ISSN:0975-9646

A Systematic Approach for News Caption Generation

PHILO SUMI
Final year M.Tech CSE
JCET, Lakkidi,
Palakkad, Kerala-679301, India

ANU.T.P
Asst Professor, *Dept of CSE*
JCET, Lakkidi
Palakkad, Kerala-679301, India

***Abstract* – Captions are essential components associated with images to make search engines to respond easily with user queries. Making appropriate captions for images is a difficult task. By making the appropriate caption will help the user to search images with long queries. But most of the images are associated with user annotated tags, captions and text surrounding the images. This paper is concerned with the task of automatic caption generation for news images in association with the related news article. In this method we will input one image and news article to the system. The system will generate most important keywords which are associated with the image in association with the image. To find the image related keywords, first we will find out the input image's features using SIFT (Scale Invariant Feature Translation) method. And using these features we will compare the image with the images which are stored in the database. After finding the best matched image we will extract the keywords associated with that image. After applying grammatical rules to the keywords an appropriate caption is generated. Here we are combing the textual modalities with the visual one. In the existing method the captions are not efficiently generated and there is no mapping between the image and the text associated with it. But we are introducing a best method for news image caption generation without costly manual involvement.**

***Keywords*- Caption generation, image annotation, summarization, topic models**

1. INTRODUCTION

A caption is an important factor for non textual media objects like image, video etc. A good caption will describe the complete content in a short manner. Without more time loss we can understand a context in a glance with an efficient caption. Generally caption generation is a difficult task. An expertise journalist can also feel difficulty in this generation task. Here we are introducing an automatic caption generation method for news images in an efficient way. This method is helpful in text summarization with the help of visual modality. Many people do summarization based on text alone, which will not have any relation with the associated image. Captions are essential components associated with images to make search engines to respond easily with user queries. Making appropriate captions for images is a difficult task. By making the appropriate caption will help the user to search images with long queries. But most of the images are associated with user annotated tags, captions and text surrounding the images. This paper is concerned with the task of automatic caption generation for news images in association with the related news article. In this method we will input one image and news article to the system. The system will generate most important keywords which are associated with the image in association with the image. To find the image related keywords, first we will find out the input image's features using SIFT (Scale Invariant Feature Translation) method. And using these features we will compare the image with the images which are stored in the database. After finding the best matched image we will extract the keywords associated with that image. After applying grammatical rules to the keywords an appropriate caption is generated.

A good caption should be informative. It must clearly identify the objects of the picture. And follow a good summarization method to summarize the textual content. And provide a relation between the textual and visual contents. Moreover lead the reader in to the article. Worthless words are avoided from a good caption. While a short caption is appropriative, but it should be informative. Following correct grammatical rule will generate a good caption. In the existing method the captions are not efficiently generated and there is no mapping between the image and the text associated with it. But we are introducing a best method for news image caption generation without costly manual involvement. We are not using any dictionaries to provide text to image relation. By using the extractive and abstractive caption generation process a good caption is generated.

Amount of images which are available on internet are huge today therefore for the better image retrieval process caption generation is very important task. Making this process automatic means it will reduce the costly manual involvement in caption generation process. There are many problems in image captioning. Because content based image retrieval uses visual similarities of images to find out the matched image. But it is suffering with the semantics information loss. Manual annotated words provides solution to this problem but it is time consuming and costly. In our system this problem is avoided with using the image to text correspondence. A good caption for image is generated in association with the associated article.

2. RELATED WORKS

Image caption generation is important task because it makes user to easily retrieve appropriate images with long queries. Image understanding works are done by many methods but less focus is on the process of caption generation for images. It actually consists of two stages [1]. First our model will create a dataset in which large amount of images and image descriptions are included. In the text annotation process we will find out the important keywords and phrases using parts of speech. After that we will apply the stop word removal and stemming process. In the image

annotation process we apply SIFT algorithm [2] to find out the local features of the image. The comparison between input image and image stored in the database are done with the Euclidian distance of their feature vectors. After that associated keywords of images are extracted. All the keywords from image and document are classified under pronoun, noun, verb etc. Then by applying extractive and abstractive caption generation process [1] automatic caption is generated.

