International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Generating Captions for Images using Machine Learning

Samiksha Lambat and Dr. S. S. Sonawane

Department of Computer Engineering, Pune Institute of Computer Technology Pune, India

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

The recent development in the field of artificial intelligence, every aspect of humans is now on the verge to be automated or observed keenly so as to make sure that be used for training or developing a machine. Keeping this in mind the machine learning algorithm is also evolving. One such ability of humans is to understand things that are there in their vicinity and describe them. So to put this aspect automatically it leads to the thought of image caption generation, where the application tries to generate description correctly based on an image. We are proposing an application that will do so by making use of the encoder-decoder method along with the convolution neural network that will help the machine to train so that it produces correct output.

Keywords: *Caption generation, machine learning, convolution neural network*

Introduction

Image captioning can be categorized under computer vision, natural language processing domain. [4]Considering it with computer vision alone earlier it was a difficult task but with the tremendous development in machine learning and deep learning algorithms and methods, this task has become easy. Image caption generation can be thought of as a process that describes the context of the image based on the person who sees it[11]. It differs from person to person, but if this process gets automated then the variation will depend on how strong we train our system. Various methods are used to provide captions for the system. The traditional system can be thought of as a store and search method, where we have a pool from which we can search for the required caption, this can be known as retrieval based image captioning. Another available method focuses on the syntax and semantics of the sentences formed named as a template-based image captioning. The latest trends that are used recently are the deep learning-based methods, and also uses the attention model.

A. Motivation

It is a good opportunity to learn more about images and language processing models, and how they work. Image caption generation is also known as image annotation which is multi-label classification and multi-label ranking tasks that are popular in the machine learning community, which gathers the attention of many. Also its various applications such as an aid to blind people, CCTV cameras, self-driving cars, etc.

B. Challenges

- Compositionality and Naturalness: For compositionality we need to take into consideration the context of objects in visual scene(Image) and same should be reflected in natural language processing. Naturalness deals with irrelevant semantics that may appear in generated captions.
- Generalization: Some objects are common and may appear in different situations based on its requirements so while training the system may get confused so as which context is correct, so for this we need to have generalization.
- Evaluation: The generation captions are evaluated based on metric but this metrics only consider the sentences to check whether it is correct or not. It does not consider image while using evaluation.

Literature Review

[1][3]proposed a CNN and LSTM based method to develop a captioning system. [2] has a multi-task system that can generate captions from images and vice versa. The system designed by [4] is based on a language CNN that is hierarchical in nature. [5] proposed a new architecture called ARNeT (Auto-Reconstructor Network) . This framework aims to improve the performance of the available traditional encoder-decoder method by reconstructing the previous hidden state with the present one.

[1][2][3][4][5] can say which different methods to generate the sentences and how efficient they would be. [6][7] focuses mainly on how to generate sequences or sentences that are multilingual in nature. [8]combined the template-based image captioning which uses semantics along with machine learning techniques such as SVM. [9] has developed an online platform that deals with a multi-keyphrase problem while forming sentences. [10] developed a stylized way to generate sentences, the style will comprise of humor and romance. [11] way of dealing with captions is similar to retrieval based image captioning, which uses the concept of the meaning of image and sentence along with potentials.

System Architecture

A. Dataset Collection

Dataset used is Flickr30k, which can be obtained from kaggle. We can say this dataset has a standard benchmark for sentence-based image description. In this dataset we have five captions each for a single image. This dataset contains images available from the Flickr website.

B. System Overview

In this paper we propose a system for generating captions from images, which can be helpful in various ways. The proposed system is based on the encoder-decoder method, which can be thought of as a way that could generate the captions more efficiently. The proposed system can be roughly divided into three modules:

Module 1: Load the available image dataset from Flickr. We cannot give an image as a direct input to our system, so we need to extract features from images. For extracting features from images, we are going to use the available Image model that is VGG16 a 16 layer model that will provide the features from images in the form of vector. The output from this is then stored in file. While using this neural network model we skip the last layer as we do not need the classification but we are interested in internal representation before classification. This module deals with the image part.

