ISSN 2347 - 3983

Volume 9. No. 8, August 2021

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter15982021.pdf

https://doi.org/10.30534/ijeter/2021/15982021

# Overview on Image Captioning Techniques

**Urja Bahety[1], Surendra Gupta[2]**

[1] Research Scholar, Department of Computer Engineering, Shri Govindram Seksaria Institute of Technology and Science, Indore, India, urjabahetyind@gmail.com

[2] Professor, Department of Computer Engineering, Shri Govindram Seksaria Institute of Technology and Science, Indore, India, sgupta@sgsits.ac.in

## ABSTRACT

Image captioning is a process to assign a meaningful title for a given image with the help of Natural Language Processing (NLP) and Computer Vision techniques. Captioning of an image first need to identify object, attribute and relationship among these in image and second is to generate relevant description for the given image. So it require both NLP and Computer vision techniques to perform image captioning task. Due to complexity of finding relationship between the attribute of the object and its feature makes it a challenging task. Also for machine it is difficult to emulate human brain however researches have shown a prominent achievement in this field and made it easy to solve such problems. The foremost aim of this survey paper is to describe several methods to achieve the same, the core involvement of this paper is to categorise different existing approaches for image captioning, further discussed their subcategories of this method and classify them, also discussed some of their strength and limitations. This survey paper gives theoretical analysis of image captioning methods and defines some earlier and newly approach for image captioning. This survey paper is basically a source of information for researchers in order to get idea of different approaches that were developed so far in the field of image captioning.

**Key words :** Computer Vision, Deep Learning, Neural Network, NLP, Image Captioning, Multimodal Learning.

## 1. INTRODUCTION

Human beings are competent enough to certainly illustrate the surrounding that we are in. At a quick glance to an image it is very easy and is usual ability for human to describe the details when we have given an image. In everyday life we encounter so many sources of images like news, internet, social media, advertisement and many more existing sources and viewers have to interpret the context behind the images and most of them do not have description with it but human have capability to understand them without any need of description but for machine it is difficult to emulate human and machine needs description to get it. In an arena of deep learning and artificial intelligence to make machine emulate human is a prolonged aim for researchers.

Though excessive progress have made in several tasks of computer vision like attribute classification [1], [2], object recognition [3], [4], image classification [5] etc., but to let machine to do a task of generation of human like caption when an image is given to it, is a fairly new task and it require both computer vision and natural language processing researches.

It is well known quest problem to caption the natural scene. There are various application areas of image captioning, as for human whether it is spoken or written, our communication is depends on natural language and define immense amount of details of our surroundings, allowing machine to depict visual world can lead immense number of applications in this field like in information retrieval, education for child [6], web searching [7], for visually disabled people [8], etc.

Many social media site, google apps also using it to locate you, to analyse what you are doing, etc. also twitter and facebook generate description like what are we wear, what we doing there, where are we etc. To understand image it needs to identify the objects, their actions and relationship among them. For machine it is no difficult task to identify the object in an image but difficult to gather and identify the features or actions and their relationship in the image. Generated description must be truthful in terms of syntactic and semantic manner. To generate good caption of an image must require understanding of natural scene in an image by obtaining salient features from the image.

In the arena of Artificial Intelligence (AI), captioning the image is very prevalent study field which deals with understanding of image and understanding language description for the image and needs to recognise object their properties and interactions and also understand location or scene type in order to produce best captions.

The chief goal of captioning of image is to produce description of an image when image is given i.e. the description which is semantically and linguistically ingenuous to the content of image. There are two main question are arrived first is linguistic processing and second is visual understanding of an image.

To make sure that generated description of an image are semantically and syntactically correct some techniques of NLP and computer vision are adopted to deal the problem arrived from modality and integrity appropriately. Several approaches have been proposed to create description of an image correctly.

This survey paper presents theoretically analysis of several different methods some of which was used earlier and some latest methods for image captioning. Below represent the overall categorization of image captioning method in structure form and detailed description of these methods describe in succeeding sections.

