RESEARCH ARTICLE OPEN ACCESS

A KERAS implementation model for Image Captioning for Object detection using Deep Learning

Dr. T. S. Suganya [1], Mrs. M. Divya [2], T. Santhosh Kumar [3] K. Prem Kumar [4]

[1],[2] Assistant Professor, Department of Computer Applications, SRM Institute of Science and Technology, Ramapuram - Chennai.[3],[4] Student, Department of Computer Applications, SRM Institute of Science and Technology, Ramapuram), Chennai.

ABSTRACT

Nowadays, people necessitate engendering captions for multiple reasons such as, posting an image on social media, creating news headlines from an image and many more. An Image Captioning system intends to produce captions for an image automatically instead of manually writing. It delivers a descriptive sentence for an image, that helps people to better understand the semantic meaning of an image. Image understanding is an essential technique to interpret semantic image data which can be implemented by VGG16. Image Captioning is an application for both Natural Language Processing and Computer Vision and can be achieved using either Traditional Machine Learning approach or Deep Learning approach. The necessity in performing the intended task is detecting objects and establishing relationships among objects. Feature Extraction is a technique for converting the image into a vector for further processing. The objects and image content are forwarded to the LSTM that will connect the words to produce a descriptive sentence. The work carried out by this thesis presents the implementation model for Object Detection based Image Captioning using Deep Learning. Kera's library, NumPy and Collab notebooks is used for making of this project. Flickr dataset and CNN is usedfor image classification.

Keywords: - **Image** captioning, LSTM,VGG16,Deep Learning, Object Detection

I. INTRODUCTION

An Image Captioning task consists of describing an image in sentence form. This requires brief knowledge of computer vision and natural language processing. The challenging task in the process of caption generation is to understand the semantics, that contains the objects and other image information, and knowledge of natural language processing. The sentence generation needs the establishment of a relationship between the extracted objects. Image information can be extracted using the technique known as Feature Extraction. An image contains various information such as objects and its meaning in the context of an image. We use deep learning techniques for image understanding, the Convolution Neural Network (CNN) is used, which is followed by Recurrent Neural Network (RNN) for generating the captions. Image Captioning has its roots in many real-life applications.

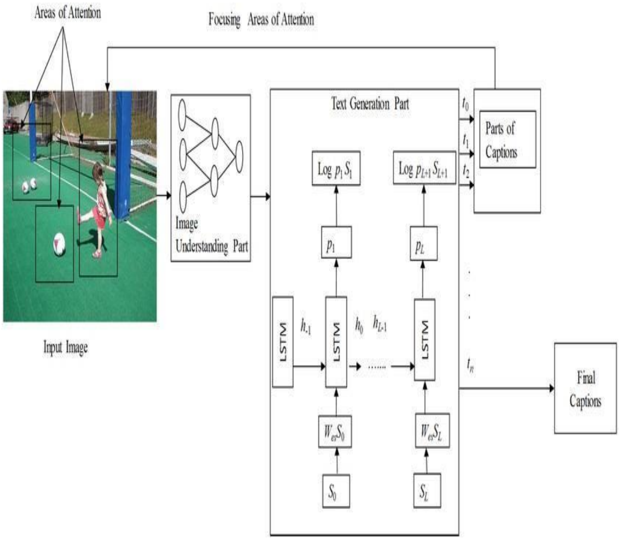
II. PROPOSED WORK

In this approach, a neural model is designed which generates descriptions for images in natural language. CNN is used as an image encoder. Firstly, pre-training is done for image classification tasks and then the RNN decoder uses this last hidden layer as input to generate the sentence. In this work, a neural framework is proposed for generating captions from images which are basically derived from probability theory. By using a powerful mathematical model,it is possible to achieve better results, which maximizes the probability of the correct translation for both inference and training.

III. SYSTEM ARCHITECTURE

The figure below shows the whole design of our working model including the components and the states occurring during the execution of the process.

