IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DEEP LEARNING APPROACH FOR TRAFFIC CONDITION PREDICTION USING 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,
Amrutvahini College of Engineering, Sangamner, India

Abstract: Using related images in conjunction with an image-based gateway is used for searching the data. It is possible to retrieve the massive of images on the web because many the images are holding without named captions on a variety of websites. It is quite easy for the users to inspect the images in accordance with their requirements. In the sense that a significant number of users are unable to recover the relevant images as a result of their failure to anticipate the appropriate inscription on their images. Based on the quality of the images, it is our responsibility to bring out a self-regulating image caption. First, the details/objects of a picture can distinctly understand, and then it will be represented by a phrase or declaration that is consistent with the grammatical rules that govern the semantic information contained in the image. Because of this, methods for combining computer vision and natural language processing are required in order to join the two distinct types of media together, which is extremely challenging. The paper aims to produce mechanized inscriptions by learning the contents of an image and applying that knowledge. Now, images are clarified only through the intervention of humans, and this proves to be an almost unthinkable task for massive databases. To contribute to a deep neural network, the picture information base is provided. A Convolutional Neural Network (CNN) works like an encoder which is used to generate a caption that extracts the best part or required part from our image, and a Recurrent Neural Network (RNN) works like decoder which is used to translate the extracted highlights from given image in order to obtain a sequential and meaningful description of the image.

***Index Terms* - Traffic dataset, Deep Learning, Image captioning, CNN, RNN**

I. INTRODUCTION

Image captioning, which aims to link images with language in order to ease research in areas such as early education, human-robot interaction and the assistance of visually impaired people, has become a popular research topic. The image captioning model is not only about notice the salient objects, their attributes, and object relationships in an image, but it is mandatory to organizing this information into a sentence that is both syntactically and semantically correct. Considering recent developments in Neural Machine Translation, recent captioning models have tended to use the various frameworks to "express" an image into a sentence, with promising results being achieved.

During the last few years, researchers have made significant advancements in a variety of areas related to understanding the computer vision, along with the picture classification, attribute classification, entity detection, scene identification, action recognition, To the contrary, having a computer impulsively result in natural language information for the image continues to be a hard and demanding task to accomplish. This task connects two very different media forms, necessitating that computer 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, which is a difficult task for computers to accomplish. Inherently difficult are the subtasks of image captioning, namely, recognizing the semantic components such as visual objects their attributes, and scenes. Additionally difficult is organizing words and phrases to express the information that has been identified, which increases the difficulty of the entire task.

II. MOTIVATION AND OBJECTIVES

Image captioning serves as a link between computer vision and natural language processing, bridging the gap between the two. Among those working in artificial intelligence, it has garnered a great deal of attention. Its goal is to generate descriptions for input images that are written in natural language. Besides being useful in applications including such human-machine interaction and content-based image retrieval, image captioning is also useful in systems that aid visually impaired people in their perception of their surroundings.

Objectives are:

i)To recognize objects and Features in the image.
ii)To generate a fluent description using natural language processing.
iii)To improve accuracy using deep learning.

III. Literature survey

J. Lu et al: Using a visual sentinel[1], the author proposes a novel and versatile consideration model for use in various situations. At each time step, our model determines whether or not to take care of the picture (and, if this is the case, which districts) or whether or not to take care of the visual sentinel. In conclusion, the model determines whether or not to look after the picture. With the COCO picture subtitling 2015 test dataset and the Flickr30K dataset, the author tested his strategy.

P. Anderson et al:[2] An integrated bottom-up and top-down consideration component is proposed in this work, which enables consideration to be determined at the level of items and other visually arresting image areas. This is the most common reason for thoughtfulness when it comes to being thought about. In this methodology, Faster R-CNN used for proposes picture areas, each of which has a related element vector, while the top-down component determines the weightings for the various elements in the picture. When this method was used to deal with picture inscribing, the results on the MSCOCO dataset achieving CIDEr/BLEU-4 scores of 117.9 and 36.9 respectively, on the assignment.

L. Chen et al:[3] A novel convolutional neural organisation named SCA-CNN is presented in this paper, which combines Spatial and Channel wise Attentions in a CNN. The SCA-CNN moderately adjust the sentence age setting in multi-layer highlight maps while encoding where (i.e., mindful spatial areas at different layers) and what (i.e., mindful channels) the visual consideration is focused on during picture inscribing. The proposed SCA-CNN design is evaluated on three benchmark picture subtitling datasets: Flickr8K, Flickr30K, and MSCOCO, according to the authors. It has been consistently demonstrated that SCA-CNN outperforms the best-in-class visual consideration-based picture inscribing techniques on a fundamental level.