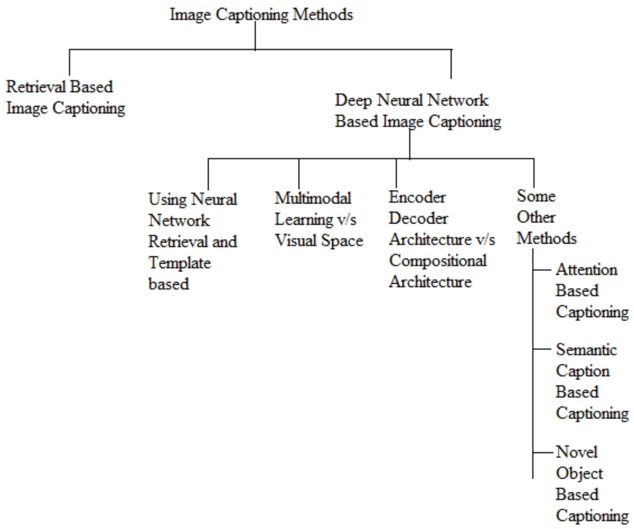
Image Captioning Methods
Retrieval Based Image Captioning
Deep Neural Network Based Image Captioning
Using Neural Network Retrieval and Template based
Multimodal Learning v/s Visual Space
Encoder Decoder Architecture v/s Compositional Architecture
Some Other Methods
Attention Based Captioning
Semantic Caption Based Captioning
Novel Object Based Captioning

**Figure 1:** Structure of Image Captioning Methods

## 2. IMAGE CAPTIONING

The foremost aim of captioning of image is to generate proper caption which is relevant to given image. Image captioning is deals with understanding of given image and description for an image, needs to identify object, attributes, properties and relationship among these and require both NLP and computer vision techniques to give good description of a given image. When an image is given as input to the image captioning model, model should give relevant description to a given image. Below is the image captioning diagram which shows the caption generation when an image is given as input to the model.

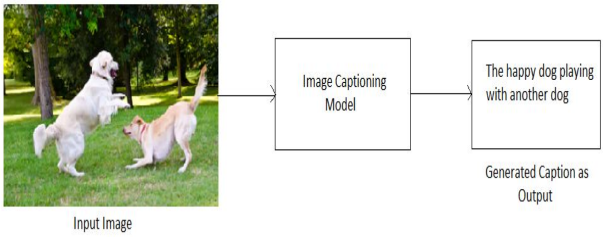
Input Image
Image Captioning Model
The happy dog playing with another dog
Generated Caption as Output

**Figure 2:** Block Diagram of Image Captioning

Image captioning has enthralled more study in the area of AI and has played remarkable role in computer vision and language processing. Captioning of image described objects attribute and their relationships. In order to understand image captioning and to produce captions for image this survey paper describes various methods to perform captioning task. This survey provides the summarised understanding of method which has been achieved earlier and the latest methods which has been achieved so far. The methods to caption an image which have surveyed are retrieval based, deep neural network based along with their types. Detailed discussion of these methods is discussed in remaining section of the paper.

Remaining paper organisation is as - Segment 3 outlined Retrieval based image captioning technique segment 4 outlined some of the deep neural network based techniques and finally discussed concluding remarks.

## 3. RETRIEVAL BASED IMAGE CAPTIONING METHOD

Earlier the kind of image captioning method which was very ordinary was retrieval based image captioning. In this technique given a query image, descriptions retrieved from set of existing descriptions or from set of predefined caption pool, from training dataset it firstly find visually alike images with their captions, the caption that was generated may be either a caption that has existed already or a caption collected from retrieved ones. Now look over direction of research which openly used retrieved description as caption of image.

A. Farhadi et al. [9] established a (object, action, scene) meaning space to associate sentences and images. "A query image has been given and they map it into the meaning space by solving Markov Random field and semantic distance between these images are determined by Lin Similarity measure [10] and each existing sentence parsed by Curran and Clark Parser [11]. The caption which is closest to the given image is considered as the caption of the query image" [12].

In order to caption an image V. Ordonez et al. [13] firstly used global image descriptors to retrieve set of images from web scale collection of captioned images. After that they use semantic content of retrieve images to do re ranking and use caption of top image as the description of the query.

Image captioning in accordance to M. Hodosh et al. [14], frames captioning for image as a task of ranking. The author gives the analysis of Kernal Canonical Correlation method [15], [16] to project text and image to common space, where training images and their captions are correlated greatly. To choose top ranked captions to act as a description of the query images, cosine similarity amongst captions and images are calculated in new common space.

