Automatic Image Caption Generation-A Survey

Kavitha S [1], Keerthana V[2], Bharanidharan A[3]
[1, 2, 3] Department of Computer Science and Engineering
[1, 2, 3]Sri Ramakrishna Engineering College

Abstract- *In many industrial, medical and scientific image processing applications, the need for various feature and pattern recognition techniques are used to match specific features in an image with a known template .Automatic caption generation from natural images is a challenging problem that has recently received a large amount of interest from the computer vision and natural language processing communities. There is a need for context based natural language description of images, however, this may seem a bit difficult but recent developments in certain fields like neural networks, computer vision and natural language processing has paved a way for describing the images accurately. In this survey, the classification of existing approaches based on how they conceptualize this problem, and also the models they used to solve this problem has been analyzed. This paper provides a detailed review of existing models, highlighting their advantages and disadvantages.*

Keywords- Artificial Intelligence, Bag Caption, Computer Vision, Convolution, Learning, Neural Networks;

I. INTRODUCTION

Automatic image caption generation is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. These two fields have seen great improvement in their respective goals of analyzing and generating text, and of understanding images as well as videos. Though both fields share a similar set of methods rooted in artificial intelligence and machine learning, they have historically developed separately, and thus their scientific communities have typically interacted very little.

Recent years, however, there is lot of interest in problems that require a combination of linguistic and visual information. A lot of day to day life tasks are of this nature, e.g., interpreting a photo in the context of a newspaper article, following instructions within a diagram or a map, understanding slides while listening to a lecture.

In addition to this, a vast amount of data is provided by the web that combines linguistic and visual information such as tagged photographs, illustrations in newspaper articles, videos with subtitles, and multimodal feeds on social media. To tackle combination of the natural language processing and vision tasks and to exploit the large amounts of multimodal data, the Computer Vision and Natural Language Processing communities have moved closer together.

II. LITERATURE SURVEY

OVERVIEW

Automatic image caption generation is a challenging task which involves preprocessing the images, extracting features and then to generate appropriate caption. Even though it is a difficult task, a lot of research is going on which explores the capability of computer vision in the field of machine learning and it helps to narrow the gap between the computer and the human beings on understanding the image. The purpose of this survey is to analyze techniques used for an automatic image caption generation using various neural network concepts.

Priyanka Jadhav, Sayali Joag, Rohini Chaure, Sarika Koli(2014) proposed a thesis that is concerned with the task of automatically generated caption for news images. This application mainly focuses on captioned images embedded in news articles, and this involves the use of both models of content selection and surface realization from data and thus avoids expensive manual annotation. Also, it is necessary that image descriptions in news articles need to be concise, focusing on the most important depicted objects or events. The caption generation process must contain two types of information, information about how the images for news corresponds to words and also the information about how these words can be combined to create a human-readable sentence. Although, the words generated for news images are admittedly noisy compared to traditional human-created keywords, this paper show that they can be used to learn the correspondences between visual and textual modalities, and also serve as a gold standard for the caption generation task.

Chenyou Fan,David J.Crandall(2016) describes about the Automatic caption generation for life logging image streams. Life logging cameras captures many images from a first person perspective, but generate so much data that it is hard for users to browse and organize their image collections effectively. This paper proposes to use automatic image caption generation algorithms to generate textual description for these image collections. Life logging photo streams are highly redundant since wearable cameras capture thousands of photos per day. Instead of simply captioning individual images, it is necessary to consider the problem of jointly captioning life logging streams, i.e., generating caption for temporally contiguous groups of photos corresponding to coherent activities or scene types.

The textual annotation for training and testing the system will be collected in two different ways. First, the online system to submit sentences for randomly selected images has been used. Second, to generate more diversity in annotators and annotation. The basic high level idea for caption generation is to learn a common feature space that is shared by both images(captured from wearable cameras)and words. Then, based on given new image as a input, the caption that are nearby in the same feature will be generated. Here, the task of mapping from image to feature space is typically done by Convolutional Neural Network and thus CNN is considered to be encoder. Then, the mapping from feature space to caption generation is done by Recurrent Neural Network and thus RNN here is considered to be decoder.

Takashi Miyazaki, Nobuyuki Shimizu(2016) presented an approach to generate a cross-Lingual caption generation. Research on image caption generation has typically focused on taking in an image as a input and to generate appropriate caption for that input image. The language for caption generation is English in many of the cases. But, it is necessary to generate caption in many languages such as Japanese. This paper has developed a Japanese version of the MS COCO caption dataset and a generative model based on deep recurrent architecture.

