International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Automatic Caption Generation for Pictures Available on Web

Dr. M Nagaratna[1], Nafees Fathima[2]

[1]Assistant Professor, Department of Computer Science, JNTU, Hyderabad, India

[2]M. Tech, CSE, JNTU, Hyderabad, India

Abstract: ***In this paper, we've an inclination to tend to introduce the novel task of automatic caption generation for news footage. The task fuses insights from laptop computer vision and tongue technique and holds promise for varied multimedia system applications, like image retrieval, development of tools supporting medium management, and for people with incapacity. It's potential to be told a caption generation model from sapless labelled knowledge whereas not costly physical involvement. Rather than physically making annotations, image captions area unit treated as labels for the image. although the caption words area unit confessedly shouting compared to ancient human-created keywords, we've an inclination to tend to suggests that they are attending to be accustomed learn the correspondences between visual and matter modalities, and in addition perform a gold customary for the caption generation task. We've got given extractive and speculative caption generation models. A key facet of our approach is to permit every the visual and matter modalities to influence the generation task.***

Keywords: Caption generation, image retrieval, Multimedia.

1. Introduction

Recent years have witnessed associate degree new growth within the quantity of digital data out there on the web. Flicker, one of the best far-famed picture sharing websites, hosts additional than three billion pictures, with approximately two.5 million pictures being uploaded each day.1 several on-line news sites like CNN, Yahoo!, and BBC publish pictures with their stories and even give pic feeds associated with current events. Browsing associate degreed finding footage in large-scale and heterogeneous collections is a necessary drawback that has attracted a lot of interest among data retrieval. Several of the search engines deployed on the net retrieve pictures while not analyzing their content, merely by matching user queries against collocated matter information. Examples embody information (e.g., the image's file name and format), user-annotated tags, captions, and, generally, text close the image. As this limits the pertinence of search engines (images that don't coincide with matter information cannot be retrieved), an excellent deal of work has cantered on the development of strategies that generate description words for an image mechanically. The literature is full of varied makes an attempt to learn the associations between image options and words exploitation supervised classification although keyword-based categorization techniques are popular and the technique of selection for image retrieval engines, there are sensible reasons for exploitation additional lingual meaningful descriptions. A listing of keywords is commonly ambiguous. A technique that generates such descriptions automatically may so improve image retrieval by supporting longer.

2. Literature Survey

2.1 Image Classification for Content-Based Indexing

User queries in content-based retrieval are typically based on semantics and not on low-level image features. Providing high-level semantic indices for large databases is a challenging problem. We have shown that certain high-level semantic categories can be learnt using specific low-level image features under the constraint that the images do belong to one of the classes under consideration

2.2 Content-Based Image Retrieval at the End of the Early Years

The paper presents a review of 200 references in content-based image retrieval. The paper starts with discuss the working environment of content-based retrieval patterns of use, types of pictures, the role of semantics, and the sensory gap. Consequent sections discuss computational steps for image retrieval systems. Stage one of the review is image processing for retrieval sorted by colour, texture, and local geometry. Features for retrieval are discussed next, sorted by accumulative and universal features, outstanding points, object and shape features, signs, and structural combinations.

2.3 Multiple Bernoulli Relevance Models for Image and Video Annotation

We have proposed a multiple-Bernoulli relevance model for image annotation, to formulate the process of a human annotating images. The results show that it outperforms, especially on the ranked retrieval task, the (multinomial) continuous relevance model and other models on both the Corel dataset and a more realistic Trec Video dataset.

2.4 Efficient Object Recognition Using Boundary Representation and Wavelet Neural Network

In this paper, an efficient object recognition method using boundary representation and the wavelet neural network is anticipated. The method employs a wavelet neural network (WNN) to characterize the singularities of the object curvature representation and to perform the object classification at the same time and in an automatic way.

2.5 Simultaneous Image Classification and Annotation

In this paper, we develop a new probabilistic model for jointly modelling the image, its class label, and its interpretation. Our model treats the class label as a global description of the image, and treats explanation terms as local descriptions of part of the image. Its underlying probabilistic assumptions naturally put together these two sources of information.

