ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 8, Issue 2, February 2020

Image Captioning Models: A Systematic Literature Survey

Shweta Dhurmekar, Prof. Deipali Gore

Department of Computer Engineering, PES Modern College of Engineering, Pune, Maharashtra, India

ABSTRACT: Image Captioning is the process of generating textual description of an image. It uses both Natural Language Processing and Computer Vision to generate the captions. Image captioning requires to recognize the important objects, their attributes and their relationships in an image. The generated sentences should be syntactically and semantically correct. Deep learning is capable of handling these complexities. Over the years there has been abundant research done on this topic. In this study a comprehensive Literature Review provides a brief overview of the existing deep learning based techniques for image captioning. The foundation of this study is to explain the common techniques and their limitations as well summarize the results from the recent researches. We also discuss the datasets and evaluation metrics.

KEYWORDS: Deep learning, Attention Model, Global Model, CNN, RNN, LSTM, Semantic, Remote Sensing, Language Model.

I. INTRODUCTION

Understanding an image and its content is an easy task for humans but for machines to do the same can be relatively difficult. Every day we come across images through various sources like Newspaper, Internet and Magazines. These sources contain images which viewers can interpret easily. Most of the images don’t have captions but humans can understand them without detailed captions. However for machines to interpret the same can be a difficult task.

To describe a scene in an image is a highly demanding task for humans and to create machines with this capability researchers have been exploring various methods. Image captioning needs more effort than image recognition because of the additional task of identifying the objects and their relationship and then creating a concise caption for that image. To name few real life applications are self-driving cars, navigation, remote sensing image retrieval, scene classification and more[14].

The science and methodology behind deep learning have been in existence for decades, but an increasing abundance of digital data and the involvement of powerful GPUs has accelerated the development of deep learning research in recent years. Convenient development libraries such as Tensor Flow and PyTorch, the open source community, large labeled datasets (e.g. MSCOCO, Flicker) and splendid demonstrations simulate the explosive growth of the deep learning field.

Image captioning has played and will continue to play a vital role in science and industry.Its applications spread to many areas, including visual recognition and scene understanding [1] to name a few. In recent years deep learning based models has positively impacted the field of image captioning. In this paper we aim to highlight the recent advances in the field of image captioning. The primary challenge towards this goal is in the design of a model that is rich enough to simultaneously reason about contents of images and their respective description in the domain of natural language.

This paper has been divided in four parts. First, we present a brief introduction about Image Captioning. Second, related work and recent research of various methods for Image Captioning. The evaluation metrics section is about how the models and generated captions are evaluated. Finally, the paper is concluded with some open questions for future studies.

DOI: 10.15680/IJIRCCE.2020. 0802012

II. LITERATURE REVIEW

Generating captions of an image is very important task in understanding the scene. Captions describes the image in brief; to generate caption machines requires, detecting, identifying and classification of objects orscenes. A relationship among the identified objects is to be formed which is an important aspect. This is a complex task to define an image or a scene.However, since human evaluation is costly and time consuming, many automatic metrics are proposed, which could serve as a proxy mainly for speeding up the development cycle of the system. The purpose of this survey is to analyze various techniques used for image caption generation.

Su Mei Xi and Young Im Cho et al (2013) proposeda novel system which generates sentential annotations for general images a weighted feature clustering algorithm is employed on the semantic concept clusters of the image regions. For a given cluster, model determines relevant features based on their statistical distribution and assign greater weights to relevant features as compared to less relevant features. In this way the computing of clustering algorithm can avoid dominated by trivial relevant or irrelevant features. Then, the relationship between clustering regions and semantic concepts is established according to the labeled images in the training set[3].

Ming Chen et al (2015) proposed two methods for image classification.The first method used off-the-shelf CNN features on full image to capture global features. The second method extracted features locally followed by feature pooling using the Fisher Vector. They conducted experiments to compare with state-of-the-art and their method outperformed many existing methods. This experiment only classifies the image and does not fully describe the image[4].

A. Karpathy and L. Fei-Fei et al (2015)proposed alignment modelbased on a novel combination of Convolutional Neural Networks over image regions, bidirectional Recurrent Neural Networks (RNN) over sentences, and a structured objective that aligns the two modalities through a multimodal embedding. This study then describe a Multimodal Recurrent Neural Network architecture that uses the inferred alignments to learn to generate novel descriptions of image regions. In this work the datasets used are Flickr8K, Flickr30K and MSCOCO datasets. However this work is subjected to few limitations where evaluating every region in isolation leads to computational inefficiency because one must forward every individual region of interest separately through the convolutional network[5].