In addition to the above mentioned technique many other techniques are used for image description generation. For example P. Hede, P.A. Moellic, J. Bourgeoys, M. Joint, and C. Thomas [3] has described usually to represent images of objects in some natural language or in a human readable form image annotation system is utilize in image base management. Text generation from images in a natural language by bearing in a mind manual database on the account of image attributes like color and texture. Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Juli Hockenmaier, David Forsyth [4] has described Creation of text from the images by simply comparing given images with documents and acquire a score to link sentence to the image. Y. Feng and M. Lapata [5] demonstrate the method to create a database of images which are naturally associated with the documents. According to B. Yao, X. Yang, L. Lin, M.W. Lee, and S. Chun Zhu [6] input image first get parsed with respect to various attributes in scenes- objects- parts and after that text generation engine generate a text. V.Ordonez, G. Kulkarni, and T.L. Berg [7] describe the method of word based model used for the generation of text from the images. M. Banko, V.Mittal, and M. Witbrock [8] use text summarization techniques for caption generation.

3. PROBLEM FORMULATION

The caption generation task can be formulated as follows. A news article and related image is given the system generates a caption related to both image and article. At the testing time we will input one news article and image the system will generate the image caption automatically.

3.1 Image Database

The database is created with huge amount of images and related news article. The dataset covers wide range of topics including national and international topics related to the field of sports, science, politics, technology etc. The images that are used in the news articles are usually around 200 pixels wide and 150 pixels height. The average length of the caption is 9.5 words, the average length of the sentence is 20.5 words and the average length of the document is 381.5 words.

The image-document input has several advantages. We can generate an efficient caption for image in association with the related article. The news article contains much more information to generate an appropriate caption. The system uses text summarization and image annotation techniques for the process.

3.2 Data Validation

The dataset which is described above is used for two purposes. First the model will learn from the image, article and caption. This will give descriptions keywords for the image. These keywords are further used for image caption generation. Second, the human authored captions will function as a gold standard for the image annotation model and for the end-to-end caption generation task. In the former method stop words are removed and stemming algorithm is used for extracting words from the article. And these words are further classified in to verbs, nouns, adjectives etc. and extractive and abstractive caption generation methods are used for automatic caption generation.

The training set of the system consists of image caption pair it is manually examined whether the content words (nouns, verbs and adjectives) present in the captions could in keywords describe the image. 90 percent of the time it was found that the captions expressed the picture's content. Figure.1 shows the proportion of caption words given a rating of 1, 2, 3, and so on. As in the figure, the majority of the words were given a rating of 4 or higher. This procedure inform as about how many words from an article describe the image. After that applying extractive and abstractive caption generation methods are applied.

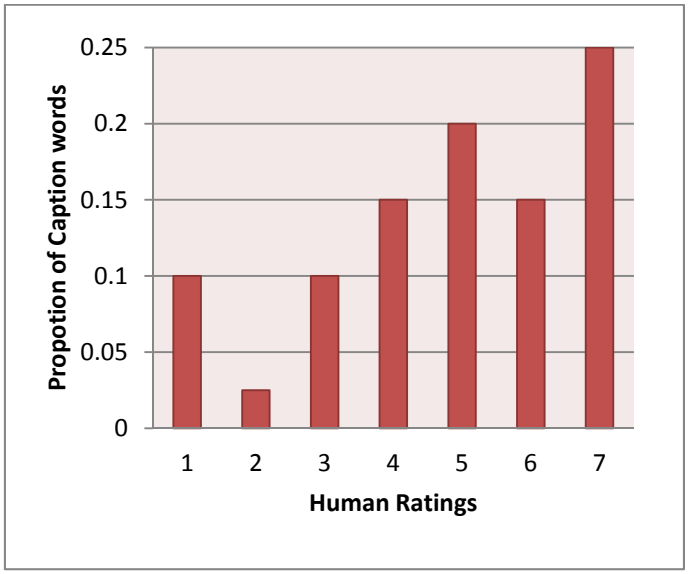


Figure.1 Proportion of caption words given a rating by human judges.

4. MODELING

Our model consists of two stages, content selection and surface realization. Content selection identifies the important keywords from the text and image. And surface realization verbalise the keywords which is extracted from the image and text.

1. The caption which is generated by this model not only identifying the objects of the image but also the events which is related with the image. We will consider Table 1as example.

Table.1 Example for an image-document-caption result.

