International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST

Image Caption Generation using Natural Language Processing

Shreya Shingade, Dr. S.A. Ubale

Department of Computer Engineering, Zeal College of Engineering & Research, Pune, Maharashtra, India

Article Info
Volume9, Issue 3
Page Number: 851-855
Publication Issue
May-June-2022
Article History
Accepted :01May2022
Published :12 May 2022

ABSTRACT

An image-based web crawler is a web crawler that searches for data using similar photos. On the web, there is a large assortment of image assets, with a significant proportion of photographs carrying both named and unidentified captions. Users must sort through the photos according to their requirements. A significant proportion of users are unable to recover the necessary images as a result of their unanticipated appropriate inscription on their photographs. The goal of our project is to create an automated photo caption depending on the image quality. To begin, a picture's content should be easily understood, followed by a statement or declaration that is consistent with the image's grammatical laws and semantical information. Computer vision and natural language processing technologies are required to merge the two forms of material, which is a difficult task. The goal of the paper is to generate mechanical inscriptions by analysing the information of an image. Currently, pictures must be removed through human involvement, which is nearly impossible in large databases. As a contribution, the picture information base is sent to a deep neural network. The Convolutional Neural Network encoder creates captions that extract the image's highlights and nuances, while the Recurrent Neural Network decoder interprets the image's highlights and articles to produce a continuous, intelligible description of the image.

Keywords: Deep Learning, part of speech, image captioning, multi-task learning

I. INTRODUCTION

Image captioning, which tries to link image with language, has become a popular study topic to aid research in areas such as cross-modal retrieval and the support of visually impaired persons. Not only must an image captioning model recognise the important objects in a picture, their qualities, and their relationships, but it must also structure this information into a syntactically and semantically correct sentence. Recent captioning models, with the advancements of Neural Machine Translation, mainly use the encoder decoder structure to "translate" a picture into a sentence, with promising results.

Researchers have achieved substantial progress in recent years in fields such as image classification, feature classification, object detection and recognition, scene recognition, action recognition, and so on. Having a machine develop natural language descriptions for an image, on the other hand, remains a complex and challenging issue. This challenge unites two very distinct media formats, needing computers to not only interpret the visual information of the image correctly and comprehensively, but also to mix and organise the semantics of the image using human language. The subtasks of picture captioning, such as identifying semantic aspects like visual objects, object properties, and sceneries, are difficult enough, but organising words and phrases to communicate this information adds to the challenge.

II. LITERATURE SUREVY

In this paper[1], author propose a novel versatile consideration model with a visual sentinel. At each time step, our model concludes whether to take care of the picture (and provided that this is true, to which districts) or to the visual sentinel. The model concludes whether to take care of the picture and where all together to extricate significant data for consecutive wordage. Author test his strategy on the COCO picture subtitling 2015 test dataset and Flickr30K.

In this work[2], authors propose a joined base up and topdown consideration component that empowers thoughtfulness regarding be determined at the degree of items and other striking image areas. This is the normal reason for thoughtfulness regarding be thought of. Inside our methodology, the base up system (in light of Faster R-CNN) proposes picture districts, each with a related element vector, while the top-down component decides highlight weightings. 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/SPICE/BLEU-4 scores of 117.9, 21.5 and 36.9, individually.

In this paper[3], Author present a novel convolutional neural organization named SCA-CNN that joins Spatial and Channel wise Attentions in a CNN. In the undertaking of picture inscribing, SCA-CNN progressively regulates the sentence age setting in multi-layer highlight maps, encoding where (i.e., mindful spatial areas at different layers) and what (i.e., mindful channels) the visual consideration is. Authors assess the proposed SCA-CNN design on three benchmark picture subtitling datasets: Flickr8K, Flickr30K, and MSCOCO. It is reliably seen that SCA-CNN fundamentally beats best in class visual consideration based picture inscribing techniques.

