International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Generating Captions for News Domain

Kalyani Zinjurde, Ena Jain

[1]Dr.BAMU University, DIETMS, Aurangabad, India

[2]Professor, Dr. BAMU University, Department of Computer Science, Aurangabad, India

Abstract: ***Many of the search engines deployed on the web retrieve images without analyzing their content, simply by matching their user queries against their collocated textual information. This limits the applicability of search engines, a great deal of work has focused on the development of generating descriptive words for an image automatically. Images are automatically captioned in two stages: Content selection and surface realization. Content selection suggests keywords for the image using image annotation model. Whereas, surface realization technique uses extractive and abstractive methods to generate caption. The approach is to analyze the performance of captioned images using phrases rather than words.***

Keywords: Image annotation Process, Text annotation Process, content selection, stemming process, stop word removal, surface realization, caption formation.

1. Introduction

Recent immense growth has witnessed in the amount of digital information available on the internet including billions of images, documents, articles, books, sound, video, and social networks many search engines display images with its respective stories. Automatic caption generation is most widely used in real life because news televisions can show the news image with the help of caption generation so that everybody can see the news images with correct information and image retrieval. A good caption must be succinct and informative that correctly identifies the subject of an image this could also assist journalists in creating descriptions. Previous papers used face detection to generate caption. In this paper image is analyzed using image processing techniques into abstract representation rendered into Natural Language description. I introduce a novel knowledge –lean framework for news image caption generation here content selection and surface realization models can be learned from weakly labeled data in an unsupervised fashion. Indeed, the output of our abstractive model compares favorably to hand-written captions and is often superior to extractive methods. Images are first segmented into objects, their signature is retrieved from the database, and a description is generated using templates. This approach can create meaningful sentences of high quality and meaningful. We focus on captioned images embedded in news articles.

1.1 Image Description and Generation

Computer vision and natural language processing (NLP) have been previously treated in isolation. The former usually deals with how to make machines see the world, while the latter mainly focuses on how to make machines understand human language. However, there is an increasing demand for bringing the two together. An image description system can help people better manage the increasing volumes of multimedia data ranging from daily life entertainment

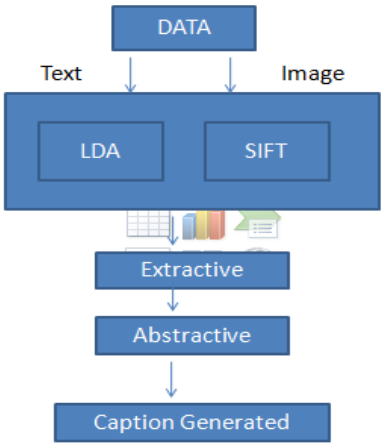


Figure 1: Working flow to generate a caption.

Besides helping people with special needs, an automatic image description generation component would also create more detailed and complete summaries for documents with multimedia representations. Current document summarization systems focus solely on textual information while ignoring pictures, graphical figures, or tables that are embedded in documents. These representations usually convey complementary information that is only implicitly described in the main text. Furthermore, these graphical representations could play an important role in determining what information is crucial for the document and should therefore be included in the summary. An image description generation module could help decide what to say in the summary and automatically render the missing information into natural language, thus enabling text summarization systems to produce more comprehensive summaries.

As far as image retrieval is concerned, automatic image description generation could help improve system accuracy and end-user experience. Although image indexing techniques based on keywords are popular and the method of choice for practical image retrieval engines, there are good reasons for using more linguistically meaningful descriptions. A list of isolated keywords is often ambiguous. An image annotated with the words “car, blue, sky” could depict a blue

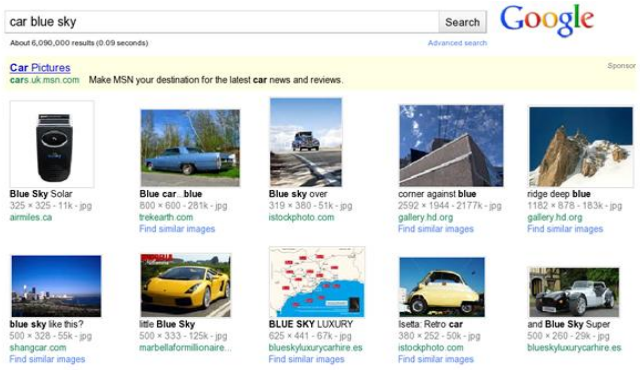
car or a blue sky, whereas the caption "a car running under the blue sky" makes the relations between the words explicit (e.g., sky is modified by blue, car is under the sky), and supplies richer underlying information usually absent from keyword lists, such as actions (e.g., running), who did what to whom, name entities and so on.