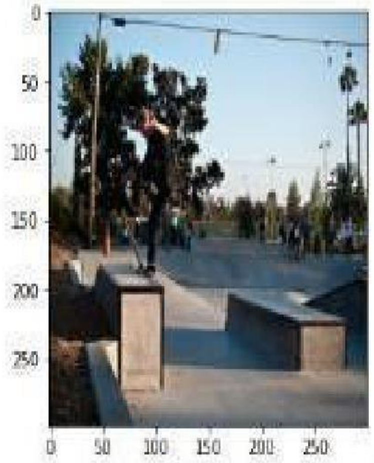
It displays the very initial process of image feeding. These images are then parsed and broken down into image vectors which separates the area of attention from other area but stores all the data regarding the image. This data is then fed to the model. This can also be done using Flickr8k datasets. These datasets consist of 8000 images through which the model can be trained. The CNN is used in the encoding and STM is used in the decoding the descriptive data which play the image again and aging develop the caption with the help of natural language processing. Initially while feeding the image, the respective descriptions is also uploaded for 6000 image and rest 2000 images are used for testing. Thus, the model also learns the description and generates on its own when the new image is given. As the last process, the output for the given image is generated.



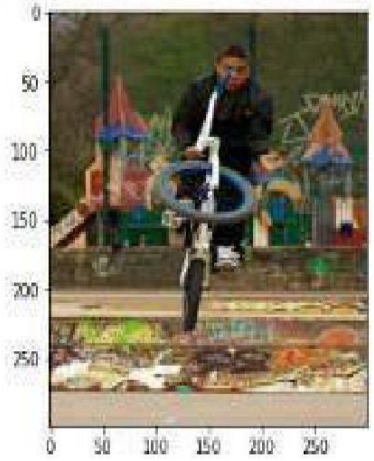
IV. RESULTS

The above implementations depict results using VGG16 and InceptionV3 as CNN in image captioning. The results were conducted on the Flickr8K Dataset with 1000 test images. The model was evaluated using CIDEr score with both the CNN variants. The resulting CIDEr score of both VGG16 and InceptionV3 are 0.3692and 0.3572 respectively. According to the training using those CNNs, it was observed that the InceptionV3 provides approximately similar results but with a greater number of epochs than that of VGG16. While using VGG16 model, the model generated required 7 epochs while the InceptionV3 requires 12 epochs to get similar results. The use of LSTM resulted in a CIDEr score of 0.39 after 7 epochs.

Predicted Caption: a brown dog is biting a brown dog



Predicted Caption: a boy is jumping off a ramp of a ramp



Predicted Caption: a man in a blue shirt is riding a bicycle

V. CONCLUSION

In this paper we have thrown some light on object detection and reviewed deep learning-based image captioning methods. We discussed different evaluation metrics and datasets with their strengths and weaknesses. A summary of experimental results is also given. We briefly outlined potential research directions in this area. Although deep learning-based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time. We have used Flickr_8k dataset which includes nearly 8000 images, and the corresponding captions are also stored in the text file. Although deep learning -based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time. The scope of image-captioning is very vast in the future as the users are increasing day by day on social media and most of them would post photos. So this project will help them to a greater extent.

REFERENCES

[1] J. Brownlee, “How to Develop a Deep Learning Photo Caption Generator fromScratch,” Machine Learning Mastery, Jun. 26, 201 https://machinelearningmastery.com/develop-adeep-learning-caption-generation-model-in-python/ (accessed Jul. 09, 2020).

[2] “[1707.07102] OBJ2TEXT: Generating Visually Descriptive Language from Object Layouts.” https://arxiv.org/abs/1707.07102 (accessed Jul. 20, 2020).

[3] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, “A Comprehensive Survey of Deep Learning for Image Captioning,” ArXiv181004020 Cs Stat, Oct. 2018, Accessed: Jul. 20, 2020. [Online]. Available: http://arxiv.org/abs/1810.04020.