T. Yao et al:[4] LSTM-A is a novel engineering that integrates ascribes into the effective Convolutional Neural Networks (CNNs) as well as Recurrent Neural Networks (RNNs) picture subtitling system by preparing them in a start to finish manner, as described in this paper. In particular, by coordinating between property relationships and incorporating them into Multiple Instance Learning, the learning of characteristics is strengthened (MIL). The author develops variations of designs by incorporating picture portrayals and properties into RNNs in various ways to investigate the shared but also fluffy connection between them in order to consolidate credits into subtitling. Broad analyses are conducted on the COCO image subtitling dataset, and our system demonstrates significant improvements when compared to cutting-edge profound models.

X. Yang et al:[5] For more human-like subtitles, the author proposes the Scene Graph Auto-Encoder (SGAE), which incorporates language inductive inclination into the encoder-decoder image subtitling structure for a more human-like subtitle appearance. Colloquial expressions and logical deductions are made in conversation as a result of the inductive inclination that we all have. For example, when we see the connection "individual on bicycle," it is natural to substitute "on" with "ride" and infer "individual riding bicycle on a street," even if the word "street" is not specified. It is necessary to use such predispositional thinking as a language earlier in order to assist the regular encoder-decoder models in becoming more outlandishly overfit to the dataset predisposition and to shine a light on thinking.

M. Cornia et al:[6] Using a saliency forecast model to predict which parts of the picture are remarkable and which are logical, the author proposes an approach to image subtitling in which a generative intermittent neural organisation can zero in on various pieces of the information image during the time period for which the inscription is being made. By conducting extensive quantitative and subjective tests on large-scale datasets, the authors demonstrate that our model achieves better execution with deference than subtitling baselines with and without saliency, as well as to a variety of best-in-class approaches that combine saliency and subtitling.

M. Yang et al:[7] A novel Multitask Learning Algorithm for cross-Domain Image Subtitling is presented in this paper by the authors, who call it "MLADIC." Image inscribing and text-to-picture combination are two targets that are being upgraded simultaneously by MLADIC, which is a perform various tasks framework. It is hoped that by leveraging the relationship between the two double undertakings, we will be able to significantly improve the picture inscribing performance in the target area. As a result, the picture inscribing task is thoroughly prepared, using an encoder-decoder model (i.e., CNN-LSTM) to produce printed representations of the information pictures. The contingent generative ill-disposed organisation (CGAN) is used in the picture blend task to incorporate conceivable pictures based on textual depictions into a final picture.\\

X. Xiao et al:[8] When it comes to picture inscription, the author proposes a novel Deep Hierarchical Encoder-Decoder Network (DHEDN). A profound progressive structure is investigated to isolate the elements of encoder and decoder in order to achieve picture inscribing. This model is capable of performing productively by utilising the portrayal limit of profound organisations to intertwine significant level semantics of vision and language in the creation of inscriptions while applying the portrayal limit of profound organisations. It is particularly important to consider visual portrayals with a high degree of deliberation at the same time, and each of these levels is associated with a particular LSTM. As an encoder for printed inputs, the LSTM is used as the base most iteration. This is accomplished by employing a centre layer in the encoder-decoder to increase the interpreting capacity of the top-most LSTM layer. More to the point, depending on whether or not the semantic upgrade module of picture highlight and the dispersion consolidate module of text include are presented, variations of our model's structures are built to investigate the effects and shared collaborations among the visual depiction, literary depiction, and the yield of the centre LSTM layer. To be more specific, the system is preparing under a fortification learning technique in order to address the presentation predisposition issue between the preparation and the testing by enhancing the arrangement slope.

J. H. Tan et al:[9] Recent work in image subtitling has demonstrated very promising crude execution, despite its early stage. However, we are aware that the majority of the encoder-decoded style networks under consideration do not scale normally to large jargon sizes, making them difficult to use on implanted frameworks with limited equipment resources and resources. This is due to the fact that the size of word and yield inserting networks grows in proportion to the size of jargon, which has an antagonistic effect on the conservatism of these organisations. As a solution to this problem, this paper introduces a brand-new concept in the field of picture inscribing: the shiny new thought. In other words, the author tackles the previously unexplored issue of the conservativeness of picture inscribing models, which has received little attention. The COMIC, reaches to the great results in various general assessment calculation using state-of-the-art approaches on the MS-COCO and InstaPIC1.1M datasets, demonstrating its superiority over existing models.

