GENERATING CAPTIONS FOR IMAGES USING NEURAL NETWORK

Samiksha Lambat[1], Dr. S. S. Sonawane[2]

Department of Computer Engineering, Pune Institute of Computer Technology 1, Pune, India,

Associate Professor, Pune Institute of Computer Technology 2, Pune, India

Abstract: *Recent studies and development have led us with a way to find out the fundamental problems in artificial intelligence. One such problem is automatic image captioning. Automatic ways to generate captions for images as humans do. This problem can be related to natural language processing and computer vision. It provides a variety of aspects which one needs to understand before implementing it. In this paper, we are proposing a system that uses a method based on CNN and LSTM networks that will generate captions automatically. This can also be known as the encoder-decoder method. We are using the Flickr8k dataset.*

Keywords: Image captioning, natural language processing, neural network.

1. INTRODUCTION

For humans to describe images or anything is easy but for machines, it's very complex. For automatically generating captions, one first needs to understand what objects are present in images, and how they are related to each other. Also to be able to generate a caption that is semantically and syntactically correct. This problem is interdisciplinary. If we divide all these tasks then the correctness of the caption can be natural language processing.

Caption: A child wearing a red jacket, playing on the slide.

Figure 1. Example of image with caption

Identifying objects in the image and extract features can be computer vision[4]. This problem is interdisciplinary. Understanding an image largely relies on acquiring image features. The strategies used for this can be widely divided into categories (1)Traditional machine learning techniques and (2) Deep machine learning techniques.

Traditional machine learning, includes Scale-Invariant Feature Transform(SIFT), the Histogram of Oriented Gradients(HOG), etc. whereas the deep machine learning techniques like CNN or any variants can be used. Similarly Image captioning can also be categorized into (1) Traditional machine learning techniques and (2) Deep machine learning techniques.

Traditional techniques can include retrieval-based, template-based techniques. And deep machine learning techniques include CNN followed by RNN, encoder-decoder

framework, attention guided technique, etc. Application for automatic image captioning system can be in aiding visually impaired persons by providing them information about the content of the image, search engines where images can be searched by sentence fragments. This can further be used in video captioning which also has many practical applications including alert systems for enhancing security. In retrieval based image captioning method**,** we have a pool of already defined or specified sentences or captions, and we have a given query image, then the retrieval-based method will try to search or generate a caption or sentences from this available pool. This can be a new sentence or an existing one. The template-based image captioning method deals with the syntactical and semantic process while having a new caption. Firstly we need to detect the visual concepts of an image and then connect it through sentence templates such as the grammar rules or optimized algorithm used to generate a sentence.

Automatic image captioning can be beneficial in various ways. With the recent surge of internet, social media, and online platforms for e-commerce, we have a lot of data available most of which is media data. Humans can understand and interact with these data easily, but if we want machines to learn from this then we need to train them in that way. As a machine learns they can outperform humans. Also its various applications such as an aid to blind people, CCTV cameras, self-driving cars, etc. CCTV cameras are used everywhere as a means used for surveillance purposes. If we can generate captions, then an alarm can be raised as soon as any malicious activity is found. This can probably reduce accidents/crimes. It can also be used in social media sites, where one can generate captions for the images posted by users and suggest it to users, as sometimes it is difficult to post a caption for our images. Self-driving cars. Aid to blind people, a product can be made that can be used while traveling, which can first generate captions for surrounding scenes, and then from this text it can be converted to audio.

This paper is divided into the following sections which include related work, the proposed methodology followed by implementation and results, and conclusion.

2. RELATED WORK

[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. [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. [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. [16] Tried to generate captions that are based on nouns, verbs, and adjectives. [10] developed a stylized way to generate sentences, the style will depict humor and romance in captions. [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. [12] for predicting the next words they provide a way that focuses on "conditioning-by-merge" and "conditioning-by-inject". [13] states that they use image captioning in service robots. Like humans can relate one word with many other words and later can remember it, similarly [14] provides a way where they relate colors in images with different objects in

the image, later generate captions. [15] tries to generate captions without using an image-caption pair dataset.