The British government could face claims it violated international humanitarian laws by allowing US arms flights to Israel to use UK airports. The Islamic Human Rights Commission (IHRC) is seeking permission to contest government bodies over what it says are crimes against the Geneva Convention. A number of US planes said to be carrying bombs to Israel refueled in the UK during the Lebanon conflict. The IHRC said it received complaints from Britons with families in Lebanon. The commission is accusing the government of "grave and serious violations" of international humanitarian law. It is seeking permission to bring its case against the Civil Aviation Authority, the Foreign and Commonwealth Office and Defence Secretary Des Brown in the High Court. The IHRC said it is bringing the case after receiving "many complaints from British citizens whose family members are in Lebanon and facing grave danger as well as acts of terror". The US aircraft believed to have refueled in the UK are said to have been carrying supplies including "bunker buster bombs".	 **US arms flights carrying bombs have been carrying supplies to Israel refueled in UK airports.**

2. The article which is associated with the image also describes the news image.

3. Detailed processing of images is not essential for this caption generation process.

4.1 Image Content Selection

Content selection is the procedure in which the keywords for the caption generation process are extracted from the image and the associated article. We will find out the important keywords using a text frequency calculator. Text frequency is measure used to figure out how much important a word in a document. This method reads one word at time through the document. And a hash table is build using each word. The hash table has two entries, one the word as the key and other is number of times that word appears in the document. And the word with higher count is considered as an important word. In order to generate efficient keywords for the particular image with respect to article made easily with the stop-word removal and stemming procedure. Using those techniques we have to remove the unnecessary letters in the document. Stemming is a process of linguistic normalisation, in which the variant forms of a word are reduced to a common form.

In order to find out the visual features we are using SIFT method. Visual features are represented in a discrete manner and each image is considered as a bag of visual words. SIFT is an algorithm [2] [9] in computer vision used to detect and describe local features in an image. It covers wide area of application including finger print recognition system, face recognition, etc. To describe features of an object it is necessary to take out required points from the object available in an image. In order to ensure the scale invariance internal representation of the original image is created. Next we will find out the interesting points of the object in the image using Laplacian of Guassian approximation. After that find out the key points that is maxima and minima among the points of interest. Edges and low contrast regions are bad key points. Eliminating these makes the algorithm efficient and robust. An orientation is calculated for each key point. Further calculations are done relative to this orientation assigned to the key points. This effectively cancels out the effect of orientation, making it rotation invariant. Finally, with scale and rotation invariance in place, one more representation is generated. This helps to uniquely identify features of the image.

The comparison between input image and image stored in the database are done with the Euclidian distance of their feature vectors. After that associated keywords of images are extracted. All the keywords from image and document are classified under pronoun, noun, verb etc. By applying stop word removal and stemming algorithm [11] unwanted letters of word are removed. Then by applying extractive and abstractive caption generation process [1] automatic caption is generated. Caption generation is done automatically without costly manual involvement. The abstractive and extractive caption generation procedures are explained below.

4.2 Extractive Caption Generation

Extractive caption generation is concerned with extracting a single sentence which contains maximum predicted keywords. It is a text summarization technique in which a sentence from the article itself is retrieved based on maximum occurrence of extracted words with that sentence.

a. Vector Space-Based Sentence Selection

The words which are extracted from the image and document and the sentences from the document is represented in a vector space to find out the occurrence of words in that sentence. Create word-sentence co-occurrence matrix to find the frequency of a word in a sentence. The word with higher frequency is considered as an important word. The Vector-Space based sentence selection is done as follows. For each keyword the number occurrence of that word in every sentence is calculated. Choose a sentence with largest number of occurrence of keywords. Retrieve that sentence as caption.

4.3 Abstractive Caption Generation

The extractive caption generation method generates a caption which is naturally grammatical but it is not possible to express a subject with a single sentence. And also the selected sentence may be long making a caption inefficient. For these reasons we will go for abstractive caption generation procedure which is more efficient than the extractive one.

a. Word-Based Caption Generation

It focuses on the probability of words that may be occurred in final caption from the given document with minimum length. The most important words from document and image are extracted and an effective caption is generated using this words.

b. Phrase-Based Caption Generation

It focuses on the combination of words (phrases) that may be occurred for generating caption from the given document. Phrases are naturally associated with function words and may potentially capture long-range dependencies.

c. Search

By reducing the size of the document search made more efficient. We can rank each sentences in terms of their relevance to the image content which is taken from the document and only the best one is considered. And also give importance to the best sentence with the neighbouring sentences under the assumption that neighbouring sentences are about the same or similar topics.

