International Journal of Emerging Trends in Engineering and Basic Sciences (IJEEBS)
ISSN (Online) 2349-6967
Volume 7, Special Issue 2 (March-April 2020), PP. 276-280

Image Captioning

Sancheti Ramteke, Kartik Wagh

Department of Computer Science, Government College of Engineering, Nagpur

Abstract: Throughourworkandresearchonthesupposed"ImageCaptioning"weareaimingtowards finally coming up with a permanent solution to a very common unnoticed problem we face on a major scale on a very regular basis. In our everyday life we often struggle to find the correct, appropriate and exact words to describe what we see, feel or imagine. Through this project, we are aiming towards creating the utility for the same. Giving an image in any format as an input and getting a proper word wise description for that particularimage.

Keywords: Beam Search algorithm, Feature Vector, Model Implementation, Image Processing

I. INTRODUCTION

Image captioning basically, in short can be described as a platform wherein the user will provide this application with an input image and it will provide the user with a proper description herein referred to as the 'caption' for the same. This simple seeming concept is in fact a very crucial tool whichcanbenefitpeoplefromdifferentbackgroundsandinsurprisinglydiversewaysaswell.These range right from the trivial applications like coming up with suitable captions for social media uploads to something major and beneficial to the society like generating audio clips for the visually impairedpopulationviatext-to-speech methodology. Inadditiontothis,thisprojectwillalsoplaya significant role in filtering the content on social media platforms and clearly helping in differentiating between what is appropriate and suitable for uploading on a social platform and what is not, thus helping in developing a more cyber secure environment for theusers.

Automatically generating a natural language description of an image is a task close to the heart of image understanding. In this paper, we present a multi-model neural network method closely related to the human visual system that automatically learns to describe the content of images. Our model consists of two sub-models: an object detection and localization model, which extracts the information of objects and their spatial relationship in images respectively; besides, a deep recurrent neural network (RNN) based on long short-term memory (LSTM) units with attention mechanism for sentences generation. Each word of the description will be automatically aligned to different objects of the input image when it is generated. This is similar to the attention mechanism of the human visual system.

II. MODULES &METHODOLOGY

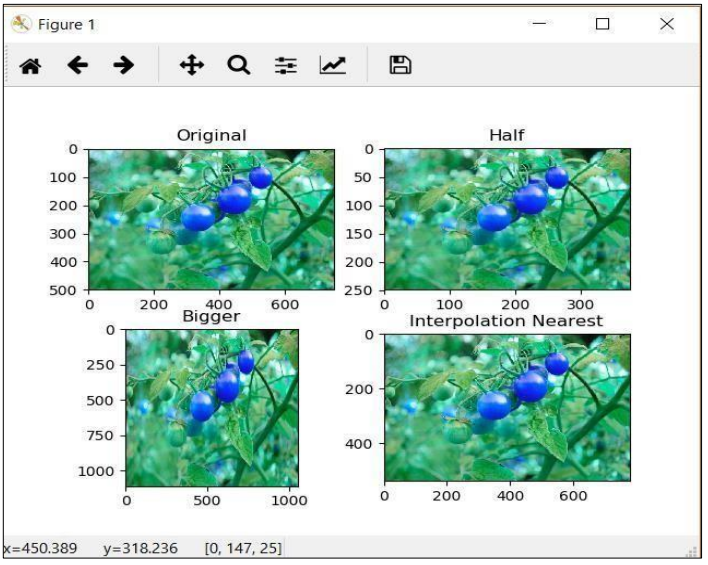
Image processing is a vital research area and the utilization of images increases in various applications. On different research areas, scientists are working on such as image compression, image restoration, image segmentation etc. to enhance the existing image processing techniques and invent new method of solving image processing problems. The latest image processing applications such as medical image processing, satellite image processing, and molecular image processing uses various image processing techniques. Conversion of color image to grayscale image is one of the image processing applications used in different fields effectively. In publication organizations' printing, a color image is expensive compared to a grayscale image. However, the conversion ofa

color image to a grayscale image requires more knowledge about the color image. So that's why we are using RGB format.

The entire captioning process can be roughly divided into two major modules which are listed below:
(i) Image Pre processing
(ii) Model Training (using Datasets)

1. IMAGEPRE-PROCESSING

This is the very first module in Image Processing System.
- Firstly, Batches of Images are created for processing.
- The images get converted to the required format.
- After which they become ready for the module
- Rigorous scanning of images through module leads to identification of objects.



A normal brief representation of image processing is shown through the picture below:

2. MODELTRAINING

Model training basically involves training or teaching the machine by providing the system with a set of images. Through the objects present in these images, the machine can learn and get trained to generate the desired output in the form of captions. However, by providing only a limited number of images in the set, we can't get the exact accuracy that we desire since the material we use for learning isn't quite sufficient. Hence, in order to overcome this problem, we generally prefer larger datasets with approximately around 10,000 images. This number is ample for the model to be trained perfectly and thus generate accurate outputs.