1.2 Extracting Image Content

The first challenge concerns identifying what the image is about (i.e., extracting its content). Given an image, an ideal image understanding system would reliably identify the depicted scene, its objects, which objects are important or prominent, and their relations etc. However, full image understanding is beyond the capability of current computer vision research. For example, most previous work [3] adopts a knowledge-rich approach, where the cross-modality correspondence is made explicit through human annotation. More recent research has placed emphasis on a relatively simpler approach, namely automatic image annotation, which can be considered as an approximation of the full image understanding problem by addressing the main objects or events instead of every object in the image [4]. Given an image, a hypothecal image annotation system is expected to automatically label it with description keywords. This task, on its own, is of significant importance for many image-based applications, such as image retrieval, picture browsing support, and story picturing [5]. Especially, since manually annotating images for a large database is a labor intensive and time consuming task. In the long run, it can also be expensive since the work has to be repeated with every new collection.

In practice, existing image retrieval systems annotate their image databases mainly by analyzing image captions (if they exist), textual descriptions found adjacent to the images, and other text-related information such as the file name of the image, metadata of the image, or user click information. For example, consider the images and their surrounding text in Figure 1.2. The short description "Blue Sky Solar Bluetooth Hands free Car Kit, include shipping." found around the first image is further used as an annotation for it even though there are many words (e.g., "solar, car, shipping") that are not directly related to the image's content. Search engine takes textual queries as input and return images with annotations most similar to them. As they do not analyze the actual content of the images, image search engines will perform poorly when retrieving pictures from unannotated collections, or with low quality annotations. The latter is common in web applications as texts found near the images are often irrelevant to their content.

To remedy this, a large number of image annotation models have been proposed recently that exploit the synergy between visual and textual modalities by learning the correspondence between image regions (or features) and keywords. These approaches follow many distinct learning paradigms, ranging from supervised classification [6] to instantiations of the noisy-channel model and methods inspired by information retrieval. Despite their differences, all these methods essentially attempt to learn the correlation between image features and words from examples of annotated images.



2. Rendering Image Content in Natural Languages

Even if we assume that we can reliably describe the image content in terms of keywords, rendering these keywords into human-readable output is far from trivial. A common framework across different image description generation methods is to rely on a domain specific background knowledge base to organize the extracted image content into a structured representation with pre-specified semantic relations, and then, to use a template-based or grammar-based surface realizer to produce sentences for this structured image content [3]. Although this framework can output grammatical sentences, the reliance on manually created knowledge bases restricts its applicability in wider domains. For instance, in an office-scene video surveillance application, a human action concept ontology is manually constructed to map a sequence of human positions and postures into abstract actions (e.g., a trajectory of head motions passing a door is interpreted as the action enter).

This knowledge base is highly related to the specific application and cannot be expected to work well out-of-domain, for instance, when applied to traffic scenes. Furthermore, manually obtaining such a background knowledge base is time consuming, costly and has to be repeated for new domains. Yao et al. (2009) [7] state that the LHI database, containing around 1 million deep segmented images together with information denoting the functional relationships among objects, is annotated by a team of 23 annotators aided by a software development team with two years full-time work. Besides the creation of knowledge bases, this framework is further limited by the substantial human involvement required in the surface realization process. The predefined sentence templates or grammars are essential parts in most realizers, but most of them are not reusable across domains. The template-filling approaches often generate repetitive and stilted text due to the limited number of predefined templates. Moreover, either template-filling or grammar based models are flexible enough to express the image content in different contexts.

2.1 The synergy between visual and textual modalities

Being multi-modal, the image description generation task must unavoidably exploit the synergy between visual and textual modalities. Many experimental studies in language

acquisition suggest that word meaning arises not only from exposure to the linguistic environment but also from our interaction with the physical world. For example, infants, from an early age, are able to form perceptually-based category representations. Perhaps unsurprisingly, words that refer to concrete entities and actions are among the first words being learned as these are directly observable in the environment. Experimental evidence also shows that children respond to categories on the basis of visual features, e.g., they generalize object names to new objects often on the basis of similarity in shape and texture. Humans can describe images effortlessly, probably because they have a common underlying representation for the two modalities.

