JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Image Descriptor Using Neural Network

Jeet Sunil Jaiswal [1], Sonali Jadhav[2], Rushikesh Khandagale[2], Prathmesh Ghate[3]

Department of E&TC, SKNCOE, SPPU, Pune

Abstract— **An image description using neural networks involves using a deep learning technique to automatically generate natural language descriptions of images. The process typically involves training a convolutional neural network (CNN) to extract meaningful features from an image and then using a recurrent neural network (RNN) to generate a sentence that describes the image. This approach has shown promising results in producing accurate and descriptive image captions, and has the potential to improve accessibility and understanding for visually impaired individuals, as well as enhancing image search and retrieval systems.**

***Keyword*: CNN Model, RNN Model, Neural Network, Machine Learning, Image Descriptor**

I. INTRODUCTION

The project here imitates the work of generating the label of the given image. The following work is achieved using different algorithms which are "Convolutional Neural Network" and "Recurrent Neural Network". We used many modules like pandas, numpy, json, pickle, keras to support this action. The project is distributed in a systematic distribution of various well-named folders that display their work. The initiative was created to help academics working in fields such as cross-modal retrieval. Image captioning, which combines a picture with words to help the blind or visually impaired, has become a popular field of study. In addition to identifying important objects, their characteristics, and linking objects in an image, the image caption model must also organize this data into a sentence that is syntactically and semantically correct. Caption models use encoder decoder architecture which helps in converting image to caption to get promising results. This is done using advanced neuron machine translation. Recently, computer vision knowledge has greatly advanced in various fields, including image classification, feature classification, object detection and recognition, scene recognition, action recognition, etc.

II. RELATED WORK

A novel versatile consideration model with a visual sentinel is proposed in paper [1]. 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.

Authors propose a joined base up and top down consideration component that empowers thoughtfulness regarding be determined at the degree of items and other striking image areas in paper [2]. 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.

Author present a novel convolutional neural organization named SCA-CNN that joins Spatial and Channel wise Attentions in a CNN in paper [3]. 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.

Algorithm named as 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 is presented in paper [4]. 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.

Scene Graph Auto-Encoder (SGAE) that consolidates the language inductive inclination into the encoder decoder image subtitling structure for more human-like subtitles is proposed in [5]. 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.

A work on 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 is implemented in [6]. 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.

A novel Multitask Learning Algorithm for cross- Domain Image Subtitling is shown in [7]. MLADIS 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.

A 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 in [8]. 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 semantics

A structure dependent on scene charts for picture inscribing is implemented in [9]. 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.

Another model dependent on the Fully Convolutional Network (FCN)- LSTM system, which can create a consideration map at a finegrained lattice astute goal in [10]. 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.

