International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 2 | ISSN : 2456-3307

Automatic Image Caption Generation

Kavitha S, Keerthana V, Bharanidharan A

Department of Computer Science and Engineering, Sri Ramakrishna Engineering College, Coimbatore, TamilNadu, India

ABSTRACT

Recent work in machine learning uses an attention-based model, which automatically learns to predict the content of image. Automated Image Caption generation uses a type of artificial intelligence called deep learning to generate appropriate caption for the images. It also provides the best words to explain the image entirely. The textual description of the image would be generated after processing the image and its visual content has been analyzed. Caption generation can talk about the features of the scene and how the people and the objects in the image interact. The shape information of an image is mostly enclosed in edges. So, it is necessary to detect edges for an input image, by using certain filters and by enhancing those areas of image which contains edges, sharpness of image will increase and image will become clearer. The filter used here is sobel operator. Edge detection is used to extract features from the image and based on these features, caption will be generated which depicts the image. To convert these features into a meaningful caption, KNN classifier is used.

Keywords : Machine Learning, KNN, MATLAB, RGB

I. INTRODUCTION

This paper is concerned with the task of generating caption for the images automatically. Various algorithms are analysed to generate the textual description associated with the images. This concept will make the machine to understand human language. In Recent years, many problems such as tagged photographs, videos with subtitles required the combination of both linguistic and visual information. To resolve these kinds of issues, the key concept of automated image caption generation has emerged. The given image acts as an image representation. The text generation involves a series of steps, content selection (choosing the desired aspect of input), Text planning (to organize the content) and surface realization (choosing the right word).

II. RELATED WORK

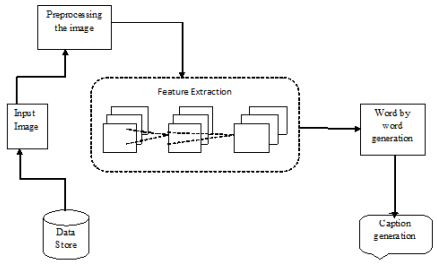
There are various research has been taken place in computer vision due to recent advances in machine learning. Priyanka Jadhav, Sayali Joag, Rohini Chaure, Sarika Koli(2014) proposed a thesis that is concerned with the task of automatically generated caption for news images.This application mainly focuses 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.

Chenyou Fan,David J.Crandall(2016) describes about the Automatic caption generation for life logging image streams. Life logging cameras capture many images from a first person perspective. So, it generate so much data that it is hard for users to browse and organize their image collections effectively.

The development in recent years has led to renewed interest in connecting vision and language. Many research groups have reported a significant improvement in generating caption using various methods. This mainly involves using a method that combines convolution neural network with recurrent neural network. 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.

III. PROPOSED SYSTEM

The proposed system deals with the generation of caption automatically for the images in the food court. Initially, various images with different activities need to be captured using camera. The images, which have been captured using camera, need to be processed. The images need to be filtered to avoid noise, so that processed images can be used for extracting the feature. Based on the feature extracted from the images, the words will be obtained. Using necessary techniques, an appropriate caption will be generated which depicts the image.



IV. ECHNICAL APPROACH

Overview

A deep recurrent architecture that automatically produces short descriptions of images has been implemented. Edge detection is used to extract features from the images. The extracted features will be given as input to KNN classifier algorithm to generate a description of the image in valid English.

Edge Detection-Feature Extractor

Edge detection has proven to be very successful framework for extracting the features from the images. It represents a method for image processing, and form a link between general feed-forward neural networks and adaptive filters. Various filters are present in edge detection which are considered as linear filters or smoothing filters. The filter used here is sobel operator, which is used to calculate edges in both horizontal and vertical direction.

KNN Classifier:

k-Nearest Neighbor
Classify (**X**, **Y**, *x*) // **X**: training data, **Y**: class labels of **X**, *x*: unknown sample
for $i = 1$ **to** m **do**
  Compute distance $d(\mathbf{X}_i, x)$
end for
Compute set I containing indices for the k smallest distances $d(\mathbf{X}_i, x)$.
return majority label for $\{\mathbf{Y}_i \text{ where } i \in I\}$

KNN is a popular classifier model in neural network, which can be used to generate appropriate words for the given image. The various user interface functions such as Selection, Reordering and Editing are used in text generation algorithm. Using selection function, a quality ordered list of possible captions will be generated. After selection, using Reordering function, it is necessary to keep track of the recognized entities so that the user can reorder them. The final UI function Editing will help to provide the exact caption for the images.

