IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A NOVEL APPROACH FOR IMAGE CAPTION GENERATOR USING CNN-RNN METHODS

[1]CH.Jeevana Jyothi, [2]G.Gayathri, [3]Ch.Bhargavi, [4]B.Jaswanth, [5]Mr.J.Karthik
[1]Student, [2]Student, [3]Student, [4]Student, [5]Assistant Professor

[1]Computer Science and Engineering,
[1]Gudlavalleru Engineering College, Gudlavalleru, India

Abstract: With the development of deep learning, the combination of computer vision and natural language process has attracted a lot of attention in the last few years. Caption embedding is representative of this file, which enables the computer to learn to use one or more sentences to understand the visual content of the image. The efficient process of producing high-definition semantics is required not only for the recognition of an object and the scene, but also for the ability to analyze the state, attributes and relationships between these objects. Although captioning is a difficult and difficult task, many researchers have found significant improvements. In this project, we describe in detail the three ways to insert image captions using deep neural networks CNN-RNN based, CNN-CNN-based framework and based on reinforcement. We then present the work representing the top three methods in a row, explain the metrics for testing and summarize the main advantages and disadvantages.

Over the past few years, the problem of making automated descriptive sentences for photographs has gained a growing interest in natural language research and computer-based research. Image captioning is an important task that requires a basic understanding of images and the ability to produce descriptive sentences with appropriate and appropriate structure. In this study, the authors propose a hybrid system that uses the multilayer Convolutional Neural Network (CNN) to produce visual vocabulary and Long Short Term Memory (LSTM) to organize sound sentences using generated keywords. A convolutional neural network compares the image to a large database of training images, and creates an accurate description using trained captions.

Index Terms-Image Caption, CNN, RNN.

I. INTRODUCTION

Often, interpreting visual content in natural languages is a basic and challenging task. It has a great potential. For example, it can help visually impaired people to better understand the content of images on the web. Also, it can provide accurate and accurate photo / video information in situations such as photo sharing on social networking or video surveillance systems. This project accomplishes this task using deep neural networks. By learning information from a couple of pictures and captions, the method can produce captions that are usually descriptive and correct in grammar.

Given a new image, the image caption algorithm should produce a description of this image at the semantic level. With the caption function of the image, people can easily understand the content of the image and put it in the form of natural language sentences according to specific needs; however, in computers, it requires the combined use of image processing, computer vision, natural language processing and other major areas of research results. The challenge of captioning an image is to design a model that can fully use image details to produce rich image captions as humans. The effective process of producing high-definition semantics in a picture requires not only the understanding of objects or image recognition in the image, but also the ability to analyze their regions, understand the relationship between them and form the appropriate mental and intellectual sentence. It is not yet clear how the brain understands imagery and how to process visual information into captions. The captioning of an image involves a deeper understanding of the world and what are the essential elements of the universe.

II. LITERATURE SURVEY

[1] Ming and Badrachalam, scientists have tried to increase the availability of inaccurate messages, as well as the net, in information analysis. We used in-depth learning models supported by direct interactions between the neural network (FNN), long-term memory (LSTM). [2] A. Kojima used case structure, practical understanding, and action patterns to produce descriptions of human activities in a consistent environment. [3] P. Hede suggested a caption technique for the image, including a series of words stored on a website called dictionaries, which could render captions with static image content, but the method failed to extract captions for real-world contexts. An image caption generator system was developed that uses deep neural networks to produce captions. [4] A. Farhadi proposed a data capture system based on the acquisition of information, in which points were made to everything in the image and points were compared to other images to produce captions. [5] UM. Hodosh has proposed a standard captioning system, where captions are generated with the help of a standard captioning system. [6] Y. Yang proposed a sentence-making strategy, which uses verbs, nouns, and adverbs to form a semantic sentence, the image is obtained with the help of trained finders and the English corpus was used to measure the image

.III . METHODOLOGY

To include a caption, there are three modes such as CNN-RNN-based framework, CNN-CNN-based framework, and reinforcement-based framework and with different accuracy effects.

