Image Captioning Using R-CNN & LSTM Deep Learning Model

Aditya Kumar Yadav
UG Scholar
PSG College of Technology

Prakash.J
Assistant Professor
Department of Computer Science & Engineering
PSG College of Technology

Abstract:- Image Captioning is the process of creating a text description of an image. It uses both Natural Language Processing (NLP) and Computer Vision to generate the captions. The image captioning task is done by combining the detection process when the descriptions consist of a single word like cat, skateboard, etc. and Image Captioning when one predicted region covers the full image, for example cat riding a skateboard. To address the localization and description task together, we propose a Fully Convolution Localization Network that processes a picture with a single forward pass which can be consistently trained in a single round of optimization. To process an image, first the input image is processed using CNN. Then Localization Layer proposes regions and includes a region detection network adopted from faster R-CNN and captioning network. The model directly combines the faster R-CNN framework for region detection and long short-term memory (LSTM) for captioning.

***Keywords**:- Image Caption, Recurrent Neural Network, Long short-term memory, Convolution Neural Network, Faster R-CNN, Natural Language Processing.*

I. INTRODUCTION

Creating captions is an important task that is relevant to both computer vision and natural language processing . Limitation of human ability for writing pictures through a machine is itself a remarkable step along the line of artificial intelligence [1-2]. The motivation of our project is to capture how objects in the image are related to each other and express them in the English language (generate captions)[3-4].

Many image captioning techniques are used to translate raw image files. These techniques do not translate the exact context of the image and therefore cannot be used to gain insight into the image files. A new system is proposed for image captioning which includes usage of contextual information and thereby accurately describe the image [5-8].

This work aims to generate captions using neural language models. The number of proposed models for the image labeling task has increased significantly since the models of the neural language, ANN and the folding neural networks (CNN) became popular [9-16]. This paper is based on one of these works, which combines a variant of a neural network from Recurrent to a CNN. The main objective of is to improve the model by making minor architectural changes and using phrases as elementary units instead of words, which can lead to better semantic and syntactic subtitles.

II. IMAGE CAPTIONING APPROACHES

2.1 Retrieval Based Image Captioning

Retrieval based image captioning technique is one of the traditional ways of producing captions for the image. An input image is given as input, retrieval based captioning methods generate a caption by retrieving one or many sentences from the pre-loaded pool of sentences. The caption that was generated can either be the exact same sentence that was pre-loaded or a sentence that is framed from the retrieved ones.

The disadvantage of this approach is that a large pool of images is required to make captions from the allocated set of sentences for those images. The output captions generated from this image captioning technique does not have any grammatical errors, but they are the sentences that were the captions for other images. In some of the cases, the captions may be irrelevant to the input image. The major disadvantage of this method is that they have limitations on their capability to describe the images in their own unique way.

2.2 Template Based Image Captioning

Its different kind of methods which is early used and it’s based on templates. Based on the templates, the captions are generated by a syntactic and semantic restricted process. Certain visual concepts must be recognized for generating a description of an image using the template-based image captioning method. The visual concepts that were recognized are then connected by a sentence. For the creation of a sentence without any grammatical errors, optimization algorithms are required.

The disadvantage of this method is that the generation of descriptions is done under the template-based framework which is strictly limited to recognized image content of visual models, typically with a low number of visual models available. There are usually limitations on the coverage, creativity, and complexity of the sentences generated.

2.3 Image Captioning Methods based on Deep Learning

2.3.1 Visual Space and Multimodal Space

Image-based subtitle methods based on learning can create subtitles from both the visual space and the image multimodal space. In visual space-based image captioning method, the features of the image and their captions are sent to the speech decoder, whereas in a multimodal space-based image captioning method, the multimodal space is shared where the device learns the image and generates captions. This process also happens through the speech decoder. With the multimodal representation, the speech decoder generates the caption for the image.

2.3.2 Supervised Learning and Deep Learning Methods

In supervised learning, the training data is given the desired output called a label and in unsupervised learning, the training data is given the desired output called the unlabeled data. Unsupervised learning uses GAN - Generative Adversarial Network. Reinforcement learning, which is the other type of learning along with the GAN approach are used for produced captions for the image.

The below-given steps are followed to produce captions.

- Captions are generated using the CNN and RNN-based combined networks.
- To evaluate the captions and to send that response to the first combined network, another CNN and RNN-based combined network is used.