2.2 Problem statement and motivation

The caption generation of word based model has a major drawback. As the image annotation model does not take function words in to account, content selection will ignore them too, at the expense of the grammaticality of the generated captions. In other words, there will be no function words to glue the content words together. Therefore Phrase Based Model used because phrases are naturally associated with function words and may potentially capture long-range dependencies. Natural language generation (NLG) is the task of producing natural language output according to certain input. The Input depends on the specific requirements of various applications. For instance, in single sentence generation, it could be a set of concepts with specified relations, or just a set of isolated keywords. And the output is expected to satisfy the input requirements, and also to be grammatical and semantically coherent. These two modules are often referred to as content selection and surface realization, Content selection usually requires a knowledge base to assist in better interpreting the input concepts.

More formally, we define the task of automatic image description generation as below:

Definition 1: Given an image I, and a related knowledge database k, create a natural language description C which captures the main content of the image under k.

Following the typical natural language generation paradigm, the task involves, first analyzing and representing the image content and then rendering it in natural language. And the knowledge base k must contain two types of information, information about how the images (or image regions) corresponds to words and information about how these words can be combined to create a human-readable sentence.

Motivations

During searching process any image can be retrieve on the basis of the collocated textual information e.g. include image file name or format name or the text surrounding that image but those images that are not associated with particular text cannot retrieved. Because of this reason a great deal of work has focused on development of one method that generates image description automatically.

A method that generates such descriptions automatically could therefore improve image retrieval by supporting longer and more targeted queries, by functioning as a short summary of the image's content, and by enabling the use of question-answer interfaces.

The ability to link images with textual descriptions would facilitate the retrieval and management of multimedia data (e.g., video and image collections, graphics) as well as increase the accessibility of the web for visually impaired(blind and partially sighted) users who cannot access the content of many sites in the same ways as sighted users can.It could also assist journalists in creating descriptions for the images associated with their articles or in finding images that appropriately illustrate their text. An automatic image caption generation module could also assist journalists in creating descriptions for the news images or videos associated with their articles. Many on-line news sites like CNN, Reuters, and BBC publish images and videos with their stories and even provide photo feeds related to current events. All tables and figures will be processed as images. You need to embed the images in the paper itself. Please don't send the images as separate files.

2.3 Content Selection

We need to deal with two main problems, namely automatic image annotation and description generation that, although closely related, have been previously studied in isolation.

When looking at previous efforts in automatic image annotation, we address the problem in terms of the training paradigm employed and their capability dealing with real-world data. Current image annotation approaches fall under two broad categories: discriminative and generative models. The former usually achieve better performance however are heavily reliant on the quality of training data which in turn influences their extendibility. In contrast, generative models can deal with low quality data more easily as well as changes in training set or even vocabulary. Recall that in this thesis we will focus on news data, where news images with associated captions and documents co-occur naturally. This type of data will be used as it is without any additional manual annotation and will allow us to treat caption generation as a summarization model. We thus examine current advances of text summarization, survey existing summarization approaches, both extractive and abstractive, and especially keep an eye on whether extra knowledge bases are employed. Extractive approaches dominate the field of automatic text summarization. The main reason is that without good linguistic analysis, it is possible to output good enough summaries both in terms of their content and grammaticality simply by deciding which sentences present the key ideas of the document. Abstractive summarization, on the other hand, is a more challenging task as sentences need not only be extracted but also rewritten. However, it has the potential of creating more human-like summaries that are more succinct and coherent.We first describe the task of automatic image description generation, and briefly review previous approaches and related applications.

3. Automatically Descriptions for Images

A handful of approaches have been proposed in the literature that automatically generates descriptions for images by examining their content. To begin with, the image is represented by image features, which are then replaced by an abstract representation, essentially a set of description words, according to a visual-to-textual representation dictionary[1][3]. The features used to represent the image content mainly include color information [2], textual features [3], detected edges [1], and so on. For certain applications, some objects are detected and recognized with prior knowledge to supply higher level features [4].

Figure: Encoding correspondence between visual features to keywords.

4. Text Summarization

Recall that extracting image content is the first step towards generating an image description. Next, we must render this content into natural language. This task mainly involves natural language generation, i.e., producing natural language outputs from non-linguistic inputs. Generally, a background knowledge base is required to structure the image content and specify relations among these contents whereas a surface realizer (either using templates or grammars), is employed to produce the description accordingly. In this thesis, we adopt a knowledge-lean approach for this task, which means that we will not utilize manually created rules, grammars or sentence templates. Recall that we focus on news image caption generation, where a news image is available together with its associated document and the task is to automatically generate a caption for the news image. Without access to the associated articles, we would be exposed to a traditional NLG problem.

