Scientific Research Publishing

Journal of Intelligent Learning Systems and Applications, 2020, 12, 14-29
https://www.scirp.org/journal/jilsa
ISSN Online: 2150-8410
ISSN Print: 2150-8402

Application of Dual Attention Mechanism in Chinese Image Captioning

Yong Zhang, Jing Zhang

College of Computer Science and Technology, Tianjin Polytechnic University, Tianjin, China
Email: 1731125392@stu.tjpu.edu.cn

How to cite this paper: Zhang, Y. and Zhang, J. (2020) Application of Dual Attention Mechanism in Chinese Image Captioning. *Journal of Intelligent Learning Systems and Applications*, **12**, 14-29.
https://doi.org/10.4236/jilsa.2020.121002

Received: May 29, 2019
Accepted: January 12, 2020
Published: January 15, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

Objective: The Chinese description of images combines the two directions of computer vision and natural language processing. It is a typical representative of multi-mode and cross-domain problems with artificial intelligence algorithms. The image Chinese description model needs to output a Chinese description for each given test picture, describe the sentence requirements to conform to the natural language habits, and point out the important information in the image, covering the main characters, scenes, actions and other content. Since the current open source datasets are mostly in English, the research on the direction of image description is mainly in English. Chinese descriptions usually have greater flexibility in syntax and lexicalization, and the challenges of algorithm implementation are also large. Therefore, only a few people have studied image descriptions, especially Chinese descriptions. **Methods:** This study attempts to derive a model of image description generation from the Flickr8k-cn and Flickr30k-cn datasets. At each time period of the description, the model can decide whether to rely more on images or text information. The model captures more important information from the image to improve the richness and accuracy of the Chinese description of the image. The image description data set of this study is mainly composed of Chinese description sentences. The method consists of an encoder and a decoder. The encoder is based on a convolutional neural network. The decoder is based on a long-short memory network and is composed of a multi-modal summary generation network. **Results:** Experiments on Flickr8k-cn and Flickr30k-cn Chinese datasets show that the proposed method is superior to the existing Chinese abstract generation model. **Conclusion:** The method proposed in this paper is effective, and the performance has been greatly improved on the basis of the benchmark model. Compared with the existing Chinese abstract generation model, its performance is also superior. In the next step, more visual prior information will be incorporated into the model, such as the action category, the relationship between the object and the ob-

ject, etc., to further improve the quality of the description sentence, and achieve the effect of "seeing the picture writing".

Keywords

Image Caption in Chinese, Dual Attention Mechanism, Richness, Accuracy

1. Introduction

Natural language processing (NLP) and computer vision (CV) are current research hotspots. NLP focuses on understanding natural language, modeling the process of text generation, implementing word segmentation, part-of-speech tagging, named entity recognition, syntactic analysis, and multi-language machine translation [1]. CV focuses on understanding images or video, enabling classification, target detection, image retrieval, semantic segmentation, and human pose estimation. The recent multimodal processing of fused text and image information has caused great interest among researchers. Image Captioning is the key technology of multi-modal processing. It can perform image-to-text multi-modal transformation and help visually impaired people understand the image content. This technique was first proposed by Farha-di *et al.*, given a binary group I, S, where I represent an image, S represents a summary sentence, and the model completes a multimodal mapping I → S from image I to a summary sentence S. This task is very easy for humans, but it poses a huge challenge to the machine, because the model not only understands the content of the image, but also produces human-readable abstract sentences [2].

The current research mainly focuses on image generation English abstracts, and there are few studies on the generation methods of Chinese abstracts [3]. Due to the rich meaning of Chinese words and the complex structure of sentences, the Chinese description of images is more difficult. Based on the data sets of flickr8k CN and flickr30k CN, an image description generation model with dual attention mechanism is proposed in this study. In each period of generating description, the model can decide whether to rely more on image or text information, so that more important information can be captured from the image, so as to improve the richness and accuracy of Chinese image description. The contribution of this paper is to build a Chinese description generation model based on dual attention mechanism, expand the application of dual attention in image description, especially in Chinese description field, and verify the good performance of this method in Chinese image description task.

