ISSN:2321-1156
International Journal of Innovative Research in Technology & Science(IJIRTS)

EXTRACTIVE AND ABSTRACTIVE CAPTION GENERATION MODEL FOR NEWS IMAGES

N.P.Kiruthika, B.E., Computer Science and Engineering, SRM Easwari Engineering College, Chennai, India.
V.Lakshimi Devi, B.E., Computer Science and Engineering, SRM Easwari Engineering College, Chennai, India.
Ms.N.D.Thamarai selvi, Assistant Professor, M.E., SRM Easwari Engineering College, Chennai, India.

Abstract --- This paper provides a model for automatically generating captions for news images, which is used to support development of news media management and many multimedia applications. In the existing method, the captions for the news images are given manually by reading the text content. Thus the caption generation task requires human involvement and hence a time consuming process. The proposed system uses a two-stage framework for automatically generating captions for news images: Content Selection and Surface realization. Content Selection identifies what the image and accompanying article are about, whereas surface realization determines how to verbalize the chosen content. The images are analyzed using the image annotation technique. It uses a multimodal vocabulary consisting of textual words and visual terms. The textual words obtained from annotation are clustered with the words in the news document. Here the extractive and abstractive models for generating short, meaningful and precise captions for the news image are used. The advantages of this model are: (1) It does not require manual annotation of images (2) It reduces the need for human supervision.

Index Terms --- Caption generation, Image annotation, Stemming, Summarization.

1 INTRODUCTION

In the recent years, there has been a wide growth of digital information available on the Internet. Images and videos play a vital role in attracting the news readers. There are many news websites such as CNN, Yahoo and BBC, which uses images to publish their news. These images are retrieved from the web without analyzing their content. It matches the user queries with the image's file name and format, captions, and text surrounding the image. Thus the caption generation task requires human intervention and supervision. The association between the image regions and the words are generated using supervised classification [1], [2]. An image can be annotated to generate a list of keywords such as *planes, bombs, airport*, whereas the caption *"planes carrying bombs are landing at the airport"* would make the relationship between the words. The image descriptions focus on the most important depicted objects or events. A system that generates these descriptions automatically can improve the image retrieval.

The standard approach to image description generation uses a two-stage framework: Content Selection and Surface Realization. The former stage analyzes the content of the image and identifies "what to say" (i.e., which events or objects are worth talking about), whereas the second stage determines "how to say it" (i.e., how to render the selected content into natural language text). Both the stages are usually manually developed. Content selection makes use of dictionaries that specify a mapping between words and image regions of features [3], [4], [5], [6], [17], and surface realization uses human written templates or grammars for producing textual output.

2 RELATED WORK

Although image understanding is a popular topic within computer vision, relatively little work has focused on caption generation. As mentioned earlier, a handful of approaches create image descriptions automatically following a two-stage architecture. The picture is first analyzed using image processing techniques into an abstract representation, which is then rendered into a natural language description with a text generation engine. A common theme across different models is domain specificity, the use of hand-labeled data, and reliance on background ontological information.

For example, He´de et al. [6] generate descriptions for images of objects shot in uniform background. Their system relies on a manually created database of objects indexed by an image signature (e.g., color and texture) and two keywords (the object's name and category). Images are first segmented into objects, their signature is retrieved from the database, and a description is generated using templates. Other work (e.g., [4], [5]) creates descriptions for human activities in office scenes. The idea is to extract features of human motion from video key frames and interleave them

with a concept hierarchy of actions to create a case frame from which a natural language sentence is generated. Yao et al. [17] present a general framework for generating text descriptions of image and video content based on image parsing. Specifically, images are hierarchically decomposed into their constituent visual patterns, which are subsequently converted into a semantic representation. The image parser is trained on a corpus, manually annotated with graphs representing image structure. A multi-sentence description is generated using a document planner and a surface realizer.

A notable exception to the use of manually crafted resources is Kulkarni et al. [7], who generate natural language descriptions for images while exploiting state of the art image recognition and generation techniques. Their image recognition system extracts visual information as a set of triples describing the depicted objects, their attributes and spatial relationships (e.g., <furry, sheep> against <green, grass>). These triples are then used to create descriptive sentences (e.g., There is a furry sheep against the green grass), while gluing words (e.g., there, is, the) are provided by a language model or templates. Farhadi et al. [18] describe a related system that can match a descriptive sentence to a given image or to obtain images that illustrate a given sentence. Their approach essentially retrieves sentences rather than composing new ones. Nonetheless, images and sentential descriptions are expressed via a shared meaning representation which also takes the form of triples describing the objects, actions, and scenes (e.g., <bus, parks, street>, <plane, fly, sky>). More recently, Ordonez et al. [19] demonstrate that this sentence retrieval task scales to a large dataset containing 1 million captioned images.