X. Li et al: In this paper[10], the author proposes a structure for picture inscribing that is based on scene charts rather than text. As a result of the way they depict object elements in pictures and the way they present pairwise connections, scene charts contain a large amount of organised data. Visible attributes and semantic data were get together for optical representation and semantic relationship best part from remarkably increased. In order to familiarise a various levelled attention-based module with learn discriminative highlights for word age at each time step, we first acquire the highlights from the previous step. The results of the tests conducted on benchmark datasets demonstrate that our strategy outperforms a small number of cutting-edge strategies.\\

Z. Zhang et al:[11] This paper proposes yet another model that can generate a consideration map at a slender lattice. Along with this, the visible attributes of each network unit was provided solely by the main piece of content. When the visual representations of various framework cells are associated with one another, the matrix shrewd marks (also referred to as semantic division) are used. This helps to traverse a large amount of "stuff" in a short period of time. This technique is capable of providing comprehensive setting data to the language LSTM decoder. Using this approach, it is possible to create a component of finely grain and semantically guided visual consideration, our model demonstrated that it is capable of producing inscriptions of high quality, with explicitly significant levels of precision and culminating in a wide range of variety.\\

M. Tanti et al: This paper[12] empirically demonstrates that the use of one architecture over another does not have a significant negative impact on efficiency. Conditioning by merger reduces the hidden state vector of the RNN by four orders of magnitude, and this combination architecture has several functional advantages. Our findings demonstrate that the visual and linguistic modalities for the sub-specified generation do not require coding by the RNN because they provide broad, memory-intensive models with little discernible performance advantage over the other modalities.

IV. Proposed Methodology

This can be broken down into two functional modules: the first is an image model that extracts the characteristics and complexities of our image; the second is a linguistic model that converts features and artefacts in our image-based model into natural expressions.

For our image-based model, we typically use a Convolutions Neural Network algorithm as a foundation (such as the encoder). We also rely on a Recurrent Neural Network for our language-dependent model, which is described below (viz decoder).

Modules Split up:

1. Image to be used as an input:
In this section, we will upload the Input Image.

2. The second step is image pre-processing.
It is in this step that we will apply image pre-processing techniques such as grayscale conversion and image noise removal.

3. Image Feature Extraction-
This step will involve the edge detection techniques and applied on the image in order to extract the image attributes.

4. Image Classification (or categorization):
In this step, we will put the image classification methods into practice.
5. As a result of this:
This step will display the result of the caption generation process.