Dong-Jin Kim ; DonggeunYoo ; BonggeunSim et al (2016)propose a method to train imagerepresentation to improve the performance of image Caption Generation. This modeltransfers the learning CNN on sentences, and extract image representation with deep Fisher kernel. With this representation system can generate sentence with gLSTM and show the improvements in performance. The performance metric is evaluated in BLEU. This model needs more training to generate more grammatically correct captions[8].

Zhenwei Shi and ZhengxiaZou et al (2017), proposed a systema remote sensing image captioning frameworkby leveraging the techniques of the recent fast development of deep learning and fully convolutional networks. The experimental results on a set of high-resolution opticalimage including Google Earth images and GaoFen-2 satellite imagesdemonstrate that the proposed method is able to generate robust and comprehensive sentence description with desirable speed performance[10].

ShiruQu, Yuling Xi, Songtao Ding et al (2017) proposed a neural and probabilistic framework which combines CNN with a special form of recurrent neural network (RNN) to produce an end-to-end image captioning. This work uses a model that takes advantage of word to vector to encode the variable length input into a fixed dimensional vector. Considering the description of the object in an image is not specific enough, we introduce an attention mechanism through visualization to show how the model is able to fix its gaze on salient objects. The datasets used are Flickr8K, Flickr30K and MSCOCO. This work has limitations where if the model is not able to identify unknown objects it generates the wrong caption[11].

Linghui Li, Sheng Tang, YongdongZhang,Lixi Deng, Qi Tian et al (2017) propose a new method called global-local attention (GLA) for generating image description. The proposed GLA model utilizes attention mechanism to integrate objectlevelfeatures with image-level feature. Through this manner, this model can selectively pay attention to objects and context information concurrently. Therefore, proposed GLA method can generate more relevant image description sentences, and achieves the state-of-the-art performance on the well-known Microsoft COCO caption dataset with several popular evaluationmetrics — CIDEr, METEOR, ROUGE-L and BLEU-1,2,3,4. This model has limitations; it cannot explicitly train the attention layer[12].

Fang Fang et al (2018) designed a novel word level attention layer to process image features with two modules for accurate word prediction. The first is a bidirectional spatial embedding module to handle feature maps, then the second module employs attention mechanism to extract word level attention which will be fed into language model. The dataset used in this study is MSCOCO dataset[13].

Zhengyuan Zhang ; Wenkai Zhang ; WenhuiDiao ; Menglong Yan ; XinGao ; Xian Sun et al (2019) proposed a Visual Aligning Attention model (VAA), is proposed. In this model, the attention layer is optimized by a well-designed visual aligning loss during the training phase. The visual aligning loss is obtained by explicitly calculating the feature similarity of attended image features and corresponding word embedding vectors. Besides, in order to eliminate the influence of non-visual words in training the attention layer, a visual vocab used for filltering out non-visual words in sentences is proposed, which can neglect the non-visual words when calculating the visual aligning loss. Experiments on UCM-Captions and Sydney-Captions prove that the proposed method is more effective in remote sensing image caption task. This model can explicitly train the attention layer[14].

WenliangCai et al (2019) proposed a Visual Question Answering (VQA) Algorithm. In this approach the model uses CNN and LSTM algorithms along with collaborative attention mechanism to generate a caption for the image related to the problem information and the combines the text information on the image description and the question to obtain an answer and give picture description as the output[15].

III. EVALUATION METRICS

Various models evaluate captions using BLEU, CIDEr, METEOR[1,2,6]. These are common metrics to evaluate different caption generation models which is useful when choosing the best model.
Table 1.shows the summary of few recent works that use the following evaluation metrics.

BLEU: Bilingual Evaluation Understudy is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. This is a precision based metrics,claims to highly correlate with human evaluation.BLEU has different n-gram based versions for candidate sentences with respect to the reference sentences.

CIDEr::Consensus-based Image Description Evaluation;measures the *n*-gram match between the captionhypothesis and the references, while the *n*-grams are weightedby term frequency–inverse document frequency (TF-IDF).

METEOr:Metric for Evaluation of Translation with Explicit ORdering is a metric for the evaluation of machine translation. It is based on the concept of unigram matching between human and machine generated translation.

ROUGE:Recall-Oriented Understudyfor Gisting Evaluation; determines the quality by comparing candidate summary with the human reference summary. Just like BLEU this evaluation also has n-gramversions.