Much work within computer vision has focused on image annotation, a task related to but distinct from image description generation. The goal is to automatically label an image with keywords relating to its content without however attempting to arrange these into a meaningful sentence or text. Despite differences in application and formulation, all previous methods essentially attempt to learn the correlation between image features and words from examples of images manually annotated with keywords. They are typically developed and evaluated on the Corel database, a collection of stock photographs, divided into themes (e.g., tigers, sunsets) each of which are associated with keywords (e.g., sun, sea) that are in turn considered appropriate descriptors for all images belonging to the same theme.

The task of generating captions for news images is novel to our knowledge. Instead of relying on manual annotation or background ontological information we exploit a multimodal database of news articles, images, and their captions. The latter is admittedly noisy, yet can be easily obtained from online sources and contains rich information about the entities and events depicted in the images and their relations. Similarly to previous work, we also follow a two-stage approach. Using an image annotation model, we first describe the picture with keywords, which are subsequently realized into a human readable sentence.

3 PROBLEM FORMULATION

The caption generation task can be formulated as follows: Given a news image and a document, the system generates a natural language caption which defines the content of both image and the document. During testing, a document and an image are given as input, for which the system generates a caption.

3.1 BBC News Database

Most of the existing systems used datasets that directly store the captions for the images. But the proposed system uses various tasks like image annotation and segmentation, object recognition and scene analysis. The datasets created by Farhadi et al. [8] and Hodosh et al. [13] contain image descriptions. However, as mentioned above, they are limited to specific object categories and scene types.

For these reasons, the proposed system uses dataset that are collected from BBC News website. The dataset covers a wide range of topics which includes national and international politics, technology, sports, education, and so on. 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 importance of the image-document input is twofold: First, the image specifies the topic about which the document speaks. Second, the document contains the rich linguistic information that is used to generate caption. The system uses text summarization and hand written grammar rules without extensive knowledge engineering.

3.2 Data Validation

The dataset described serves two purposes. First, the system uses the images, captions and associated articles as training data to learn an image annotation model that will provide description keywords for the picture. 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 case, the stop words are removed and the caption is treated as a bag of content words,

ie., noun, verbs and adjectives. As the image annotation plays a key role in this generation process, it is important to access the quality of the captions as labels and whether they do indeed capture some of the image's content. There is no point in learning an image annotation model on labels that are extremely noisy and plainly wrong.

The training set of the system consists of image-caption pairs, that is manually examined whether the content words (nouns, verbs and adjectives) present in the captions could in principle describe the image. It was found that the captions expressed the picture's content 90 percent of the time. Figure.1 shows the proportion of caption words given a rating of 1, 2, 3, and so on. As can be seen, the majority of the words were given a rating of 4 or higher. This task involved assessing how well the caption words captured the image's content.

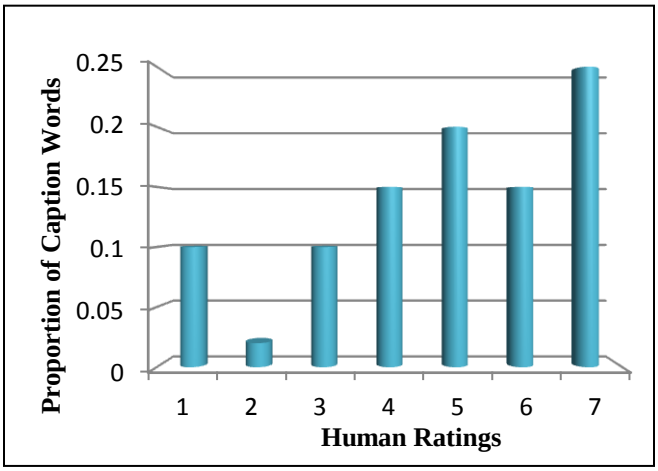

Figure.1 Proportion of caption words given a rating by human judges.

4 MODELING

The proposed model consists of two stages - Content selection and Surface realization. Content selection identifies the image and its related document, whereas surface realization determines how to verbalize the chosen content. The assumptions of the caption generation task are summarized as follows:

1. The caption depicts not only the content of the image but also its objects, attributes and events. For example, in Table.1 the caption mentions the *WalMart* shown in the picture and also the fact that it has fallen in its quarterly profits.

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

The world's largest retailer, Wal-Mart, has reported a 22% drop in quarterly profit and given a weaker-than-expected earnings forecast for the coming year. Wal-Mart said tough winter weather, cuts to government benefits and higher taxes contributed to the fall. Wal-Mart's UK supermarket business, Asda, said like-for-like sales fell 0.1% in the final quarter of 2013.	 **Wal-Mart reports fall in quarterly profits**

2. The accompanying article describes the content of the image. These images conventionally depict the events, objects or people mentioned in the article.