CNN-RNN-based framework:

In the human eye, the image contains different colors to create different scenes. But from a computer standpoint, most images are painted in pixels on three channels. However, in a neural network, various data modes all go to great lengths to create a vector and perform subsequent tasks on these features. It has been convincingly demonstrated that CNNs can produce rich image representation by inserting it into a vector of fixed length, so that this representation can be used for a variety of viewing functions such as object recognition, detection and segmentation. Therefore, captioning methods based on encoder-decoder frameworks often use CNN as an image encoder. The RNN network acquires historical information through continuous coverage, with better training capabilities and can do better than mining in-depth knowledge of languages such as semantics and syntax information entered in alphabetical order. With the interdependence between different local names in historical information, a recurring neural network can easily be represented in a hidden layer state. In the captioning function of an image based on the encoder-decoder framework, the coding part of the CNN model is for extracting image features. It can use models like AlexNet, VGG, GoogleNet and ResNet. In the decoder section, the frame adds the word vector expression to the RNN model. For each word, it is first represented by a single hot vector, and then using an embedded model, it becomes the same size as the image element. The captioning problem of the image caption can be explained by the binary form (I, S), where I represent the graph and the S sequence of identified words, S = {S1, S2 ···} and Si the word from the data set to extract. The goal of the training is to increase the limitations of the target definition p (S | I) for the purpose of the produced statement and the target statement is very similar.

Mao et al developed a multimodal Recurrent Neural Network (m-RNN) model that skillfully integrates the CNN and RNN model to solve the problem of image captioning. Due to the disappearance of the gradient and the limited memory RNN memory problem, the LSTM model is a special type of RNN model structure that can solve the above problems. Adds three units of control (cell), which is the input gate, the output and the forgotten. As the data enters the model, the data will be judged by cells. Information that meets the rules will be omitted, and non-compliant information will be forgotten. However, there are also two major barriers to image capturing work, which encourage further important research. The first is that the metrics used for testing and training loss are different. Second is that in training, the input of each step comes from the actual caption and when done, each word made is based on a pre-formed word; If the word does not produce correctly, it can be far from the truth of the ground.

CNN-CNN-based framework:

This framework contains three key elements such as the RNN process. First and last word embedding elements in both cases. However, while part of the facility consists of LSTM or GRU (Gated Recurrent Unit) units in the RNN case, encrypted conferences are compiled in a CNN-based manner. This, unlike the RNN, goes ahead with the feed without repetitive function. Aneja et al has shown that the CNN-CNN framework has a faster training time with a certain number of parameters, but the loss is higher for CNN than RNN. The reason for the accuracy of the CNN model is that CNN is charged with producing a possible distribution of a very high profile name. However, the smallest distribution is not so bad, where the guesswork of many words is possible to predict the various captions, shown in Figure.1

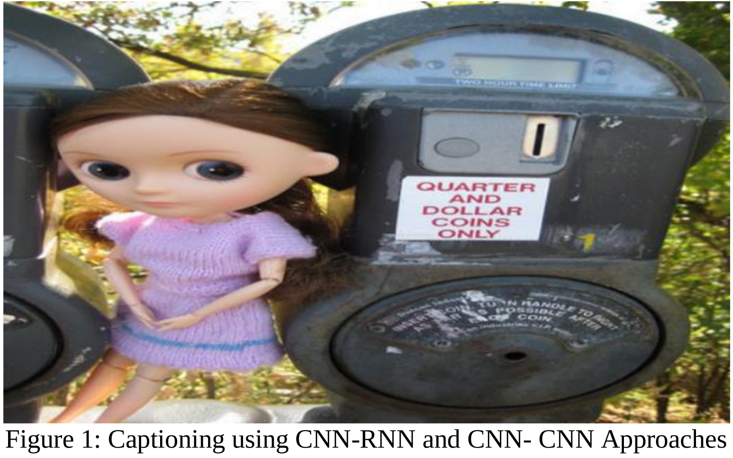


Figure 1: Captioning using CNN-RNN and CNN- CNN Approaches

In the above figure are captions from CNN-RNN "Parking meter with mark" and captions from CNN-CNN say "Doll is sitting next to the parking meter", but the world truth says "Doll with joints said looking at its space between meters two parking lots"

Reinforcement framework:

Learning reinforcement is widely used in sports, theory control, etc. Control or play problems have specific goals for natural improvement, and defining a proper use policy is not essential for wording. When you apply a photo-enhancing learning enhancement, the production model (RNN) can be viewed as an agent, interacting with the external environment (words and context vector as always input). The agent's parameters define the policy, the function of which causes the agent to select an action.

