A Survey on Automatic Image Captioning

Gargi Srivastava(✉) and Rajeev Srivastava

Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi 221005, U.P., India
{gargis.rs.cse16,rajeev.cse}@iitbhu.ac.in

Abstract. Automatic image captioning is the process of providing natural language captions for images automatically. Considering the huge number of images available in recent time, automatic image captioning is very beneficial in managing huge image datasets by providing appropriate captions. It also finds application in content based image retrieval. This field includes other image processing areas such as segmentation, feature extraction, template matching and image classification. It also includes the field of natural language processing. Scene analysis is a prominent step in automatic image captioning which is garnering the attention of many researchers. The better the scene analysis the better is the image understanding which further leads to generate better image captions. The survey presents various techniques used by researchers for scene analysis performed on different image datasets.

Keywords: Image captioning · Scene analysis · Computer vision

1 Introduction

Automatic image captioning is the process of providing natural language captions for images automatically. The area is garnering attention from researchers because of the huge unorganized multimedia data pouring in every second. Automatic image captioning is a step ahead of automatic image tagging where images are tagged with relevant keywords related to the contents in the image. Various researchers have come up with the definition of automatic image captioning. In [1], authors in their work define automatic image captioning as the process by which a computer system automatically assigns metadata in the form of captioning or keywords to a digital image. Mathews et al. [12] in their paper define it as automatically describing the objects, people and scene in an image. Wang et al. [21] in their paper give the definition as recognition of visual objects in an image and the semantic interactions between objects and translate the visual understanding to sensible sentence descriptions. Liu et al. [22] mention that the grammar must be error-free and fluent. For summing up image captioning can be defined as generating short descriptions representing contents (object, scene and their interaction) of an image in human-like language automatically.

D. Ghosh et al. (Eds.): ICMC 2018, CCIS 834, pp. 74–83, 2018.
https://doi.org/10.1007/978-981-13-0023-3_8

Automatic image captioning is viewed as an amalgamation of computer vision and natural language processing. The computer vision part is about recognizing the contents of an image and the natural language processing part is about converting the recognition into sentences. Research has flourished in both the fields. Computer vision researchers try to better understand the image and natural language processing research try to better express the image. Because of this integration, automatic image captioning has come out as an emerging field in artificial intelligence.

1.1 Applications

Automatic image captioning is an interesting area because of its application in various fields. It can be used in image retrieval system to organize and locate images of interest from a database. It is also useful for video retrieval. It can be used for the development of tools that aid visually impaired individuals to access pictorial information. It finds application in query-response interfaces. Journalists find the application useful in finding and captioning images related to their articles. Human-machine interaction systems can also employ the results of automatic image captioning. Such systems are also helpful in locating images verbally. It can also be used for military intelligence generation, surveillance systems, goods annotation in warehouse and self-aware systems (Fig. 1).

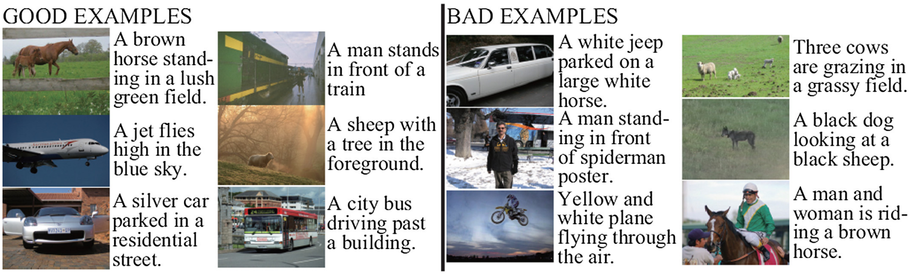
GOOD EXAMPLES
A brown horse standing in a lush green field.
A jet flies high in the blue sky.
A silver car parked in a residential street.
A man stands in front of a train
A sheep with a tree in the foreground.
A city bus driving past a building.
BAD EXAMPLES
A white jeep parked on a large white horse.
A man standing in front of spiderman poster.
Yellow and white plane flying through the air.
Three cows are grazing in a grassy field.
A black dog looking at a black sheep.
A man and woman is riding a brown horse.

Fig. 1. Example of automatically captioned images [3].

1.2 Scene Analysis

Scene analysis is a module in automatic image captioning and has gained importance recently. In image captioning, generally, the output is the main object in the image without caring about what the background of the image is. This negligence makes the description of the image very vague and unclear.

