2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC)
GLA University, Mathura, UP, India. Feb 28-29, 2020

Image Captioning: A Comprehensive Survey

Himanshu Sharma
Department of Computer Engineering and Applications
GLA University Mathura
himanshu.sharma@gla.ac.in,

Manmohan Agrahari
Department of Computer Engineering and Applications
GLA University Mathura
manmohan.agrahari_cs16@gla.ac.in

Sujeet Kumar Singh
Department of Computer Engineering and Applications
GLA University Mathura
sujeet.singh_cs16@gla.ac.in,

Mohd Firoj
Department of Computer Engineering and Applications
GLA University Mathura
mohd.firoj_ec16@gla.ac.in

Ravi Kumar Mishra
Department of Computer Engineering and Applications
GLA University Mathura
ravi.mishra_cs16@gla.ac.in

2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) 978-1-7281-6575-2/20/$31.00 ©2020 IEEE 10.1109/PARC49193.2020.236619

***Abstract*: The primary purpose of image captioning is to generate a caption for an image. Image captioning needs to identify objects in image, actions, their relationship and some silent feature that may be missing in the image. After identification the next step is to generate a most relevant and brief description for the image that must be syntactically and semantically correct. It uses both computer vision concepts for identification of objects and natural language processing methods for description. It's difficult for a machine to imitate human brain ability however researches in this field have shown a great achievement. Deep learning techniques are enough capable to handle such problems using CNN and LSTM. It can be used in many intelligent control systems and IOT based devices. In this survey paper, we are presenting different approaches of image captioning such as retrieval based, template based and deep learning based as well as different evaluation techniques.**

Keywords: Computer Vision; Natural language Processing; IOT; Control System

I. INTRODUCTION

There are so many sources of images such as television, internet, news and many more existing sources whereas most of images do not contain description however human are capable enough to interpret themselves without having a description whereas this is very tough for machine to interpret. Machines need description to understand.

In the era of artificial intelligence, captioning of a natural scene is a well-known research problem that helps to give a description of an image. This research is crucial for various reasons even big wings like Facebook, Google they are using it, for locating where are you, what are you doing and many more such activities. Image understanding need to identity objects, action and relationships. This is not more difficult to identity direct objects in image however this is tough for machine to identity some silent features like (people are waiting for train, however train is not on platform). Generated sentence must be syntax wise and semantic wise correct. A good understanding of an natural scene is based on the obtained features of a given natural scene. The technique used for this is extensively separated into two sets (1) Conventional Machine Learning Based Methods (2) Deep-Learning Based Methods [1, 2, 3, 4].

CNN-RNN framework based image captioning technique have two drawbacks in training phase. First drawback is each caption gets equal importance without their individual importance and second drawback is during caption generation objects may not be correctly recognized. Encoder-Decoder framework, in this paper we have also suggested called Reference based long short term memory (R-LSTM) that main aim is by implementing reference information to give more descriptive caption for a query image. According to relation between image and words during the training phase, different weights are assigned. In addition to maximizing the agreement-score among the captions produced through the captioning methods and the reference data from the adjoining images of the intentional images that can limit the issue of not recognize correctly an image.

Fig. 1. Generated Caption can be "A soft-drink Company Pepsi is sponsoring a cricket match"

The crucial benefit of huge labeled datasets, like ImageNet and deep learning, broadly speaking deep convolutional neural networks (CNN) is very useful. For an image which automatically generates a sentence description, has enthralled more research focus in artificial intelligence is called image captioning, it has played a significant position in computer vision, i.e., allowing computer systems to recognize images, that can be helpful in various purposes, together with childhood education, video tracking, sentiment evaluation and visible impairment rehabilitation. The generator must be able to locate their states, apprehend relations amongst them, and deduce the semantic data in natural language. Image captioning efforts particularly undertake the template-based methods, that requires describing the diverse elements for example direct or indirect objects in addition to their relationships and attributes.

These techniques are mainly based on the encoder-decoder pipeline that includes two simple steps. Firstly, with the assist of CNN Image features are deduced to encode the image into a hard and fast period embedding vectors. Secondly, generating a language description usually a recurrent neural network is used as decoder. A sincere thanks to the characteristics representation competence of CNN and

the temporal modeling of RNN, the neural network-based methods are more adoptable that can deduce new sentences.

