ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR

International Journal OF Engineering Sciences & Management Research

AUTOMATIC CAPTION GENERATION FOR NEWS IMAGES USING EXTRACTIVE AND ABSTRACTIVE MODELS

Sushma Patwardhan, Harjeet Kaur
Student, Assistant Professor, Department of E&TC, Indira College of Engineering and Management, Pune, India

Keywords: Caption Generation, Image retrieval, Multimedia.

ABSTRACT
Automatic image caption generation is of great interest to many image related applications. Now a day's, whenever retrieving images from the search Engines that retrieves images without analyzing their content, simply by matching user queries against the image's file name and format, user-annotated tags, captions, and, generally, text surrounding the image. Also the retrieved image does not contain any textual data along with the images. We introduced the task of automatic caption generation for news images. The task fuses insights from computer vision and natural language processing and holds promise for various multimedia applications, such as image retrieval, development of tools supporting news media management, and for individuals with visual impairment. It is possible to learn a caption generation model from weakly labelled data without costly manual involvement. Instead of manually creating annotations, 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. 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.

INTRODUCTION
Recent years have witnessed an unprecedented growth in the amount of digital information available on the Internet. Flickr, one of the best known photo sharing websites, hosts more than 3
billion images, with approximately 2.5 million images being uploaded every day.

Figure 1. Images uploaded on websites

Many online news sites like CNN, Yahoo!, and BBC publish images with their stories and even provide photo feeds related to current events. Browsing and finding pictures in large-scale and heterogeneous collections are an important problem that has attracted much interest within information retrieval. Many of the search engines deployed on the web retrieve images without analysing their content, simply by matching user queries against collocated textual information. Examples include metadata (e.g., the image's file name and format), user-annotated tags, captions, and, generally, text surrounding the image. As this limits the applicability of search engines (images that do not coincide with textual data cannot be retrieved), a great deal of work has focused on the development of methods that generate description words for a picture automatically. The literature is littered with various attempts to learn the associations between image features and words using supervised classification [1], [2], instantiations of the noisy-channel model [3], latent variable models [4], [5], [6], and models
inspired by information retrieval [7], [8].Automatic image caption generation is of great interest to many image related applications. Examples include image search engines and tools for helping people with visual impairment to access multimedia information in the same way as sighted people. However, relatively little work has focused on the interplay between visual and linguistic information in literature. Existing efforts often follow a two-step natural language generation framework consisting of content selection and surface realization. This paper is concerned with the task of automatically generating captions for images, which is important for many image related applications. Examples include video and image retrieval as well as the development of tools that aid visually impaired individuals to access pictorial information. Our approach leverages the vast resource of pictures available on the web and the fact that many of them are captioned and collocated with thematically related documents. Our model learns to create captions from a database of news articles, the pictures embedded in them, and their captions, and consists of two stages. Content selection identifies what the image and accompanying article are about, whereas surface realization determines how to verbalize the chosen content. We approximate content selection with a probabilistic image annotation model that suggests

keywords for an image. The model postulates that images and their textual descriptions are generated by a shared set of latent variables (topics) and is trained on a weekly labelled dataset (which treats the captions and associated news articles as image labels). Inspired by recent work in summarization, they propose extractive and abstractive surface realization models. Experimental results show that it is viable to generate captions that are pertinent to the specific content of an image and its associated article, while permitting creativity in the description. Indeed, the output of our abstractive model compares favourably to handwritten captions and is often superior to extractive method. In this paper, we explore the feasibility of creating captions, using annotation keywords, for images associated with news documents. The availability of the accompanying news documents in our dataset enables us to formulate the generation module so that it resembles text summarization. We then propose both extractive and abstractive caption generation models. The backbone for both approaches is the probabilistic image annotation model that suggests content for an image given this image and its associated document. We can then simply identify the sentences in the document that share these keywords or create a new caption that is potentially more concise but also informative and fluent. Specifically, for extractive models, we examine how to establish criterions for selecting sentences that are similar to the image content. We also present abstractive caption generation models that operate over image description keywords and document phrases. Their combination gives rise to many caption realizations which we select probabilistically by taking into account dependency and word order constraints. Experimental results show that both approaches generate readable captions with little human involvement. Our abstractive model defined over phrases yields more grammatical output than word-based methods.

RELATED WORK

[1] *Yansong Feng and Mirella Lapata et al.,* In paper we tackle the problem of automatic caption generation for news images. Our approach leverages the vast resource of pictures available on the web and the fact that many of them are captioned. Inspired by recent work in summarization, we propose extractive and abstractive caption generation models. They both operate over the output of a probabilistic image annotation model that pre-processes the pictures and suggests keywords to describe their content. Experimental results show that an abstractive model defined over phrases is superior to extractive methods.

[2] *Girish Kulkarni et al.,* demonstrated that visually descriptive language offers computer vision researchers both information about the world, and information about how people describe the world. The potential benefit from this source is made more significant due to the enormous amount of language data easily available today. We present a system to automatically generate

natural language descriptions from images that exploits both statistics gleaned from parsing large quantities of text data and recognition algorithms from computer vision. The system is very effective at producing relevant sentences for images. It also generates descriptions that are notably more true to the specific image content than previous work.