Compared to general NLG systems, summarization models rely less on human labor in many aspects. Firstly, the source text provides a fertile field for lexical choices. Word morphology and necessary function words are often manually imported into an NLG system as rules. With respect to grammaticality, most traditional NLG systems either adopt sentence-templates or rely on predefined grammars to create human-readable sentences. In summarization, the source text naturally supplies grammatical sentences or phrases that can be used to produce grammatical summaries. Additionally, sentence templates or similar paradigms are not suitable for general purpose generation applications due to their inability of generating diverse, expressive and flexible sentences in previously unseen domains.

5. Proposed Work

The availability of appropriate image datasets is therefore crucial for our task. An ideal dataset should (1) be representative of real-world data, (2) relatively easy to collect as we hope to rely on minimal or no human involvement, (3) include images with annotations that will potentially supply visual-textual correspondences, (4) contain auxiliary information that could allow us to mine related linguistic information in order to help us create human readable descriptions, and (5) contain gold standard captions for evaluating the output of our system.

The first process of our proposed work is to take the dataset which contains more number of images and their documents.In annotation process first we check the wordings in those documents with the help of part-of-speech.After that we perform stop-word removal and stemming processes. In annotation process first, we use SIFT algorithm and LDA.Visual features receive a discrete representation and each image is treated as a bag of visual words. In order to does this use the Scale Invariant Feature Transform (SIFT) algorithm [1]. The general idea behind the algorithm is to first sample an image with the difference-of-Gaussians point detector at different scales and locations. Each detected region is represented with the SIFT descriptor [2], which is a histogram of directions at different locations in the detected region and scale. Importantly, this descriptor is, to some extent, invariant to translation, scale, rotation, and illumination changes. SIFT features have been shown to be superior to other descriptors and are considered state of the art in object recognition.Further quantize the SIFT descriptors using the K-means clustering algorithm to obtain a discrete set of visual terms which form our visual vocabulary V [3] . 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. More formally, each image I is expressed in a bag-of-words format vector, $[W_{v1},W_{v2}$,. . . , $W_{vL}]$, where W_{vi}= n only if I has n regions labeled with vi.

Banko et al. (2000) (see also Witbrock and Mittal 1999) [8] propose a bag-of-words model for headline generation. Following the traditional NLG paradigm, their model consists of a content selection and surface realization component. 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 from other words in the headline. Despite its simplicity, the caption generation model has a major drawback. Bear in mind that most image annotation models consider only content words — it does not make sense to output function words as they are not descriptive of the image content. This means that the content selection component will naturally tend to ignore these non-descriptive words. This will seriously impact the grammaticality of the generated captions, as there will be no appropriate function words to glue the content words together. One way to remedy this is to revert to a content selection model that ignores the image and simply estimates the probability of a word appearing in the caption given the same word appearing in the document. At

the same time, we modify our surface realization component so that it takes note of the image annotation probabilities. Intuitively, we hope the new language model will prefer words that have high image annotation probabilities while are likely to appear in a sentence according to the background language model.

$$P(w1, w2, \ldots, wn) = \prod_{i=1}^{n} P(wi \in H | wi \in D) \cdot P(len(H) = n) \cdot \prod_{i=2}^{n} P(wi \mid wi-1) \ldots\ldots\ldots\ldots (2)$$

where wi is a word that may appear in headline H, D the document being summarized and P(len(H)=n) a headline length distribution model. The above model can be easily adapted to our image caption generation task. Content selection is now the probability of a word appearing in the caption given the image and its associated document which we obtain from the output of our image annotation model.

$$P(w1, w2, \ldots, wn) = \prod_{i=1}^{n} P(wi \in C | I, D) \cdot P(len(C) = n) \cdot \prod_{i=3}^{n} P(wi \mid wi-1, wi-2) \ldots\ldots\ldots (3)$$

Where C is the caption, I the image, D the accompanying document.

6. Implementation Details

1) Input image and article: The image and the article are the inputs
2) Content Extraction: Only relevant content is extracted.
3) Summarization : Abstractive summarization is benefitial to create words or phrases for the entire article with its respective image.
4) Text annotation model: Finds out important keywords and phrases using parts of speech.
5) Image annotation model: learns to automatically label images under assumption that the images and the surrounded text are shared set of latent variables.Our annotation model takes these topic distributions into account by searching frequrnt keywords for that image and its relevant document.I have compared LDA and its alternative perceptron.And as per the theory, it is found that LDA is best suitable.Also I have compared 3 different types of LDA namely, word overlap, standard vector space model, and Txt LDA and finally found that mix LDA is better.Also I can use here SIFT algorithm, which is a part of LDA.
6) Caption Generation: After comparing extractive and abstractive caption generation processes I conclude that abstractive is better one since it provides a word based or phrase based caption whereas, extractive results in a single sentence and there is a probability of generating a single sentence that may not clearify the whole article.Also they are not concise and they are not catchy.
   a) Word based : Content selection is used as the probability of a word appearing in the headline given that the similar word occurs in the corresponding document.Surface realizations also take the length of a caption into account to generate output of reasonable length.
   b) Phrase based : In word based generation process there is no guarantee that the caption will be compatible.to overcome this I need phrases to capture long-range dependencies.