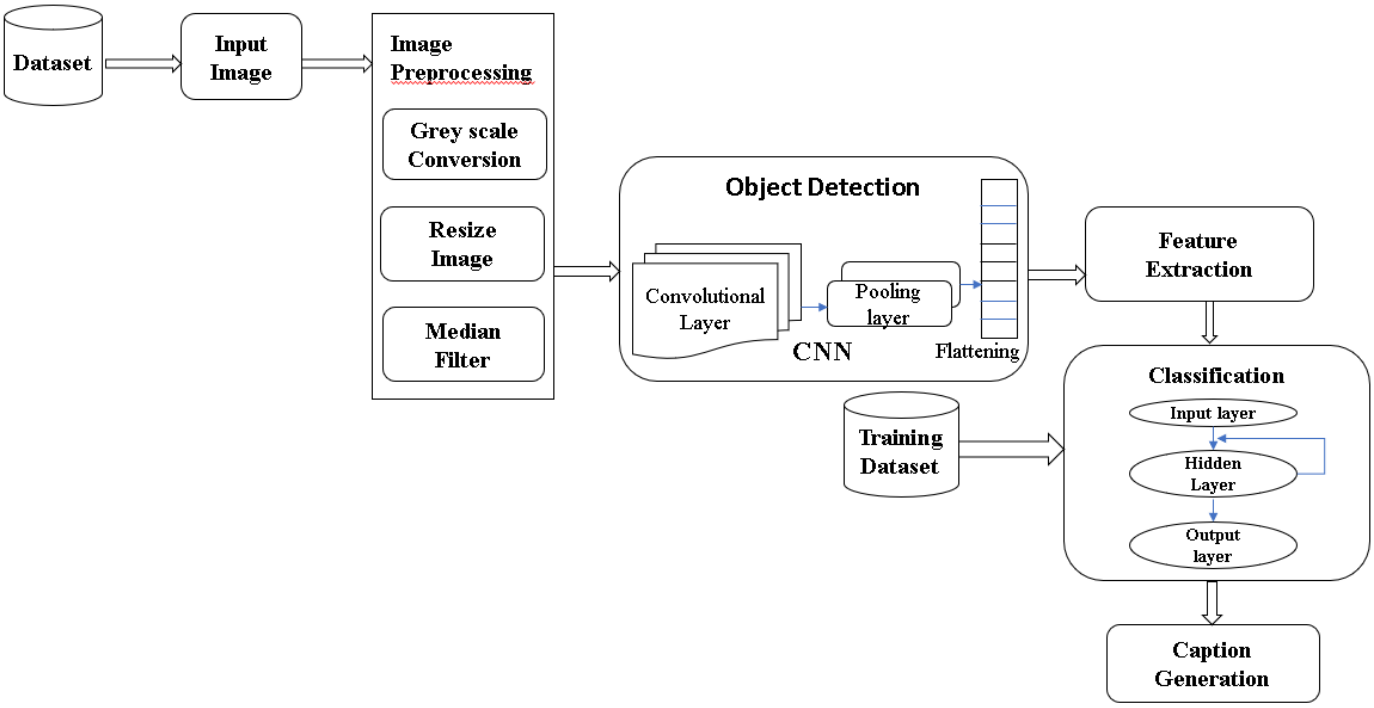

Fig1.System Architecture

V. ALGORITHMS

CNN (Convolution neural network)

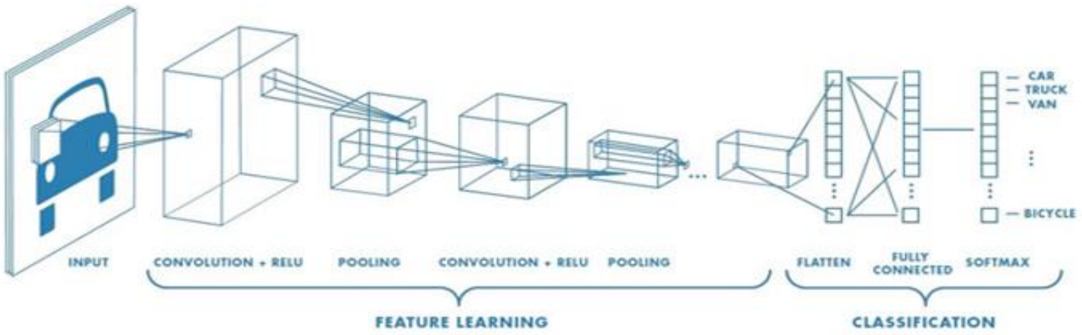

Fig2. Convolutional Neural Network

Convolution Layer -

Convolution (image) is the first layer that is used to remove features from an image (image). Convolution is a technique for learning image attributes that preserves the relationship between pixels by using small input data squares. Applied filters, such as a sharp box and a Gaussian whirlwind filter, can be used to generate an image with various filters, such as recognition, edge detection, sharpening, and image bleeding, among others.

Pooling Layer -

It is possible that the number of parameters will be reduced if the number of photographs is excessive. Spatial pooling, also known as sampling, is a technique for reducing the dimensionality of each map while still retaining relevant information.\\

Fully Connected Layer -

The map matrix for this layer has been vector converted (x1, x2, x3,...). These characteristics have been combined to create a model with layers that are completely interconnected.\\

Softmax Classifier -

Finally, we have the softmax or sigmoid activation to categorise the outputs.\\

RNN (Recurrent neural network)

The recurrent neural network (RNN) works as follow: The previous phase output serves as an input for the next step. The inputs and outputs of regular neural networks are autonomous, but the words that came before them are important in situations where it is appropriate to forecast the next word in a phrase, and so the words that came before them must be kept in mind. This resulted in the development of the RNN and the resolution of the problem through the use of a secret layer. The primary and most significant feature of RNNs is the presence of a hidden state that retains some information about the sequence under consideration. \\
Steps:

1)The network receives a single phase of the input.
2)Calculate the current condition using the current input set and the previous state.
3)For the next step, the current ht is ht-1.
4)You may take as many time measures depending on the issue and include input from all the previous countries.
5)Upon completion of all the time stages, the final current state is used to determine the result.
6)The output is then compared with the real output (i.e. the destination output and error).
7)The failure is then retransmitted back to the network, so the network (RNN) is qualified.

