International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.8, Issue.4, Apr 2020 E-ISSN: 2347-2693

A Comparative Study of Various Deep Learning Techniques Based on Automatic Image Captioning

Anurag[1*], Naresh Kumar[2]

[1]Department of Electronics and Communication Engineeering, SGT University,Gurugram, India
[2]Department of Electronics and Communication Engineeering, SGT University, Gurugram, India

DOI: https://doi.org/10.26438/ijcse/v8i4.156160 | Available online at: www.ijcseonline.org

Received: 6/Apr/2020, Accepted: 23/Apr/2020, Published: 30/Apr/2020

***Abstract*:** Generating a description of an image is called image captioning. Image captioning requires recognizing the important objects, their attributes, and their relationships in an image. This process has many potential applications in real life. A noteworthy one would be to save the captions of an image so that it can be retrieved easily at a later stage just on the basis of this description. In this survey article, we aim to present a comprehensive review of existing deep-learning-based image captioning techniques. We discuss the foundation of the techniques to analyze their performances, strengths, and limitations. We also discuss the datasets and the evaluation metrics popularly used in deep-learning-based automatic image captioning.

***Keywords*:** Image Captioning, Deep Learning, Encoder, Decoder

I. INTRODUCTION

Image captioning requires recognizing the important objects, their properties and attributes, and their relationships in an image. This process has many potential applications in real life. A noteworthy one would be to save the captions of an image so that it can be retrieved easily at a later stage just on the basis of this description. Image captioning is important for many reasons. For example, it can be used for automatic image indexing. Image indexing is important for content-based image retrieval (CBIR), and therefore, it can be applied to many areas, including biomedicine, commerce, the military, education, digital libraries, and web searching. Understanding an image largely depends on obtaining image features. The techniques used for this purpose can be broadly divided into two categories: (1) traditional machine-learning-based techniques and (2) deep machine-learning-based techniques. In this survey article, we aim to present a comprehensive review of existing deep-learning-based image captioning techniques. We discuss the foundation of the techniques to analyze their performances, strengths, and limitations. We also discuss the datasets and the evaluation metrics popularly used in deep-learning-based automatic image captioning.

II. IMAGE CAPTIONING METHODS

Image captioning methods can be categorized into three main categories:

- Template Based Image Captioning
- Retrieval Based Image Captioning
- Novel Image Caption Generation

Template-based approaches have fixed templates with a number of blank slots to generate captions. In these approaches, different objects, attributes, and actions are detected first and then the blank spaces in the templates are filled[1].

In retrieval-based approaches, captions are retrieved from a set of existing captions. Retrieval-based methods first find the visually similar images with their captions from the training dataset. These captions are called candidate captions. The captions for the query image are selected from this captions pool.

Deep Learning Image Captioning methods mostly belong to novel image caption generation method.

- Encoder-Decoder Architecture
- Compositional Architecture
- Multimodal Space
- Dense Captioning
- Attention Based Image Captioning
- Semantic Based Image Captioning

© 2020, IJCSE All Rights Reserved

a) Encoder-Decoder architecture-based methods use a simple CNN and a text generator for generating image captions[3].

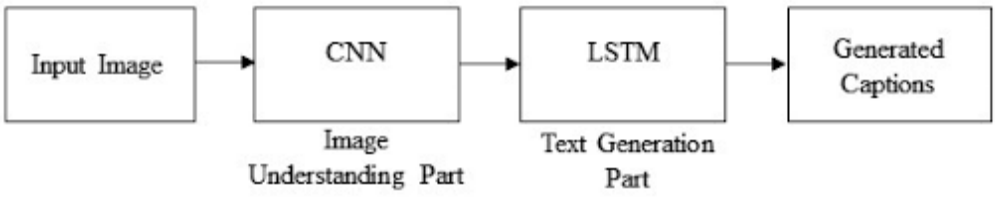

Fig.1 A block diagram of simple Encoder-Decoder architecture-based image captioning

b) Compositional Architecture-Based Image captioning. Compositional architecture-based methods composed of several independent functional building blocks: First, a CNN is used to extract the semantic concepts from the image. Then a language model is used to generate a set of candidate captions. In generating the final caption, these candidate captions are re-ranked using a deep multimodal similarity model[7].