[3] *Amitkumar Kohakade et al., p*roposed a new approach for generating caption for such images. Approach presented here focuses on important terms occurring in news like named entities, using term weighting find out weighted terms which helps in describing news. On other hand by image processing we find out who's in picture as it helps in making accurate caption by using face recognition and it will increase image retrieval. Some of experiments presented here shows performance of face recognition algorithms on standard datasets and also on own developed face dataset, also we train NER model on Indian names which gives better results. As it covers text and image content it helps in generating better caption and also for improving image retrieval accuracy.

[4] *Dr. M Nagaratna et al.,* propose 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.

[5] *Vini Varghese et al.,* introduced the task of automatic caption generation for news images. The task fuses insights from computer vision and natural language processing and holds promise for various multimedia applications, such as image retrieval, development of tools supporting news media management, and for individuals with visual impairment. It is possible to learn a caption generation model from weakly labelled data without costly manual involvement. Instead of manually creating annotations, 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. 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.

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR

International Journal OF Engineering Sciences & Management Research

[6] *Priyanka Jadhav el at.,* focuses on use of classification techniques using neural network to reduce the data traffic from the node and thereby reduce energy consumption. The sensor data is classified using ART1 Neural Network Model. Wireless sensor network populates distributed nodes. The cooperative routing protocol is designed for communication in a distributed environment. In a distributed environment, the data routing takes place in multiple hops and all the nodes take part in communication. This protocol has been designed for wireless sensor networks. This ensures uniform dissipation of energy for all the nodes in the whole network. Directed diffusion routing protocol is implemented to carry out performance comparison. The paper discusses classification technique using ART1 neural network models. The classified sensor data is communicated over the network using two different cases of routing: cooperative routing and diffusion routing. Ptolemy-II-Visual Sense is used for modelling and simulation of the sensor network. Lifetime improvement of the WSN is compared with and without classification using Cooperative routing and diffusion routing.

EXISITING 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
- 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.

LIMITATION -

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

Figure 2. Output of Google Images for the query "car, blue, sky". Images are ranked based on their relevance to the query.

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. Figure2 shows the output from Google Images1, a popular image retrieval engine. Given a query, search engines usually retrieve relevant pictures by analyzing the image caption (if it exists), textual descriptions found adjacent to the image, and other text-related factors such as the file name of the image and click through data (Weston et al., 2010). However, to our best knowledge, since they do not analyze the actual content of the images, search engines cannot be used to retrieve pictures from unannotated collections. As an example, we submitted the query "car, blue, sky" in the hope of finding pictures describing "a car running under the blue sky". The search engine returned the images shown in Figure 2, ranked based on their relevance to the query. Only three images capture the scene, "a car under the blue sky", while the rest are about blue sky, or just a car. The example illustrates that existing image retrieval engines could benefit from captioned image databases, which would provide a more natural and accurate search experience for end-users, e.g., by supporting longer and more targeted queries and enabling the use of question-answer interfaces. 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

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR
International Journal OF Engineering Sciences & Management Research

publish images and videos with their stories and even provide photo feeds related to current events. Journalists and editors have to manually create captions for these images. The latter must be informative, clearly identify the subject of the pictures, provide context for them, establish their relevance to the news articles, or sometimes establish their relation with previous events. This task is difficult even for humans as it requires both general real-world knowledge and awareness of the specific news events being depicted. An automatic image description generation model can help produce sentences that describe the news image itself or relate the image to current or previous relevant news events. Journalists could then select suitable sentences from this output according to specific requirements.

PROPOSED SYSTEM

In this paper, we tackle the related problem of generating captions for news images. Our approach leverages the vast resource of pictures available on the web and the fact that many of them naturally co-occur with topically related documents and are captioned. We focus on captioned images embedded in news articles, and learn both models of content selection and surface realization from data without requiring expensive manual annotation. At training time, our models learn from images, their captions, and associated documents, while at test time they are given an image and the document it is embedded in and generate a caption. Compared to most work on image description generation, 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.

ADVANTAGES:

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

CRITERIA FOR GOOD CAPTION:

There are several criteria for a good caption.
1. Clearly identifies the subject of the picture, without detailing the obvious.
2. Is succinct.
3. Establishes the picture's relevance to the article.
4. Provides context for the picture.
5. Draws the reader into the article.

Different people read articles in different ways. Some people start at the top and read each word until the end. Others read the first paragraph and scan through for other interesting information, looking especially at pictures and captions. Those readers, even if the information is adjacent in the text, will not find it unless it is in the caption. However, it is best not to tell the whole story in the caption, but use the caption to make the reader curious about the subject.