In the future, the rapid development of artificial intelligence will make many people face the risk of unemployment. In the future, robots in the field of image description can replace the work of newspaper news writers. Because computers are obviously superior to humans in image processing, the photos captured from the front of news can be quickly written into corresponding press releases, which improves the timeliness of news. At present, robot medical treatment is also a

hot topic, so image description will be very popular in the field of medicine Great help, we go to the hospital to see a variety of test sheets and ultrasound images, and no longer need to queue up for doctors to see the cause of disease. The robot will analyze and describe these medical images, and then generate scientific diagnosis results, so image description has a great development prospect in the future.

2. Related Works

For the generation model of image description, the current common method is based on the extension of neural network, most of which are composed of encoder and decoder. The image is encoded using a pre-trained deep convolutional neural network (CNN), and then the image is embedded into a recurrent neural network (RNN), and the corresponding description sequence is finally output as a description. Mao *et al.* proposed a neural network based image summary generation model m-RNN, which uses CNN to model images, uses RNN to model sentences, and uses multimodal space to correlate images and text [4]. Vin-yals *et al.* proposed the Google NIC model, which projects images and words into a multimodal space and generates summaries using long and short time memory networks (LSTM) [5]. Jia *et al.* proposed the model gLSTM, which uses semantic information to guide long and short-term memory networks to generate summaries [6]. Xu *et al.* introduced an attention mechanism into the decoding process, enabling the digest generation network to capture local information of the image [7]. Li *et al.* constructed the first Chinese image summary dataset Flickr8k-CN and proposed the Chinese abstract generation model CS-NIC [8], which uses Google Net to encode images and model the digest generation process using long and short time memory networks.

In addition to the good progress made on the model, the Chinese dataset of image descriptions is also evolving, such as Flickr8k-CN, Flickr30k-CN, MSCOCO Chinese datasets and the recent AI Challenger image Chinese description dataset. The number of images to be described is increased from a few thousand to hundreds of thousands, and the description of the image is also from a simple object description to a rich motion and expression description, thereby improving the quality of the image in the field of Chinese description [9].

3. Model for Image Captioning

The basic expression of the basic attention model can be understood as follows: When we are looking at one thing, we must always pay attention to somewhere in the things we are currently looking at. In other words, when we look at when you are elsewhere, attention shifts with the movement of your eyes, which means that when people notice a certain target or a scene, the attention distribution inside the target and in each spatial position within the scene is not the same. This is also true in the following situations: When we try to describe a thing, the

words and sentences we are talking about at the moment are the first to correspond to a certain segment of the thing being described, and the other parts follow the description. Relevance is constantly changing [10]. The addition of the attention mechanism allows the image to be more focused on finding useful information related to the current output in the input data during the generation of the description, thereby improving the quality of the output. The ultimate goal of the attention model is to help a framework like codec to better learn the interrelationships between multiple content modalities, so as to better represent this information and overcome the drawbacks that are difficult to design because of its inability to interpret.

Encoder-decoder framework is used in basic image captioning model with an attention mechanism. However, when the model generates the description, the decoder should have different attention strategies for different words. For example, the words "注视着(watching)", "正在(is)", or "在...旁边(beside…)", such words are called non-visual words, and the model should rely more on semantic information when generating such phrases. Not visual information. Moreover, in the process of generating the description, the attention gradient can mislead or reduce the validity of the visual information because the non-visual phrase features cannot be learned from the image during the training process. Therefore, this paper proposes an image description generation model with a dual attention mechanism of the neural network classifier, which can decide whether to rely more on images or text information at each time period of generating the description.

3.1. Basic Model for Image Captioning in Chinese

For a given image I, our goal is to automatically predict the Chinese sentence S to briefly describe the visual content of the picture [11]. This sentence is a sequence of n Chinese words. $S = \{c_1, \cdots, c_n\}$. Usually, a method of acquisition is to use pre-trained image description model in English and translate its output from English to Chinese through machine translation. Contrary to this post-translation strategy, we discuss how to build a Chinese model directly from a Chinese multimedia corpus in this section. Record as $C = \{(I_i, S_i, 1, \cdots, S_i, m_i)\}$, which the i_{th} training image is accompanied by m_i sentences.

