Image Semantic Description Based on Deep Learning with Multi-attention Mechanisms

Jian Yang(✉) and ZuQiang Meng

College of Computer, Electronics and Information,
Guangxi University, Nanning 530004, China
yangjianl015@foxmail.com

Abstract. In the era of big data, cross-media and multi-modal data are expanding, and data processing methods fail to meet corresponding functional requirements. Aiming at the characteristic of large expression gap of multi-model data, This paper proposes a multimodal data fusion method based on deep learning, which combines the advantages of deep learning in the field of image detection, text sequence prediction, and the multi-attention mechanism. The BLEU algorithm is used to calculate the similarity of four levels of description statements of model output and image. Training and testing were conducted in the Flickr8K data set. Comparing with the traditional single mode state image description method, the experiments show that under the BLEU index, the multi-AM model can achieve better results.

Keywords: Mulit-model data · Deep learning · Attention mechanism
Image semantic expression

1 Introduction

With the emergence of various high-tech electronic products, the carrier of important information is no longer a traditional single text, video or audio, but a variety of media. Cross-media and multi-modal data mainly showed the underlying data the characteristics of heterogeneous and high-level semantic expression similar to the "semantic gap" problem, in recent years, more and more scholars and researchers have been involved the study of multi-modal data [1, 2].

Image scene description is a important research direction in the field of image understanding, its object of study is by the color, texture, shape representation of image information, such as the target task is to get a accurate description of image content text sequences, across the two modal images and text information expression. Traditional image description task mainly has two key issues - image content representation and classification judgment, namely to find the most representative image scene, and then to learning and training of scene, scene category classification model. Image content description often requires human visual feature extracting, design visual dictionary, such as complex work, and need additional researchers' prior knowledge, this part of the work has a great influence for the classification effect. The main sources of problems or image and text data in the underlying data expression differences, how to eliminate the differences, and for a variety of modal data fusion is the core content of this paper.

Z. Shi et al. (Eds.): IIP 2018, IFIP AICT 538, pp. 356–362, 2018.
https://doi.org/10.1007/978-3-030-00828-4_36

Hinton since 2006, put forward the concept of deep learning, a large number of papers published about deep learning, in-depth study has been successfully applied to computer vision, speech recognition, natural language processing, and other fields. In multiple modal data fusion, image and text, for example, the depth of the neural network can use different models of two kinds of modal data feature extraction, and image and text are similar characteristics of space vector modeling method, therefore, for the fusion of images and text on the feature space, image semantic expression based on deep learning can be realized.

2 Related Work

The study of semantic description of traditional image scenes is mainly based on single-modality images. To reduce the "semantic gap", an image analysis layer such as a visual dictionary is constructed between the low-level visual features and high-level semantic information. Image semantic information description must establish the mapping relationship between low-level visual features and high-level semantics. In recent years, more and more scholars have begun to pay attention to the research field of multimodal data fusion. Sawant et al. (2011) thinks it needs to capture events, locations, and personalities in addition to the visual features of the human beings in the traditional object monitoring and scene interpretation [3]. The abstract concept of a reference, Ma et al. (2011) carried out related research on image and text fusion methods, and proposed a data fusion framework based on image content and tags-a new random walk model that uses fusion parameters to balance content and labels between the impact [4]. Hollink et al. (2005) describes a semi-automatic image annotation algorithm in specific fields, the goal is to identify domain features to increase semantic understanding [5].

In recent years, the intensified research on deep neural networks has remained high and breakthroughs have also been made. This has benefited from convolutional neural networks and recurrent neural networks [6], and most researchers are based on these two types of deep neural network models and have done a lot of optimization work. Karpathy and Li (2015) and Xu et al. (2015) did a lot of work in the field of image understanding [7, 8]. Li et al. propose an image semantic alignment model, extract the key information in the image and align it with the keywords in the sentence description. However, they all use the RNN structure in the language model, and there is a lack of semantic relevance.