The remaining paper is outlined as follows: In Sec. 2 we give an overview of retrieval based image captioning methods. In Sec.3 gives an overview of deep neural network based image captioning methods. In Sec. 4, different image caption evaluation metrics are discussed. In Section 5, results of image captioning methods on Flickr8k [6] and Flickr30k [7] benchmark datasets are presented. Finally, we discuss some promising directions and our concluding remarks.The remaining paper is outlined as follows: In Sec. 2 we give an overview of retrieval based image captioning methods. In Sec.3 gives an overview of deep neural network based image captioning methods. In Sec. 4, different image caption evaluation metrics are discussed. In Section 5, results of image captioning methods on Flickr8k [6] and Flickr30k [7] benchmark datasets are presented. Finally, we discuss some promising directions and our concluding remarks.

II. Retrieval-Based Image Captioning

Earlier, Retrieval based captioning was a very common approach. In this approach caption for a query image is retrieved from an existing caption pool. Generated caption may be directly retrieved one from caption pool or caption may be composed one. Retrieval based image captioning, first identify visually close image to the query image from the training dataset .There are some researches which consider caption of image is decided by directly retrieved caption from caption pools .A query image have been given, they plot it into the meaning space by solving a Markov Random Field, and the semantic distance between these images is deduced by Lin similarity measure [8] and each existing sentence is parsed with the help of Curran and Clark parser [9]. The caption which is closest to the given image will be considered as a caption of the query image. Ordonez et al. [10] firstly used global image computing to extract a group of images from a web-scale; basically web-scale is combination of captioned images.

In image captioning, according to Hodosh et al. [11] captioning of an image he frames as a task of ranking. The instigator gives the Analysis of Kernel Canonical Correlation method [12, 13] that correlates maximum training images and their captions to get down image and texting of material into a familiar area. In the novel general area, similarities of cosines between images and phrases are computed for taking out top positioned phrases to describe the query images. To minimize the effects of estimation of noise visual techniques that are based on retrieval of an image for image captioning, Charniak and Mason firstly use visually likely which is used to select a group of images captioned for an image query. From the snaps of the retrieved images, they calculate a probability of density of a word based on the querying of the image. The probability of density of word based technique is used to rank the current captions to retrieve the largest score of the query caption. The former principals have determined that given a query image always exist a phrase that is permissible to it. This evaluation is perfectly true. So, without using selected sentences as descriptions of images directly, similarly, sentences which are retrieved are useful for the composition of image query. Stanford Core NLP is used by Gupta et al toolkit to measure phrase in the dataset to deduce a sentence list for every picture. For a query picture, to deduce a narrative, generation of an image is processed based on global characteristics of an image to select a group for image query. To generate phrases combined with retrieved images, a model is taught to predicate phrase significant is used. Therefore, a narrative sentence is generally based on the chosen important sentences.

This retrieval base image captioning is not completely accurate. There are certain limitations of this method. For a given query image, such methods transfer well-formed phrases written by human. However, the outcomes are grammatically correct; conditioning description of the image to sentences that already exists cannot associate to new object mixtures. Resulted captions might not be important to image scene, under certain conditions. There are many disadvantages to their capability to deduce images for these methods.

III. Captioning Methods based on Deep Neural Networks

Deep neural networks are significantly implemented for handling the captioning of an image. Different techniques are based on different structures. Deep neural network based captioning methods are categorized based on the main structures.

A. Retrieval and template based Methods utilizing Neural Networks

Using deep neural networks we can also do image captioning. The problems of embedding and ranking can be solved with the retrieval based methods, to make captioning of the image as multi-modality for use in the deep modeling suggested by researchers. For retrieval of description, the phrases or sentences as compositional vectors are represented by dependency tree recursive neural network proposed by Socher et al. [14]. A max-margin is used to map the acquired multimodal feature into a common space. By using deep neural networks, consequently performance captioning of image techniques is enhanced due to advancement and adaptation of new models. Disadvantages of sentences composed by methods of retrieval and template based did not extirpate.

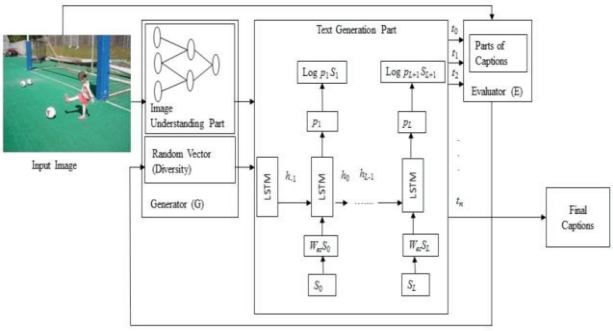

Fig. 2. Deep Learning-Based Image Captioning Methods

B. Multimodal learning based Image Captioning

The adaptation of Image captioning techniques such as retrieval based as well as template base faced few limitations during sentence generation phase. However deep neural networks approaches used in image captioning, they never believe in pre-existing captions and assumptions about structure of sentence during the caption generation phase. These techniques can produce better, flexible and expressive

sentences with better structure. To generate caption for an image the multi-model neural networks believe in pure learning. Common configuration of multimodal learning based image captioning methods is shown in Fig.