5. EVALUVATION

In this section we will evaluate our model based on the performance and output. First we will learn our system with image-article-caption pair. At the testing time we will input one image with the related news article to generate an effective caption. In the text annotation process we will find out the important keywords and phrases using parts of speech. After that we will apply the stop word removal and stemming process. In the image annotation process we apply SIFT algorithm to find out the local features of the image. The comparison between input image and image stored in the database are done with the Euclidian distance of their feature vectors. After that associated keywords of images are extracted. All the keywords from image and document are classified under pronoun, noun, verb etc. Then by applying extractive and abstractive caption generation process automatic caption is generated.

5.1Image Annotation

5.2 Caption Generation

We trained our model with a good dataset. After applying various methods keywords are generated from the input image and news article. Documents and captions were parsed in order to obtain dependencies for the phrase-based abstractive model.

a. Evaluation

The generated by our system strictly follows the grammatical rules and establishes a relation between image and the associated document. Table.2 reports mean ratings for the output of the extractive system, the word-based caption, phrase-based caption, and the human-authored caption. We performed an Analysis of Variance (ANOVA) to examine the effect of system on the generation task. After the analysis performance test is performed on the mean values. The word based caption generation is the worst one than the others. Because of that proposed system is based on the phrase-based caption generation method. It produces grammatically aligned captions very effectively.

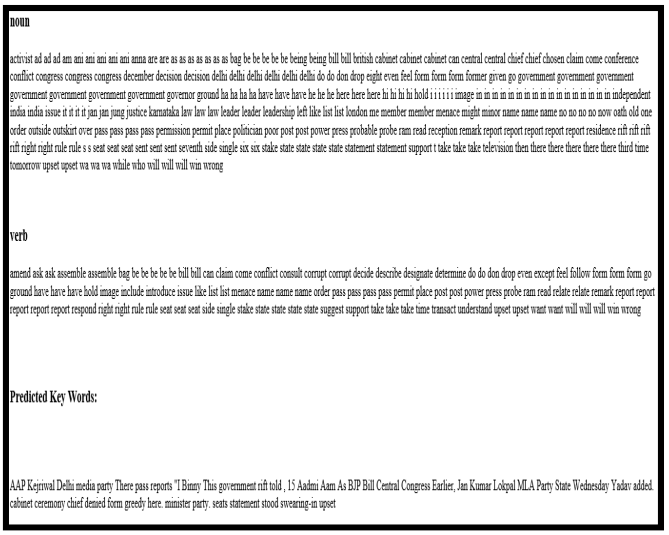
a. Evaluation

We created one dataset with huge amount of image covering almost all topics. The features of the input image is calculated with the SIFT method and using this features then used K-means to quantize these features into a discrete set of visual terms. These visual terms are converted to keywords using a visual vocabulary.

b. Result

Our model takes input as an image and corresponding document. We input image of size 620 X 420 pixel. Figure 2.a and 2.b shows the input image and generated keywords from the image and document.

Figure.2.a Input Image



noun

activist ad ad an ani ani ani ani anna are are as as as as as as bag be be be be be being being bill bill british cabinet cabinet cabinet can central central chief chief chosen claim come conference conflict congress congress congress december decision decision delhi delhi delhi delhi delhi delhi do do don drop eight even feel form form form former given go government government government government government government government government governor ground ha ha ha ha have have have he he he here here here hi hi hi hi hold i i i i i i image in in in in in in in in in in in in in in in in in in in in in independent india india issue it it it it jan jan jan jung justice karnataka law law law leader leader leadership left like list list london me member member menace might minor name name name no no no no now oath old one order outside outskirt over pass pass pass pass permission permit place politician poor post post power press probable probe ram read reception remark report report report report report residence rift rift rift rift right right rule rule s s seat seat seat sent sent sent seventh side single sox sox stake state state state state statement statement support t take take take television then there there there there there third time tomorrow upset upset wa wa wa while who will will will win wrong

verb

amend ask ask assemble assemble bag be be be be be bill bill can claim come conflict consult corrupt corrupt decide describe designate determine do do don drop even except feel follow form form form go ground have have have hold image include introduce issue like like list list menace name name name order pass pass pass pass permit place post post power press probe ram read relate relate remark report report report report report respond right right rule rule seat seat seat side single stake state state state state suggest support take take take time transact understand upset upset want want will will will win wrong