In this work, we study the neural image description (NIC) model for generating Chinese sentences, as shown in Formula (1). NIC is a probability model, which uses LSTM neural network to calculate the posterior probability of a given input image sentence. So the image will be annotated with sentences that produce the maximum probability. For model θ, given the input picture I, the probability of the model automatically generating sequence S is given.

$$P(S \mid I;\theta) = \prod_{t=0}^{N} P(S_t \mid S_0, S_1, \cdots, S_{t-1}, I;\theta) \tag{1}$$

After decomposition into the form of continuous multiplication, the problem becomes the conditional probability of modeling $P(S_t \mid S_0, S_1, \cdots, S_{t-1}, I;\theta)$.

Usually, RNN is the first choice, because it can theoretically retain all the above information (LSTM is used to alleviate the "long-term dependence" problem), instead of taking only one window like n-gram or CNN. Logarithmic likelihood function is obtained by taking logarithmic likelihood function.

$$\log P(S \mid I;\theta) = \sum_{t=0}^{N} \log p(S_t \mid S_0, S_1, \cdots, S_{t-1}, I;\theta) \tag{2}$$

where $S_0 = START$ and $S_{t+1} = END$ are two special tokens indicating the beginning and the end of the sentence.

The training objective of the model is to maximize the logarithmic likelihood sum of all training samples.

$$\theta^* = \arg\max_{\theta} \sum_{(I,S)} \log P(S \mid I;\theta) \tag{3}$$

where (*I*, *S*) training samples. This method of maximum likelihood estimation is equivalent to empirical risk minimization using logarithmic loss function.

Of course, it is not feasible to calculate the probability of all sequences and then select the most probabilistic sequence, because each location has a word with thesaurus size as a candidate, and the search size will increase exponentially with the length of the sequence, so we need to use beam search to reduce the search space.

Conditional probabilities in Equation (1) are estimated by the LSTM network in an iterative manner. The network maintains a cell vector c and a hidden state vector h to adaptively memorize the information fed to it. The embedding vector of an image, obtained by applying an affine transformation on its visual feature vector, is fed to the network to initialize the two memory vectors. In the t_{th} iteration, new probabilities P_t over each candidate word are re-estimated given the current chosen words. The word with the maximum probability is picked up, and fed to LSTM in the next iteration. The recurrent connections of LSTM carry on previous context per iteration. We apply beam search to maintain k best candidate sentences. The iteration stops once the END token is selected. To express the above process in a more formal way, we write

$$x_{-1} := W_e \cdot CNN(I), \tag{4}$$

$$x_t := W_s \cdot w_t, t = 0, 1, \cdots, \tag{5}$$

$$p_0, c_0, h_0 \leftarrow LSTM(x_{-1}, 0, 0), \tag{6}$$

$$p_{t+1}, c_{t+1}, h_{t+1} \leftarrow LSTM(x_t, c_t, h_t). \tag{7}$$

The parameter set θ consists of W_e, W_s, and affine transformations inside LSTM. While DeViSE directly takes a pretrainedword2vec model to construct W_s, NIC optimizes W_e and W_s simultaneously via maximum-likelihood estimation. The structure of neural network image description generation model is shown in **Figure 1**.

$$\arg\max_{S} P(S \mid I;\theta^*) \tag{8}$$

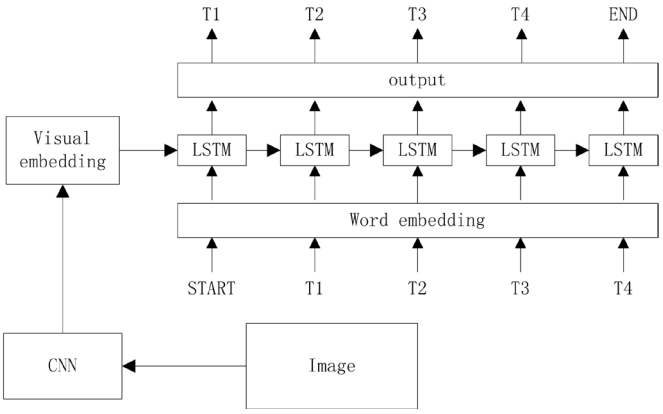

Figure 1. The Chinese-captioning model.