To lighten the effect of estimation of noise visual techniques that are based on retrieval for image captioning, Charniak and Mason [17] at first for query image use visual similarity to retrieve set of captioned images. From the captions of retrieved images they calculate a word probability density conditioned on querying image. To retrieve the largest score as caption of the query the word probability density method is used to rank recent caption. The prior principal determined that given a query image always there exist a sentence which is relevant to it. This hypothesis is barely accurate in practise so in place of using selected sentences as description of query image directly, in other line of retrieval based research, retrieved sentences are used for the composition of new description for image query.

Dataset is provided with paired imaged and captions Stanford coreNLP toolkit is used by Y. Verma et al. [18] to process sentences in dataset to derive a sentence list for every single image. First image retrieval is achieved based upon global features of an image to retrieve a set of images for the query in order to produce a description for a query image. A model is trained to predicate phrase relevance is used to select phrases from the ones related with retrieved images subsequently. Lastly a description is produced which is based upon selected relevant phrases.

There are certain limitations also of this method. However this method creates general and syntactically correct captions for image i.e. captions generated by this method are grammatically correct and simple but this method is unable to create image specific captions and the captions which are semantically correct i.e. in some conditions generated captions may be unrelated to the content of the image.

## 4. DEEP NEURAL NETWORKS BASED IMAGE CAPTIONING

Due to excessive evolution made in arena of deep learning, work start to relay on deep neural networks to captioning of image. Deep neural network is now taken on broadly to handle image captioning task. Categorisation is done on deep neural network based approaches to subclasses on the basis of key structures & framework.

### 4.1 Using Neural Networks Retrieval and Template Based Method

Increasing the progresses in the arena of deep neural network we can perform captioning task with ease. The problem of ranking and embedding can be resolved with retrieval based method, to make image captioning as multimodality for use in deep modelling recommended by researchers. “To retrieve a description for a query image Socher et al. [19] proposed to use dependency tree recursive neural network to represent sentences or phrases as compositional vectors. Another deep neural network model is used as visual model to extract the features from the images. Obtained multimodal features are mapped to a common space using max margin objective function. Correct sentence and image pair in common space has larger inner product after the training” [20]. At the end sentence retrieval is done which is based upon similarities between representation of sentences and images in common space.

Significantly by using deep neural network the execution of image captioning is enhanced. However using deep neural network in template and retrieval based methods does not overthrow their weaknesses so limits of caption produced by these techniques are not taken out completely.

### 4.2 Captioning of Image Based on Multimodal Learning v/s Visual Space

Captioning based on deep neural network can create captions from visual and multimodal space. Dataset of image captioning have captions in text form for corresponding image. In multimodal space methods a shared multimodal space is learn from images and matching captions, then this multimodal depiction is pass to language decoder. In contrast in a visual space based method the matching captions and image feature are passed to language decoder independently. Several image captioning technique use visual space to generate captions for image. This approach is discussed briefly in section 4.4.

#### 4.2.1 Multimodal Space

Multimodal space method architecture containing vision part, language encoder, language decoder and multimodal space part. To extract image feature vision part use deep Convolutional Neural Network (CNN) as extractor of feature and language decoder extracts word features and learn dense feature embedding for every word then it passed to recurrent layer. Multimodal space maps features of image with word features to a common space, then resultant feature map are handed to a language decoder that produce captions by decoding the feature map.

Initially work is done in this field by Kiros et al. [21] the technique uses a CNN to take out the feature of image to generate the captions and uses multimodal space that represent text and image both together for multimodal learning and generation of caption, also it propose multimodal neural network model that was conditioned on image input to create images captions. In their approach log bilinear language model is adapted to multimodal space. This method depends on high level image feature and representation of word learned from multimodal neural language modal and deep neural network.

“A Karpathy et al. [22] proposed deep, multimodal model, embedding of natural language and image data for bidirectional image and caption embedding task previously mention multimodal based approach use common embedding space that maps image and caption directly, however this method work at finer level and embed fragments of captions and fragments of images this method breaks down images into number of objects and captions into dependency tree relation. This approach achieves some improvement in retrieval task as compared to other methods. This approach has some drawback also like in term of modelling the dependency tree can model relation easily but they are not always correct” [23].