3. Since our images are implicitly labeled, we do not assume that all objects can be enumerated in the image.

4.1 Image Content Selection

We define a probabilistic image annotation model based on the assumption that images and their surrounding text are generated by a shared set of latent variables or topics. Our annotation model takes these topic distributions into account while finding the most likely keywords for an image and its associated document. The keywords are used for generating a caption that is related to both the news document and the image.

Words and images are distinct modalities, but both modalities are on same level as they describe same objects. The first step is the segmentation of the picture into regions, using image segmentation algorithm. Regions are then described by a standard set of features, including color, texture and shape. The visual features receive a discrete representation and each image is treated as a bag of words. To achieve this, Scale Invariant Feature Transform (SIFT) algorithm is used [14], [15]. The idea behind this algorithm is to sample the image with difference-of-Gaussian point detector at different scales and locations. Each detected region is represented with SIFT descriptor, which is a histogram of directions at different locations in the detected regions. The SIFT descriptors are further quantized using *K-means* clustering algorithm to obtain a discrete set of visual terms which form our visual vocabulary. Each entry in this vocabulary represents a group of image regions which are similar in content or appearance and assumed to originate from similar objects.

4.2 Extractive Caption Generation

The idea behind Extractive caption generation is to create a summary simply by identifying and subsequently concatenating the most important sentences in a document, independently of style, text type and subject matter. For the task of caption generation, only the extraction of a single sentence is required. This sentence must be maximally similar to the description keywords generated by the annotation model.

4.2.1 Vector Space-Based Sentence Selection

The keywords and sentences are represented in vector space and computing the similarity between the two vectors representing the image keywords and documents sentences, respectively. A word-sentence matrix must be created, where each row represents a word, each column represents a sentence and each entry the frequency with which the word appears within the sentence. The ***Vector-Space based sentence selection*** algorithm is defined as:

1. For each word W_i,
   i. From sentence S_1 to S_n, Find the number of occurrence of the word W_i.
   ii. Choose the sentence that has the largest number of occurrence.

2. Retrieve all the sentences for all the keywords.

4.3 Abstractive Caption Generation

There is often no single sentence in the document that uniquely describes the image's content. In most cases the keywords are found in the document but interspersed across multiple sentences. The selected sentences make for long captions, which are not concise. For these reasons, we turn to abstractive caption generation technique.

4.3.1 Word-based Caption Generation

Content selection is modeled as the probability of a word appearing in the headline given that the same word appears in the corresponding document and is independent of other words in the headline. The likelihood of different surface realizations is estimated using a *bigram* model. Since the individual words cannot frame a meaningful caption, the phrase-based caption generation technique is used.

4.3.2 Phrase-Based Caption Generation

Phrases are naturally associated with function words and may potentially capture long-range dependencies. The retrieval of phrases are done using the ***trigram selection model.***

1. For each keyword W_i, Choose the words W_{i-1} and W_{i-2}
2. Form a cluster of W_i, W_{i-1} and W_{i-2}
3. Retrieve the most commonly occurring cluster as a phrase.

4.3.3 Search

Search can be made more efficient by reducing the size of the document. Using the extractive generation method as mentioned in section 4.2, we can rank its sentences in terms of their relevance to the image content and consider only the best ones. We can also consider the single most relevant sentence together with its surrounding context under the assumption that neighboring sentences are about the same or similar topics.

5 EVALUATION

In this section, we evaluate our caption generation model and discuss the results that are produced. Initially, we use the training data to study the image-document-caption model and then produce captions for a given input image and a news document. The image annotation model highlights the important objects or events depicted in the image. Using the extractive caption generation method, the frequent words in the document are obtained. The phrase-based caption generation method is then applied to generate an abstract and relevant caption for the image.

5.1 Image Annotation

5.1.1 Evaluation

The experiments for our caption generation system were conducted on the BBC dataset as described in Section 3.1. We used 1,344 image-caption-document tuples for training, 140 tuples for development, and 140 for testing. We excluded for the vocabulary low frequency words (appearing fewer than five times) and words other than nouns, verbs and adjectives. We preprocessed the images as follows: We first extracted SIFT key-points with descriptors from each image and 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.

5.1.2 Results

Our image annotation model takes the images of 620 X 420 pixels as input. Our caption generation model performs image annotation in order to detect the objects in the given input image. Once the objects are detected, the system produces keywords for the image for which the caption is to be generated. The Figures 2(a) and 2(b) represent the input image and the generated keywords respectively.