Predicted Key Words:

AAP Kejriwal Delhi media party There pass reports "I Binny This government rift told , 15 Aadmi Aam As BJP Bill Central Congress Earlier, Jan Kumar Lokpal MLA Party State Wednesday Yadav added. cabinet ceremony chief denied form greedy here. minister party. seats statement stood swearing-in upset

Figure.2.b Predicted Keywords

Table.2 Mean Ratings on Caption Output

Model	Grammatically	Relevance based
Extractive caption generation	6.42	4.10
Abstract words	2.08	3.20
Abstract phrases	4.80	4.96
Human authored caption	6.39	5.55

b. Result

The result of the system is shown in figure 3. The system generates grammatically aligned and topic relevant caption is generated.

Corruption is big issue for Aam Aadmi Party (AAP), Aam Aadmi Party leader designate Arvind Kejriwal said.

Figure.3 System output

6. CONCLUSION

This paper is concerned with the task of automatically generating captions for images, which is important for many image related applications. Here using Content selection and surface realization stages to generate data without requiring expensive manual annotation. We can also use the following additional algorithms to improve the efficiency of caption, they are: stemmer and text frequency calculator. In information retrieval, stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root form generally a written word form. Text frequency is a numerical statistic which reflects how important a word is to a document in a collection or corpus. And the not only generates caption according to the image but also the related article is taken in to account. By making the appropriate caption will help the user to search images with long queries. It makes an efficient image retrieval process.

We can perform several improvements to the proposed system. Currently we are using SIFT method to calculate the local features of the image. Visual words are generated more efficiently making the features global. And also we could allow an infinite number of topics and develop a nonparametric version that learns how many topics are optimal.

REFERENCES

[1] Yansong Feng, Member, IEEE, and Mirella Lapata, Member, IEEE "*Automatic Caption Generation for News Images*,"IEEE Transactions on Pattern Analysis and Machine Intelligence, VOL. 35, NO. 4, APRIL 2013

[2] D. Lowe, "*Object Recognition from Local Scale-Invariant Features*," Proc. IEEE Int'l Conf. Computer Vision, pp. 1150-1157,1999.

[3] P. He´de, P.A. Moe¨llic, J. Bourgeoys, M. Joint, and C. Thomas, "*Automatic Generation of Natural Language Descriptions for Images*," Proc. Recherche d'Information Assiste´e par Ordinateur, 2004.

[4] A. Farhadi, M. Hejrati, A. Sadeghi, P. Yong, C. Rashtchian, J.Hockenmaier, and D. Forsyth, "*Every Picture Tells a Story: Generating Sentences from Images*," Proc. 11th European Conf.Computer Vision, pp. 15-29, 2010.

[5] Y. Feng and M. Lapata, "*Automatic Image Annotation Using Auxiliary Text Information*," Proc. 46th Ann. Meeting Assoc. of Computational Linguistics: Human Language Technologies, pp. 272-280, 2008.

[6] B. Yao, X. Yang, L. Lin, M.W. Lee, and S. Chun Zhu, "*I2T: Image Parsing to Text Description*," Proc. IEEE, vol. 98, no. 8, pp. 1485-1508, 2009.

[7] V. Ordonez, G. Kulkarni, and T.L. Berg, "*Im2Text: Describing Images Using 1 Million Captioned Photographs*," Advances in Neural Information Processing Systems, vol. 24, pp. 1143-1151, 2011.

[8] M. Banko, V. Mittal, and M. Witbrock, "*Headline Generation Based on Statistical Translation*," Proc. 38th Ann. Meeting Assoc. for Computational Linguistics, pp. 318-325, 2000.

[9] http://www.aishack.in/2010/05/sift-scale-invariant feature transform/ author: Uthkarsh

[10] http://videolectures.net/mlss09uk_blei_tm/ author: David Blei , Computer Science Department, Princeton University

[11] http://snowball.tartarus.org/algorithms/porter/stemmer.html

Philo Sumi is a final year MTech Computer Science and Engineering student at Jawaharlal College of Engineering and Technology, Lakkidi.

Anu T P is a professor at Jawaharlal College of Engineering and Technology, Lakkidi.