Consider an image where a person is standing and in the background there is a river and another image where the background is a desert. If the focus is only on the object, both the images will be captioned as a person. If the background scene is taken into consideration, the first image may be captioned as a person standing in front of a river and the second image may be captioned as a person in desert. Suppose a journalist wants a sample image for her article and sends a query in the image database as keywords person, river. In first case of image annotation both the images will be retrieved whereas in second case only the first image will be retrieved. Thus, scene analysis is very important for proper image captioning which leads to better image retrieval results (Figs. 2 and 3).

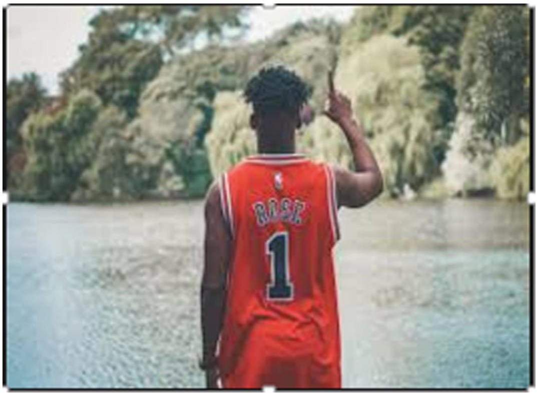

Fig. 2. Without scene analysis: a person. With scene analysis: a person in front of river (Sample image taken from internet)

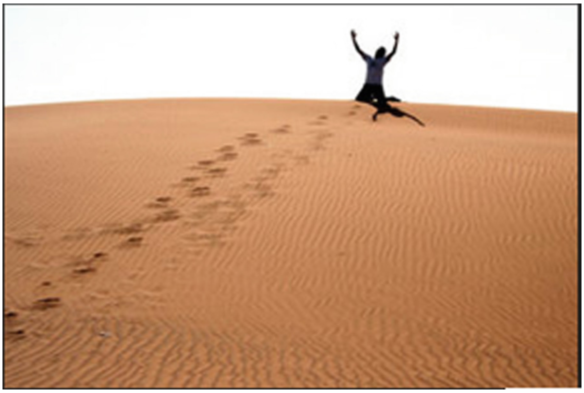

Fig. 3. Without scene analysis: a person. With scene analysis: a person in a desert (Sample image taken from internet)

For scene analysis, the image needs to be broken down into segments for understanding. This leads to the inclusion of another image processing field - image segmentation. Various segmentation techniques exist and several are coming up as to segment the images in a way that the machine understands the image better and can generate better captions. Another field included in scene analysis is object recognition which in itself is a very broad research area.

Object detection can be enhanced by adding contextual information. Scene analysis provides the required contextual information. As the number of scenes is finite, scene analysis is also considered as scene classification problem. Since objects are related to the scenes, the probability distribution of each object over different scenes is different. Convolutional neural networks have been trained over 25 million images of Places dataset to predict approximately 200 different scene-types.

In a nutshell, scene analysis of an image is very important. Without this there is no scope for meaningful captioning of images.

2 Related Works

A lot of research has been done in the field of automatic image captioning. The whole procedure of generating image captions by machines follow a common framework which is discussed below.

2.1 Framework

On the whole, the entire procedure can be subdivided into 2 parts: image processing and language processing. Image processing part includes: image segmentation, feature extraction and classification. Feature extraction and classification can be together referred to as object recognition.

After the object recognition, we obtain the keywords corresponding to the identified objects in the images. These keywords are then fed to language processing unit which results in forming meaningful captions for images.

Each of the three modules are independent and can be researched upon individually. Techniques applied for one of them does not affect the one used for the other module. It is beneficial as each module can be studied and analyzed in isolation (Fig. 4).

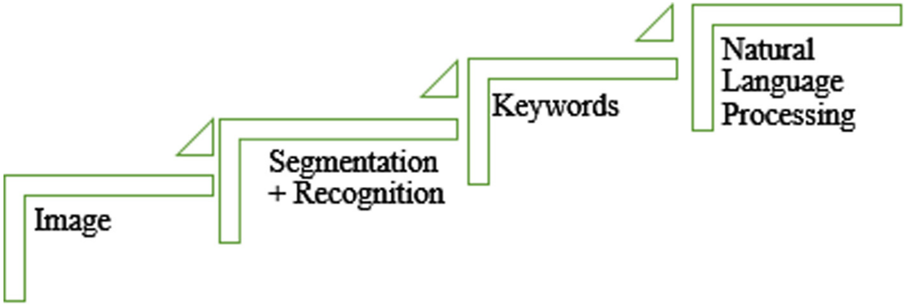

Fig. 4. Steps in automatic image captioning

2.2 Approaches

For segmentation and recognition various techniques can be used: supervised learning, unsupervised learning, neural networks or probabilistic methods (Fig. 5).