The model Training process is closely associated with the concept of Neural Networks which work in combination so as to generate the expected sentence or a phrase (caption).

The two neural networks are named Convolution Neural Network (CNN) which is responsible for the detection and identification and Recurrent Neural Network (RNN). Both of these networks are described below:

2.1 Object Detection:

As the foundation of image understanding, object detection has been investigated for years.Convolutional Neural Networks [1] were first introduced by the R-CNN [2] for object detection. It processes the regions of interest independently, which is time-consuming. Then SPP-net [3], Fast R-CNN [4] were proposed to share convolutional layers among regions in classification. Ren, et al. proposed Faster R-CNN by utilizing CNN to do region proposal [5]. YOLO [6] and SSD [7] shared more convolutional layers for region proposal and region classification and made detection even faster. ViP-CNN is based on Faster R-CNN due to its superior performance.

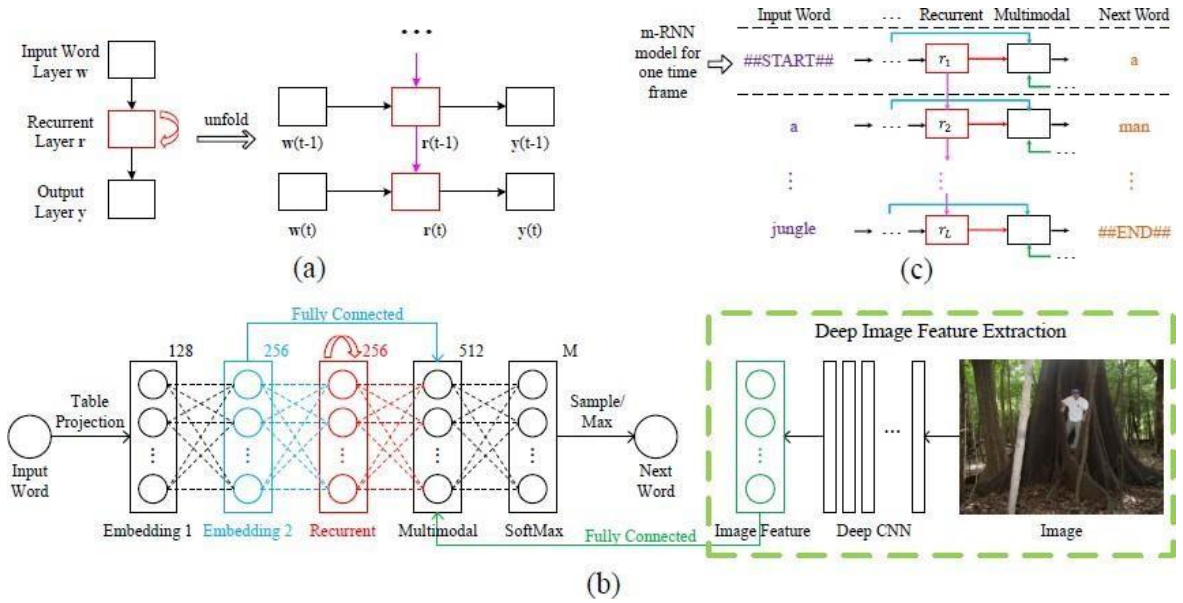
However, it is not trivial to arrange the subject, predicate and object into an end-to-end framework. We propose PMPS and corresponding training scheme that uses the entire phrase to guide learning. The entire design is based on careful analysis of the specific problem.

2.2 Image Caption:

Describing image with natural language have been explored for many years [8, 9, 10, 11, 12, 13]. Recently, using the visual features from CNN, Recurrent Neural Networks (RNNs) [14, 15] have been adopted to generate captions because of its success on processing natural language. Combining the RNN and CNN becomes a standard pipeline on solving the Image Captioning problems [16, 17, 18, 19, 20, 21]. However, the pipeline does not fit for visual relationship detection due to the difference between the sentence for image captioning and phrase for visual relationship detection. Compared to sentences, the phrase has fixed structure of subject-predicate-object. In addition, most of the related works focus on the whole image or image region, while the visual relationship detection targets on the region and its sub regions.

2.3 Recurrent Neural Network (RNN)

We briefly introduce the simple Recurrent Neural Network (RNN) or Elman network [22] that is widely used for many natural language processing tasks, such as speech recognition [23,24]. Its architecture is shown in Figure.