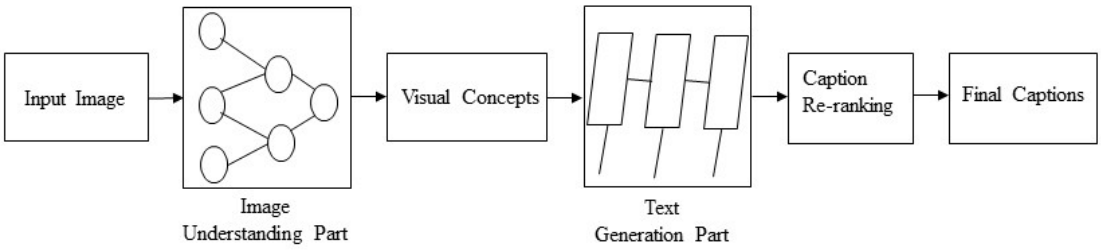

Fig.2 A block diagram of a compositional network-based captioning.

c) Multimodal space based image captioning

The architecture of a typical multimodal space-based method contains a language encoder part, a vision part, a multimodal space part, and a language decoder part. The vision part uses a deep convolutional neural network as a feature extractor to extract the image features. The language encoder part extracts the word features and learns a dense feature embedding for each word. It then forwards the semantic temporal context to the recurrent layers. The multimodal space part maps the image features into a common space with the word features[4].

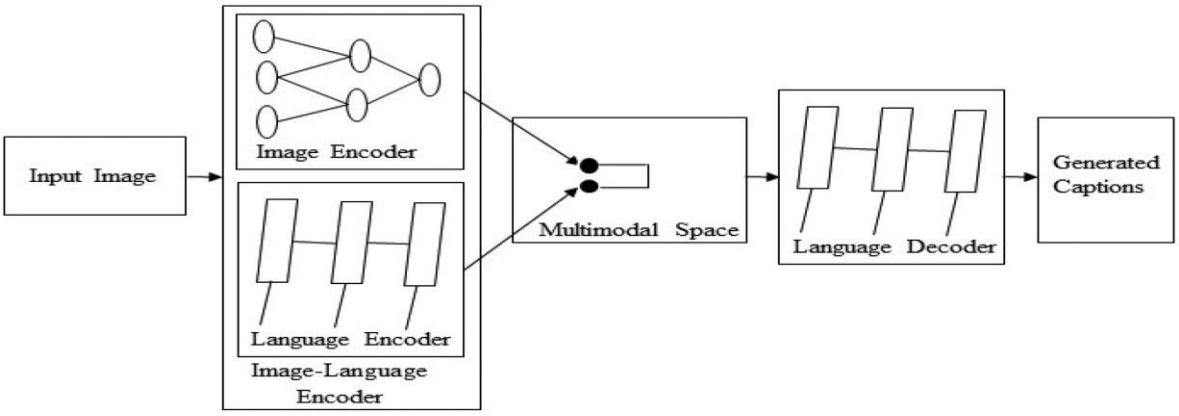

Fig.3

d) Dense captioning

This method localizes all the salient regions of an image and then generates descriptions for those regions.
A typical method of this category has the following steps:

- Region proposals are generated for the different regions of the given image.
- CNN is used to obtain the region-based image features.
- The outputs of Step 2 are used by a language model to generate captions for every region.