3. Existing System

Many of the search engines deployed on the web retrieve images without analyzing content, simply by matching user queries against collocated textual information. Examples include metadata (e.g., the image's file name and format), user-annotate tags, captions, and, generally text neighbouring the image. As this limits the applicability of search engines (images that do not coincide with textual data cannot be retrieved), a great arrangement of work.

4. Limitations

- The web retrieve images without analyzing their content, simply by corresponding user queries against collocated textual information.
- Images that do not match with textual data cannot be retrieved

5. Proposed System

In this paper, we tackle the related problem of generating captions for news images. Our advance leverages the vast source of pictures available on the web and the fact that many of them naturally co-occur with topically related documents and are captioned. It focuses on captioned images surrounded in news articles, and learns both models of satisfied selection and surface realization from data without requiring expensive manual footnote. At training time, our models learn from images, their captions, and linked documents, while at test time they are given an image and the document it is embedded in and produce a caption. Compared to most work on image description making, our approach is shallower, it does not rely on dictionaries specifying image-to-text correspondences, nor does it use a human-authored grammar for the caption creation task. It uses the document co-located with the image as a proxy for linguistic, visual, and world-knowledge. Our innovation is to exploit this implicit information and treat the surrounding document and caption words as labels for the image, thus reducing the need for human supervision.

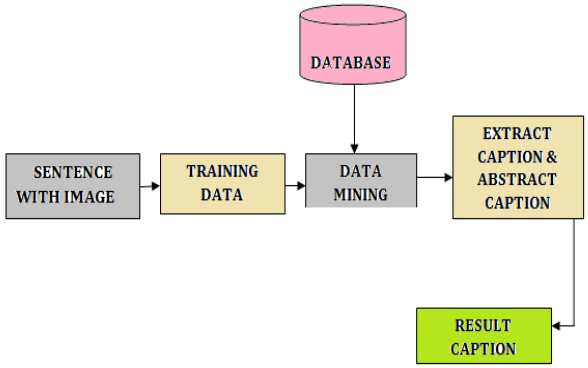

Figure 1: Architecture Diagram

6. Advantages

- Content determination and surface realization from data without requiring expensive manual annotation.
- It does not rely on dictionary specifying image-to-text correspondence, nor does it use a human-authored grammar for the caption creation task.
- It reduces the need for human management.

7. System Implementation

7.1 Data Collection

We created our own dataset by downloading articles from the intelligence websites. The dataset covers a wide range of topics including national and international policy, technology, sports, education, and so on. Information articles normally use colour images which are around 200 pixels wide and 150 pixels high. The captions tend to use half as many words as the document sentences and more than 50 percent of the time contain words that are not attested in the document.

7.2 Input preparation

The document should contain the necessary background information which the image describes or supplements. And also we can exploit the rich linguistic information inherent in the text and address caption generation with methods relative to text summarization without extensive knowledge engineering. The caption generation task is not constrained in any way, words and syntactic structures are chosen with the aim of creating a good caption rather than rendering the task acceptable to current vision and language generation techniques.

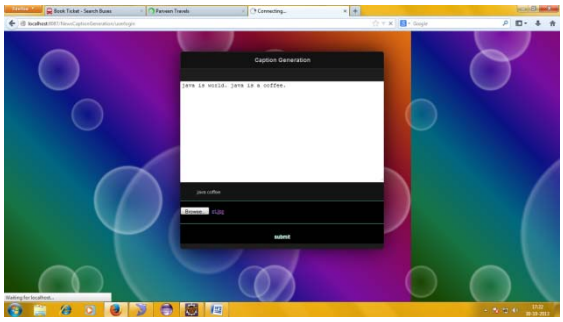

Figure 1: Input Preparation

7.3 Abstractive Caption

We turn to abstractive caption generation and present models based on single words but also phrases. 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 of other words in the headline. They also take the distribution of the length of the headlines into account in an attempt to relative to the model toward generating output of reasonable length.