SYSTEM ARCHITECHTURE

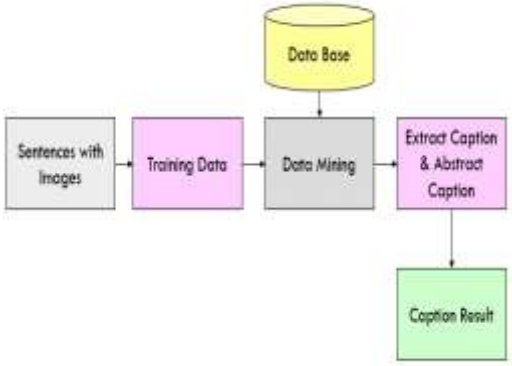
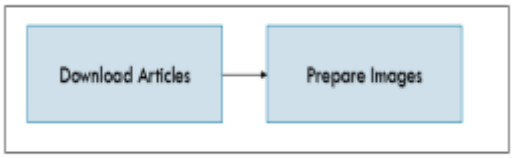


Figure 3 Architecture Diagram

ISSN 2349-6193
Impact Factor (PIF): 2.243

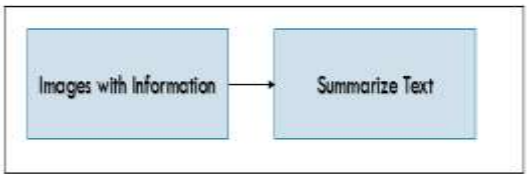
IJESMR
International Journal OF Engineering Sciences & Management Research

DATA COLLECTION



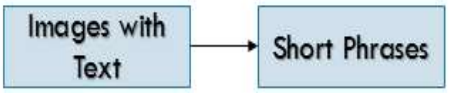
Data collection usually takes place early on in an improvement project, and is often formalised through a data collection plan which often contains the following activity.
1. Pre collection activity — agree on goals, target data, definitions, methods
2. Collection — data collections
3. Present Findings — usually involves some form of sorting analysis and/or presentation.
We created our own dataset by downloading articles from the News websites. The dataset covers a wide range of topics including national and international politics, technology, sports, education, and so on. News 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.

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.

ABSTRACTIVE CAPTION



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.

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.

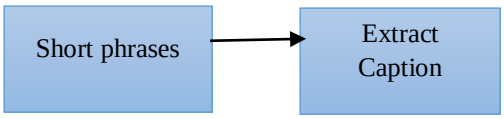
PHRASE-BASED CAPTION GENERATION

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 Wi, Choose the words Wi-1 and Wi-2
2. Form a cluster of Wi, Wi-1 and Wi-2
3. Retrieve the most commonly occurring cluster as a phrase.

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR
International Journal OF Engineering Sciences & Management Research

Extractive Caption



A phrase may refer to any group of words In linguistics, a phrase is a group of words (or sometimes a single word) that form a constituent and so function as a single unit in the syntax of a sentence. A phrase is lower on the grammatical hierarchy than a clause. 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 document. Without a great deal of linguistic analysis, it is possible to create summaries for a wide range of documents, independently of style, text type, and subject matter. For our caption generation task, we need only extract a single sentence. And our guiding hypothesis is that this sentence must be maximally similar to the description keywords generated by the annotation model.

CAPTION GENERATION

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

Figure 4. Input image for generating the image caption

Article: Tremors felt across Indian Subcontinent

An earthquake measuring 7.4 with its epicenter near Kathmandu, Nepal was felt across the Indian subcontinent Tremors of were felt in as far as Chennai. Bihar, West Bengal experienced quakes of magnitude 7.1. Assam was hit by a 7.3 magnitude earthquake. Odisha and Jharkand also felt the massive earthquake. “Earthquake which had its epicentre in Nepal region, was felt here in the city and other parts of the Eastern region. The quake measured 7.1- 7.2 on Richter scale. We are still waiting for more details,” D K Das, a senior official of Kolkata Metrological department said. Hundreds of people ran out of their homes, offices, metro stations and high rise buildings and assembled on the streets. People ran out of Metro stations and high rise buildings. The quake lasted for nearly a minute. No loss of life or property has been reported so far. The quake with epicentre in Nepal was on latitude 27.6 degree north and 86.6 degree east at a depth of 18 km, according to the Central Seismological Observatory.

RESULT

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

An earthquake measuring 7.4 with its epicenter near Kathmandu, Nepal was felt across the Indian subcontinent.

ISSN 2349-6193
Impact Factor (PIF): 2.243

IJESMR

International Journal OF Engineering Sciences & Management Research

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.

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. P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," Proc. Seventh European Conf. Computer Vision, pp. 97-112, 2002.
4. D. Blei, "Probabilistic Models of Text and Images," PhD dissertation, Univ. of Massachusetts, Amherst, Sept. 2004.
5. 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.
6. C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous Image Classification and Annotation," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 1903-1910, 2009.
7. 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.
8. 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.
9. 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.
10. 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, Los Angeles, California. Association for Computational Linguistics.
11. Feng, Y. and Lapata, M. (2010a). How many words is a picture worth? Automatic caption 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.