© 2020, IJCSE All Rights Reserved

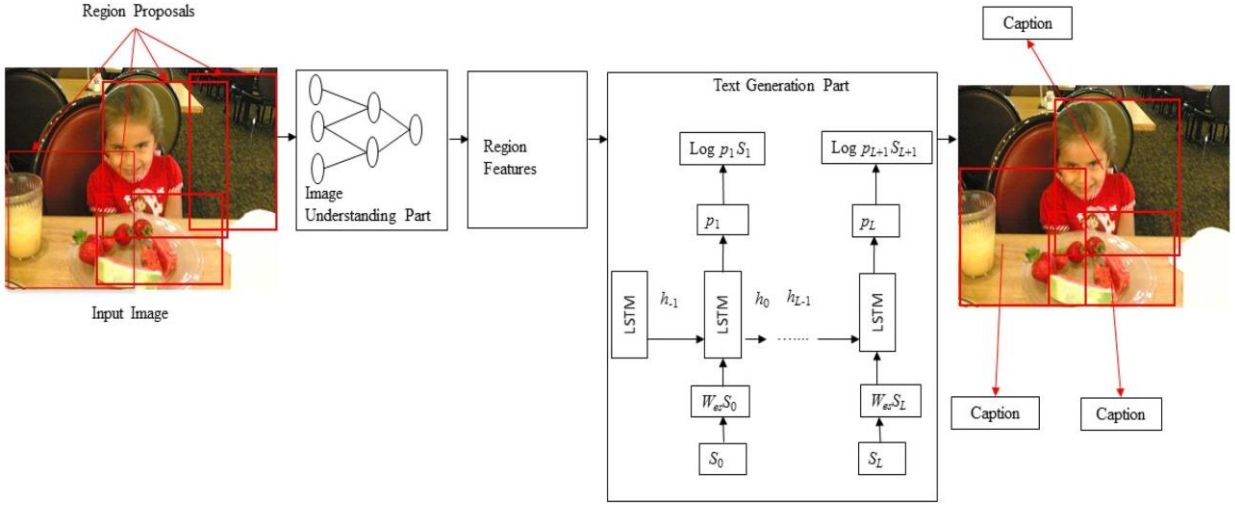

Fig.4

e) Attention-based image captioning methods focus on different salient parts of the image and achieve better performance than encoder-decoder architecture-based methods.

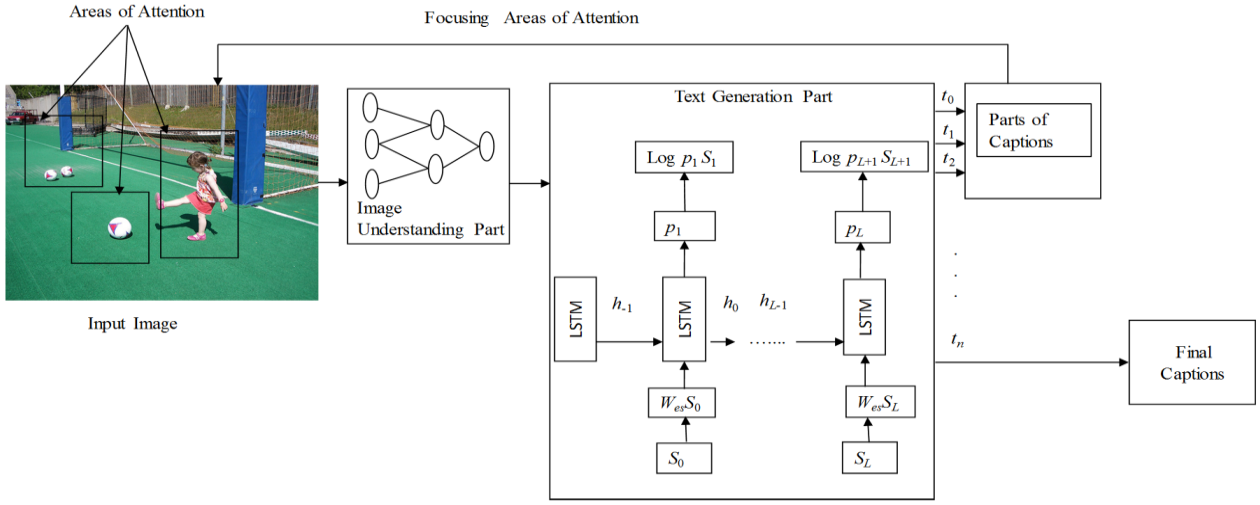

Fig.5 A block diagram of a typical attention-based image captioning technique[8].

f) Semantic concept-based image captioning methods selectively focus on different parts of the image and can generate semantically rich captions[9].