3 Deep Neural Network Model Based on Multi-attention Mechanism

In this paper, we propose a deep neural network based on Convolution Neural Network (CNN) and Long Short Term Memory Network Network (LSTM) to describe the image in sentences, and introduce multi-attention mechanism on this basis. When we describe an image, we need to pay attention to the content of the image as well as the language foundation. When we get the word "cat," we focus on the cat part of the

image and ignore the rest. The prediction of a word requires not only the introduction of attention mechanism in the language model, but also in the image.

3.1 CNN for Feature Extraction

Russakovsky et al. (2015) summarized the competition results of each team in the ImageNet competition and the brief description of the algorithm in the last five years [9]. Among them VGG network performance is particularly outstanding. Simonyan and Zisserman (2014) believed that the filter of 7 × 7 could be decomposed into a number of 3 × 3 filters, the smaller filter means more flexible channel [10]. Figure 1 shows the convolution structure of vgg-16 network, We used his convolutional layer to extract image features as input to the image part.

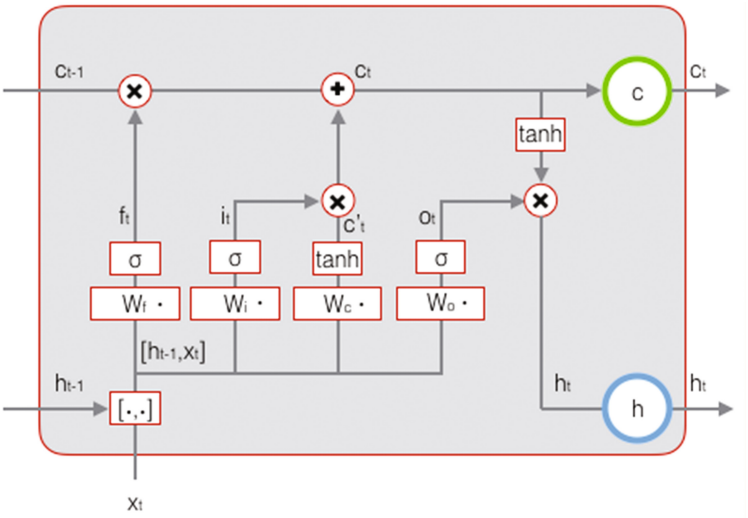

Fig. 1. Calculation of a single LSTM node

3.2 LSTM for Sequences' Predict

DNN may be able to extract excellent characteristic expression, but LSTM is better than DNN in word prediction. LSTM is an improved recurrent neural network [11], general RNN cannot save too much information, there is only one state in the hidden layer, it is very sensitive to short-term input, if we add another state C to save long-term information, the problem will be solved. LSTM uses gate to control long-term status C. The gate can be expressed as:

$$\boldsymbol{g}(\mathbf{x}) = \boldsymbol{\sigma}(\boldsymbol{W}\mathbf{x} + \mathbf{b}) \tag{1}$$

W is the weight vector of the gate and b is the bias term. $\boldsymbol{\sigma}$ is the sigmoid function and the range is (0,1), so the state of the gate is half open and half closed. The structure

of a single node in the LSTM network is shown in Fig. 1. The final output of the LSTM is jointly controlled by the output gate and cell state:

$$\boldsymbol{h_t} = \boldsymbol{o_t} \circ \boldsymbol{tanh}(\boldsymbol{c_t}) \tag{2}$$

Because of the control of oblivion gate, it can save the information of a long time ago, and because of the control of the input door, it can avoid the current inconsequential content from entering the memory.