2.4 Dense Captioning Methods

The image captioning techniques discussed earlier can generate only a single caption for the given input image. Using different areas in the image, they get information about different objects. Though they have information about many regions of the image, they do not generate a separate caption for all the regions. The dense captioning method generates captions for all the regions that were detected in the image.

The following steps are used by the dense captioning technique.

- For different regions of the image, region suggestions are created.
- To obtain information about the detected regions, CNN is used.
- The result obtained from the step (2) is used by NLP to generate captions for each region.

III. IMAGE CAPTIONING APPROACH

Various stages of captioning the image using the deep learning model is illustrated in the Fig.1 where the given input image is loaded and features are extracted from the image then the caption text file is loaded and data pre-processing and tokenization is done. After tokenization, mapping of the image with its caption is stored in a separate file. Then the image is encoded using a pre-trained model which is used for image classification (VGG model). After encoding, train the model and then evaluate it. After this, display the most matched caption as the output.

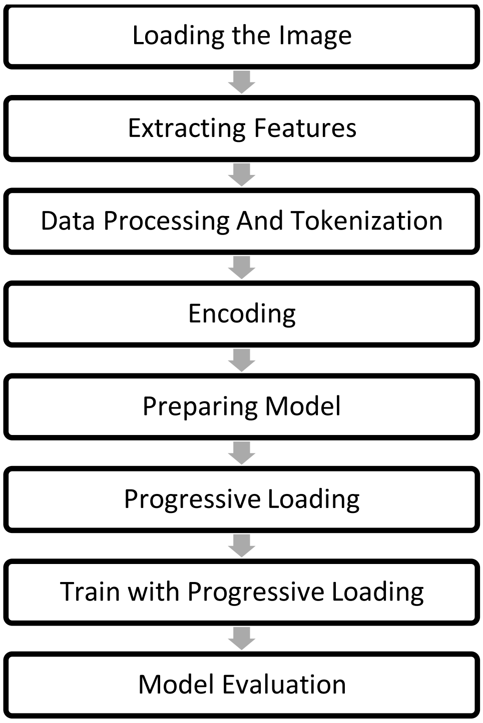

FIG 1. STAGES OF CAPTIONING THE IMAGE USING THE DEEP LEARNING MODEL

3.1 Loading the Image

Loading of the image is done using keras. It provides a function through which we load the images as a matrix, and then the matrix is transformed to a NumPy array through keras and loads the images in a fixed dimension as specified in a VGG model.

3.2 Extracting Features

Next stage is to extract features from the loaded images. Pre-trained model is used to understand the key elements of the photos. VGG model is used to obtain the required weights and Feature extraction takes place.

3.3 Data Processing and Tokenization

Descriptions required be tokenizing and handling easily. Cleaning can be done by converting to lowercase, removing punctuations, removing unnecessary words like 'a', removing numbers. If vocabulary is small, then it model trains faster but it is not expressive. The image identifiers and descriptions are saved to a new file.

3.4 Encoding

For processing, the words need to be encoded, since we are going to use probability, Keras is used to prepare the descriptions. Firstly, we map the image identifiers to the present descriptions. At the end of this process, we get shape of data for the model that is 8092 photos, 28 words (length of description) and 4485 words(vocabulary dataset).

3.4 Preparing Model

A simple model that generates the sentence word by word is used. The input for the image is **the image and the just predicted word** and model is called recursively it uses the already predicted words to generate the new words and it uses input and output pairs and the new words are predicted by probability.

3.5 Progressive Loading

If the computing facility doesn't have enough processing capability, the photos and descriptions can be loaded in a progressive way. Using Keras, we can load in a progressive way through a function. It is used to create batches of samples for the training of model and the generator gives an array in the form of input and output for the model. And the input is the input images and encoded word sequences which are in form of array. The hot encoded words are the output.

3.6 Train with Progressive Loading

Data generator along with a function on the model is used to train it. Model is saved after each epoch (Epoch is one complete representation of data set), Models are created and the model with lowest loss is selected.

3.7 Evaluation:

Once the model is created we evaluate the model. BLEU (Bilingual Evaluation Understudy Score) [8] is used for evaluating the model. It gives information about the quality of text. It compares the generated sentence to the reference sentence. The Bilingual Evaluation Understudy Score of the model is determined and the model is resulting in better score. The BLEU score obtained is provided in Fig.2.

Bilingual Evaluation Understudy Score			
BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**
0.548790	0.296703	0.195520	0.082717

FIG 2. BILINGUAL EVALUATION UNDERSTUDY SCORE FOR THE MODEL

The sample test result is provided in Fig.3, where an input image of a dog walking on the water is provided to the model and the model predicted the caption as dog is running through the water.