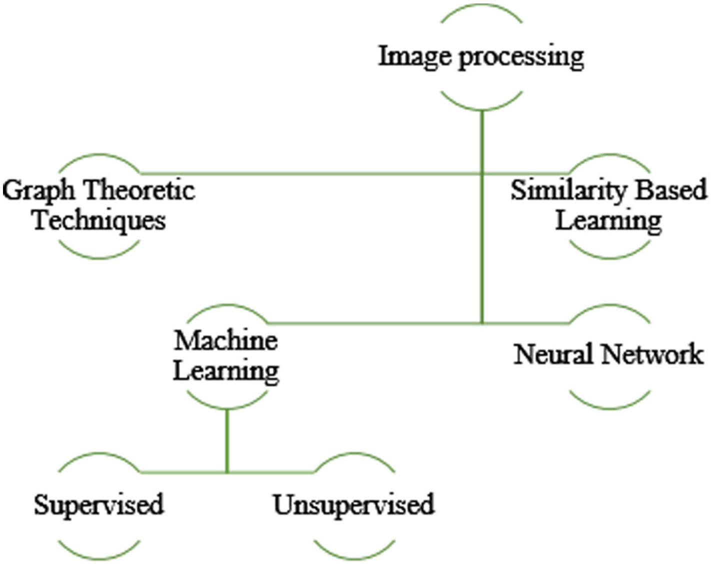

Fig. 5. Various approaches applied for image processing part of automatic image captioning

3 Comparative Study

See Table 1.

Table 1. A comparative study of different works in automatic image captioning

S. No.	Year and Author	Abstract	Result	Datasets	Limitations	Merits
1	Sumathi and Hemalatha [1]	Approach: treat annotation as a retrieval problem. Features: low level image features Simple combination of basic distances using JEC to find the nearest neighbor. Classification: SVM	Precision: 77% Recall: 35% F1 score: 51%	Flickr	Only annotations no captions	Simple distance-based technique

(*continued*)

Table 1. (*continued*)

S. No.	Year and Author	Abstract	Result	Datasets	Limitations	Merits
2	Yu and Sein [2]	Approach: intensity invariant approach. Preprocessing steps: gray scale converting, noise filtering, image enhancing. Segmentation: based on color intensity	No comparative results mentioned	Not mentioned	Only annotation tags are generated	Simple graph technique
3	Ushiku et al. [3]	Approach: generate a sentential caption for the input image by summarizing captions	BLEU: 0.63 NIST: 0.82	PASCAL sentence	Caption accuracy is sensitive to the retrieval precision of training samples	Instead of generating captions from scratch, they have used summarization technique
4	Federico and Furini [4]	Approach: automatic speech recognition with caption alignment mechanism	No comparative results mentioned	Video lectures of university professors	Since the behavior of ASR depends on the acoustic and language models, the audio markup insertion is likely to affect the performance of the speech analyzer	Cost effective solution
5	Feng and Lapata [5]	Probabilistic image annotation model for content selection. Extractive and abstractive surface realization model	Translation edit rate: 1.77. Grammaticality: 6.42. Relevance: 4.10	BBC dataset	1. Finite topics 2. Very little linguistic knowledge 3. Local features	Extremely helpful to journalists
6	Xi and Im Cho [6]	Features: weighted feature clustering based on statistical distribution. Annotation: maximal conditional probability	Precision: 40–60% Recall 40–60%	Corel	Dominated by weakly relevant features	Gives an insight to information gain theory
7	Horiuchi et al. [7]	Approach: collect general phrases for generating image descriptions	17/20 of image descriptions are scored higher than the image descriptions selected by 1 million. 6/7 of image descriptions are scored higher than the image descriptions of Integer Linear Programming	PASCAL visual object classes challenge	1. Image descriptions too concise 2. Image descriptions affected by quality of image retrieval system	Selecting phrases based on frequency results in fewer errors and more relevant descriptions

(*continued*)

Table 1. (*continued*)