3.3 Multi-attention Mechanism

The AM model is one of the most important developments in the NLP field in the past few years, which appear in most of the current papers are attached to the Encoder-Decoder framework. But the AM model can be used as a general idea, When a general RNN model generates a language sequence, the predicted next word is only related to its first n words, and n is generally less than 5.

$$\boldsymbol{y_i} = \boldsymbol{f}(\boldsymbol{y_{i-1}}, \boldsymbol{y_{i-2}}, \boldsymbol{y_{i-3}} \cdots, \boldsymbol{C}) \tag{3}$$

The above formula C represents the semantic encoding, Obviously the semantic semantics of each word is C is unreasonable, and a word will not only be related to the most recent words, So we give each word a probability distribution that expresses its relevance to other words, and Replace C with C_i.

$$\boldsymbol{C_i} = \sum_{j=1}^{T_x} \boldsymbol{\alpha_{ij}} \boldsymbol{h_j} \tag{4}$$

$\boldsymbol{T_x}$ indicates the number of other words associated with $\boldsymbol{C_i}$, $\boldsymbol{\alpha_{ij}}$ represents the probability of interest between two words. $\boldsymbol{h_j}$ is the information of the word itself. The question now is how to calculate the probability distribution. It is usually to calculate the similarity between the current input information $\boldsymbol{H_i}$ and the previous information $\boldsymbol{h_j}$. In the word2vec model, it is the distance between two vectors.

After going through a multi-layer convolutional structure, the image information is compressed into a vector I. When predicting each word, you need to associate some information in the vector. The attention parameter W for the image is what we need to train to get.

$$\boldsymbol{A_i} = \boldsymbol{I} \circ \boldsymbol{W_i} \tag{5}$$

Finally, the functional relationship of each predicted word is as follows and The model is shown in Fig. 2.

$$\boldsymbol{y_i} = \boldsymbol{f}(\boldsymbol{A_i}, \boldsymbol{C_i}, \boldsymbol{y_{i-1}}, \boldsymbol{y_{i-2}}, \boldsymbol{y_{i-3}}, \cdots) \tag{6}$$

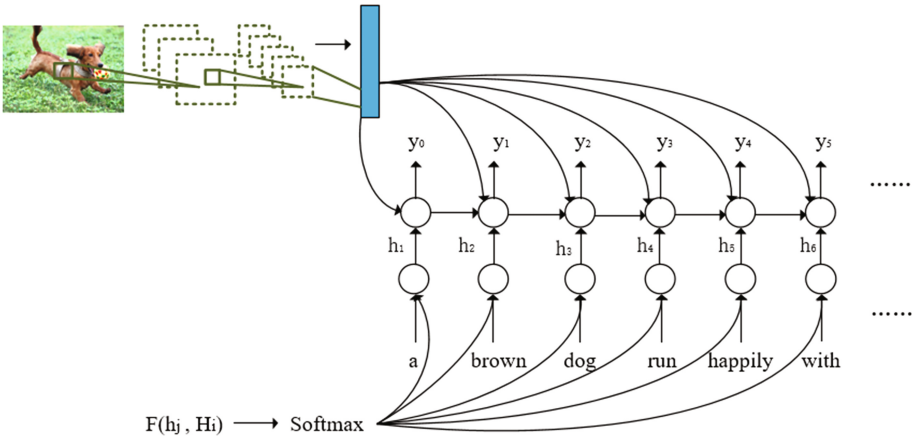

Fig. 2. Deep neural network of multi-attention mechanism

4 Experiments

In order to compare with the prior art, we conducted a large number of experiments using the BLEU metrics to evaluate the effectiveness of our model. Experiments used the Flickr 8K data set, contrast traditional image scene expression method.

4.1 Experimental Environment and Parameter Deployment

Experiments using the current popular TensorFlow framework, and VGG-16 model to extract image features, In the whole model, set the batch-size to 8 to relieve the pressure of GPU memory. Fully connected layer functions as a "classifier" throughout the convolutional neural network, while the convolutional layer, the pooled layer, and the activation function layer can be viewed as mapping the image to the feature space. Deeper layers will inevitably bring better results.