3.2. Model with Attention

Firstly, we propose a context vector c_t for spatial attention model, which is defined as:

$$c_t = g\left(V, h_t\right) \tag{9}$$

where g is Attention function, $V = \left[v_1, \cdots, v_k\right], v_i \in R^d$ is Spatial image features, each feature is a d-dimension, representing a part of the image. It is the hidden state of T-Time cyclic neural network. Given LSTM's spatial image features $V \in R^{d*k}$ and hidden states $h_t \in R^d$, we feedback them through a single layer neural network, and then use the SOFTMAX function to generate the attention distribution on k regions of the image [12]:

$$z_t = w_h^{\mathrm{T}} \tanh\left(W_v V + \left(W_g h_t\right) I^{\mathrm{T}}\right) \tag{10}$$

$$a_t = \mathrm{softmax}\left(z_t\right) \tag{11}$$

where $I \in R^k$ is a vector with all elements set to 1, $W_v, W_g \in R^{k*d}$ and $W_h \in R^k$ are the parameters to learn. $\alpha \in R^k$ is the attention weight of the feature in V. Based on attention distribution, context vectors c_t can be obtained in the following ways:

$$c_t = \sum_{i=1}^{k} \alpha_{ti} v_{ti} \tag{12}$$

Finally, c_t and h_t combine to predict the next word, as shown in the following figure:

$$\log p\left(y_t \mid y_1, \cdots, y_{t-1}, I\right) = f\left(c_t, h_t\right) \tag{13}$$

where f is a non-linear function of output probability y_t, c_t is a visual context vector at t-time extracted from image I, and h_t is a hidden state of RNN at t-time. The structure of attention model is shown in **Figure 2**.

3.3. Double-Attention Model

3.3.1. Model Framework

Figure 3 illustrates the framework of a Chinese image captioning model based

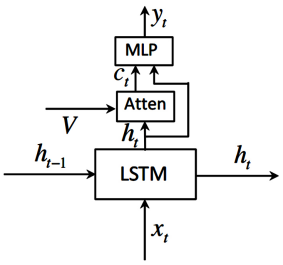

Figure 2. Attention model.

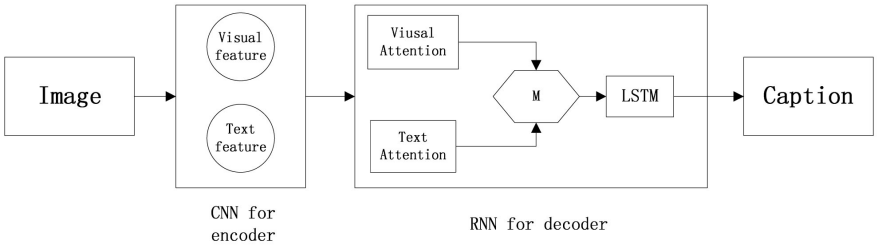

Figure 3. The framework of a Chinese image captioning model based on dual attention mechanism.

on dual attention mechanism. Because this paper adopts a dual attention mechanism, it describes that attention strategies should be adopted for visual and textual features respectively in the process of generation. Encoder consists of two neural networks: one is visual feature extraction network and the other is text feature extraction network. Visual feature $V_I \in R^n$ is the hidden layer output of convolution network, which depicts the deep visual features of images, focuses on visual information, and uses real vector coding. Text feature $W_I = \{w_1, w_2, \cdots, w_m\}, w_i \in [0,1]$ is the output layer result of convolution network, which reflects the probability of visual-related combinations appearing in the description. It focuses on text information and uses probability vector coding. The decoder is composed of LSTM Chinese description generation network. The visual and text features are fused by the neural network classifier when generating the description. The output is the Chinese description of the image. Encoder models feature based on convolutional neural network and decoder models sequence based on long-term and short-term memory network. Because the data set of Chinese abstracts is limited, the generalization of neural networks is reduced by using a cooperative training method, so the three neural networks are trained separately on different data sets. For sufficiently large summary data sets, collaborative training is an ideal choice.