A multimodal recurrent neural network (MRNN) language model is used by J. Mao et al. [24] to create novel captions for image. There are two subnetwork in this method deep Recurrent Neural Network (RNN) used for captions & deep CNN used for images both of these networks interacts with one another in multimodal space to form MRNN model. In this method both fragments of captions and images are given as input and it computes probability distribution in order to produce subsequent word of captions, in this model five more layers are there- a recurrent layer, two layers of word embedding, a multimodal layer and a softmax layer. Kiros [21] proposed a technique that is based on log bilinear model and alexnet for feature extraction. This MRNN approach is some interrelated to method of kiros [21] which uses fixed length context where as in "this method temporal context is stored in recurrent architecture which allows variable context length. Two word embedding layer use vector in order to generate a dense word representation. It encodes both semantic and syntactic meaning of words. Semantically relevancy of word can be get by calculate Euclidian distance between the two dense word vectors in embedding layer. Most of the image sentence based multimodal methods use pre computed word embedding vectors in order to initialise their model. On the other hand this method randomly initialises word embedding layer and learn from training dataset which helps to generate good caption for image" [23]. Below figure shows multimodal based image captioning-

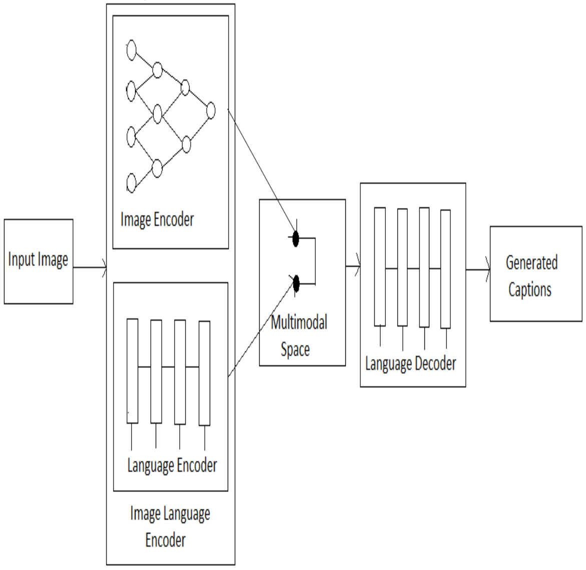

**Figure.4.2.1:** Multimodal Based Image Captioning

## 4.3 Captioning of Image Based on Encoder- Decoder Architecture v/s Compositional Architecture

### 4.3.1. Captioning Based on Encoder Decoder Architecture

In this image features are extracts from activation of convolutional neural network and to create sequence of words it then fed to Long Short Term Memory (LSTM). Basically CNN is used get scene type i.e. to notice objects and relationship between them and output of this are then used by language model to transform them into words and combined phrases that create caption of image.

Vinyals et al. [25] describes "a method called neural image caption generator, this method uses LSTM for caption generation and CNN for image representation. This CNN uses novel method for batch normalisation and output of its last layer is used as input to the LSTM decoder. LSTM is able to keep track of objects that already described using text" [23] and neural image caption generator trained on maximum likelihood estimation.

To generate a caption for image, image statistics is included to initial state of LSTM and the succeeding word which would be generated is based on preceding hidden state and current time step. Till it catches the end of the token of sentence this procedure is continues. Below figure shows encoder decoder architecture-

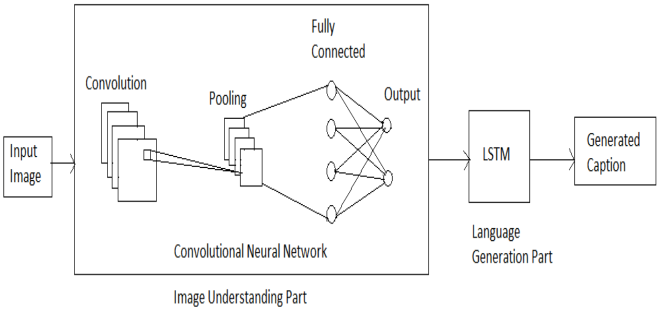

**Figure 4.3.1:** Encoder-Decoder Architecture based Image Captioning

### 4.3.2 Image Captioning Based on Compositional Architecture

This method is composition of various separate functional building blocks. Convolutional neural network is used to get the semantic meaning from the image first afterword language model is used to create a set of candidate captions and these candidate captions are re ranked with the help of deep multimodal similarity model in order to generate final caption. Below figure shows compositional architecture based captioning-

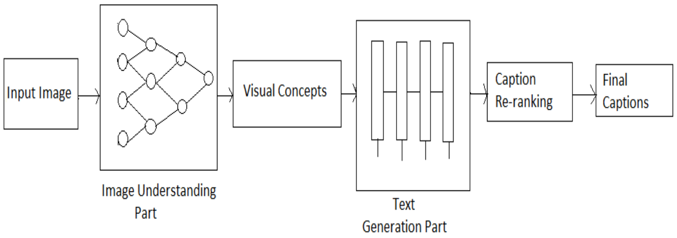

**Figure 4.3.2:** Compositional Network Based Image Captioning

All the process of this methods follows steps like first image features are obtained using convolutional neural network then attributes are gotten from visual features, multiple captions are created by language model in order to follow above two steps, at last captions that are generated is re-ranked using a model called deep multimodal similarity model to select great excellence captions of image.

