iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** XII **Month of publication:** December 2023

DOI: https://doi.org/10.22214/ijraset.2023.56796

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue XII Dec 2023- Available at www.ijraset.com

Image and Video Captioning Using Machine Learning

Dhruv Kumar[1], Dhawal Kamble[2], Sayuja Kute[3], Narayan Chavan[4]
Department of Information Technology, R.M.D Sinhgad School of Engineering

Abstract: ***Image based web crawler is the way toward looking through data by utilizing related images. The tremendous assets of images are accessible on the web in that a large number of the images are contain as with named and without named caption. The users are needed to look through the images relying upon their necessities. In that a significant number of the users can't recover the pertinent images as a result of their unpredicted appropriate inscription on their images. Our task is to generate an automatic caption for the images based on the image content. To produce an image caption, firstly, the content of the image should be fully understood; and then the semantic information contained in the image should be described using a phrase or statement that conforms to certain grammatical rules. Thus, it requires techniques from both computer vision and natural language processing to connect the two different media forms together, which is highly challenging. The paper targets producing mechanized inscriptions by learning the contents of the image. At present images are clarified with human intercession and it turns out to be almost unthinkable task for tremendous databases. The picture information base is given as contribution to a deep neural network Convolutional Neural Network encoder for creating caption which extricates the highlights and subtleties out of our image and Recurrent Neural Network decoder is utilized to interpret the highlights and articles given by our image to acquire consecutive, meaningful description of the image.***
Keywords: Deep Learning, part of speech, image captioning, multi-task learning

I. INTRODUCTION

To facilitate researches in areas such as cross-modal retrieval and the assistance of visually impaired people image captioning which aims to link image with language has become a hot research topic. An image captioning model needs to not only recognize the salient objects, their attributes, and object relationships in an image, but also organize these types of information into a syntactically and semantically correct sentence. With the advances of Neural Machine Translation, recent captioning models generally adopt the encoder decoder framework to “translate” an image into a sentence, and promising results have been achieved.

In recent years, researchers have made significant advances in some areas of computer vision understanding, such as image classification, feature classification, object detection and recognition, scene recognition, action recognition, etc. However, having a computer automatically generate natural language descriptions for an image remains a difficult and challenging task. This task connects the two quite different media forms, requiring that computers not only have a correct and comprehensive understanding of the visual content in the image, but also use human language to combine and organize the semantics of the image. The subtasks of image captioning, i.e., identifying semantic elements such as visual objects, object attributes, scenes, are inherently challenging, and organizing words and phrases to express these identified information adds even more difficulty to the entire task.

II. RELATED WORK

J. Lu et al*:* In this paper, 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.

P. Anderson et al: In this work, 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.

L. Chen et al: In this paper, 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.

T. Yao et al: In this paper, 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.

X. Yang et al: Author 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.

M. Cornia et al: In this work, 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.

M. Yang et al: In this paper, 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.

X. Xiao et al: Author propose 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.

J. H. Tan et al: Late works 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.

X. Li et al: In this paper, 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.
Z. Zhang et al: this paper proposes another model dependent on the Fully Convolutional Network (FCN)- LSTM system, which can create a consideration map at a finegrained lattice astute goal. Additionally, the visual component of every network cell is contributed simply by the chief article. By embracing the matrix shrewd marks (i.e., semantic division), the visual portrayals of various framework cells are associated to one another. With the capacity to go to huge territory "stuff", our strategy can additionally sum up an extra semantic setting from semantic marks. This technique can give thorough setting data to the language LSTM decoder. In this way, a component of fine-grained and semantic-guided visual consideration is made, which can precisely interface the significant visual data with each semantic significance inside the content. Shown by three trials including both subjective also, quantitative examinations, our model can produce inscriptions of high caliber, explicitly significant levels of precision, culmination, what's more, variety.
M. Tanti et al: In this paper, author empirically show that it is not especially detrimental to performance whether one architecture is used or another. The merge architecture does have practical advantages, as conditioning by merging allows the RNN's hidden state vector to shrink in size by up to four times. Our results suggest that the visual and linguistic modalities for caption generation need not be jointly encoded by the RNN as that yields large, memory-intensive models with few tangible advantages in performance; rather, the multimodal integration should be delayed to a subsequent stage.

