ISSN:0975-9646

Automatic Caption Generation for News Images

Priyanka Jadhav, Sayali Joag, Rohini Chaure, Sarika Koli

Information Technology,Pune University,
NDMVP COE,Nasik-13,Maharashtra,India

Abtract- **This thesis is concerned with the task of automatically generating captions for images, which is important for many image related applications. Our model learns to create captions from publicly available dataset that has not been explicitly labelled for our task. A dataset consists of news articles, the pictures embedded in them, and their captions, and consists of two stages. First stage consists of content selection which identifies what the image and accompanying article are about, whereas second stage surface realization determines how to put the chosen content in a proper grammatical caption. For content selection, we are using probabilistic image annotation model that suggests keywords for an image. This model postulates that images and their textual descriptions are generated by a shared set of latent variables (topics) and is trained on a weakly labeled dataset (which treats the captions and associated news articles as image labels). The abstractive surface realization model generates captions that are favorable to human generated captions.**

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

1.INTRODUCTION

Many search engines available on the web retrieve images without analyzing their content, simply by matching user queries against colocated textual information. For example metadata,user-annotated tags, captions, and, generally, text surrounding the image. But this has major disadvantages, they challenge the applicability of search engines. So there is a need of focussing on the development of methods that generate description words for a picture automatically. In our project, we tackle the related problem of generating captions for news images. We focus on captioned images embedded in news articles, and make use of both models of content selection and surface realization from data and thus avoid expensive manual annotation.For example,an image annotated with the words blue, sky, car could depict a blue car or a blue sky, whereas the caption "car running against the blue sky" would make the relations between the words explicit. Also, image descriptions need to be concise, focusing on the most important depicted objects or events. A method that generates such descriptions automatically could also assist journalists in creating descriptions for the images associated with their articles or in finding images that appropriately illustrate their text. Also the linking of images with textual descriptions would facilitate the retrieval and management of multimedia data .It could also assist journalists in creating descriptions for the images associated with their articles or in finding images that appropriately illustrate their text.

2.BLOCK DIAGRAM

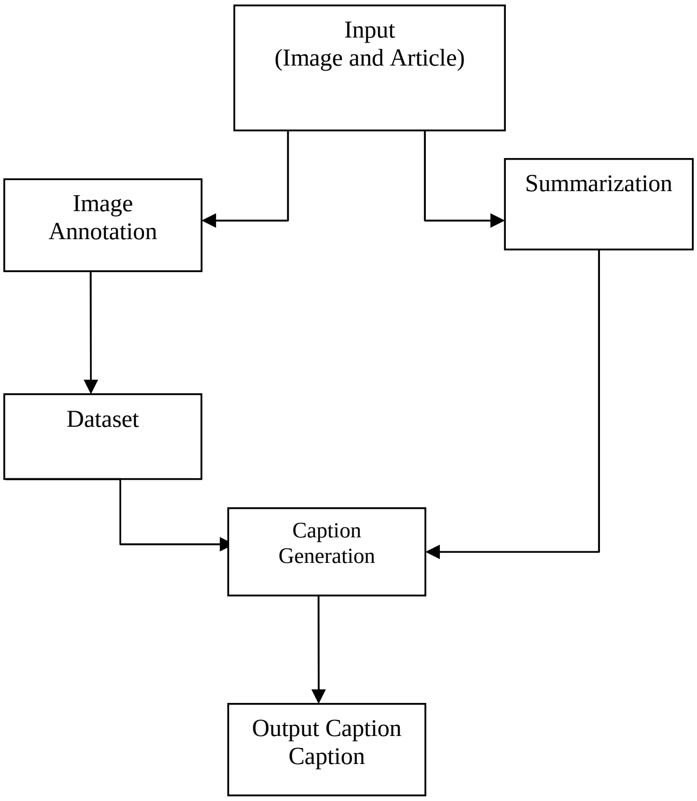

Fig 1.1:Block Diagram of Proposed work

3.RELATED WORK

All previous methods 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.

1. Automatic Image Description Generation

This applications follws a two stage architecture. The image 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 indexed by image signature(e.g. Colour and texture), and reliance on background ontological information.

2.***Automatic Description for human activities in video***
The idea is to extract features of human motion from video keyframes and interleave them with a concept hierarchy of actions to create a case frame from which a natural language sentence is generated.

4.PROPOSED WORK

Instead of relying on manual annotation or background ontological information we exploit a multimodal database of news articles, images, and their captions. Using an image annotation model, we first describe the picture with keywords, which are subsequently realized into a human readable sentence. Our experiments used news articles accompanied by captioned images.
We formulate the image caption generation task as follows: Given a news image I and its associated documentD, create a natural language caption C that captures the image's content given D. The training data thus consists of document-imagecaption tuples . During testing,we are given a document and an associated image for which we must generate a caption. And the knowledge base 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.

5.IMPLEMENTATION DETAILS

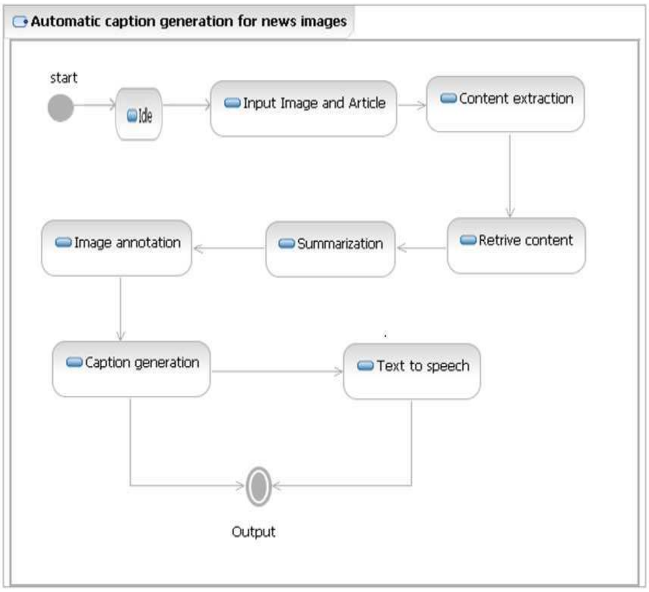

Fig1.2 State Diagram

5.1 Input image and article

The image and the article are the inputs of the project.

5.2 Content extraction

In this only the related content is extracted and from this the relevant content is retrived in retrieve content phases.

5.3Summarization

In this phase the only focus is on textual information while ignoring pictures ,graphical figures, or tables that are embedded in documents to produce more comprehensive summaries. In this we use abstractive summarization module which produces more human-like summaries. In this, the source text naturally supplies grammatical sentences or phrases that can be used to produce grammatical summaries. The abstractive summarization first identify the key content of documents in the form of constituents,e.g., words or phrases, which are then organized into a grammatical sentence.

5.4Image annotation

In this we annote the image from associated news document. For this we propose a probabilistic image annotation model that learns to automatically label images under the assumption that images and their surrounding text are generated by a shared set of latent variables or topics. We use Latent Dirichlet Allocation, a probabilistic model of text generation, we represent visual and textual meaning jointly as a probability distribution over a set of topics. Our annotation model takes these topic distributions into account while finding the most likely keywords for an image and its associated document. We have compared the mix LDA with the other three types of LDA namely word overlap, standard vector space model and Txt LDA. And as per the results, its found that the Mix LDA significantly ($p < 0:01$) is better than these models by a wide margin and has accuracy of 57.3% .As compared with other models it has accuracy of 31.0% over Txt LDA, 38.7% over the vector space model, and accuracy of 31.3% over the word overlap.For this we also use the Scale Invariant Feature Transform (SIFT) algorithm,which is the part of LDA.