Neural language model is suggested by Kiros et al. [15] for an image caption. In their approaches, a log-bilinear language is used. However, for an image to generate a caption, in multimodal cases Mao et al. [16] used a RNN language model for directly modeling the probability of getting a word that is termed on a given image and earlier produced words. In their techniques, a deep convolutional network is utilized for extracting features of image and RNN along with a multimodal part is used to model word distribution termed on image features and context words. RNN language model mainly consists of an output layer, input layer as well as a recurrent layer.

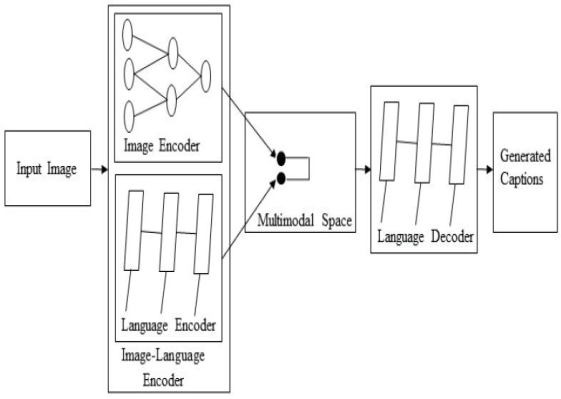

Fig. 3. Multimodal-Based Image Captioning Methods

C. Encoder-Decoder Framework based Image Captioning

The image In order to caption natural scenes, image captioning research Kiros et al. [15] implemented this framework which is called encoder and decoder based to combined use image-text embedding model and multi-modal sentence generation models, for a query image, a description which is an output produced word by word similar to language translation. For encoding textual data a special RNN they used called Long Short Term Memory (LSTM) and for encoding visual data a deep Convolutional neural network used. Then, by using optimizing a pair-wise ranking loss, the visual content is encoded then is fed into an embedding space extended by Long Short Term Memory hidden states which encode textual data. In the embedding space, for decoding image features conditioned on feature vector of background word a structure content neural language is utilized that permit to generate sentence word after word. By the exact motivation from neural machine translation, Vinyals et al. [17], first for encoding image as an encoder adopted deep CNN [18] and next for decoding purpose as a decoder he used LSTM [19] in RNN that will help in generating description from image features.

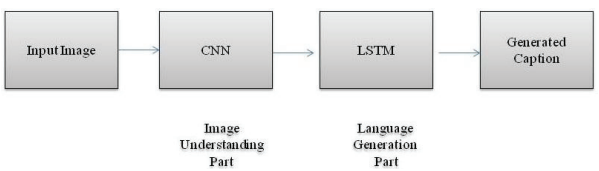

Fig. 4. Common Image Captioning model as encoder-decoder architecture

IV. Evaluation Metrics

For every caption, we append two extra symbols to In this part will explore different type of evaluations metrics use for image captioning problem. It is difficult to evaluate image captioning methods. For comparing image captioning system competence in term closeness of sentences with human generated sentences and semantic correctness, there are many evaluation metrics created. The universally employed evaluation metrics are BLEU [20], ROUGE [21], METEOR [22], etc.

A. BLEU (Bilingual Evaluation Understudy)

It is one of the techniques use to evaluate the quality of generated text, for this purpose bleu uses metric where each text is match against set of reference texts composed by human itself. To determine the closeness of machine generated text with ground truth and a score evaluated for each of them however there is no need to give attention on syntactical correctness. Finally, an average score is computed to evaluate the overall quality of generated text. The performance of BLEU metric depends on the generated text size and number of references text. There are certain limitations, bleu scores only be good if generated text is precise and in few cases high bleu score will not give assurance that value of produced text will be good.

B. ROGUE (Recall-Oriented Understudy for Gisting Evaluation)

The quality of text summary is evaluated by Rouge which is nothing but a set of metrics. Rouge match pair of words, sequences of words and n-gram with human composed reference summaries. There are many different tasks specific rouge are available like ROUGE-W, ROUGE-SU, ROUGE-1,2 where in small summaries ROUGE-SU and ROUGH-2 gives better performance and for single document evaluation ROUGH-1and ROUGE-W is good. It has also limitation to compute multi-document text summarization.

C. METEOR (Metric for Evaluation of Translation with Explicit ORdering)

METEOR is a different metric that helps to compute machine generated language. A generalized unigram match is done between machine generated text and human composed reference. And based on similarity a score evaluated in the case of many references, most excellent score is selected from individually evaluated ones..