In this paper[4], authors present Long Short-Term Memory with Attributes (LSTM-A) a novel engineering that coordinates ascribes into the effective Convolutional Neural Networks (CNNs) additionally Recurrent Neural Networks (RNNs) picture subtitling system, via preparing them in a start to finish way. Especially, the learning of characteristics is fortified by coordinating between property relationships into Multiple Instance Learning (MIL). To consolidate credits into subtitling, Author develop variations of designs by taking care of picture portrayals and properties into RNNs in various manners to investigate the shared yet additionally fluffy connection between them. Broad analyses are led on COCO image subtitling dataset and our system shows clear upgrades when contrasted with cutting edge profound models.

Author[5] propose Scene Graph Auto-Encoder (SGAE) that consolidates the language inductive inclination into the encoder decoder image subtitling structure for more human-like subtitles. Instinctively, we people utilize the inductive inclination to make collocations and logical deduction in talk. For instance, when we see the connection "individual on bicycle", it is normal to supplant "on" with "ride" and surmise "individual riding bicycle on a street" even the "street" isn't clear. In this way, misusing such inclination as a language earlier is required to help the regular encoder-decoder models more outlandish overfit to the dataset predisposition and spotlight on thinking.

In this work[6], Author propose an image subtitling approach in which a generative intermittent neural organization can zero in on various pieces of the information image during the age of the inscription, by abusing the molding given by a saliency forecast model on which parts of the picture are remarkable and which are logical. Authors show, through broad quantitative and subjective tests for enormous scope datasets, that our model accomplishes better execution with deference than subtitling baselines with and without saliency and to various best in class approaches consolidating saliency and subtitling.

In this paper[7], author present "MLADIC", a novel Multitask Learning Algorithm for cross-Domain Image Subtitling. MLADIC is a perform various tasks framework that all the while upgrades two coupled targets through a double learning component: image inscribing and text-to-picture combination, with the expectation that by utilizing the relationship of the two double undertakings, we can upgrade the picture inscribing execution in the target area. Solidly, the picture inscribing task is prepared with an encoder-decoder model (i.e., CNN-LSTM) to create printed depictions of the info pictures. The picture blend task utilizes the contingent generative ill-disposed organization (CGAN) to integrate conceivable pictures dependent on text depictions.

Author propose[8] novel Deep Hierarchical Encoder-Decoder Network (DHEDN) is proposed for picture inscribing, where a profound progressive structure is investigated to isolate the elements of encoder and decoder. This model is able to do productively applying the portrayal limit of profound organizations to intertwine significant level semantics of vision and language in creating inscriptions. In particular, visual portrayals in high degrees of deliberation are at the same time considered, and every one of these levels is related to one LSTM. The base most LSTM is applied as the encoder of printed inputs. The use of the center layer in encoder-decoder is to upgrade the interpreting capacity of top-most LSTM. Moreover, contingent upon the presentation of semantic upgrade module of picture highlight and dispersion consolidate module of text include, variations of structures of our model are built to investigate the effects and shared collaborations among the visual portrayal, literary portrayals and the yield of the center LSTM layer. Especially, the system is preparing under a fortification learning technique to address the presentation predisposition issue between the preparation and the testing by the arrangement slope enhancement.

Late works[9] in image subtitling have demonstrated very promising crude execution. In any case, we understand that the majority of these encoder-decoded style networks with consideration don't scale normally to huge jargon size, making them hard to utilize on implanted framework with restricted equipment assets. This is on the grounds that the size of word and yield inserting networks develop relatively with the size of jargon, antagonistically influencing the conservativeness of these organizations. To address this impediment, this paper presents a shiny new thought in the space of picture inscribing. That is, author tackles the issue of conservativeness of picture inscribing models which is heretofore unexplored. Proposed model, named COMIC, accomplishes tantamount outcomes in five basic assessment measurements with state-of-the-workmanship approaches on both of the MS-COCO and InstaPIC1.1M datasets.

In this paper[10], author propose a structure dependent on scene charts for picture inscribing. Scene charts contain plentiful organized data since they portray object elements in pictures as well as present pairwise connections. To use both visual highlights and semantic information in organized scene charts, we extricate CNN highlights from the jumping box counterbalances of article elements for visual portrayals, and concentrate semantic relationship highlights from significantly increases (e.g., man riding bicycle) for semantic portrayals.