III. SYSTEM ARCHITECTUR

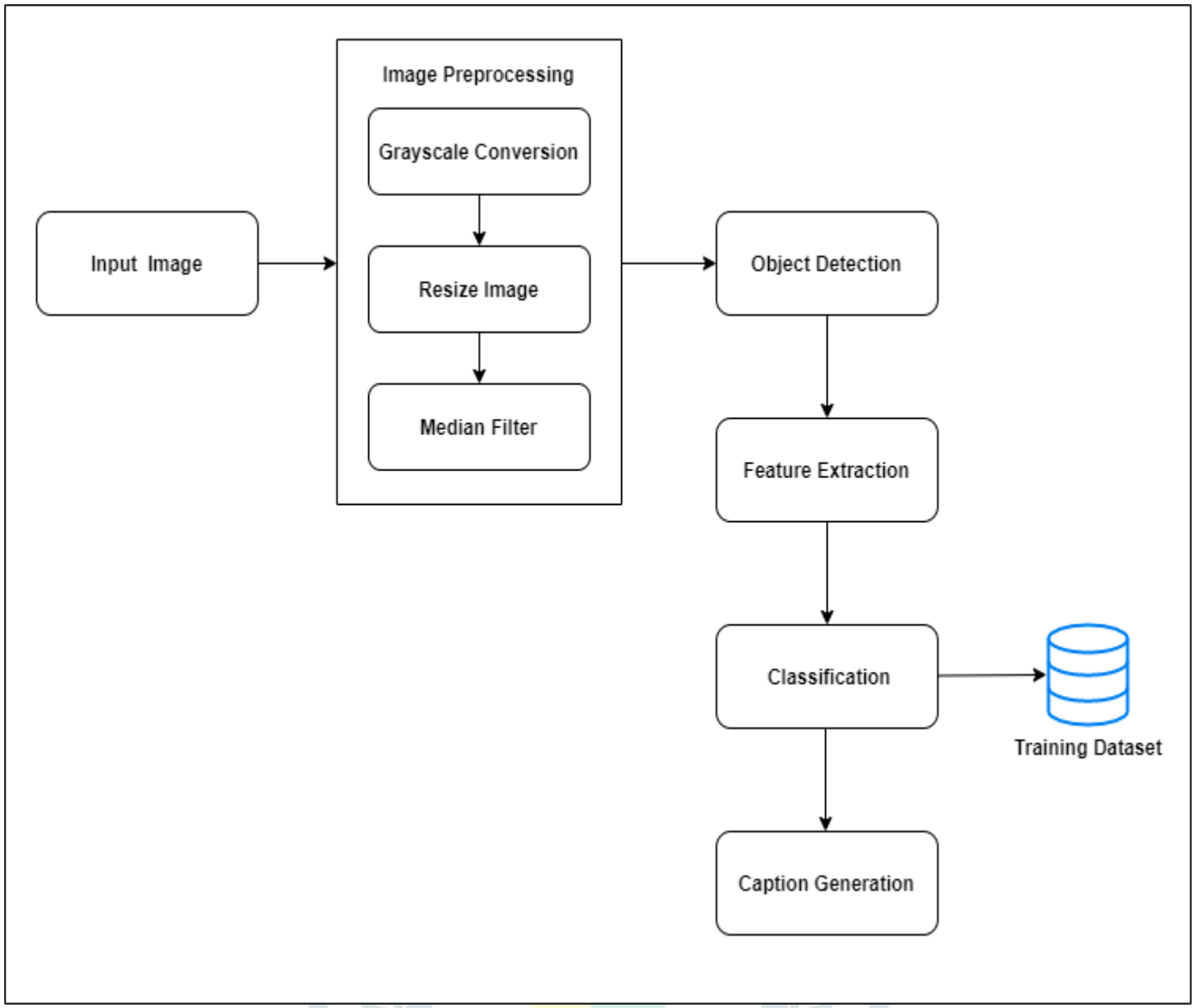

Fig. 1 System Architecture

A. Input Image-

This is the most initial step of the process, in this we provide the image dataset to algorithm for which we want to generate the description

B. Greyscale Conversion-

This step is one of the sub step of image preprocessing, were we perform the initial necessary checks on the image, gray scaling the image helps the algorithm to distinctly identify the object

C. Resize Image-

In this step we resize all the image present in the dataset to a fix size.

D. Median Filter-

We generally perform this step to remove the excess noise and signal from the image. Helps in clearly identification of the image.

E. Object Detection-

As this step suggests all the main objects are identified with performing all the image preprocessing, further also helps in feature extraction.

F. Feature Extraction-

Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data.

G. Classification-

Classification algorithms used in machine learning utilize input training data for the purpose of predicting the likelihood or probability

H. Caption Generation-

In this step, after the all the necessary processing on the image the caption is generated using the algorithm

IV. USECASE

Image descriptor using neural networks is a powerful technology that allows machines to analyze and understand visual content. By training neural networks on vast amounts of image data, it is possible to create accurate and reliable models that can recognize and classify objects, scenes, and other features of images. The applications of image descriptors using neural networks are vast, and they have the potential to revolutionize a wide range of industries, from e-commerce to healthcare. By automating the process of image analysis, image descriptors can save time and resources, improve accuracy, and unlock new possibilities for innovation.

A. Assistive Technology:

Providing descriptions of images can be extremely helpful for visually impaired individuals, allowing them to better understand and interact with visual content.

B. E-commerce:

Automatically generating product descriptions for images can save time and resources for online retailers, while also improving the shopping experience for customers.

C. Social Media:

Describing images can improve accessibility on social media platforms, enabling more people to engage with and understand visual content.

D. Content Management:

Automatically generating image descriptions can also help with content management, making it easier to organize and search through large collections of images.

V. CONCLUSION

A deep neural network (NDNN) model for improving image captioning methods. NDNN explores the relationship in visual attention and learns the attention transfer mechanism through an adapted LSTM model, where a memory cell in the form of a matrix stores and propagates visual attention, and the output gate is reconstructed to filter the attention values. Combined with the language model, both the generated words and visual attention areas acquire memory in space. We embedded the NDNN model into three classical attention-based image description frameworks, and adequate experimental results on MS COCO and Flicker dataset demonstrate the superiority of the proposed NDNN.

ACKNOWLEDGMENT

I would like to express my gratitude to the developers and researchers who have contributed to the development of neural network models for image description. Their dedication and hard work have made it possible to achieve accurate and detailed image descriptions using automated techniques. I would also like to acknowledge the availability of open-source frameworks such as TensorFlow and PyTorch, which have made it easier for developers to experiment with and improve upon these models. Additionally, I would like to thank the large and diverse community of researchers, developers, and enthusiasts who continue to contribute to this field and push the boundaries of what is possible with neural networks.

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 captioningwith 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-domainimage 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," *IEEETransactions 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

[13] G. Hoxha, F. Melgani and B. Demir, "Retrieving Images with Generated Textual Descriptions," *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, Yokohama, Japan, 2019.

[14] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton and C. J. Taylor, "A minimum description length approach to statistical shape modeling," in *IEEE Transactions on Medical Imaging*.

[15] M. Cavazza, R. Green and I. Palmer, "Multimedia semantic features and image content description," *Proceedings 1998 MultiMedia Modeling. MMM'98 (Cat. No.98EX200)*, Lausanne, Switzerland, 1998.

[16] W. Liu and M. Zhang, "Multiple description image coding method based on balanced multiwavelet transformation," *2010 3rd International Congress on Image and Signal Processing*, Yantai, China, 2010.

[17] Shao Hong, Cui Wen-cheng and Tang Li, "Medical Image Description in Content-Based Image Retrieval," *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, 2013.

[18] D. Liu, D. Gao and G. Shi, "A new multiple description image coding scheme based on Compressive Sensing," *2011 IEEE 13th International Conference on Communication Technology*, Jinan, China.

[19] C. Ates, Y. Urgun, B. Demir, O. Urhan and S. Erturk, "Polyphase downsampling based multiple description image coding using optimal filtering with flexible redundancy insertion," *2008 International Conference on Signals and Electronic Systems*, Krakow, Poland, 2012.

[20] J. van de Weijer and C. Schmid, "Blur Robust and Color Constant Image Description," *2006 International Conference on Image Processing*, Atlanta, GA, USA, 2015.