S. No.	Year and Author	Abstract	Result	Datasets	Limitations	Merits
8	Ramnath, et al. [8]	Approach: uploading photo to a cloud service and running parallel modules to generate captions	Of 2385 ratings, 49.6% were very good	Personal photos	1. Few recognition capabilities	Keyword based search for personal photos
9	Sivakrishna Reddy, et al. [9]	Features: SIFT. Annotations: clustering	No comparative results mentioned	Not mentioned	Not generating new captions	Reusing existing caption
10	Shivdikar et al. [10]	Approach: combination of feature detection algorithms, context-free grammar	F1 Score: 94.33% BLEU: 0.75	Flickr8K Flickr30K COCO SBU	1. Changing nature of output by single layer learning 2. Smaller n-grams	Forms the base for increasingly complex and accurate neural network algorithms
11	Mathews [11]	Predicting human-like names for visual objects Sentences expressing a strong positive/negative sentiment	Among the 2633 visual concepts the method improves upon the Most frequent name baseline for 1222 concepts	ImageNet	Accuracy not good for ambiguous names	Introducing style is good for generating customized caption
12	Mathews et al. [12]	Predicting basic-level names using a series of classification and ranking tasks	Precision: 34.7%	ImageNet-Flickr	Only picture-to-word, no captions	Naming visual concepts is important part of automatic image captioning
13	Plummer et al. [13]	Coreference chains. Manually annotated bounding boxes	Recall: 76.4%	Flickr30k	1. No attempt to match regions in a query image and phrases in a candidate matching sentence 2. Does not care about the nature of the auxiliary data	Technique helps to localize entities in an image which is helpful for continued progress in image captioning
14	Vijay and Ramya [14]	Create captions for news images	No comparative results mentioned	News articles	No auxiliary information used	Easy understanding of news articles
15	Shahaf et al. [15]	Influence of the language of cartoon captions on the perceived humorousness of the cartoons	The classifier picked funnier of the two captions 64% of the time	Crowdsourced cartoon captions	1. Brief context and anomaly analysis 2. Ignored visual concept of cartoons 3. Cannot identify weaknesses of captions	Opens scope for caption understanding
16	Li et al. [16]	Generate Chinese sentence descriptions for unlabeled images	Machine translated neural image captioning is more suited for Chinese captioning	Flickr8K	Not much improvement observed	Expanding the scope to languages can help build the system for people from different linguistic backgrounds

(*continued*)

Table 1. (*continued*)

S. No.	Year and Author	Abstract	Result	Datasets	Limitations	Merits
17	Jin and Nakayama [17]	Approach: forms image annotation task as a sequence generation problem predict proper length of tags	Precision: 36% Recall: 37% F1: 34% N+: 267	Corel 5K, ESP Game, IAPR TC12	Nave approach to decide order	Order of tags in training phase has a great impact on the final annotation performance
18	Shi and Zou [18]	Fully convolutional networks	Precision: 95.3% Recall: 94.1%	Google Earth, GaoFen-2	1. Different geographical levels are not considered 2. Lesser ground features	Expansion of scope to remote sensing images
19	Shetty et al. [19]	Approach: augment CNN features with scene context features	CIDEr Score: 0.954	MSCOCO	1. Fails to learn relationship between object and image 2. Cannot count objects properly 3. Vocabulary size is small	Employing scene analysis gives better information about the image
20	Li et al. [20]	Scene oriented CNN	BLEU: 0.68 METEOR: 22.8	MSCOCO	Scope for sentiment addition to captions	Including scene information
21	Wang et al. [21]	Deep CNN Approach: use of history and future context information Data augmentation techniques: multi-crop, multi-scale, vertical mirror	BLEU: 0.67 METEOR: 19.5 CIDEr: 66.0	Flickr8K Flickr30K MSCOCO	Less focus on language representation	Focus on language representation to generate better caption
22	Liu et al. [22]	Approach: formulate image captioning as a multimodal translation task, Represent the input image as a sequence of detected objects	No comparative results mentioned	Not mentioned	No detailed methods mentioned	High-level features enrich the visual part
23	Blandfort et al. [23]	Approach: deep convolutional neural network for detecting adjective noun pairs graphical network. Architecture: concept and Syntax Transition Network	31.5% of the captions generated were reported as more human-like in comparison to the original caption. In 62.5% of images atleast one subject chose their caption over the original one	YFCC100M	1. Grammar is given less weightage. 2. Considering concept scores only for thresholding not for ranking 3. Generation of similar sentences 4. Scope for network optimization	Includes sentiment factor
24	Tariq and Foroosh [24]	Approach: extract contextual cues from available sources of different data modalities and transforms them into a probability space	METEOR: 0.053 TER: 1.75	TIME magazine	Only annotations no captions	Importance of weighted auxiliary information

4 Issues and Challenges

A number of open research issues and challenges have been identified in this field. A few of them are listed below:

1. Large collections of digital images exist without annotations.
2. The quality and quantity of training set becomes an important factor in determining the quality of captions that are generated.
3. Images with low resolution, low contrast complex background and texts with multiple orientation, style, color and alignment increase the complexity of image understanding.
4. The training set must include as much variety as possible.
5. Searching the optimal method for each of them is very expensive and it has a major effect on the performance of the overall system.
6. Capturing sentiments in the captions is a major challenge as not many datasets are available that include sentiment based annotations.
7. Few datasets are available that provide captions in different languages and moreover machine translation results are not always relevant.