V. Results

TABLE I. Statistics of Datasets

Name of Dataset	volume			Overall Images
	Training	Valid.	Test	
Flickr8k [6]	6000	1000	1000	8091
Flick30k [7]	28000	1000	1000	31,783

TABLE II. EVALUATION OF VARIOUS MODELS ON FLICKR8K DATASET BLUE(B) METRIC

Model	B-1	B-2	B-3	B-4
Mao et al. [16]	56.5	38.6	25.6	17.0
Jia et al. [23]	64.7	45.9	31.8	21.6
Xu et al. [24]	67.0	45.7	31.4	21.3
Wu et al. [25]	74.0	54.0	38.0	27.0
Karpathy et al. [26]	51.0	31.0	52.0	-
Kiros et al. [15]	65.6	42.4	27.7	17.7
Vinyals et al. [17]	63.0	41.0	27.0	-

TABLE III. EVALUATION OF VARIOUS MODELS ON FLICKR30K DATASET ON BLUE(B) METRIC

Model	B-1	B-2	B-3	B-4
Mao et al. [16]	60.0	41.0	28.0	17.0
Jia et al. [23]	64.6	46.6	30.5	20.6
Xu et al. [24]	66.9	43.9	29.6	19.9
Wu et al. [25]	74.0	54.0	38.0	27.0
Kiros et al. [15]	60	38	25.4	17.1
Vinyals et al. [17]	66.3	42.3	27.7	18.3

VI. CONCLUSION

In this paper, we have discussed various image captioning models. We have also presented the limitations of the discussed approaches. Also different evaluation metrics are also presented and discussed. We have shown the results of various methods are performed on Flickr8k and Flickr30k datasets. In future, models using reinforcement learning and unsupervised learning will be highly accepted for the captioning of natural scenes. Integration of textual cues with the visual information will definitely enhance the image captioning task to great extent.

REFERENCES

[1] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate (2014), arXiv preprint arXiv: 1409.0473 .

[2] K. Cho , B. Van Merriënboer , Ç. Gülçehre , D. Bahdanau , F. Bougares , H. Schwenk , Y. Bengio , Learning phrase representations using rnn encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empir- ical Methods in Natural Language Processing (EMNLP), Association for Compu-tational Linguistics, Doha, Qatar, 2014, pp. 1724–1734 .

[3] I. Sutskever , O. Vinyals , Q.V. Le , Sequence to sequence learning with neu- ral networks, in: Advances in neural information processing systems, 2014, pp. 3104–3112

[4] C. Szegedy , W. Liu , Y. Jia , P. Sermanet , S. Reed , D. Anguelov , D. Erhan , V. Van- houcke , A. Rabinovich , Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9 .

[5] H. Liu and P. Singh. ConceptNet - A practical commonsense reasoning toolkit. BT technology journal, 22(4):211–226, 2004.

[6] Rashtchian C, Young P, Hodosh M, Hockenmaier J. (2010) Collecting image annotations using Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk. Association for Computational Linguistics, pp 139–147

[7] Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans Assoc Comput Linguist 2:67–78

[8] D. Lin, An information-theoretic definition of similarity, in: Proceedings of the Fifteenth International Conference on Machine Learning, pp. 296–304.

[9] 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.

[10] V. Ordonez, G. Kulkarni, T. L. Berg., Im2text: Describing images using 1 million captioned photographs, in: Advances in Neural Information Processing Systems, 2011, pp. 1143–1151.

[11] 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) 853–899.

[12] F. R. Bach, M. I. Jordan, Kernel independent component analysis, Journal of Machine Learning Research 3 (2002) 1–48.

[13] D. R. Hardoon, S. R. Szedmak, J. R. Shawe-Taylor, Canonical correlation analysis: An overview with application to learning methods, Neural Computation 16 (2004) 2639–2664.

[14] [14] 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) 207–218.

[15] Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. In: International conference on machine learning, pp 595–603

[16] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). In International Conference on Learning Representations (ICLR).

[17] 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.

[18] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations (ICLR), 2015.

[19] Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Computing 9(8):1735–1780

[20] K. Papineni, S. Roukos, T. Ward, W. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Meeting on Association for Computational Linguistics, Vol. 4.

[21] C.-Y. Lin, F. J. Och, Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics, in:Meeting on Association for Computational Linguistics, 2004.

[22] A. Lavie, A. Agarwal, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: The Second Workshop on Statistical Machine Translation, 2007, pp. 228–231.

[23] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In Proceedings of the IEEE International Conference on Computer Vision. 2407–2415.

[24] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In International Conference on Machine Learning. 2048–2057.

[25] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. IEEE transactions on pattern analysis and machine intelligence 40, 6, 1367–1381.

[26] Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3128–3137.