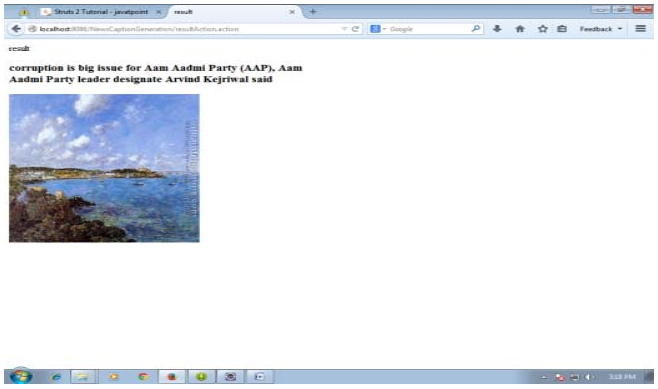

Figure 2: Caption generation

7.4 Extractive Caption

This Extractive caption mostly focuses on sentence extraction. The idea is to create a summary simply by identifying and subsequently concatenating the most important sentences in a article. Without a great arrangement of linguistic analysis, it is possible to create summaries for a wide range of documents, originally of style, text type, and subject matter. For our caption creation task, we need only extract a single sentence. And our guiding assumption is that this sentence must be maximally similar to the description keywords generated by the annotation model.

8. Conclusion

In this paper, we have a tendency to introduce the novel task of automatic caption generation for news pictures. The charge fuses insights from pc apparition and language process and holds promise for numerous multimedia system applications, like image and video renovation development of tools supporting fourth estate administration, and for people with visual destruction. As a departure from previous work, we approach this task in a very knowledge-lean fashion by investment the immense resource of pictures out there on the net and exploiting the actual fact that a lot of of those accompany matter info (i.e., captions and associated documents). Our results show that it doable to be told a caption generation model from infirm labelled information while not expensive manual involvement. The dataset we have a tendency to use that contains real-world pictures and exhibits an outsized vocabulary as well as each concrete object names and abstract keywords; rather than manually making annotations, image captions are treated as labels for the image. Though the caption words are confessedly screeching compared to ancient human-created keywords, we have a tendency to show that they will be wont to learn the correspondences between visual and matter modalities, and conjointly function a gold normal for the caption generation task. Moreover, this news data set contains a singular element, the news document, that provides each info concerning to the image's content and wealthy linguistic info needed for the generation procedure.

9. Future Scope

Our caption generation model adopts a two-stage approach wherever the image process and surface realization area unit meted out consecutive. A additional general model ought to integrate the 2 steps in an exceedingly unified framework. Indeed, Associate in Nursing avenue for future work would be to outline a phrase-based model for each image annotation and caption generation, e.g., by exploiting recent add police work visual phrases. we have a tendency to additionally believe that our approach would have the benefit of additional elaborate linguistic and non-linguistic info. As an example, we have a tendency to might experiment with options associated with document structure like titles, headings, and sections of articles, and additionally exploit syntactical info additional directly. The latter is presently utilized in the phrase primarily based model by taking attachment possibilities into consideration. We could, however, improve grammaticality additional globally by generating a grammatical tree (or dependency graph).

References

[1] Vailaya, M. Figueiredo, A. Jain, and H. Zhang, "Image Classification for Content-Based Indexing," IEEE Trans. Image
[2] Processing, vol. 1s0, no. 1, pp. 117-130, 2001.
[3] A.W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years,"
[4] IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 22, no. 12,pp. 1349-1380, Dec. 2000.
[5] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for aFixed Image Vocabulary," Proc. Seventh European Conf. Computer
[6] Vision, pp. 97-112, 2002.
[7] D. Blei, "Probabilistic Models of Text and Images," PhD dissertation, Univ. of Massachusetts, Amherst, Sept. 2004.
[8] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M.Jordan, "Matching Words and Pictures," J. Machine Learning Research, vol. 3, pp. 1107-1135, 2002.
[9] Wang, D. Blei, and L. Fei-Fei, "Simultaneous Image Classificationand Annotation," Proc. IEEE Conf. Computer Vision and PatternRecognition, pp. 1903-1910, 2009.
[10] V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures," Proc. 16th Conf. Advances in Neural Information Processing Systems, 2003.
[11] S. Feng, V. Lavrenko, and R. Manmatha, "Multiple Bernoulli Relevance Models for Image and Video Annotation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1002-1009, 2004.
[12] L. Ferres, A. Parush, S. Roberts, and G. Lindgaard, "Helping People with Visual Impairments Gain Access

to Graphical Information through Natural Language: The igraph System," Proc. 11th Int'l Conf. Computers Helping People with Special Needs, pp. 1122-1130, 2006.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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] 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.