3. PROPOSED METHODOLOGY

The encoder-decoder framework is a widely used approach. In this paper also the encoder-decoder framework is used to generate caption. CNN and LSTM were used. The system is trained and tested on the Flickr8k dataset, and the caption is generated for a new image.

Feature extraction: The input is the image dataset from the Flickr8k dataset. For this purpose transfer learning was used, a way, where a model developed for a task, is reused as the starting point for a model on a second task. A pre-trained model is used to extract the features from images. The VGG16 model, which is a convolutional neural network model proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" is used. Then these extracted features are stored in a file so that it can be reused instead of recomputing. The last layer is removed from the loaded VGG16 model since we do not require the results from VGG16. The output is extracted features of 1-dimensional 4,096 element vector.

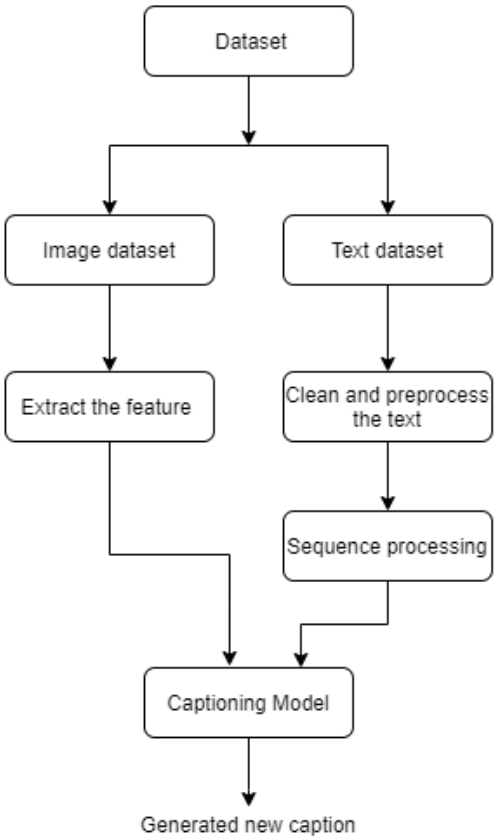

Figure 2. System overview

Sequence processing: In this step, the input is the text dataset from the Flickr8k dataset. Here we have 5 captions for a single image, so it needs some cleaning. After cleaning, the summarize vocabulary of text is created. The dictionary of image identifiers and descriptions is created. After cleaning captions the vocabulary size is 8,763 words. Following are the text cleaning tasks:

1. Convert all words to lowercase.
2. Remove all punctuation.

3. Remove all words that are one character or less in length (e.g. ‘a’).
4. Remove all words with numbers in them.
Later the Recurrent neural network(RNN) type that is Long Short Term Memory(LSTM) is used to generate caption one word at a time. Here we specify a maximum length for caption, like 34 words.

LSTM caption generator

LSTM function LSTM (x_t) returns p_{t+1} and tuple (m_t , c_t) passed as current hidden state to next hidden state[17].

$$i_t = \sigma(W_{ix} x_t + W_{im} m_{t-1}) \quad (1)$$
$$f_t = \sigma(W_{fx} x_t + W_{fm} m_{t-1}) \quad (2)$$
$$o_t = \sigma(W_{ox} x_t + W_{om} m_{t-1}) \quad (3)$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(W_{cx} x_t + W_{cm} m_{t-1}) \quad (4)$$
$$m_t = o_t \odot c_t \quad (5)$$
$$p_{t+1} = \text{Softmax}(m_t) \quad (6)$$

where,
f_t = forget gate that selectively ignore past memory cell,
i_t = input gate that selectively ignores part of the current input,
o_t = output gate that allows to filter current memory cell for its final hidden state.