It has three types of layers in each time frame: the input word Neural Network (m-RNN) architecture. (a). The simple RNN. (b). Our m-RNN model. The input of our model is an image and its corresponding sentences (e.g. the sentence for the shown image is: amanatagiant tree in the jungle). The model will estimate the probability

distribution of the next word given previous words and the image. This architecture is much deeper than the simple RNN.(c). The illustration the unfoldedm-RNN. The model parameters are shared for each temporal frame of the m-RNN model.

layerw, the recurrent layer r and the output layer y. The activation of input, recurrent and output layersattimetisdenotedasw(t),r(t),andy(t)respectively.w(t)istheone-hotrepresentationofthe currentword.Thisrepresentationisbinary,and hasthesamedimensionofthevocabularysizewith only one non-zero element. y(t) can be calculated asfollows:

x(t) = [w(t) r(t - 1)]; r(t) = f1(U x(t)); y(t) = g1(V r(t)); (1)

where x(t) as a vector that concatenates w(t) and r(t -1), f1(.) and g1(.) are element-wised sigmoid and softmax function respectively, and U, V are weights which will be learned.

ThesizeofRNNisadaptiveto thelengthoftheinputsequenceand therecurrentlayersconnectthe sub-networks in different time frames. Accordingly, when we do the backpropagation, we need to propagate the error through recurrent connections back in time[25].

III. CONCLUSION

Through this Research paper on Image captioning we have tried to briefly demonstrate the general idea behind the working system of Image captioning. We have tried to explain the reader what exactly is the basic skillset required to get this system running. Image captioning, as mentioned in the methodology section is majorly a two-step task. These steps are image pre-processing as well as machine training with the help of neural networks (Convolution and Recurrent neural Networks).As far as the practical real life applications of this system are concerned, it can be extended in the most diverse ways possible. This means it can be used for most minimalistic purposes like generating a proper aesthetic caption for any chosen picture for social media platforms and thus also acting as a proper filter for uploading the right kind of content on these platforms. The application can be as huge as being the backbone of the system developed for the visually impaired population to understand and thus see images through the proper text to speech technique that has been developed earlier. This application is constructively working towards the welfare of the society by helping a considerably statistically large part of the humankind across the globe. We are however lacking faster and more time efficient training techniques and thus further hoping to develop a solution for thesame.

IV. ACKNOWLEDGMENT

This research was supported by Department of Computer Science, Government College Of Engg, Nagpur.We thank Professor MrsMuktaWagh, Department of Computer Science, Government College Of Engg,Nagpur for assistance and for comments that greatly improved the manuscript.

We would also like to show our gratitude to the NiharikaKalambe and YadneshWadichar, Department of Computer Science, Government College OfEngg, Nagpur for sharing their pearls of wisdom with us during the course of this research.

REFERENCES

Journal Papers:

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.3

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.3

[3] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *CVPR*, 2014.3

[4] R. Girshick. Fast r-cnn. In ICCV, 2015.3

[5] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*,2015.3

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv*preprint arXiv:1506.02640, 2015.3

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed.Ssd: Single shot multibox detector. *arXiv*preprint arXiv:1512.02325, 2015.3

[8] K. Barnard, P. Duygulu, D. Forsyth, N. d. Freitas, D. M. Blei,and M. I. Jordan. Matching words and pictures. JMLR, 2003.3

[9] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, 2010.3

[10] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In ICCV, 2011.3

[11] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi,A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. TPAMI, 2013.2

[12] P.Kuznetsova,V.Ordonez,A.C.Berg,T.L.Berg,and Y.Choi.Generalizingimagecaptionsforimage-textparallelcorpus.InACL,2013.3

[13] R. Socher and L. Fei-Fei. Connecting modalities: Semisupervised segmentation and annotation of images using unaligned text corpora. In *CVPR*, 2010. 3

[14] P. J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. Neural Networks,1988.3

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation,1997.3

[16] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov,R. S. Zemel, and Y. Bengio. Show, attend andtell: Neural image caption generation with visual attention.*arXiv*preprint arXiv:1502.03044, 2015. 3

[17] J. Johnson, A. Karpathy, and L. Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning.*arXiv*preprint arXiv:1511.07571, 2015.3

[18] X. Chen and C. L. Zitnick. Learning a recurrent visual representation for image caption generation. *arXiv*preprint arXiv:1411.5654, 2014.3

[19] J. Donahue, L. Anne Hendricks, S.Guadarrama,

[20] M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell.Long-term recurrent convolutional networks for visual

recognition and description. In *CVPR*, 2015. 3.

[21] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng,P. Doll´ar, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From captions to visual concepts and back. In *CVPR*, 2015.3

[22] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*,2015.3

[23] J. L. Elman. Finding structure in time. Cognitive science, 14(2):179–211,1990.

[24] T. Mikolov, M. Karafi´at, L. Burget, J. Cernock`y, and S. Khudanpur. Recurrent neural network based language model. In INTERSPEECH, pages 1045–1048,2010.

[25] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur. Extensions of recurrent neural network language model. In ICASSP, pages 5528–5531, 2011.

[26] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors.Cognitive modeling, 1988.