In the sequence generation sequence, the verb refers to predicting the next word in sequence each time. After taking the action the agent restores its internal state (hidden units of RNN). When the agent reaches the end of the sequence, he sees the prize. In such a framework, the RNN decoder acts as a stochastic policy, in which the choice of action is tantamount to producing the next word. During training the PG method selects actions according to the current policy and looks only at reward at the end of the sequence (or after the maximum length of the sequence), by comparing the sequence of actions from the current policy against the correct action sequence. The goal of training is to find agent parameters that increase the expected reward.

The idea of using a PG (gradient policy) to further the diverse objectives of captioning was first suggested in MIXER paper, by holding a standard sentence selected as a reward signal in a reinforced learning environment. In the MIXER method, since the problem-creating set of text production has too much action to make the problem difficult to read about in the first random policy, it takes steps to train RNN with cross-entropy loss of a few epochs using True Sequence. This is a new form of training that combines MLE with the goal of strengthening.

IV. IMPLEMENTATION

To model first we must collect the data and perform the data processing in advance. Later we have to split the database into a train and a test set. Also use algorithms in it and predict the outcome.

4.1 Data Collection:

We use Flicker8k images and text data to create our data sets. Data sets have a rich set of images and each image has at least five captions. We download all the images and save them to google drive for further processing of the project.

4.2 Image Featuring: Contains

1) Image processing: Image files are stored on google drive and we upload images to our google colab operating environment. In image processing, we turn the image into a finished shape and size. An image is identified by 3 channels because all images must have 3 channels.

2) Extract Feature: In the case of automatic production of definition, the main feature is the content of the object in our image. To remove the features from the image, we use a pre-trained VGG16 model design and remove the last layer that is fully connected and after that the rendering of the VGG16 model layer is a key element of our project.

3) Encoder: When the image is transferred to the CNN layer, we receive the enclosed image form. CNN can produce an input image presentation by inserting it into a long vector. Such a presentation can be applied to various computer viewing activities.

4.3. TEXT PLANNING: Contains

1) Text Processing: In text processing, we clean up unnecessary data. We remove the special character / token contained in the caption and end with a combination of text and a number like 'sunana44'. All words in the title are included.

2) Token and Vectorize: We split the caption text with space and find all the different words for each topic. We should measure vocabulary size to save memory size. We include two tokens for each topic 'start' and 'end'. All caption tokens are vectorized.

3) Decoder: The LSTM (Long-Term Memory) network is used as a decoder to provide the RNN (Recurrent Neural Network) token shown to indicate the beginning of a sentence and the token indicates the end of a sentence. The model learns to predict captions which are targeted variables. Captions are predicted word for word. Therefore, each word is entered into a computer of a fixed size.

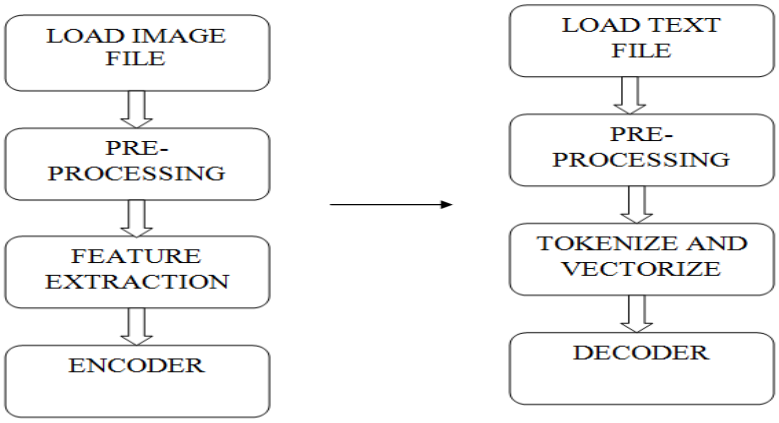


Figure 2: Image and Text processing

V. RESULTS AND DISCUSSIONS

The results and discussions in many ways are as follows

CNN–RNN Method:

By CNN-RNN, the problem of long-term dependence on the neural network can be solved. Vinyals et al. has developed a NIC (Neural Image Caption) model that takes an image as embedding in the encoder component and produces descriptive definitions for LSTM networks in the decoder section. The model effectively solves the problem of finding natural language sentences very well. It is very important to use computers that interact with the native language, which makes computer processing no longer remain at the simplest level of simulation, but progresses to the level of semantic comprehension. At this point, image elements are considered as powerful portable elements combined with weight information. The first method of attention was proposed in 2001, suggesting “soft care” which means selecting regions based on different weights and “hard attention” that pays attention to a particular point of view. Test results obtained using deep care-based networks have achieved amazing results. The use of attention-grabbing equipment enables the model to produce each word according to the corresponding image region as shown in figure