After acquiring these highlights, we acquaint a various leveled attention based module with learn discriminative highlights for word age at each time step. The test results on benchmark datasets show the predominance of our strategy contrasted and a few cutting edge strategies.

III. PROPOSED METHODOLOGY

The process of captioning images can be broken up functionally into two modules, one is an image model that extracts the characteristics and complexities of our image and the other is a linguistic model, converting features and artefacts that are converted into a natural expression in the image based model. Typically using the Convolutional Neural Network algorithm for the image-based model (such as the encoder). And it depend on a Recurrent Neural Network for the language dependent model (viz decoder). Semantic attention has been shown to be effective in improving the performance of image captioning. The core of semantic attention based methods is to drive the model to attend to semantically important words, or attributes. In previous works, the attribute detector and the captioning network are usually independent, leading to the insufficient usage of the semantic information. Also, all the detected attributes, no matter whether they are appropriate for the linguistic context at the current step, are attended to through the whole caption generation process. This may sometimes disrupt the captioning model to attend to incorrect visual concepts. To solve these problems, we introduce two end-to-end trainable modules to closely couple attribute detection with image captioning as well as prompt the effective uses of attributes by predicting appropriate attributes at each time step.

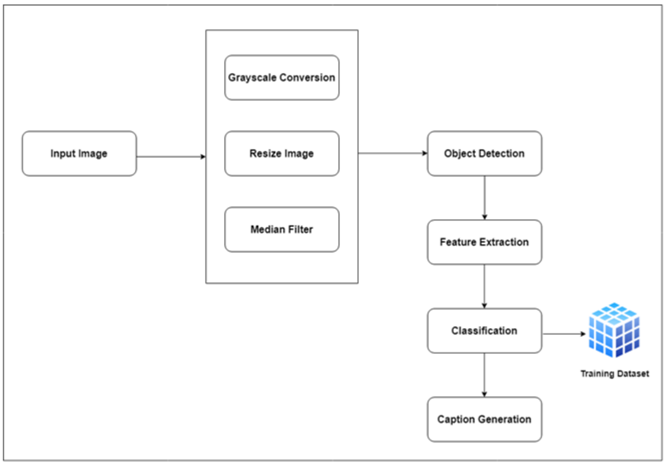

Figure 1. System Architecture

IV. RESULTS AND DISCUSSION

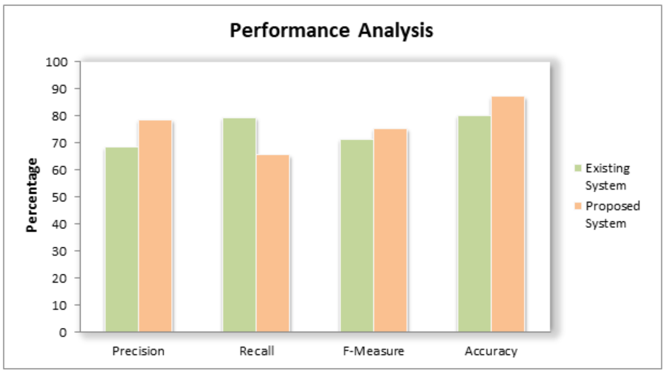

Figure 2. Classification Result

	Existing System(Naïve Bayes)	Proposed System(CNN)
Precision	68.45	78.70
Recall	79.44	65.64
F-Measure	72.11	74.31
Accuracy	80.29	87.26

Table 3. Classification Table

V. CONCLUSION

In this paper, A proposal of a novel deep neural network(NDNN) model to improve the image captioning methods. The NDNN explores the spatio-temporal relationship in the visual attention and learns the attention transmission mechanism through a tailored LSTM model, where the matrix-form memory cell stores and propagates visual attention, and the output gate is reconstructed to filter the attention values. Combined with the language model, both of the generated words and the visual attention areas obtain memory in the space. The embedding of the NDNN model in three classical attention-based image captioning frameworks, and adequate experimental results on the MS COCO and Flicker dataset demonstrate the superiority of the proposed NDNN.

VI. 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.