III. EXISTING SYSTEM/OPEN ISSUES

Image captioning's objective is to automatically provide descriptions for a given image, i.e., to capture the relationships between the objects seen in the picture, produce natural language expressions (see an example in Fig. 1), and assess the calibre of the resulting descriptions. that neither do they provide an end-to-end mature general model to address this issue, nor do they provide sensible feature observations on objects or actions in the image.

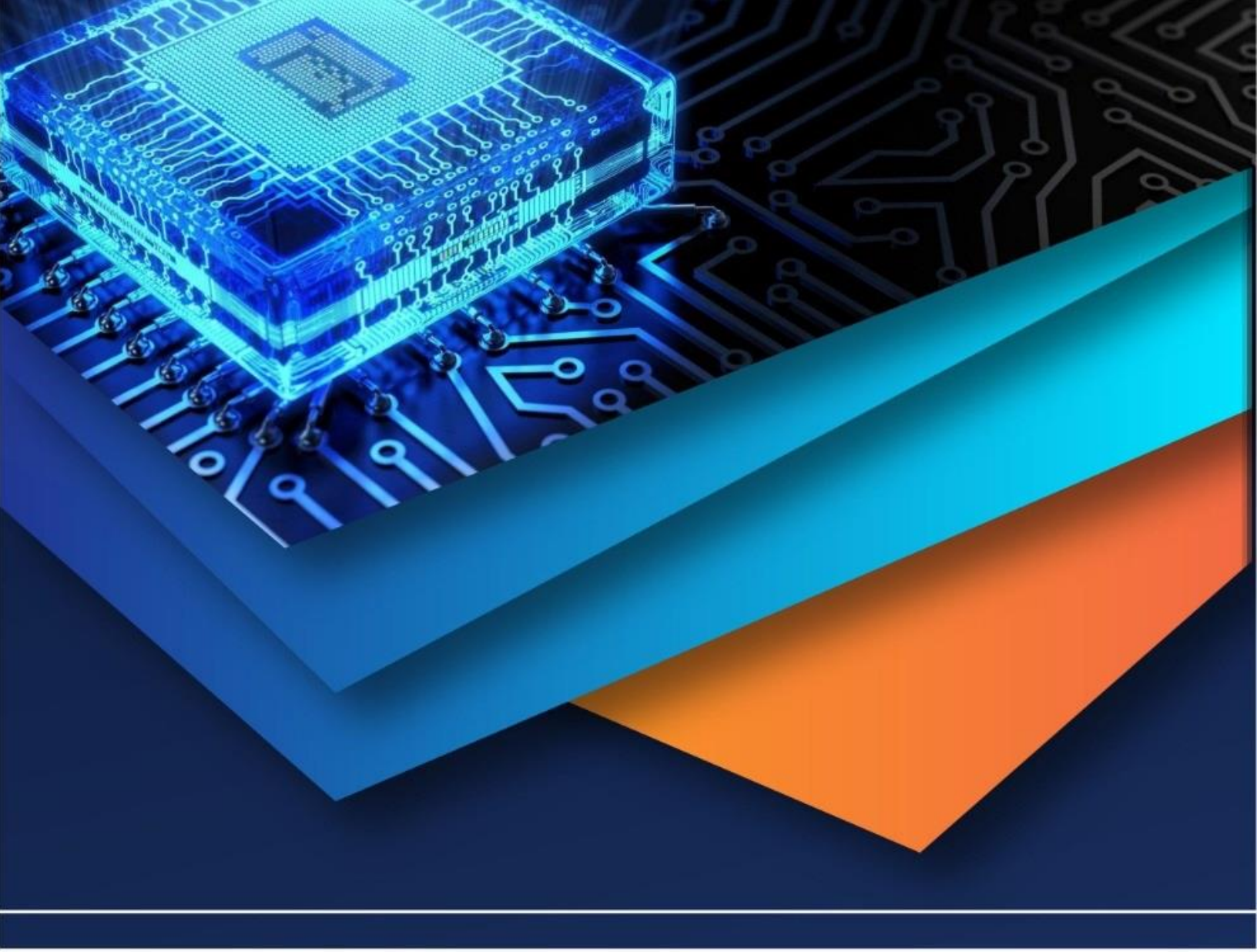
IV. CONCLUSION

In this paper, we propose a novel neural network(NN) model to improve the image captioning methods. The NN explores the relationship in the visual attention and learns the attention transmission mechanism through a tailored 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. We embed the NN 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 NN.

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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY