IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A SURVEY ON DEEP LEARNING TECHNOLOGIES USED FOR IMAGE CAPTIONING

[1]Ms.Komal Thorat, [2]Prof.R.L.Paikrao

[1]PG Student, Department of Computer Engineering, [2]Associate Professor, Department of Computer Engineering
[1]Amrutvahini College of Engineering, Sangamner, India

Abstract: Image captioning is the process of generating textual information about an image. Nowadays it is attracting more and more attention, as it is a newly emerged research area. To accomplish the goal of image captioning, semiology information of images needs to be recorded and conveyed in natural languages. Both Natural Language Processing and Computer Vision are used to generate the captions, so it becomes a quite challenging task. Numerous techniques have been proposed to tackle this problem. In this paper, we present a survey on different approaches in image captioning research. Based on the techniques and models adopted, we classify image captioning. In this paper, we first drew attention to methods used earlier, and then we focus our main attention on neural network-based methods and used models, which give state-of-the-art results.

***Index Terms* - Image captioning, Attention mechanism, Deep Learning, Encoder-decoder framework, Multi-Task Learning**

I. INTRODUCTION

Human beings can easily narrate the surroundings they are in. Taking into consideration an image, it is ordinary for a human to narrate a massive number of details about that image with a quick look. This is one of the primary mortal intelligence. Building computers to mimic humans’ ability to transcribe the visual world has been an eternal goal of researchers within the area of artificial intelligence.

Enabling the computers to explain the visual world will cause a good number of possible applications, like producing natural human-robot interactions, babyhood education, information retrieval, and visually impaired assistance. During this advancing generation, the Driving assist system uses image captioning which eases the driver to identify the changes in road scenes and fend off the road accidents. Lane signs are a great aspect to drive safely. Identifying lane signs accurately at an apparent time is remarkable for any motorist to secure an altogether pleasant journey.

While looking into the captioning model then it is a necessity to not only realize the noticeable entity, entity correlation in a picture, and their features but also to systematize these varieties of findings into accurate order. There are two key areas involved in image captioning, i.e. visual understanding and verbal processing, within the previous couple of years, researchers have made a remarkable plow in an exceedingly few areas of supercomputer vision perception, like picture classification, attribute classification, entity detection, scene identification, action recognition, etc. to make sure an end in decisions are grammatically and semantically correct, methods of computer vision and linguistic communication processing are purported to be chosen and for this, various techniques are proposed.

Initially, techniques used are retrieval and template-based, in which uncompromising hard-coded rules and time-consuming features are utilized. Outputs of such methods have restraint. In this look-over, we will focus mainly on various neural network-based methods.
This paper is put in order as follows. We first write up retrieval and template-based image captioning approaches in Sections II. Section III is concerned with neural network-based methods. State of art methods and benchmark datasets are discussed in Section IV. The conclusion will be given in Section V.

II. RETRIEVAL & TEMPLATE BASED CAPTIONING

Retrieval-based captioning is one kind of image captioning method that was common in early times. Given a question image, retrieval-based methods produce a caption for it by fetching one or a collection of sentences from a discriminant sentence pool. The result in the caption can either be a sentence that has already existed, or a sentence collected from the retrieved ones.

Template-based approaches generate caption templates whose slots are filled in based on outputs of entity detection, feature classification, and scene identification. Farhadi et al. [9] infer a triplet of scene elements which is converted to text using templates. Kulkarni et al. adopt a Conditional Random Field (CRF) to jointly reason across objects, attributes, and prepositions before filling the slots. It uses more powerful language templates such as a syntactically well-formed tree and adds descriptive information from the output of attribute detection.

III. NEURAL NETWORK METHODS

Due to great improvements in the area of deep learning, recent work begins to believe in deep neural networks for impulsive image captioning. In this section, we will analyze such methods. Even though deep neural networks are now widely adopted for tackling the image captioning task, different methods may be based on different frameworks. Therefore, we classify deep neural network-based methods into the main framework they used.

Spatial Attention Model and Adaptive Attention Model

In general, Author Long Chen et al.[3] propose that a caption word only relates to partial regions of an image. For example, if there is a picture of a girl who is holding the cake in her hand, in this scenario when we want to predict cake, only image regions that contain cake are useful. Therefore, applying a global image feature vector to generate a caption may lead to sub-optimal results due to the irrelevant regions. Instead of considering each image region equally, the spatial attention mechanism attempts to pay more attention to the semantic-related regions. Author Jiasen Lu et al.[1]propose a unique spatial attention model for extracting spatial image attributes and stated the below function:

The context vector ct is defined as:
$ct = g(V , ht)$
where g is nothing but the attention function, $V = [v1,..., vk]$, $vi \in Rd$ is called the spatial image features, during which d is the dimensional representation corresponding to an element of the image, ht is the hidden state of Recurrent Neural Network at time t.

While spatial attention-based decoders are demonstrated to be effective for image captioning, they can't determine when to depend upon visual signals and when on the language model so, J.Lu et al.[1] they propose an adaptive attention model supported by the visual sentinel, to compute the context vector. That the new adaptive context vector is defined as cˆt, which is modeled as a combination of the spatially attended image features (i.e. context vector of spatial attention model) and also the visual sentinel vector. This trades off the quantity of new information the network considers from the image with what it already knows within the decoder memory (i.e., the visual sentinel).

In comparison with the regular image captioning models, Author Jia Huei Tan et al. [9] proposed COMIC which has extensively minor learnable parameters, which are resulting in the reduced requirement on GPU memory and storage. A closely related work to this is LightRNN but with a few differences - i) COMIC requires only a one-word embedding matrix (as against two in LightRNN); ii) COMIC doesn't necessitate any changes within the model architecture (LightRNN requires a word embedding table), and iii) LightRNN is applied for language modeling only. On the opposite hand, COMIC is orthogonal to compression.

Using this approach improves the state-of-the-art on BLEU-4 from 0.325 to 0.332, METEOR from 0.251 to 0.266, and CIDEr from 0.986 to 1.085.

Author Peter Anderson et al.[2]put forward the integrated bottom-up and top-down visual attention techniques. In this work, they define spatial regions about bounding boxes and implement bottom-up attention using Faster R-CNN. Faster R-CNN identifies objects in two stages. The primary stage, described as a Region Proposal Network (RPN), predicts the entity. Within the second stage, region of interest (RoI) pooling is used to extract a tiny feature map for every box proposal. This perspective allows attention to be calculated more easily at the level of objects and other salient regions. Applying this approach to image captioning and visual question answering achieves state-of-the-art results in both parts. Applying this way to deal with picture inscribing, outcomes on the MSCOCO test worker set up another best in class for the assignment, accomplishing CIDEr/BLEU-4 scores of 117.9 and 36.9 individually.

Xu Yang et al.[5] presented a unique unsupervised learning method: Scene Graph Auto-Encoder (SGAE) which could be a sentence self-reconstruction network, that uses scene graphs [5] to bridge the gap between the two worlds. A scene graph (G) could be a unified representation that connects-1) the objects 2) their attributes, and 3) their relationships in a picture or a sentence by directed edges.The model can do a replacement state-of-the-art score among all the compared methods in terms of CIDEr-D, 127.8 which is a completely 7.2 points uplift with an upgraded version. Similarly Xiangyang Li et al.[10]propose a model based on scene graphs for developing natural language descriptions. Because scene graphs represent object entities and pairwise relationships, they provide rich information for narrating images.

Min Yang et al.[7] talk about the Multi-task Learning Algorithm for cross-Domain. MLADIC is a multitasking system that concurrently optimizes two integrated objectives along with a dual learning mechanism that is image captioning and text-to-image synthesis. This approach achieves the CIDEr/BLEU-4 scores of 119.6, 36.1individually. Xinyu Xiao et al.[8] designed the Hierarchical EncoderDecoder Network (DHEDN), which resulted in the transformation of the caption to image (top) and image to caption (bottom) successfully. They build a deep hierarchical encoder-decoder to merge the visual and textual semantics before decoding. CIDEr/BLEU-4 scores of 117.0, and 36.7 individually.

IV. STATE OF THE ART METHOD AND BENCHMARK DATASET COMPARISON

Below are the image captioning evaluation metrics that give a state of the art results-

The evaluation metrics include 1. BLEU, 2. METEOR, 3. ROUGE-L, and 4. CIDEr. BLEU, METEOR, and ROUGE-L are initially designed to judge machine translation grades.

Bilingual evaluation understudy (BLEU) is used to varying lengths of phrases of an applicant sentence to match against reference sentences written by human beings to measure their closeness.
ROUGE-L is intended for the capability and fluency of machine translation. this metric automatically includes the lengthy in-sequence common n-grams, the sentence-level structure will be naturally captured.
CIDEr is a standard that uses human consensus to evaluate the quality of image captioning. This criterion measures the similarity of a sentence generated by the captioning method to the majority of ground truth sentences put down by human beings.

In detail, three benchmark datasets are extensively used to evaluate image captioning methods. The datasets are Flickr8K, Flickr30k, and Microsoft COCO Caption dataset.

Flickr8K involves 8,000 images extracted from Flick. The images in this dataset mainly enclose the human and animal pictures.
Flickr30k is a dataset that is enlarged from the Flickr8k dataset. In Flickr30k there are 31,783 commented images. Each image is linked with five sentences wilfully written for it. The images in this dataset are mainly about humans engaged in day-to-day activities and events.
Microsoft COCO Caption dataset is designed by collecting the images of complex daily incidents with common objects in their natural text.

V. CONCLUSION

In this paper, we introduce the survey on image captioning. We systematize the image captioning approaches and supported techniques in each method. Our main attention is targeted at neural network-based methods, which provide a state-of-the-art results. After that, we discussed the benchmark datasets.

VI. ACKNOWLEDGMENT

I thank my guide Prof.R.L.Paikrao for analyzing the data and for advising on all aspects related to this survey.

REFERENCES

[1] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3242–3250
[2] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 6077–6086.
[3] L. Chen et al., "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5659–5667.
[4] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 4904–4912.
[5] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 10685–10694.
[6] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying more attention to saliency: Image captioning with saliency and context attention," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 14, no. 2, p. 48, 2018.
[7] M. Yang, W. Zhao, W. Xu, Y. Feng, Z. Zhao, X. Chen, and K. Lei, "Multitask learning for cross-domain image captioning," IEEE Transactions on Multimedia, vol. 21, no. 4, pp. 1047–1061, 2018.
[8] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep hierarchical encoder-decoder network for image captioning," IEEE Transactions on Multimedia, 2019.
[9] J. H. Tan, C. S. Chan, and J. H. Chuah, "Comic: Towards a compact image captioning model with attention," IEEE Transactions on Multimedia, 2019.
[10] X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," IEEE Transactions on Multimedia, 2019.
[11] Z. Zhang, Q. Wu, Y. Wang, and F. Chen, "High-quality image captioning with fine-grained and semantic-guided visual attention," IEEE Transactions on Multimedia, vol. 21, no. 7, pp. 1681–1693, 2018.
[12] M. Tanti, A. Gatt, and K. P. Camilleri, "Where to put the image in an image caption generator," Natural Language Engineering, vol. 24, no. 3, pp. 467–489, 2018.