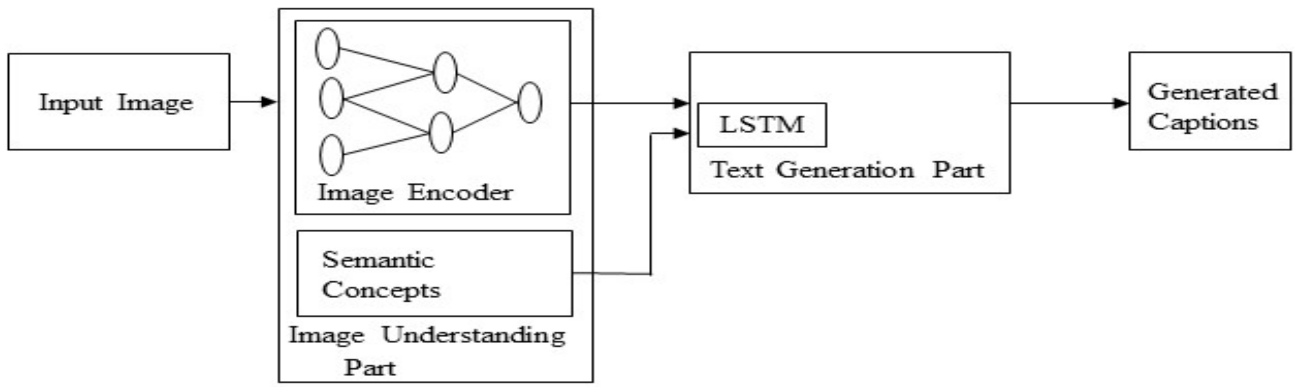

Fig.6 A block diagram of a semantic concept-based image captioning.

© 2020, IJCSE All Rights Reserved

DATASETS:

- MSCOCO
- Flickr30k
- Flickr8k

These dataset are common and popular datasets used for image captioning[11].

MSCOCO dataset is very large dataset and all the images in these datasets have multiple captions. Flickr8k contains 8K images in dataset and Flickr30K contains 30K images in the dataset. Visual Genome dataset is mainly used for region based image captioning [1].

EVALUATION METRICS: Different evaluation metrics are used for measuring the performances of image captions[12].

ROUGE: It is a set of metrics that are used for measuring the quality of text summaries.

METEOR: It is a metric that is used to measure the machine translated language.

SPICE: It is a metric that measures the image captioning based on semantics.

BLEU: It is a metric that is used to measure the quality of machine generated text.

III. PERFORMANCE OF DIFFERENT METHODS

Dataset: Flickr30K

Table.1

Category	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
MS, SL, WS	0.600	0.410	0.280	0.190	-
VS, SL, WS, EDA	0.646	0.466	0.305	0.206	0.179
VS, SL, WS, EDA, AB	0.669	0.439	0.296	0.199	0.184
VS, SL, WS, EDA, SCB	0.730	0.550	0.400	0.280	-

Dataset: Flickr8K

Table.2

Category	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
MS, SL, WS	0.565	0.386	0.256	0.170	-
VS, SL, WS, EDA	0.647	0.459	0.318	0.216	0.201
VS, SL, WS, EDA, AB	0.670	0.457	0.314	0.213	0.203
VS, SL, WS, EDA, SCB	0.740	0.540	0.380	0.270	-

From the above table, it can be deduce that MSCOCO dataset is showing effective results on various combinations of image captioning methods. However there are different approaches for different kinds of applications for image captioning. So, as per the requirement, the effective approach may be used for image captioning.

Dataset: MSCOCO

Table.3

Category	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
MS, SL, WS	0.670	0.490	0.350	0.250	-
VS, SL, WS, EDA	0.670	0.491	0.358	0.264	0.227
VS, SL, WS, EDA, AB	0.718	0.504	0.357	0.250	0.230
VS, SL, WS, EDA, SCB	0.740	0.560	0.420	0.310	0.260

© 2020, IJCSE All Rights Reserved

IV. FUTURISTIC DIRECTIONS

Generation-based methods can generate novel captions for every image. However, these methods fail to detect prominent objects and attributes and their relationships to some extent in generating accurate and multiple captions[16].

Working on an open-domain dataset will be an interesting avenue for research in this area.
Supervised learning needs a large amount of labeled data for training. Therefore, unsupervised learning and reinforcement learning will be more popular in the future in image captioning.