Author	Architecture	Evaluation
A. Karpathy and L. Fei-Fei et al, 2015	CNN and Bidirectional RNN	MS-COCO BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR CIDEr 62.5 45.0 32.1 23.0 19.5 66
Dong-Jin Kim ; DonggeunYoo ; BonggeunSim ; In So Kweon et al, 2016	CNN and gLSTM	BLEU-1 BLEU-2 BLEU-3 BLEU-4 66.8 46.9 3 1.5 20.8
Fang Fang ; Hanli Wang ; Pengjie Tang et al, 2018	CNN, LSTMand Bidirectional RNN	MSCOCO CIDEr BLEU-3 BLEU-4 METEOR ROUGE-L 98.7 41.7 31.4 25.2 53.4
Linghui Li, Sheng Tang, YongdongZhang,Lixi Deng, Qi Tian et al, 2017	Global and Local feature based attention VGG/Faster R-CNN and LSTM Attention based	FLICKR 30k BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR CIDEr 56.8 37.2 23.2 14.6 16.6 36.2
ShiruQu, Yuling Xi, Songtao Ding et al, 2019	Visual Aligning Attention CNN and LSTM Encoder and decoder	UCM Captions BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR CIDEr 81.92 75.11 69.27 63.87 43.80 3.39 Sydney Caption BLEU-1 BLEU-2 BLEU-3 BLEU-4 METEOR CIDEr 74.31 66.46 60.29 54.95 0.39 0.69

Table1: The summary of few models for Image Captioning.

IV. CONCLUSION

In this work various Image captioning models were presented. Over the years how the models and designing approach have shaped is observed. This paper gives an overview of the some of the best models as well as concludes among all the models proposed the attention based models have shown the potential to outperform the rest.Attention based techniques are the best techniques chosen when it comes to generate captions. To evaluate the performance various automated metrics are there some of them are BLEU, METEOR,CIDEr and ROUGE.

This paper systematically represents the research done over the years to give an overview of how the technology has evolved. There's more research going on in this field of study so amount of information is constantly increasing.

REFERENCES

[1] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proc. 40th Annu. Meeting AssociationComputational Linguistics, 2002.

[2] S. Banerjee and A. Lavie."METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proc. ACL Workshop onIntrinsic and Extrinsic Evaluation Measures for Machine Translation and/orSummarization, 2005.

[3] Su Mei Xi and Young Im Cho, "Image Caption Automatic Generation Method Based on Weighted Feature" in International Conference on Control, Automation and Systems (lCCAS 2013), Oct 2013.

[4] Ming Chen ; Lu Zhang ; Jan P. Allebach "Learning deep features for image emotion classification" in IEEE International Conference on Image Processing (ICIP), 2015.

[5]A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proc. CVPR'15, Jun. 2015, pp. 3128–3137.

[6] R. Vedantam, L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in Proc. European Conf. Computer Vision, 2015,

[7] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in Proc. European Conf. Computer Vision, 2016.

[8] Dong-Jin Kim ; DonggeunYoo ; BonggeunSim ; In So Kweon," Sentence Learning on Deep Convolutional Networks for Image Caption Generation" in International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Oct 2016.

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, TimoRehfeld, Markus Enzweiler, Rodrigo Benenson, UweFranke, Stefan Roth, and BerntSchiele, "The cityscapes dataset for semantic urban scene understanding," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213– 3223.

[10] Zhenwei Shi and ZhengxiaZou , "Can a Machine Generate Humanlike Language Descriptions for a Remote Sensing Image" in IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, 2017.

[11]ShiruQu, Yuling Xi, Songtao Ding, "Visual attention based on long-short term memory model for image caption generation" in 29th Chinese Control And Decision Conference (CCDC) 2017.

[12] Linghui Li, Sheng Tang, YongdongZhang,Lixi Deng, Qi Tian, "GLA: Global-local Attention for Image Description" in IEEE Transactions on Multimedia, Sept 2017.

[13]Fang Fang ; Hanli Wang ; Pengjie Tang "Image Captioning with Word Level Attention" in 25th IEEE International Conference on Image Processing (ICIP),2018

[14] Zhengyuan Zhang ; Wenkai Zhang ; WenhuiDiao ; Menglong Yan ; XinGao ; Xian Sun"VAA: Visual Aligning Attention Model for Remote Sensing Image Captioning" in IEEE transaction , Sept 2019.

[15]WenliangCai ; GuoyongQiu, "Visual question answering algorithm based on image caption" in IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), 2019