The experiment uses the Flickr8k dataset, which includes Flickr8k_Dataset and Flickr8k_text. We compressed the image to a size of 300 × 300 pixels so that we could train more samples; In Flickr8k_text, each image corresponds to 4 different types of descriptions. We use Word2vec to vectorize the words. In terms of semantically similar words, the Euclidean distances of their vectors are often very close. Then we only select the first 200 words of frequency, and the rest are replaced by [UNK], which makes the load of the model smaller.

4.2 Evaluating Indicator

The popular automatic evaluation method is the BLEU algorithm proposed by IBM [12] ,The BLEU method first calculates the number of matching n-grams in the reference sentence and the generated sentence, and then calculates the ratio of the number of n-grams in the generated sentence as an evaluation index. It focuses on the accuracy

of generating the word or phrase in the sentence. The accuracy of each order N-gram can be calculated by the following formula:

$$P_n = \frac{\sum_i \sum_k \min\left(h_k(c_i), max_{j \in m} h_k\left(s_{ij}\right)\right)}{\sum_i \sum_k \min(h_k(c_i))} \tag{7}$$

The upper limit of N is 4, which means that only the accuracy of 4-gram can be calculated.

4.3 Results and Discussion

There are two comparisons in the experiment. One is to compare the traditional basic model, and the other is to compare the current mainstream model. As can be seen from Table 1, the accuracy of the Multi-AM model in the keyword is better than the traditional model.

Table 1. Comparison with other methods on the flicker8K dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Object-based	54	28.0	14.5	13.1
Scene-based	57.9	38.3	17.4	16.0
Multi-AM	**68.2**	**42.3**	**23.5**	**18.3**

From the results of Table 2, the accuracy of one and two keywords is better than other mainstream algorithms, but the effect of obtaining more keywords is not good. but the effect of obtaining more keywords is not good.

Table 2. Comparison with other mainstream model on the flicker8K dataset

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
LogBilinear	54	28.0	14.5	13.1
Multi-RNN	57.9	38.3	**24.5**	16.0
R-CNN	63.0	35.5	21.8	**20.2**
Multi-AM	**68.2**	**42.3**	23.5	18.3

5 Conclusions

Semantic description of images is a complex task, and the framework based on deep learning has become the current mainstream method. This paper proposes a multi-attention mechanism based on deep neural network, we need to understand the grammar of the sentence generation, as well as the content in the image. Words have different levels of attention to image content and different levels of attention from context. Experimental results show that the multi-attention mechanism can achieve

higher scores under the BLEU evaluation criteria, which shows that the method can extract more keywords.

Future research should attempt to lack the sentence description data set and improve the semantic matching of sentences. We need to accurately position the predicted word in the image, which is very difficult.

References

1. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: consensus-based image description evaluation. In: Computer Vision and Pattern Recognition IEEE, pp. 4566–4575 (2015)
2. Vinyals, O., et al.: Show and tell: a neural image caption generator. In: Computer Vision and Pattern Recognition, pp. 3156–3164. IEEE (2015)
3. Sawant, N., Li, J., Wang, J.Z.: Automatic image semantic interpretation using social action and tagging data. Multimed. Tools Appl. **51**(1), 213–246 (2011)
4. Ma, H., et al.: Bridging the Semantic gap between image contents and tags. IEEE Trans. Multimed. **12**(5), 462–473 (2010)
5. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53
6. Hollink, L., Little, S., Hunter, J.: Evaluating the application of semantic inferencing rules to image annotation. In: International Conference on Knowledge Capture, pp. 91–98. ACM (2005)
7. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. IEEE Show Tell Neural Trans. Pattern Anal. Mach. Intell. **39**(4), 664–676 (2017)
8. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: Computer Science, pp. 2048–2057 (2015)
9. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. **115**(3), 211–252 (2015)
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Computer Science (2014)
11. Jia, X., et al.: Guiding the long-short term memory model for image caption generation. In: IEEE International Conference on Computer Vision, pp. 2407–2415. IEEE (2016)
12. Papineni, K.: A method for automatic evaluation of machine translation. In: Proceedings of ACL (2002)