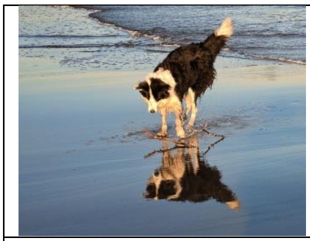	Dog is running through the water
Test Input Image	**Predicted Caption**

FIG 3. PREDICTED CAPTION FOR THE GIVEN INPUT IMAGE

IV. CONCLUSION

The caption for the given input image is produced using deep learning model. Initially the images were preprocessed and the text in order to train a deep learning model. And designed and trained a deep learning image caption generation model. Then, evaluated the train caption generation model using which produced captions for new images that are given as input apart from the loaded images.

REFERENCES

[1]. Jiuxiang Gu, Gang Wang, JianfeiCai, and Tsuhan Chen. 2017. An empirical study of language cnn for image captioning. In Proceedings of the International Conference on Computer Vision (ICCV).

[2]. Rajesh, R., & Mathivanan, B. (2017). Predicting Flight Delay using ANN with Multicore Map reduce framework, Communication and Power Engineering. In Communication and power engineering. Walter de Gruyter GmbH & Co KG

[3]. Kumar, Akshi. (2017). A survey of evolution of image captioning techniques. International Journal of Hybrid Intelligent Systems. 14. 1-19. 10.3233/HIS 170246.

[4]. Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image sentence embeddings using large weakly annotated photo collections. In European Conference on Computer Vision. Springer.

[5]. DzmitryBahdanau, Kyunghyun Cho, and YoshuaBengio. 2015. Neural machine translation by jointly learning to align and translate. In International Conference on Learning Representations (ICLR).

[6]. Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. 2011. Learning photographic global tonal adjustment with a database of input/output image pairs. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE.

[7]. Chuang Gan, ZheGan, Xiaodong He, JianfengGao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[8]. Abhaya Agarwal and AlonLavie. 2008. Meteor, m-bleu and m-ter. In Proceedings of the Third Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2008, 115–118.

[9]. R Sandhya, J Prakash, B Vinoth Kumar (2020). Comparative Analysis of Clustering Techniques in Anomaly Detection in Wind Turbine Data. Journal of Xi'an University of Architecture & Technology, Vol. 12 No.3 2020, pp. 5684-5694.

[10]. Prakash J., & Vinoth Kumar B. (2021). Exploration of computational intelligence insights and data analytics to combat COVID-19. Advances in Data Mining and Database Management, 373-383. https://doi.org/10.4018/978-1-7998-3053-5.ch018

[11]. Sivanandhini, P., & Prakash, J. (2020). Crop Yield Prediction Analysis using Feed Forward and Recurrent Neural Network. International Journal of Innovative Science and Research Technology, 5(5), 1092–1096. doi:10.38124/volume5issue5

[12]. Prakash J. (2018). Enhanced Mass Vehicle Surveillance System, J4R, Volume 04, Issue 04 ,002, 5-9, June2018.

[13]. Prakash, J., & Joy, E. J. (2020). A comparison of different surrogate models for delamination detection in composite laminates using experimental modal analysis. PROCEEDINGS OF ADVANCED MATERIAL, ENGINEERING & TECHNOLOGY. https://doi.org/10.1063/5.0019366

[14]. Sivanandhini, P., & Prakash, J. (2020). Comparative Analysis of Machine Learning Techniques for Crop Yield Prediction. International Journal of Advanced Research in Computer and Communication Engineering, 9(6), 289-293. https://doi.org/10.17148/IJARCCE.2020.964

[15]. Prakash, J., Vinoth Kumar, B., & Shyam Ganesh, C. R. (2020). A comparative analysis of deep learning models to predict dermatological disorder. Journal of Xi'an University of Architecture & Technology, 12(11), 630-639. https://www.xajzkjdx.cn/gallery/64-nov2020.pdf

[16]. Prakash J., Kumar B.V. (2021) An Empirical Analysis of Hierarchical and Partition-Based Clustering Techniques in Optic Disc Segmentation. In: Sharma H., Saraswat M., Kumar S., Bansal J.C. (eds) Intelligent Learning for Computer Vision. CIS 2020. Lecture Notes on Data Engineering and Communications Technologies, vol 61. Springer, Singapore. https://doi.org/10.1007/978-981-33-4582-9_7