Automatic Image Caption Generation System

Satyabrat Mandal:
Student in Department of Information Technology,
Smt. Kashibai Navale College Of Engineering,
Savitribai Phule Pune University,
Ambegaon, Pune, Maharashtra, India.

Nachiket Lele:
Student in Department of Information Technology,
Smt. Kashibai Navale College Of Engineering,
Savitribai Phule Pune University,
Ambegaon, Pune, Maharashtra, India.

Chinmay Kunawar:
Student in Department of Information Technology,
Smt. Kashibai Navale College Of Engineering, Savitribai Phule Pune University,
Ambegaon, Pune, Maharashtra, India.

Abstract:- Computer vision has become omnipresent in our society, with uses in several fields. In this project, we specialize in one among the visually imparting recognition of images in computer vision, that is image captioning. The problem of generating language descriptions for images is still considered a problem which needs a resolution and this has been studied more regressively within the field of videos. From past few years more emphasis has been given to still images and their descriptions with human understandable natural language. The task of detecting scenes and object has become easier due studies that have taken place in last few years. The main motive of our project is to train convolutional neural networks and applying various hyper parameters with huge datasets of images like Flicker 8k and Resnet, and combining the results of these images and their classifiers with a recurrent neural and obtain the desired caption for the image. In this paper we would be presenting the detailed architecture of the image captioning model.

Keywords:- *Computer Vision, Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Xception, Flicker 8K, LSTM, Preprocessing.*

I. INTRODUCTION

In the past few years the field of AI namely Deep Learning has developed a lot because of its impressive leads to terms of accuracy in comparison with the already existing Machine learning algorithms. It might be a difficult task to get a meaningful sentence from an image but if done successfully, it can have a huge impact, as an example helping the visually impaired to possess a better understanding of images.

Image captioning is considered a bit more difficult in comparison with image classification, which has been the main focus point within the computer vision community. The task to find the relationship between the objects in the image is the most important factor to consider. In addition to the visual understanding of the image, the above semantic knowledge has got to be expressed during a tongue like English, which suggests that a language model is required. The attempts made within the past have all been to stitch the both two models together.

In the model proposed we attempt to combine this into one model which consists of Convolutional Neural Network (CNN) encoder which usually creates image encodings. We use the Xception architecture with some modifications. These encodings are then passed to a LSTM network layer which are a kind of Recurrent Neural Network. The specification used for the LSTM network add similar way because the ones utilized in machine translators. We then use Flickr8k dataset to train and coach the model. The model generates a caption as an output that is to be supported by the dictionary which is formed from the tokens of caption within the training set.

II. PROBLEM DEFINATION

Image caption generation has been considered as a challenging and significant research area that is constantly following advancements in statistical language modelling and image recognition system. Caption generation can benefit many like helping the visually impaired by aiding them by enabling automatic captions of the millions of images uploaded to the internet every day which will help them understand the World Wide Web.

III. PROBLEM SOLUTION

In our perception the main components of image captioning are CNN and RNN. And then merging them both to get the captions as follows for the images.

Convolution Neural Network (CNN)

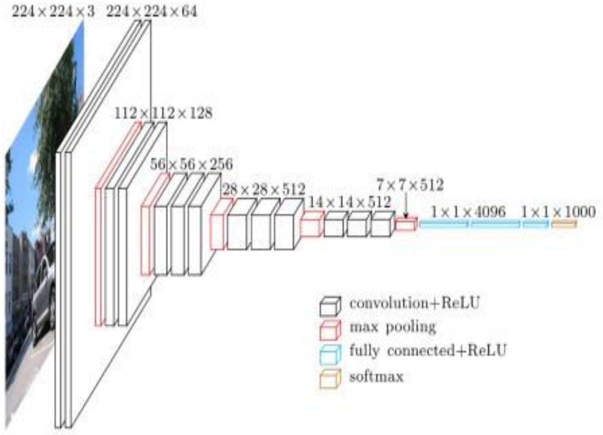


Fig.1: CNN Architecture

Convolutional Neural Network (CNN) has been an important factor for the improvement in image classification. Image net Large Scale Visual Recognition competition (ILSVRC) have various open source deep learning frameworks like ZFnet, Alexnet, Vgg16, Resnet, Xception, etc. which do have great ability to classify images. And for encoding our images we are using Xception in our model. The image used for classification needs to be a 224*224 image. The one and only preprocessing done is by subtracting the mean RGB values from each pixel determined from the training images. The CNN layer of 3*3 filters and the stride length is fixed at 1. Max pooling is done using 2*2-pixel window having stride length of 2. Images need to be converted into 224*224-dimensional image. The output of the encoder would thus be a 1*1*4096 encoded and which is then passed to the language generating RNN. We do have many other frameworks which are successful in this field like Resnet but they are very expensive computationally since the number of layers in Resnet is very high as compared to Xception therefore it requires a very powerful system.