Multiple UUs in one single scenario, with the help of GANs and reinforcement learning to optimize our models.
Automatic understanding of multimodal information is still an unsolved research problem.
Increasing vertical depth of encoder-decoder for better image captioning

V. CONCLUSIONS

In this article, we have reviewed deep-learning-based image captioning methods. We have given a taxonomy of image captioning techniques, shown generic block diagrams of the major groups, and highlighted their pros and cons. We discussed different evaluation metrics and datasets with their strengths and weaknesses. A brief summary of experimental results was also given. We briefly outlined potential research directions in this area. Although deep-learning-based image captioning methods have achieved remarkable progress in recent years, a robust image captioning method that is able to generate high-quality captions for nearly all images is yet to be achieved.

REFERENCES

[1]. MD. ZAKIR HOSSAIN, FERDOUS SOHEL, MOHD FAIRUZ SHIRATUDDIN, and HAMID LAGA, "A Comprehensive Survey of Deep Learning for Image Captioning", ACM Computing Surveys, Vol. 51, No. 6, Article 118, February 2019.
[2]. Zhihong Zeng, Xiaowen Li, "Application of human computing in image captioning under deep learning", Springer Nature 2019, May 2019.
[3]. Xianhua Zeng, Li Wen, Banggui Liu, Xiaojun Qi, "Deep Learning for Ultrasound Image Caption Generation based on Object Detection", *Neurocomputing* (2019), doi: https://doi.org/10.1016/j.neucom.2018.11.114, Nov 2018.
[4]. Christian Otto, Matthias Springstein, Avishek Anand, Ralph Ewerth, "Understanding, Categorizing and Predicting Semantic Image-Text Relations", ICMR '19, Ottawa, ON, Canada , June 10–13, 2019.
[5]. Xinyu Xiao, Lingfeng Wang, Kun Ding, Shiming Xiang, and Chunhong Pan, "Deep Hierarchical Encoder-Decoder Network for Image Captioning", DOI 10.1109/TMM.2019.2915033, IEEE Transactions on Multimedia, 2019.
[6]. Yuting Su, Yuqian Li, Ning Xu, An-An Liu, "Hierarchical Deep Neural Network for Image Captioning", Springer Science+Business Media, LLC, Springer Nature , 2019.
[7]. CHENG WANG, HAOJIN YANG, and CHRISTOPH MEINEL," Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning", *ACM Trans, Multimedia Comput. Commun., Appl.* 14, 2s, Article 40, April 2018.
[8]. Xiaoxiao Liu, Qingyang Xu, Ning Wang, "A survey on deep neural network-based image captioning", https://doi.org/10.1007/s00371-018-1566-y, Springer Nature, 2018.
[9]. Vasiliki Kougia, John Pavlopoulos, Ion Androutsopoulos, "A Survey on Biomedical Image Captioning", arxiv:1905.13302v1, May 2019.
[10]. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering", IEEE Explore, 2019.
[11]. Justin Johnson, Agrim Gupta, Li Fei-Fei, "Image Generation from Scene Graphs", IEEE Explore, 2019.
[12]. Yang Feng, Lin Ma, Wei Liu, Jiebo Luo, "Unsupervised Image Captioning", IEEE Explore, 2019.
[13]. Songtao Ding , Shiru Qu , Yuling Xi , Arun Kumar Sangaiah , Shaohua Wan, " Image caption generation with high-level image features", Pattern Recognition Letters 123 (2019) 89–95, Mar 2019.
[14]. Lin Ma, Wenhao Jiang, Zequn Jie, Yu-Gang Jiang, and Wei Liu, "Matching Image and Sentence with Multi-faceted Representations", DOI 10.1109/TCSVT.2019.2916167, IEEE Transactions on Circuits and Systems for Video Technology, 2019.
[15]. Lun Huang, Wenmin Wang, Gang Wang, "IMAGE CAPTIONING WITH TWO CASCADED AGENTS", ICASSP 2019, IEEE, 978-1-5386-4658-8/18, 2019.
[16]. Alexander G Schwing Jyoti Aneja, Aditya Deshpande. 2018. Convolutional image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 5561–5570.
[17]. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In European Conference on Computer Vision. Springer, 382–398.
[18]. Xinlei Chen and C Lawrence Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the IEEE conference on computer vision and pattern recognition. 2422–2431.
[19]. Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. In Association for Computational Linguistics. 103–111.
[20]. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

© 2020, IJCSE All Rights Reserved