[4] G. Nishad, “Automatic Image Captioning: Building an image-caption generator from scratch!” Medium, Mar. 12, 2019.

[5] “deep learning - What’s the commercial usage of ‘image captioning’?,” Artificial Intelligence Stack Exchange. https://ai.stackexchange.com/questions/10114/whats-the- commercialusage-of-image-captioning (accessed Jul. 20, 2020).

[6] D. Hutchison et al., “Every Picture Tells a Story: Generating Sentences from Images,” in Computer Vision – ECCV 2010, vol. 6314, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 15–29.

[7] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, “Composing Simple Image Descriptions using Web-scale N-grams,” p. 9.

[8] M. Hodosh, P. Young, and J. Hockenmaier, “Framing Image Description as a Ranking Task Data, Models and Evaluation Metrics Extended Abstract,” p. 5.

[9] P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi, “Collective Generation of Natural Image Descriptions,” in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Jeju Island, Korea, Jul. 2012, pp. 359– 368, Accessed: Dec. 24, 2019. [Online]. Available: https://www.aclweb.org/anthology/P12- 1038.

[10] N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman, and J. Yuvaraj, “Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach,” in 2019 5th International Conference on Advanced Computing Communication Systems (ICACCS), Mar. 2019, pp. 107–109, doi: 10.1109/ICACCS.2019.8728516.

[11] C. Amritkar and V. Jabade, “Image Caption Generation Using Deep Learning Technique,” in 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Aug. 2018, pp. 1–4, doi: 10.1109/ICCUBEA.2018.8697360.

[12] P. Shah, V. Bakrola, and S. Pati, "Image captioning using deep neural architectures," in 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS), Mar. 2017, pp. 1–4, doi: 10.1109/ICIIECS.2017.8276124.

[13] F. Fang, H. Wang, and P. Tang, "Image Captioning with Word Level Attention," in 2018 25th IEEE International Conference on Image Processing (ICIP), Oct. 2018, pp. 1278– 1282, doi: 29 10.1109/ICIP.2018.8451558.

[14] D.-J. Kim, D. Yoo, B. Sim, and I. S. Kweon, "Sentence learning on deep convolutional networks for image Caption Generation," in 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Aug. 2016, pp. 246–247, doi: 10.1109/URAI.2016.7625747.

[15] A. Poghosyan and H. Sarukhanyan, "Short-term memory with read-only unit in neural image caption generator," in 2017 Computer Science and Information Technologies (CSIT), Sep. 2017, pp. 162–167, doi: 10.1109/CSITechnol.2017.8312163.

[16] M. Tanti, A. Gatt, and K. P. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?" ArXiv170802043 Cs, Aug. 2017, Accessed: Jul. 20, 2020. [Online]. Available: http://arxiv.org/abs/1708.02043.

[17] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning," ArXiv161201887 Cs, Jun. 2017, Accessed: Jul. 20, 2020. [Online]. Available: http://arxiv.org/abs/1612.01887.

[18] M. Nguyen, "Illustrated Guide to LSTM's and GRU's: A step by step explanation," Medium, Jul. 10, 2019. https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru- s- a-step-bystep-explanation-44e9eb85bf21 (accessed Jan. 01, 2020).

[19] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," 2002, pp. 311–318.

[20] J. Hui, "Real-time Object Detection with YOLO, YOLOv2 and now YOLOv3,"Medium, Aug. 27, 2019. https://medium.com/@jonathan_hui/real-time-object-detection- with- yoloyolov2-28b1b93e2088 (accessed Jul. 20, 2020).

[21] "Yolo Framework | Object Detection Using Yolo," Analytics Vidhya, Dec. 06, 2018. https://www.analyticsvidhya.com/blog/2018/12/practical-guide-object-detection- yoloframewor-python/ (accessed Jul. 20, 2020).