Recurrent Neural Network(RNN)

Recurrent neural networks are types of artificial neural network where the connections between units are formed by a directed cycle. Recurrent neural can also be termed as networks with loops where the information usually persist in networks. Recurrent neural network can be considered as multiple copies of same network with each network passing the message to its successor. One of the problems with RNNs is that they do not take long-term dependencies into account. To surpass the problem which usually occurs due to of “long term dependencies”, Hochreiter and Schmidhuber put forward a term called the Long Short-Term Memory. The key and the importance that backs the LSTM network is the horizontal line that is running on the top which is known as the cell state. All the repeating modules are supported by the cell states and every module is modified with the help of gates. All These things lets LSTM network to persist all the available information.

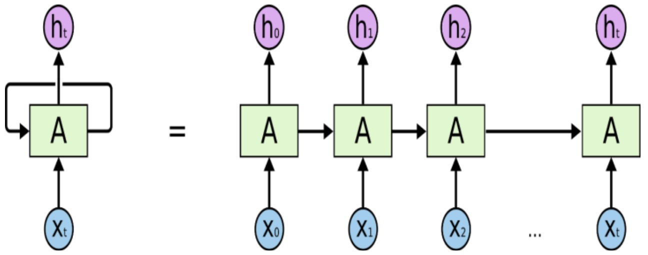


Fig. 2: A simple neural network unrolled into simple neural network.

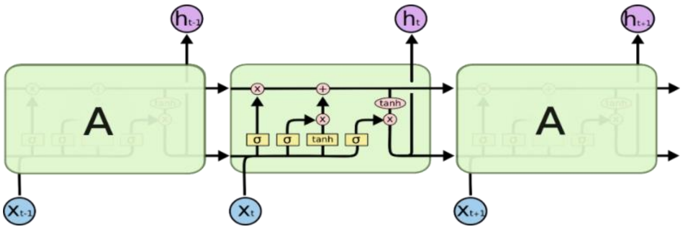
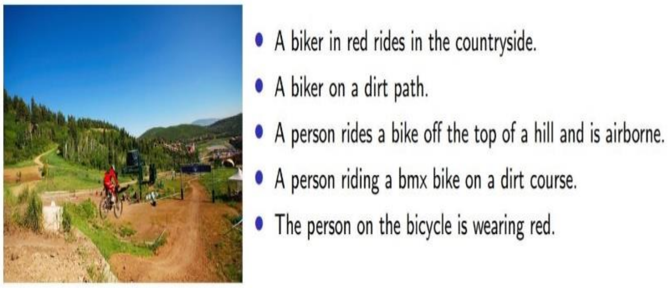


Fig. 3: Four interacting layers in a LSTM layer

Datasets to be used

For the task of image captioning we use Flickr8k dataset. The dataset contains 8000 images with 5 captions per image. The dataset by default is split into image and text folders. Each image has a unique id and the caption for each of these images is stored corresponding to the respective id.

The dataset contains 6000 training images, 1000 development images and 1000 test images.



- A biker in red rides in the countryside.
- A biker on a dirt path.
- A person rides a bike off the top of a hill and is airborne.
- A person riding a bmx bike on a dirt course.
- The person on the bicycle is wearing red.

Fig. 4: Sample photo with captions from the Flickr8k dataset

Tokenizing Captions

The Recurrent Neural Network (RNN) segment is trained on the captions that are given in the Flicker 8K dataset. We are supposed to train the RNN to forecast the succeeding word of a sentence that is inspired from the foregoing words. Because of t this we are supposed to alter the captions linked with the images that are in the list of tokenize words. This can turn any string into a list of integers.

Firstly, we go through all the captions that are trained and then generate a dictionary that plots all the distinctive words to a numerical index format. So, each and every word that we pass through will have an integer value accordingly that we would be able to see in this dictionary. The words of these dictionaries are referred to as our vocabulary. It remold each and every word into a caption and then it is converted to a vector format that is desired to be used. After this step, we are supposed to train the RNN that can help to figure out the next word in a sentence.

IV. RESULTS AND OBSERVATION

We did test our system by testing around 300 images of different category, and we observed that for about 178 images we got perfect captions and these object basically were having very few objects that is around one or two but when the image object wears a multicolor shirt it couldn't recognize the colors and determines the brightest color like red as the main color like in fig 5. And this also cannot determine moving and still objects and also doesn't determine multiple same objects like in fig.6.

We did find a precision of 63% from our observation which was better than other dataset which was used before it.