5 Conclusion and Future Work

Automatic image captioning is an emerging area in the field of artificial intelligence and computer vision. The area has real life applications in various fields. It is an ensemble of various modules which opens a lot of area for exploration. Better captions can be generated with proper segmentation. Enhanced descriptions can be made using sentiment addition, activity recognition, background identification and scene analysis. Moreover the areas of deep learning for faster and accurate results can also be explored further. If the hardware resource cost is a limitation, traditional machine learning algorithms can also be investigated for the purpose.

References

1. Sumathi, T., Hemalatha, M.: A combined hierarchical model for automatic image annotation and retrieval. In: International Conference on Advanced Computing (2011)
2. Yu, M.T., Sein, M.M.: Automatic image captioning system using integration of N-cut and color-based segmentation method. In: Society of Instrument and Control Engineers Annual Conference (2011)
3. Ushiku, Y., Harada, T., Kuniyoshi, Y.: Automatic sentence generation from images. In: ACM Multimedia (2011)
4. Federico, M., Furini, M.: Enhancing learning accessibility through fully automatic captioning. In: International Cross-Disciplinary Conference on Web Accessibility (2011)
5. Feng, Y., Lapata, M.: Automatic caption generation for news images. IEEE Trans. Pattern Anal. Mach. Intell. **35**(4), 797–811 (2013)

6. Xi, S.M., Im Cho, Y.: Image caption automatic generation method based on weighted feature. In: International Conference on Control, Automation and Systems (2013)
7. Horiuchi, S., Moriguchi, H., Shengbo, X., Honiden, S.: Automatic image description by using word-level features. In: International Conference on Internet Multimedia Computing and Service (2013)
8. Ramnath, K., Vanderwende, L., El-Saban, M., Sinha, S.N., Kannan, A., Hassan, N., Galley, M.: AutoCaption: automatic caption generation for personal photos. In: IEEE Winter Conference on Applications of Computer Vision (2014)
9. Sivakrishna Reddy, A., Monolisa, N., Nathiya, M., Anjugam, D.: A combined hierarchical model for automatic image annotation and retrieval. In: International Conference on Innovations in Information Embedded and Communication Systems (2015)
10. Shivdikar, K., Kak, A., Marwah, K.: Automatic image annotation using a hybrid engine. In: IEEE India Conference (2015)
11. Mathews, A.: Captioning images using different styles. In: ACM Multimedia Conference (2015)
12. Mathews, A., Xie, L., He, X.: Choosing basic-level concept names using visual and language context. In: IEEE Winter Conference on Applications of Computer Vision (2015)
13. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: International Conference on Computer Vision (2015)
14. Vijay, K., Ramya, D.: Generation of caption selection for news images using stemming algorithm. In: International Conference on Computation of Power, Energy, Information and Communication (2015)
15. Shahaf, D., Horvitz, E., Mankoff, R.: Inside jokes: identifying humorous cartoon captions. In: International Conference on Knowledge Discovery and Data Mining (2015)
16. Li, X., Lan, W., Dong, J., Liu, H.: Adding Chinese captions to images. In: International Conference in Multimedia Retrieval (2016)
17. Jin, J., Nakayama, H.: Annotation order matters: recurrent image annotator for arbitrary length image tagging. In: International Conference on Pattern Recognition (2016)
18. Shi, Z., Zou, Z.: Can a machine generate humanlike language descriptions for a remote sensing image? IEEE Trans. Geosci. Remote Sens. **55**(6), 3623–3634 (2016)
19. Shetty, R., Tavakoli, H.R., Laaksonen, J.: Exploiting scene context for image captioning. In: Vision and Language Integration Meets Multimedia Fusion (2016)
20. Li, X., Song, X., Herranz, L., Zhu, Y., Jiang, S.: Image captioning with both object and scene information. In: ACM Multimedia (2016)
21. Wang, C., Yang, H., Bartz, C., Meinel, C.: Image captioning with deep bidirectional LSTMs. In: ACM Multimedia (2016)
22. Liu, C., Wang, C., Sun, F., Rui, Y.: Image2Text: a multimodal caption generator. In: ACM Multimedia (2016)
23. Blandfort, P., Karayil, T., Borth, D., Dengel, A.: Introducing concept and syntax transition networks for image captioning. In: International Conference on Multimedia Retrieval (2016)
24. Tariq, A., Foroosh, H.: A context-driven extractive framework for generating realistic image descriptions. IEEE Trans. Image Process. **26**(2), 619–631 (2017)