RNNs, this repeating module will have a very simple structure, such as a single tanh layer.

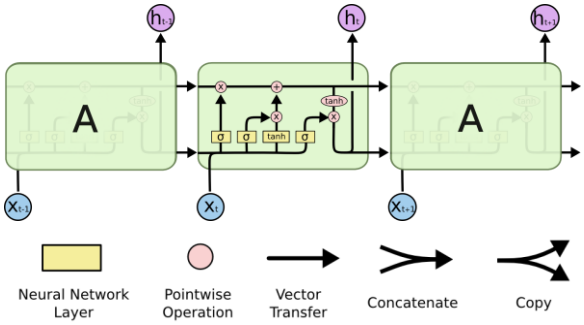

Fig. 2. LSTM approach

In the above diagram, each line carries an entire vector, from the output of one node to the inputs of others. The pink circles represent pointwise operations, like vector addition, while the yellow boxes are learned neural network layers. Lines merging denote concatenation, while a line forking denotes its content being copied and the copies going to different locations.

The advantage of an LSTM cell compared to a common recurrent unit is its cell memory unit. The cell vector has the ability to encapsulate the notion of forgetting part of its previously stored memory, as well as to add part of the new information. To illustrate this, one has to inspect the equations of the cell and the way it processes sequences under the hood.

3. Results and Discussion

A. Dataset

To train a model for image captioning, several datasets are available which include, MS-COCO Dataset, Flickr 8K, and Flickr30k Dataset. For the purpose of this model, Flickr 8k

CANDIDATE	man	in	black	wetsuit	is	surfing									
REFERENCE 1	a	man	dressed	in	black	is	surfing	on	a	large	blue	wave			
REFERENCE 2	a	man	rides	a	surfboard	through	rough	white	and	blue	water				
REFERENCE 3	a	surfer	catches	a	nice	wave	in	the	ocean	and	tries	to	pull	himself	up
REFERENCE 4	a	surfer	is	riding	a	wave									
REFERENCE 5	a	surfer	rides	the	wave										

Dataset is used to help with the computation time. Flickr 8k Dataset consists of 8k Labelled Images with 5. The dataset consists of the images as well as the matching captions. The dataset provides 5 image captions for each image.

B. Evaluation

In this paper, the BLEU (Bilingual evaluation understudy) algorithm is used to evaluate the quality of sentences generated by the model. BLEU is an evaluation metric for evaluating the quality of text generated by the model, to the natural language. The quality is determined by the closeness between machine generated text and human samples. Both text and fluency are major factors in evaluation, which determines the score. Scores are usually calculated by comparing the segments to a set of quality reference, the average of every segment determines the overall performance. The score ranges from 0 to 1. BLEU algorithm is a modified version of precision evaluation. It compares the predicted text to multiple references. A unigram precision evaluates predicted text by searching each and every word of predicted text in any reference. The loophole in this unigram precision is that poor translation can also achieve a high precision, because it doesn't evaluate the fluency of the predicted text. In practice, a unigram score is not adequate to determine the quality of a text. BLEU also computes the same metric using n-gram, it can be defined as the highest length of text similar to any reference text. Unigram score refers to the amount of information retained, whereas n-gram score shows the fluency of the text. For sample, we've used the image given below for testing purposes. The image belongs to the test dataset from FLickr8k dataset. This image contains five different references describing the image. The first row in the above table contains the predicted text.

Fig. 3. Test image

The predicted sentence is short. Also it has been observed that BLEU score is biased towards short sentences because longer n-grams are not possible. Hence the possibility of shorter text to good scores are high. We've evaluated till 2-grams as our predicted text is short. The result for 1-gram is 0.625 which indicates that the text retains most of the important information. The result for 2-gram is 0.422577, which indicates that the fluency is moderate.

BLEU-1	0.625
BLEU-2	0.422577

4. Conclusion

With the help of AI there's a possibility to assist visually impaired people using computer vision and natural language processing. All the state-of-art tech can be compressed into a single standalone IoT device, which will be able to assist its user to navigate in their surroundings. The IoT device can be fitted in day-to-day accessories such as spectacles or caps. The BLEU score for our model is 0.625, which is not enough to be implemented in the real world, but it increases the possibility of a solution which would help visually impaired people.

5. Limitations

This study has some potential limitations. Some major limitations were noticed during the development of the model. One such factor is the lack of element of intelligence which can determine multiple actions leading towards a single output or multiple outputs defining a scenario. Furthermore, there is also a concern for recognising high-speed objects and predicting their course path to alert the user to take action. One real-life implementation concerns the fact that there are a lot of events happening every second, most of such events are irrelevant to the user. Both computing of the extra information and feeding it to the user will be a waste of resources and will also not be user-friendly. During the model training phase, although LSTM solved some of the major issues of RNN such as vanishing gradient, still it is only capable of a remembering sequence of 100s, not more than that. Adding on to that, LSTM requires intensive hardware resources. One more limitation covers the hardware aspect of the study that deals with the deployment of the proposed deep learning model. Four basic components required for the working of this model are processing unit, camera, speakers, and battery. Depending on the size of the product, it is really difficult to fit all these components in a handy device. Since, size and weight is a major factor, there's a trade-off between compute power and size.

References

[1] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, Jiebo Luo; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4651-4659.
[2] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, Tao Mei; Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4894-4902.
[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks" NIPS, pages 1097–1105, 2012.
[4] Sonu Pratap Singh Gurjar, Shivam Gupta and Rajeev Srivastava. "Image Content Description using LSTM Approach".
[5] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image."
[6] descriptions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128-3137, 2015.
[7] Kulkarni, Girish, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, AlexanderC. Berg, and Tamara L. Berg. "Babytalk: Understanding and generating simple image descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12 (2013):2891-2903
[8] Elliot, Desmond and Frank Keller. "Image Description using Visual Dependency Representations," in EMNLP, vol. 13, pp. 1292-1302, 2013.
[9] Mao, Junhua, Wei Xu, Yi Yang, Jiang Wang and Alan L. Yuille. "Explain images with multimodal recurrent neural networks," 2014.
[10] Ting Yao, Yingwei Pan, Yehao Li, Tao Mei; Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 684-699.
[11] Ruturaj R. Nene. "Caption Generation for Images Using Deep Multimodal Neural Network."
[12] Francois Chollet, "Xception: Deep Learning with Depth Wise Separable Convolutions."
[13] Records for Blind people: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment
[14] Hodosh, Micah, Peter Young, and Julia Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics." Journal of Artificial Intelligence Research, vol. 47, pp. 853-889, 2013.