This model will take the image as a input and uses the Japanese version of dataset to generate a caption in Japanese. Usually, the Japanese portion of corpus is small, thus the proposed model in this paper was designed to transfer the knowledge representation obtained from English portion into Japanese portion. The resulting experiments showed that bilingual comparable corpus has better performance than a monolingual corpus. This indicates that image understanding using a resource-rich language benefits a resource-poor language. First, it is necessary to overcome the language barrier, to create a connection between source and target languages. Second, to develop an appropriate knowledge transfer approach.

Yansong Feng(2011) proposed an idea of generating caption for the news images. Most of the previous work has focused on generating descriptions for domain specific images, the task of caption generation is novel to our knowledge. The proposed caption generation model comprises two steps, namely content selection, which operationalize as image annotation, and then surface realization. the main aim in this thesis is to develop a knowledge-lean approach to automatically generating descriptions for news images that requires minimal supervision and does not rely on manually created resources. News data is huge and they are publicly available, although noisy. The task of generating caption for these news images is a challenging task and the approach used here will outline these challenges.

Some of the challenges faced by caption generation for news images are Extracting Image Content, Rendering Image Content in Natural Language, The Synergy between the Visual and Textual Modalities. These challenges need to be faced with necessary efficiency. The concept of automatic caption generation for news images can be used in news channels and also in newspapers.

III. CONCLUSION

The major task in automatic image caption generation lies in extracting features and to generate appropriate caption for the input image. Each image contains various features and there are various methods to generate these features. It is necessary to implement the accurate method to extract features from the images. Once various features from the image have been extracted, the appropriate caption for the particular image will be generated. The proposed model of Automatic Caption generation can be applied in many applications to explain the particular image exactly. The key aspect of this approach is to allow both the visual and textual information to influence the caption generation task.

ACKNOWLEDGMENT

The authors expresses their sincere gratitude to the principal, director Academics and Head of the Department of Computer Science and Engineering of Sri Ramakrishna Engineering College for giving constant encouragement and support to complete the work.

REFERENCES

[1] Aswathy K S, Gnana Sheela K, “Survey on Feature Extraction of Images for Appropriate Caption Generation”, International Journal of Engineering Research and General Science Volume 4, Issue 1, January-February, 2016 ISSN 2091-2730.

[2] D. D. Sapkal, Pratik Sethi, Rohan Ingle, Shantanu Kumar Vashishtha, Yash Bhan, “A Survey on Auto Image Captioning”, Vol. 5, Issue 2, February 2016.

[3] Thirupathi Podeti , Bhanu Prasad A, “Dynamic Caption Generation with Image Annotation”, Thirupathi Podeti et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (4) , 2014, 5221-5224,ISSN:0975-9646.

[4] Vini Varghese, J Saravanan,” A Systematic Approach for News Caption Generation”, International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014), Vol. 2, Issue 2, Ver. 1 (April - June 2014).

[5] Priyanka M. Kadhav, Pragati Patil,” Efficient Phrase-Based Model for Automatic Caption Generation of Images”.International journal of innovative Research and Development, Vol 2 Issue 11, November, 2013.

[6] Amitkumar Kohakade, Emmanuel M,” Content based Caption Generation for Images Embedded in News Articles”, International Journal of Computer Applications (0975 – 8887) Volume 100– No.11, August 2014.

[7] Deshmukh Sonali Dattatray,Ugale Pravin Chandrakant Walzade Amit Balasaheb,Kshirsagar Jayesh Prabhakar , Prof. S.B.Gote,” The Review of The Automatic Caption Generation for News Articles and Personal Photos”, International Journal of Advance Engineering and Research Development, Volume 3, Issue 3, March -2016, e-ISSN (O): 2348-4470 p-ISSN (P): 2348-6406.

[8] Zenghai Chen, Hong Fu, Zheru Chi and David Dagan Feng. 2012. “An Adaptive Recognition Model for Image Annotation”, IEEE Transactions on Systems, Man, and Cybernetic Part C: Applications and Reviews. Vol.42. Issue 6. pp.1120-1127.

[9] Yansong Feng, Member and Mirella Lapata. 2013. “Automatic Caption Generation for News Images”. IEEE Transactions on Pattern Analysis and Machine Intelligence.Vol.35. Issue 4. pp.797-812 .

[10] Man Lan, Chew Lim Tan, Jian Su. 2009. “Supervised and Traditional Term Weighting Methods for Automatic Text Categorization”. IEEE Transactions on Pattern Analysis and Machine Intelligence. Vol.31. Issue 4. pp. 21-735.

[11] Allan Hanbury. 2008. “A survey of methods for image annotation”. Journal of Visual Languages and Computing Elsevier. Vol. 19, Issue 5. pp. 617–627.

[12] Omesh kalambe, Shubhangi Giripunje, ”Caption generation for Image with Efficient Document Retrieval”, International Journal for Scientific Research & Development Vol. 3, Issue 02, 2015 ISSN (online): 2321-0613.