Module 2: The second module deals with text descriptions available for generating new sentences. Load the available text dataset from Flickr. This dataset contains a huge number of sentences that will be input for the system so we need to first clean this data. We need to follow some steps:

- Need to convert all the words to lowercase.
- Remove all punctuation.
- Remove all words that are single character or less in length (e.g. 'a').
- Remove all words with numbers in them.

The whole cleaned text is then used to create a vocabulary which is smaller in size. If it is small then time required for generation will be lower.These two modules can be performed in any sequence it does not matter.

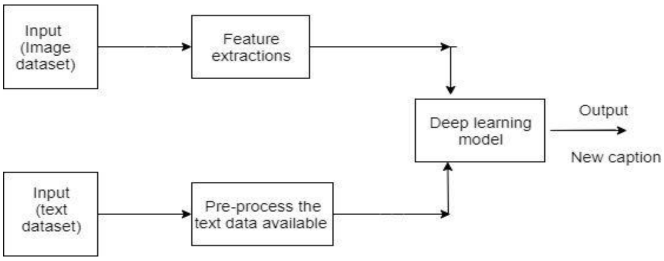


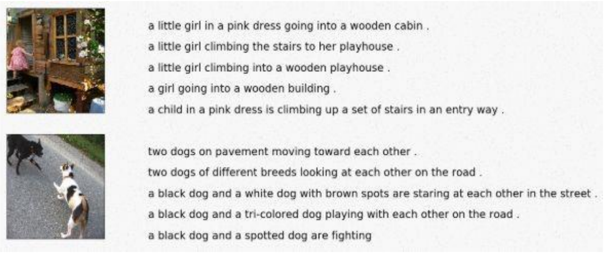
Fig.1, Block diagram

Module 3: This is the deep learning model, which has input from module 1 and module 2, as in fig.1. The input to this module acts as an encoder. In this phase we will be using LSTM (Long Short Term Memory) to generate a sequence of words known as a sentence. LSTM is a type of RNN(Recurrent Neural Network).[4] Used mainly for the sequential processing task, and gives good results. It remembers information for longer periods of time, which can be used to build our data. This phase acts as a decoder phase where we get the newly generated caption for the image. The output is the result of predictions. Both the feature extractor and text processing output a fixed-length vector. These are then merged together and processed by a Dense layer to make a final prediction. This module will perform following task:

1. Loading Data: The output data from module 1and module 2.
2. Defining the Model.
3. Fitting the Model.
4. Generate the output.

Result and Discussions

The Flickr dataset that is available, as in fig.2, shows that there are 5 different caption present for a single image.



a little girl in a pink dress going into a wooden cabin .
a little girl climbing the stairs to her playhouse .
a little girl climbing into a wooden playhouse .
a girl going into a wooden building .
a child in a pink dress is climbing up a set of stairs in an entry way .

two dogs on pavement moving toward each other .
two dogs of different breeds looking at each other on the road .
a black dog and a white dog with brown spots are staring at each other in the street .
a black dog and a tri-colored dog playing with each other on the road .
a black dog and a spotted dog are fighting

Fig.2, Input dataset

From the available caption we need to find out the most frequent words and identify whether they are useful or not as in fig.3. It shows that, 'a','.','in', are the words which when discarded would not affect the sentences available.

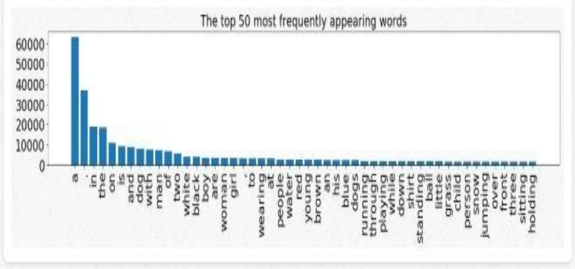


Fig.3, Top 50 most occuring words

Following are the the 50 least occurring words which are then useful for generating captions fig.4.

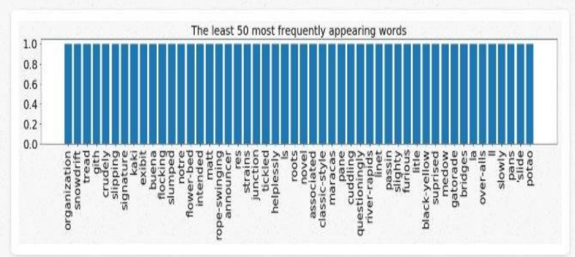


Fig.4, Least 50 occurring words

After performing the cleaning task that is removing punctuation, single character and numeric values we again find out the top 50 most frequently occuring words fig.5, and 50 least frequently appearing words fig.6.