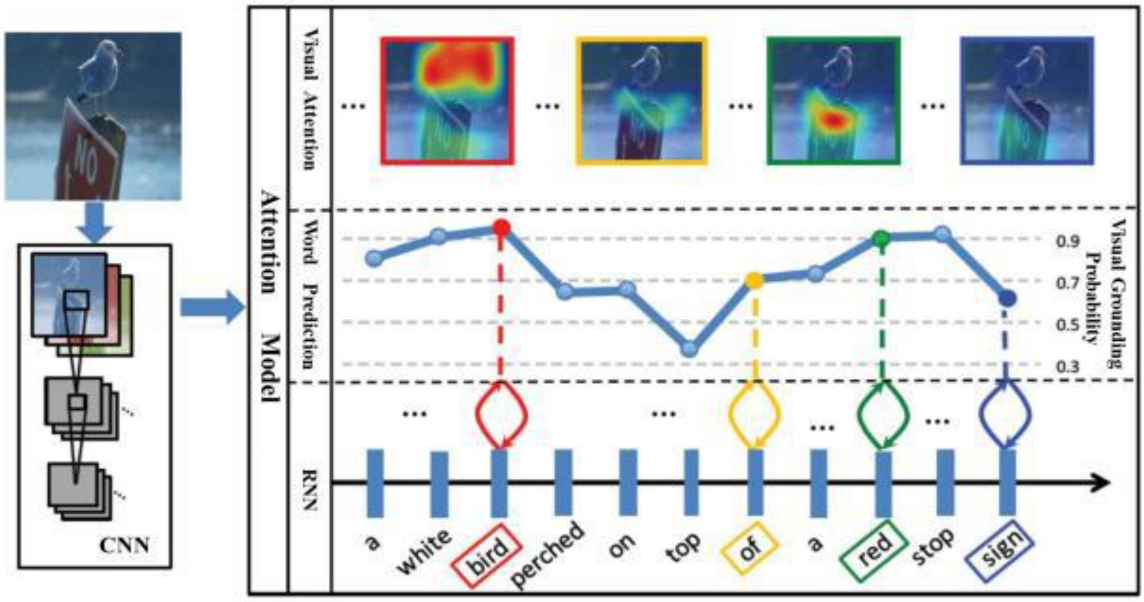


Figure 3 Illustration of the Attention Model

CNN–CNN Method:

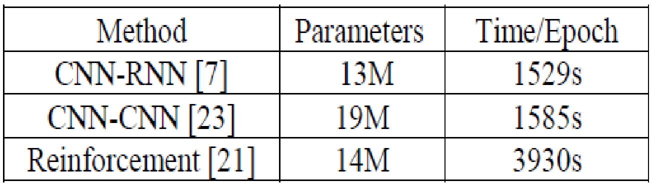
In this case, a separate release of convolution and a triple gate of recurrence play a similar role. Although the methods vary, the goal is to overlook minor content and highlight important content. Therefore, precisely, there is no significant difference between a convolutional model and a repetitive model. But the fact is that CNN is faster than RNN training that is easy to understand and does not contradict. The inevitable outcome is affected by two factors.

- Conversions can be processed similarly, and duplicates can only be processed sequentially. Having multiple machines trained with convolutional models simultaneously is much faster than training a repetitive secondary model.
- GPU chip can be used to accelerate convolution model training, and currently no hardware to accelerate RNN training.

The CNN-CNN based framework is a match between CNN and RNN in the field of machine translation and image captioning. In the recent years, CNN turns out to be a wide application given their effectiveness in computer vision and a lot of researches have been studied in the machine translation. In the same way, this convolutional model development can be applied to image captions. As the CNN-CNN framework for capturing image captions began to be promoted in 2017, and there are many improvements being used in this framework in machine translation that can also be used in image interpretation

Reinforcement Approach:

The reinforcement learning model is driven by visual semantic embedding, which works well on different test metrics without retraining. Visual semantic embedding, which provides a measure of similarity between pictures and sentences, can measure similarities between pictures and sentences, the accuracy of the captions produced and provide a comprehensive global goal of providing image captions to enhance learning. Instead of studying the successive loop model to greedily find the next word, the decision-making network uses the "policy network" and "value network" jointly to determine the next loop in each step. The policy network provides the confidence to predict the next word in the current context. The value network assesses the reward value of all existing state extensions.