### 4.4 Some Other Methods

These are some categories of methods which are independent of other methods like semantic concept based, novel object based, and attention based captioning are all assemble in this category.

### 4.4.1 Attention Based Captioning

Basically in captioning CNN is used to get features from image and RNN is used as decoder to transform this representation word by word. In that method they do not deliberate spatial aspects of image that is pertinent to the part of captions and captions are generated by considering a scene as a whole. This method address this limits because it focus upon several part of input image dynamically whereas output sequences are being generated. Firstly use CNN so that information of image is acquired that is derived from whole scene, based on output of above step language generation phase generate words, after this main areas of given image are concentrated on every time step of language generation model which is based upon words that are generated, and till the end state of language generation model dynamically the captions are reorganised.

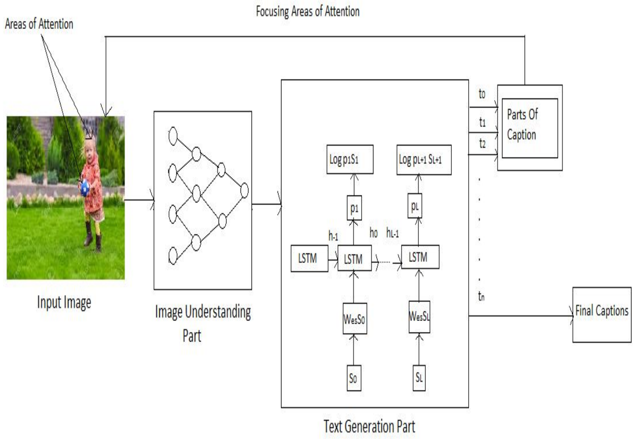

**Figure 4.4.1:** Attention Based Image Captioning

### 4.4.2 Semantic Caption Based Captioning-

“This method selectively attend to a set of semantic concept proposal extracted from the image, these concepts are then combined to a hidden states and output of RNN. In this method first CNN based encoder is used to encode semantic concepts image feature. After this image feature are given as input to language generation model then semantic concepts are added to different hidden states of language generation model. At last language generation model will generate captions with semantic concepts” [23]. Below figure shows semantic caption based captioning-

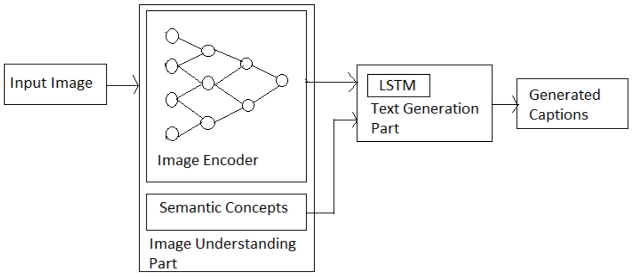

**Figure 4.4.2:** Semantic Caption Based Image Captioning

### 4.4.3 Novel Object Based Captioning

“Deep learning based captioning methods have achieved good results and they highly depend on paired image and sentence caption dataset. These types of method only generate description of object within the context therefore methods need large set of training dataset consisting of image-text pairs. Novel object based captioning method generate description of novel object that are not there in paired image caption datasets” [23]. In this method a language model and a separate lexical classifier are trained upon unpaired text and data of image. On paired image caption data a deep captioned model is trained, at last both the models are joined together to train combinely to produce caption for novel object. Below figure shows novel object based captioning-

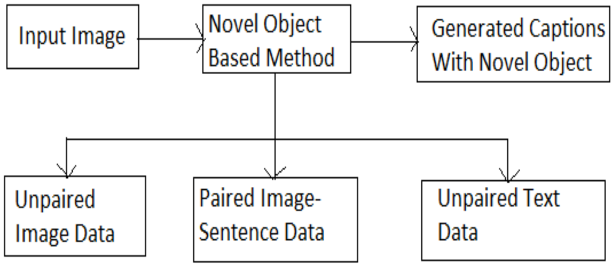
Input Image
Novel Object Based Method
Generated Captions With Novel Object
Unpaired Image Data
Paired Image-Sentence Data
Unpaired Text Data