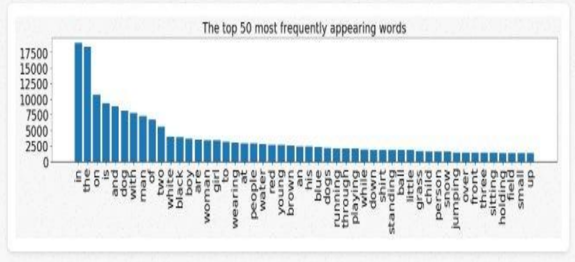


Fig.5, Top 50 most occuring words

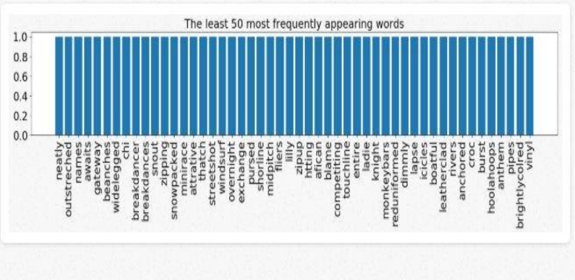


Fig.6, Least 50 most occuring words

The results shown above are used for the sequence generation from available text, where then the model learns to generate new sentences.

Conclusion

The proposed system will develop a way to generate captions for images, this currently is done using the encoder-decoder method. This system aims to reduce the problem where multiple people may have different views on the same image so by using his system we can say that we are having a generalized sentence given by a machine that can satisfy everyone. For future work, we can use different combinations of neural networks to enhance the system.

Acknowledgment

I want to express my thanks to Prof. M. S. Takalikar, HOD of Computer Engineering Department, PICT, for her encouragement and providing the best facilities. I thank all the staff members for their indispensable support and for the most valuable time lent as and when required. I thank all the people who are directly or indirectly involved in this project.

References

[1]. J. Aneja, A. Deshpande, A. G. Schwing- Convolutional Image Captioning in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018. [2]. M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei- Multitask Learning for Cross-Domain Image Captioning in IEEE Transaction on Multimedia, Vol.21, NO. 4, April 2019.

[3]. O. Vinyals, A. Toshev, S. Bengio, D. Erhan- Show and Tell: A Neural Image Caption Generator in arx1411.4555v2, 20 April 2015.

[4]. J. Gu, G. Wang, J. Cai, T. Chen- An Empirical Study of Language CNN for Image Captioning in IEEE International Conference of Computer Vision,2017.

[5]. X. Chen, L. Ma, W. Jiang, J. Yao, and W. Liu- Regularizing RNNs for Caption Generation by Reconstructing the Past with the Present in IEEE/CVF Conference on Computer Vision and Pattern Recognition,2018.

[6]. I. Sutskever, O. Vinyals, Q. V. Le- Sequence to sequence learning with neural network in Proceedings of the Advances in Neural Information Processing Systems, 2014.

[7]. N. Kalchbrenner, P. Blunso- Recurrent Translation Models in Proceedings of the Conference on Empirical Methods in Natural Language Processing 2013.

[8]. Y. Yang, C. L. Teo, H. Daume, Y. Aloimono- Corpus-Guided sentence generation of natural images in Proceedings of the Conference on Empirical Methods in Natural Language Processing 2011, pp. 444-454. [9]. Y. Ushiku, T. Harada, Y. Kuniyoshi- Efficient image annotation for automatic sentence generation in Proceedings of the 20th ACM International Conference on Multimedia, 2012.

[10]. C. Gan, Z. Gan, X. He, J. Gao, L. Deng- StyleNet: Generating Attractive Visual Captions with Styles in IEEE Conference on Computer Vision and Pattern Recognition,2017.

[11]. A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchain, J. Hockenmaier, M. D. Forsyth- Every picture tells a story: Generating sentences from images in Proceedings of the European Conference on Computer Vision, 2010, pp.15-29.