Figure.2(a) Input image for generating the image caption

Figure.2(b) Keywords generated from the image

5.2 Caption Generation

Our caption generation experiments were conducted on the same BBC News dataset used for image annotation and using the same training, development, and test set partitions. In addition, documents and captions were parsed in order to obtain dependencies for the phrase-based abstractive model.

5.2.1 Evaluation

The system output is mainly based on two dimensions, grammaticality and relevance. 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. Performance tests were carried out on the mean of the ratings shown in Table.2 (for grammaticality and relevance). The word-based caption yields the least grammatical output. It is significantly worse than the phrase-based caption, the extractive system, and the human written caption. Thus the word-based caption generation technique is considered as an inefficient system and hence the proposed system makes use of phrase-based caption generation technique.

Table.2 Mean Ratings on Caption Output

Model	Grammaticality	Relevance
Extractive Caption	6.42	4.10
Abstract Words	2.08	3.20
Abstract Phrases	4.80	4.96
Human-authored caption	6.39	5.55

5.2.2 Result

The result caption from the news document and the image annotated model using Phrase-based caption generation technique is shown in Figure.3

Figure.3 Output image with caption

6 CONCLUSION

In this paper, we introduced the novel task of automatic caption generation for news images. This becomes useful for various multimedia applications, such as image and video retrieval and development of tools supporting news media management. We have presented extractive and abstractive caption generation models. A key aspect of our approach is to allow both the visual and textual modalities to influence the generation task. This is achieved through an image

annotation model that characterizes pictures in terms of description keywords that are subsequently used to guide the caption generation process.

Our results show that the visual information plays an important role in content selection. Simply extracting a sentence from the document often yields an inferior caption. Our experiments also show that a probabilistic abstractive model defined over phrases yields promising results. It generates captions that are more grammatical than a closely related word-based system.

The model presented here could be further improved in several ways. First, we could allow an infinite number of topics and develop a nonparametric version that learns how many topics are optimal. Second, our model is based on word unigrams, and in this sense takes very little linguistic knowledge into account. Recent developments in topic modeling could potentially rectify this, e.g., by assuming that each word is generated by a distribution that combines document-specific topics and parse-tree specific syntactic transitions. Third, our model considers mostly local features for representing the images. A better representation would also take global feature dependencies into account.

7 ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers who with their comments, helped increase the quality of the paper.

REFERENCES

[1] Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image Classification for Content-Based Indexing", *IEEE Trans. Image Processing*, vol. 10, no. 1, pp. 117-130, 2001.

[2] A.W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349-1380, Dec. 2000.

[3] Abella, J.R. Kender, and J. Starren, "Description Generation of Abnormal Densities Found in Radiographs," *Proc. Symp. Computer Applications in Medical Care, Am. Medical Informatics Assoc.*, pp. 542-546, 1995.

[4] Kojima, T. Tamura, and K. Fukunaga, "Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions," *Int'l J. Computer Vision*, vol. 50, no. 2, pp. 171-184, 2002.

[5] Kojima, M. Takaya, S. Aoki, T. Miyamoto, and K. Fukunaga, "Recognition and Textual Description of Human Activities by Mobile Robot," *Proc. Third Int'l Conf. Innovative Computing Information and Control*, pp. 53-56, 2008.

[6] 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.

[7] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A.C. Berg, and T.L. Berg, "Baby Talk: Understanding and Generating Image Descriptions," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1601-1608, 2011.

[8] 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.

[9] M. O ¨ zcan, L. Jie, V. Ferrari, and B. Caputo, "A Large-Scale Database of Images and Captions for Automatic Face Naming," *Proc. British Machine Vision Conf.*, pp. 1-11, 2011.

[10] M. Corio and G. Lapalme, "Generation of Texts for Information Graphics," *Proc. Seventh European Workshop Natural Language Generation*, pp. 49-58, 1999.

[11] S. Elzer, S. Carberry, I. Zukerman, D. Chester, N.Green, and S.Demir, "A Probabilistic Framework for Recognizing Intention in Information Graphics," *Proc. 19th Int'l Conf. Artificial Intelligence*, pp. 1042-1047, 2005.

[12] Aker and R. Gaizauskas, "Generating Image Descriptions Using Dependency Relational Patterns," *Proc. 48th Ann. Meeting Assoc. for Computational Linguistics*, pp. 1250-1258, 2010.

[13] M. Hodosh, P. Young, C. Rashtchian, and J. Hockenmaier, "Cross- Caption Coreference Resolution for Automatic Image Understanding," *Proc. 14th Conf. Computational Natural Language Learning*, pp. 162-171, 2010.

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

[15] Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, no. 2, pp. 91-110, 2004.

[16] Sadeghi and A. Farhadi, "Recognition Using Visual Phrases," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1745-1752, 2011.

[17] 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.

[18] A. 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.

[19] 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.