**Figure 4.4.3:** Novel Object Based Image Captioning

## 5. CONCLUSION

The proposed survey paper describes the captioning for image models and techniques adopted in each methods. The survey paper has also described some strength and limitations of discussed approaches. First this paper describes primary captioning for image workings which are template and retrieval based also it described that retrieval methods are good in generating simple and grammatically correct captions but also has some limitation i.e. unable to generate semantically correct caption. Afterword the key attention is dedicated on the approaches which are based on deep neural network which gives prominent results in generating good captions, then further divide them in subcategories since distinct agendas are used in deep neural network based techniques and some other methods related to image captioning.

## REFERENCES

1. T. Yang, C. Gan, B. Gong, Learning attributes equals multisource domain generalization, in: IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2016, pp. 87–97.
2. H. Nickisch, C. H. Lampert, S. Harmeling, Learning to detect unseen object classes by between class attribute transfer, in: IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009, pp. 951–958.
3. P. F. Felzenszwalb, R. B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) 2010, pp. 1627–1645.
4. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 2014, pp. 580-587.
5. P. Lakshmi Prasanna, D. Raghava Lavanya, T. Sasidhar, B. Sekhar Babu, Image Classification Using Convolutional Neural Network, in: International Journal of Emerging Trends in Engineering Research, October 2020, pp. 6816-6820.
6. S.Vimala, P.Haritha, S.Malathi, A Systematic literature review on story telling for kids using image captioning-deep learning, in: 4th International Conference on Electronics, Communication and Aerospace Technology(ICECA), 2020.
7. Shubham Chaturvedi, S. Iyer, Tirthraj Dash, Image captioning based image search engine: An alternative to retrieval by metadata: SocProS 2017, pp. 181-191.
8. B. Makav, V. Kilic, A new image captioning approach for visually impaired people, in: 11th International Conference on Electrical and Electronics Engineering(ELECO), 2019.
9. A.Farhadi et al., Every picture tells a story: Generating sentences from images, in: European Conference on Computer Vision, 2010, pp. 15–29.
10. D. Lin, An information-theoretic definition of similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296–304.
11. J. Curran, S. Clark, J. Bos, Linguistically motivated large-scale nlp with cc and boxer, in: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, pp. 33–36.
12. Himanshu Sharma, Manmohan Agrahari, Mohd Firoj, Image Captioning: A Comprehensive Survey, in: International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control (PARC), 2020, pp. 325-328.
13. V. Ordonez, G. Kulkarni, T. L. Berg., Img2text: Describing images using 1 million captioned photographs, in: Advances in Neural Information Processing Systems, 2011, pp. 1143–1151.
14. M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: data, models and evaluation metrics, Journal of Artificial Intelligence Research 47 2013, pp. 853–899.
15. F. R. Bach, M. I. Jordan, Kernel independent component analysis, Journal of Machine Learning Research 3 2002, pp. 1–48.
16. D. R. Hardoon, S. R. Szedmak, J. R. Shawe-Taylor, Canonical correlation analysis: An over view with application to learning methods, Neural Computation 16 2004, pp. 2639–2664.
17. R. Mason, E. Charniak, Nonparametric method for data driven image captioning, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014.
18. A. Gupta, Y. Verma, C. V. Jawahar., Choosing linguistics over vision to describe images, in: AAAI Conference on Artificial Intelligence,Vol.5, 2012.
19. R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, A. Y. Ng, Grounded compositional semantics for finding and describing images with sentences, TACL 2 2014, pp. 207–218.
20. Shuang Bai , Shan An, A Survey On Automatic Image Caption Generation, Article in Neurocomputing, May 2018, pp. 291-304.
21. Kiros R, Salakhutdinov R, Zemel R, Multimodal neural language models. In: International conference on machine learning, 2014, pp. 595–603.
22. A Karpathy, A Joulin, and FeiFeiFLi. Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems, 2014, pp. 1889–1897.
23. Md. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput. Surv. Vol. 51, 2019, pp. 1-36.
24. Junhua Mao et al., Deep captioning with multimodal recurrent neural networks (m-rnn). In International Conference on Learning Representations (ICLR), 2015.
25. O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164.