Decoder: The Decoder model merges the vectors from both input models using an addition operation. This it is fed to the output Dense layer that makes a softmax prediction over the entire output vocabulary for the next word in the sequence.

4. Dataset

Dataset used is Flickr8k. 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. The images presented were chosen from six different Flickr groups and tend not to contain any well-known person. Images in this dataset do not contain any famous person or place so that the entire image can be learned based on all the different objects in the image.

5. Results

The implementation was done using Google Colab. The reason for using Google Colab is, it provides a python jupyter like notebook environment, that runs entirely in the cloud. It is easy to use and access and also provide suitable hardware to run the system, because of which speed of execution increases. Google Colaboratory integrates PyTorch, TensorFlow, Keras, OpenCV, and free cloud service with free GPU. For our implementation purpose, we have 12.72Gb of RAM, disk space of 68.42GB, and GPU runtime. The GPUs available in Colab are Nvidia K80s, T4s, P4s, and P100s, which GPU to allocate depends on your usage.

The dataset used for implementation was the Flickr8k dataset. This dataset has two different files one is all 8091 images and the other is captions for those images. This dataset provides 5 captions per image. So we have 40,455 total captions available. The system consists of three basic modules which are listed and explained below in detail. For the following images, we can see how our model generates caption for them.

Table 1. Following table shows the images and their respective captions generated by the system.

Sr.No.	Image	Generated caption	Category
1		Dog is running through the water.	Correct caption
2		Skier is standing on a snowy mountain.	Correct caption
3		Man on bike jumping off the dirt.	Somewhat correct caption
4		Two children playing on the swing.	Incorrect caption

6. Conclusion

We proposed a system to generate a caption for an image using deep learning techniques. We trained and tested on the available Flickr8k dataset, then generated a caption for any random new image. We can say that fluency is neglected. But we include only objects, attributes, and relations in the candidate caption for better score and understanding. As we need to train our system to generate results, they cannot always be correct we may not always get the accurate caption. The category we see in the above table is given manually.

Our system generates captions on images, which we can categorize them as correct, incorrect, somewhat correct captions like [3]. We need to keep in mind that if we train our model on the animal dataset, then we cannot expect our model to generate a correct caption for natural images, eg. Waterfall, sunset, etc.. So to get proper results, we should see what type of dataset we are using and what type and nature of images that dataset has.

REFERENCES

[1] J. Aneja, A. Deshpande, A. G. Schwing, "Convolutional Image Captioning", 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", 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", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3156-3164.

[4] J. Gu, G. Wang, J. Cai, T. Chen, "An Empirical Study of Language CNN for Image Captioning", 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", 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", Proceedings of the Advances in Neural Information Processing Systems, 2014.

[7] N. Kalchbrenner, P. Blunso, "Recurrent Translation Models", 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", 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", 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", 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", Proceedings of the European Conference on Computer Vision, 2010, pp.15-29.

[12] M. Tanti, A. Gatt, and K. P. Camilleri, "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator", Proceedings of the 10th International Natural Language Generation conference, pages 51–60, September 4-7 2017.

[13] R. C. Luo, Y.T. Hsu, Y.C.Wen and H. J. Ye, "Visual Image Caption Generation for Service Robotics and Industrial Applications", IEEE International Conference on Industrial Cyber Physical Systems (ICPS), 2019.

[14] M. Konno, K. Suzuki, and M. Sakamoto, "Sentence Generation System Using Affective Image", 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and the 19th International Symposium on Advanced Intelligent Systems (ISIS).

[15] S. Venugopalan, L. A. Hendricks, and M. Rohrbach, "Captioning Images with Diverse Objects", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5753-5761.

[16] H. Fang, S. Gupta, F. Iandola and R. K. Srivastava, "From Captions to Visual Concepts and Back", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1473-1482.

[17] Moses Soh, "Learning CNN-LSTM Architectures for Image Caption Generation", Dept. Computer Science, Stanford University, Stanford, CA, USA, 2016, cs224d.stanford.edu.