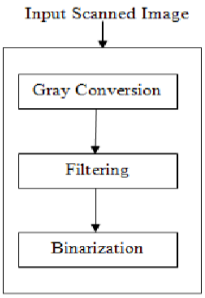
Module Description

Module 1: Preprocessing The Image

Preprocessing the image normally involves removing low-frequency background noise and reflection, normalizing the intensity of the individual particles, and masking portions of images. Image preprocessing is the technique of enhancing data images prior to computational processing. Pre-processing is a common name for operations with images at the lowest level of abstraction — both input and output are intensity images. These iconic images are of the same kind as the original data captured by the sensor, with an intensity image usually represented by a matrix of image function values (brightnesses).

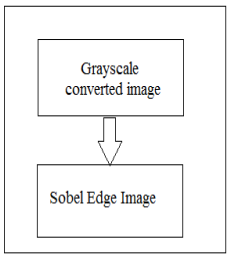
There are various types of images, which has been collected and organized as dataset. The necessary words used to describe the images are feed into the system as dataset. Using the MATLAB R 2013a software, the images are loaded .The filename and pathname are specified in order to read image file for which caption need to be generated. The input images are processed and the filtering technique is applied in order to remove the noise in the image. It is necessary to remove the noise from the image, because boundaries in images will remain sharp and do not blur. Special function is used to filter the noise and average of RGB is obtained. The RGB image is converted into grayscale image using RGB to bin methods, where average of color is derived to find dominant color. The value fixed for R=1, G=2, B=3.

Grayscale = (R + G + B / 3)



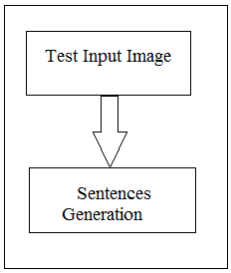
Module 2: Feature Extraction

As a result of processing the image, RGB to gray scale conversion image is obtained. The converted grayscale image is used for edge detection. Fourier descriptors are used to identify boundaries of the loaded image, it also includes aspect ratio which are used to find the width and height of the image. Here sobel edge operator has been used. The purpose is to determine the existence and location of edges in a picture.



Module 3: Caption Generation

The images from the food court will be given as input and the expected output for these images are the generation of caption which depicts the image. The automatic generation of caption will be done using KNN classifier. When edges are detected, certain condition or parameters are applied to check whether the value generated matches with the input value. Based on the features from the images, the key phrase, which suits the image, will be extracted. The extracted key phrases were from a standard database and this augmented database was used to segment the training documents into word sequences based on a maximum matching principle, which favours the longer words and phrases. The extracted words will be converted to form a meaningful caption, which is done using KNN classifier-nearest neighbourhood algorithm.



V. CONCLUSION AND FUTURE ENHANCEMENT

In this paper recent advance in automatic image description and closely related problems has been analyzed. The modern automated captioning techniques could work well enough to be used in many practical applications. The dataset contains the real world pictures and exhibits an outsized vocabulary automatically rather than manually making annotation, image captions are treated as labels for the image. The experimental results showed that it is possible to learn the correlations between visual and textual information from the dataset even if it is not explicitly annotated in any way. The key aspect of this approach is to allow both the visual and textual information to influence the word generation task. It is expected that the modularity of this approach combined with attention to have useful application in other domains.

VI. FUTURE WORK

The dataset discussed in this paper can be further refined according to specific applications to eliminate some of the noise. Most modern mobile phones are able to capture photographs, making it possible for the visually impaired people to make images of their environments. These images can then be used to generate captions that can be read out loud to the visually impaired, so that they can get a better sense of what is happening around them. Future work can include this text-to-speech technology, so that the generated descriptions are automatically read out loud to visually impaired people. In addition, future work could focus on translating videos directly to sentences instead of generating captions of images.

VII. ACKNOWLEDGEMENT

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.

VIII. 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.

.