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

7. Evaluation

As this paper explores the feasibility of automatic caption generation for images in news domain it is having some limitations like images that don't co-inside with textual data can't be retrieved. We need to follow some evaluation techniques to get oue work done without any manual involvement and human authored grammar. At testing time we will input 1 news articles and image, then system will generate caption. 1) Image Database: Database created with huge amount of images & articles. The dataset also covers wide range of topics including national and international topics related to science , sports, politics, technology, etc.

1) Content selection : Identifies important keywords from text and images. Text frequency calculator is used to extract these keywords which will also find out the importance of that keyword in that article. One word is read at a time and accordingly hash table is built.
2) Surface Realization: Verbalize the keywords extracted from text & image.
3) Stemming process is used to reduce the variant forms of a word into a common single word.
4) Stop word removal is used to remove most frequently occurring meaningless words.
5) SIFT: Describe local features in an image.

8. Result

In this paper I have explained the possibility of generating captions automatically for particularly news domain because the dataset for news domain is easily available.Image annotation model converts image features into self understandable keywords which are frequently used to guide the generation process.Finally sentence is generated by combining phrases generated by abstractive method. Following image and respective article is taken from the standard dataset:

Image:

Article

They denounced the government violation of a May peace agreement and expressed concern for the safety of civilians. This is the first time since the peace deal that the Sudanese army has been confirmed as fighting in Darfur. The attack was against the rebel Justice and Equality Movement, (JEM) which did not sign the deal. Some 2m people have fled their homes in Darfur since conflict began in 2003, and tens of thousands of people are reported to have been killed in ongoing violence. Both government and militia troops had been observed massing near the western town of Geneina before the attack on Friday. An assortment of armed groups that remained outside of the peace agreement, including Chadian elements, are known to be taking shelter in the Jebel Moon mountains. The attack is confirmation that Darfur's conflict has changed in nature, the BBC's Jonah Fisher reports from Sudan. The two signatories to the peace agreement - the government and the Minni Minnawi faction of the Sudan Liberation Movement - are using the agreement as a springboard to attack those outside the deal. The SLM Minnawi has launched a wide-ranging offensive against their former rebel allies and supporters, leaving at least 80 people killed and thousands displaced. JEM spokesman Ahmed Hussein Adam said the Sudanese government was systematically attacking groups who had refused to sign for peace. SLM Minnawi was the main rebel group in the Darfur conflict, and the only one that signed the deal on 5 May with the Sudanese government.

9. Conclusion

I have introduced the task of automatic caption generation for news domain. This concept virtualizes the inheritance of computer vision and NLP. Already it is available on internet resource to retrieve images with respect to the user queries. Here the results shows caption generated by weakly labeled data without any costly manual involvement. Captions are treated as labels for images. These captions can be used to learn the correspondances between textual and visual modalities and act as a gold standard. We built our dataset from resources that are publicly available on the internet without manual post processing.

10. Future Scope

A news in the form of video involves key frames from streaming video data. So instead I can use this strategy for video purpose. Currently I treat the image regions or ROI as bags of words, which could be extended to bigrams according to their spatial relations. Also I could however, improve grammaticality more globally by generating a dependency graph. In this work, images were preprocessed by extracting primarily local feature representations (e.g., color, texture, corners, SIFT features, etc.), without considering more global representations, such as spatial relationship among different regions. An obvious extension would be taking spatial information into account when dealing with image representations. Currently, we treat the image regions or detected regions of interest as bags-of-words, which could be extended to bigrams according to their spatial relations.

The image caption generation task has been formulated as a two-step approach, where the image content extraction and caption generation are carried out sequentially. A more general model should integrate the two steps in a unified framework. Indeed, an avenue for future work would be to define a phrase-based model for both image annotation and caption generation.

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.

Author Profile

Kalyani Zinjurde received the B.E degree in Computer Engineering from PESCOE in 2010 and persuing M.E degree in computer Engineering ,Working as a lecturer in DIETMS, Aurangabad (Mah.), India.