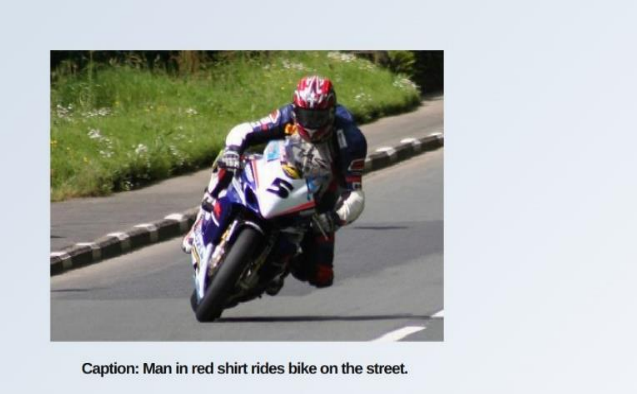
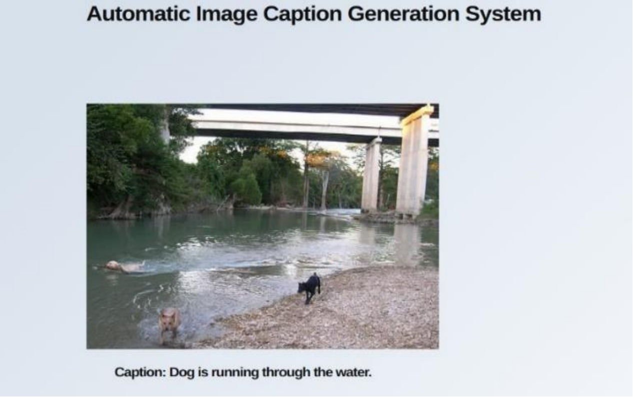

Caption: Man in red shirt rides bike on the street.

Fig.5: snapshot of the output


Automatic Image Caption Generation System

Caption: Dog is running through the water.

Fig.6: snapshot of the output

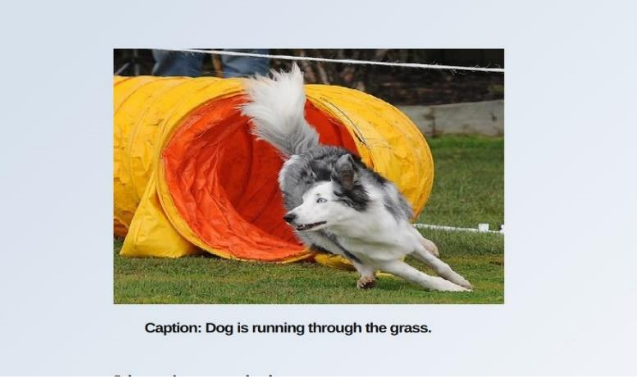

Caption: Dog is running through the grass.

Fig.7: snapshot of the output

V. CONCLUSION

Image caption generation involves Convolutional Neural Network and Long short-term Memory to detect objects and captioning the images. Image caption generation has many advantages, we discussed a convolutional approach for image caption generation. Even though automatically generating captions for images is a complex task, with the help of models and powerful deep learning networks, it is possible to obtain good results.

In the future scope we further can extend our project in the next higher level by modifying our model for generating captions even for the live video. Currently our model generates captions only for the image, which itself a difficult task and captioning live video is much more complex to create. This is completely GPU based and captioning live video cannot be possible with the general CPUs. Captioning video is a popular research area in which it is going to change the way of life of the people with the use cases being widely usable in almost every domain. It automates the major tasks like video surveillance and other security tasks.
Also, we can extend our work by enhancing our model to develop a voice clip for the caption that is generated by the system this will help the visually impaired people to get an idea about the image.

ACKNOWLEDGEMENT

We are very grateful to all the teachers of our college who have helped us with their valuable guidance towards the completion of our project entitled "Automatic Image Caption Generation System" as this is a part of our syllabus of Bachelor of engineering (B.E) course. We convey our genuine regards towards our department who have helped us with the essential guidance.

We want to convey our special thanks to Prof.M.V.Raut for providing us with all the necessary instructions and guidance, solving our problems, giving us an insight on each and every step and contributing with her knowledge and experience in making this project come true. We are also very thankful to Prof. R. H. Borhade, Prof. L.V. Patil for their invaluable support. The acknowledgement will be incomplete without mentioning our Principal Prof. Dr. A. V. Deshpande, whose constant assistance encouragement has been highly important in making our project.

We would like to express our gratitude towards our parents & friends for their kind cooperation and encouragement which help us in finishing this project.

REFERENCES

[1]. Fang, Hao, et al. "From captions to visual concepts and back." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
[2]. Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International Conference on Machine Learning. 2015.

ISSN No: 2456-2156

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

[4]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 (2017).

[5]. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., & Ng, A.Y. (2014). Grounded Compositional Semantics for Finding and Describing Images with Sentences. TACL, 2, 207-218.