5.5 Caption generation

For this, we are using the abstractive caption generation approach. As we know, 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 are spread across multiple sentences. Secondly, the selected sentences make for long captions which are not concise and overall not as catchy as well. For these reasons we turn to abstractive caption generation and in this we are using two models, firstly the word based model and secondly the phrase based model.
5.5.1Word-basedCaptionGeneration
In this approach of caption generation ,Content selection is modelled 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. The likelihood of different surface realizations is estimated using a bigram model. They also take the distribution of the length of the headlines into account in an attempt to bias the model towards generating output of reasonable length.
5.5.2 Phrase based caption generation
In the word based caption generation, there is no guarantee that these will be compatible with their surrounding context or that the captions will be globally coherent. Thus to avoid these problems, we turn our attention to phrases which are naturally associated with function words and may potentially capture long-range dependencies.

6.RESULT

In this , we explored the possibility of automatically generating captions for news Images. This software allows both visual and textual information to influence automatic caption generation task. Here,image annotation model converts features of an image into self explanatory keywords which are subsequently used to guide the generation process. Given an accompanying document our abstractive model generates an sentence. Both outputs are processed and caption has been generated.The news and the article are as follows:

Image:

Article:

Eighteen people have been injured in the incident, which sent smoke billowing into the city sky.More than 250 firefighters are tackling the blaze at the scene near 116th Street and Park Avenue. All train services in and out of Grand Central terminal have been halted following the incident near its tracks. New York City Mayor Bill de Blasio said in a news conference from the scene that the gas leak had been reported to the utility company just 15 minutes before the blaze.Mr de Blasio said the "major explosion" had destroyed two buildings and heavily damaged other structures.A number of individuals were still missing in the area, he added. Generated caption from our project: Two dead as the newyork city buildings collapse after a gas blast.

7.CONCLUSION

We have introduced the task of automatic caption generation for news images. The task fuses insights from computer vision and natural language processing. Alike from previous work, we have approached this task in a knowledge-lean fashion by using the vast resource of images available on the Internet and exploiting the fact that many of these co-occur with textual information (i.e., captions and associated documents). Our results show that it is possible to learn a caption generation model from weakly labeled data without costly manual involvement. Image captions are treated as labels for the image. Although the caption words are admittedly noisy compared to traditional human-created keywords, we show that they can be used to learn the correspondences between visual and textual modalities, and also serve as a gold standard for the caption generation task. Moreover, this news dataset contains a unique component, the news document, which provides both information regarding to the image’s content and rich linguistic information required for the generation procedure.

8.FUTURE SCOPE

As video processing usually involves processing key frames (images) from streaming video data, it is also possible to adapt existing models and applications from images to video (e.g., automatic video summarization).The dataset discussed in this thesis can be further refined according to specific Applications to use the news document to increase the annotation keywords by identifying synonyms or even sentences that are similar to the image caption. 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. For instance, we could experiment with features related to document structure such as titles, headings, and sections of articles and also exploit syntactic information more directly. We could, however, improve grammaticality more globally by generating a well-formed tree (or dependency graph).

ACKNOWLEDGEMENT

We take this opportunity to thank Prof. Vaishali S. Pawar for their valuable guidance and for providing all the necessary support to accomplish this research. We would like to extend our gratitude towards our beloved Principal Dr. Jayant T. Pattiwar for his great support.

REFERENCES

1. Feng, Y. and Lapata, M. (2010c). Visual information in semantic representation. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 91–99, LosAngeles, California. Association for Computational Linguistics.
2. Feng, Y. and Lapata, M. (2010a). How many words is a picture worth? automaticmcaption generation for news images. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1239–1249, Uppsal, Sweden. Association for Computational Linguistics.
3. Holub,A. (2007). Information Fusion for Multidocument Summarization: Paraphrasing and Generation. PhD thesis, California Institute ofTechnology
4. Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization.Technical report, Microsoft Research.