Formula for calculating current state:

$$h_t = \int (h_{t-1}, x_t) \qquad \ldots(1)$$

where,

h_t =current state
h_{t-1}=Previous state
x_t = Input state

Formula for applying Activation function:

$$ht = \text{activation}\,(W_h h\; h_t - 1 + W_x h\; X_t) \qquad ..(2)$$

where,
W_h h= Weight at recurrent neuron
W_xh = Weight at input neuron

Formula for calculating output:

$$Y_t = W_h Y\; h_t \qquad \ldots(3)$$

where,

Y_t =Output
W_hY =Weight at output layer

VI. RESULT ANALYSIS AND DISCUSSION

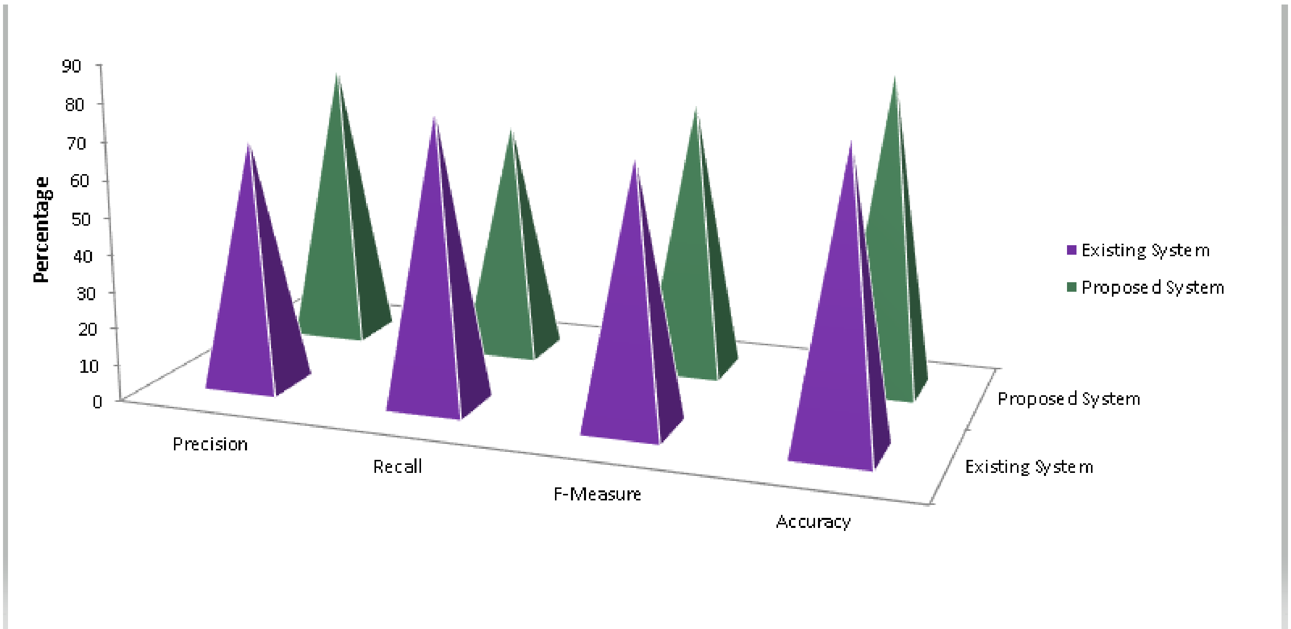

Fig3.Performance Analysis

Evaluation Metrics	Random Forest	Neural Network
Precision	69.05	79.19
Recall	81.44	63.64
F-Measures	75.11	79.31
Accuracy	81.02	89.11

Table 1: Comparison based on Evaluation Metrics

VII. CONCLUSION

To improve the captioning methods for images, we propose a novel deep neural network (NDNN) model in this paper. With the help of a LSTM model, decoding of images becomes easier. This paper embeds the novel deep neural network model based on CNN and RNN and state of the art results on various benchmark with the help of MS COCO and Flicker datasets.

VIII. ACKNOWLEDGEMENT

I thank my guide Prof. R.L. Paikrao for analyzing the data and for advising on all aspects related to this paper, along with other concepts too.

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.