Method	Parameters	Time/Epoch
CNN-RNN [7]	13M	1529s
CNN-CNN [23]	19M	1585s
Reinforcement [21]	14M	3930s

Table1 Training time for one mismatch

DISCUSSIONS

In Figure 4, we showed the best results of the above three methods of the five test metrics. We can see that both CNN-RNN is derived, and the reinforcement methods can get better performance than the CNN-CNN-based framework, which significantly improves the speed of training without significantly affecting accuracy. Other than that, the reinforcement framework works very well, because the objective work is very logical as we present.

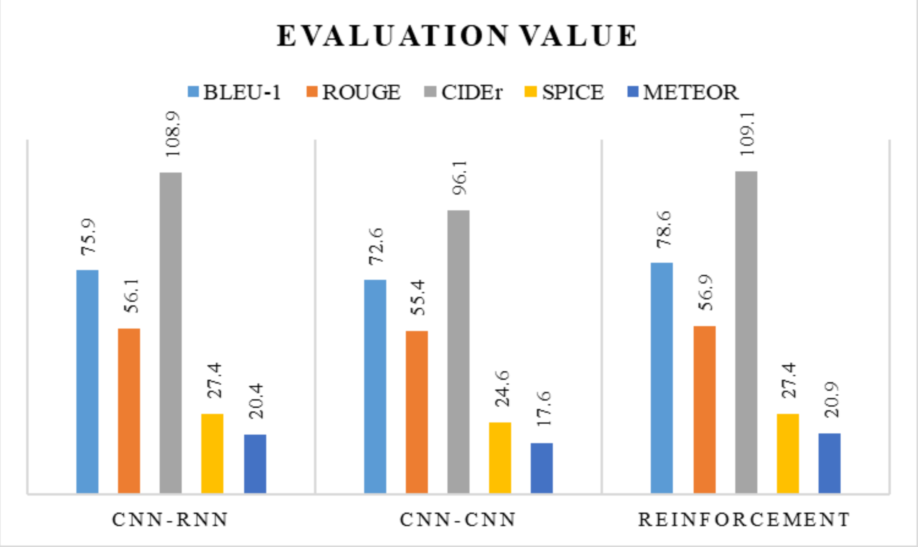


Figure 4 Evaluation Index of three methods

VI. CONCLUSION

For this project, we have used an in-depth learning method for capturing photo captions. Experimental experiments show that the proposed model is capable of producing beautiful captions for images automatically. We have introduced a single auto-filtering image for ResNet50 and LSTM based software maintenance. The proposed model was designed with a single encoder-decoder structure. We have adopted ResNet50, a flexible neural network, as the encoder makes the image a coherent presentation as a graphical feature. Subsequently, the LSTM language model was selected as a decoder to produce descriptive sentences. Meanwhile, we have combined a soft focus model with LSTM so that learning focuses on a specific part of the image to improve performance. The whole model is fully trained using a stochastic gradient drop which makes the training process easier. Experimental experiments show that the proposed model is capable of producing beautiful captions for images automatically.

REFERENCES

[1] Deepak D, Chitturi Bhadrachalam Deep neural approach to fake-news identification International conference on computational intelligence and data science (ICCIDS 2019), Procedia computer science, Vol. 167 (2020).

[2] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human ac-tivities from video images based on concept hierarchy of actions, Int. Comput. Vis. 50 (2002) 171-184 .

[3] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic atten-tion, in:Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4651–4659.

[5] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hocken-maier, D. Forsyth, Every picture tells a story: Generating sentences from images, in: Proceedings of the European Conference on Computer Vision, 2010, pp. 15–29.

[6] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a rank-ing task: data, models and evaluation metrics, J. Artif. Intell. Res. 47 (2013) 853–899.

[7] Y. Yang, C. L. Teo, H. Daume, Y. Aloimono, Corpus-guided sentence generation of natural images, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2011, pp. 444–454 .

[8] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, A. Y. Ng, Grounded composi-tional semantics for finding and describing images with sentences, TACL 2 (2014) 207–218 .

[9] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, S3how and tell: a neural image cap-tion generator, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3156–3164 .

[10] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic atten-tion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4651–4659 .