3.3.2. Dual Attention Mechanism

For visual feature V_I and text feature W_I, the description generation network RNN (V_I, W_I) based on dual attention mechanism completes the mapping of $V_I, W_I \to S_I$, where S_I is the Chinese description of image *I*. In this paper, long-term and short-term memory network is used to model the abstract generation process. The calculation process of network t-time is

$h_t, c_t = LSTM\left(x_{t-1}, h_{t-1}, c_{t-1}\right)$, as follows: $x_t \in R^d$ is the input of t-time, $c_t \in R^d$ is the cell state, and $h_t \in R^d$ is the hidden cell state. $LSTM\left(\bullet\right)$ function has been given in 3.1, and we will not elaborate on it here.

In this paper, 3.2 introduces the ordinary encoder-decoder framework, which is not discussed here. But the context vector c_t is defined in this paper. For the model without attention mechanism, c_t is the feature map extracted from the image after CNN, which is invariable. For the model with attention mechanism, based on hidden state, decoder will pay attention to different regions of the image. c_t is the feature map extracted from the region after CNN.

As shown in **Figure 4**, we set up a neural network classifier C in the model to judge the attention distribution of the attention model at t-time,

$$\beta_t = w_h^{\mathrm{T}} \tanh\left(W_s s_t + W_g h_t\right) \tag{14}$$

$$C = \beta_t s_t + \left(1-\beta_t\right) c_t \tag{15}$$

where $\beta_t \in [0,1]$ is used to control the degree to which the model focuses on visual feature s_t and textual features c_t. When $\beta_t \geq 0.5$, the visual features of the image account for a large proportion, so the classifier guides the model to focus on the visual features of the image.

At the same time, attention distribution α_t of K regions in the attention model is also extended to α_t by splicing an element after z_t:

$$\hat{\alpha}_t = \mathrm{softmax}\left(\left[z_t, \beta_t\right]\right) \tag{16}$$

where z_t is the attention distribution over the k regions of the image, and β_t is the classifier vector at time t.

Finally, the probability over a vocabulary of possible words at time t can be calculated as:

$$\log p\left(y_t \mid y_1, \cdots, y_{t-1}, I\right) = f\left(C, h_t\right) \tag{17}$$

where f is a non-linear function of output probability y_t, C is a visual context vector at t-time extracted from image I, and h_t is a hidden state of RNN at t-time.

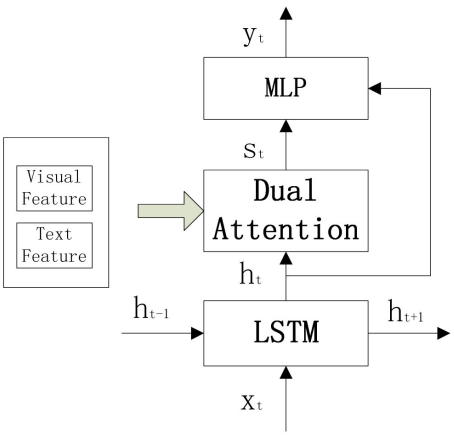

Figure 4. Dual attention mechanism with neural network classifier.

4. Experiment

The experiment in this paper uses a single Tesla K80 GPU to test on the Deep Learning Framework Tensorflow. In order to verify the validity of the model based on dual attention mechanism, this paper carries out experiments on Flickr8k-CN dataset, and evaluates the quality of description generation of multiple models. The models compared in the experiments include Google NIC, CS-NIC of Renmin University of China, CNIC of Chinese Academy of Sciences and DA-NIC proposed by us. These three models are the latest research results in this field in the past two years.

4.1. Dataset

The dataset used in this paper is the Flickr8k-CN [13], which is the most abundant, the most descriptive language and the largest image Chinese dataset. The image data comes from Baidu and Google's image database. The image description dataset released this time is mainly based on Chinese description sentences. Compared with the English data sets common to similar scientific research tasks, Chinese descriptions usually have greater flexibility in syntax and lexicalization, and the challenges of algorithm implementation are also greater. The data set contains 300,000 images and 1.5 million sentences in Chinese. This article follows an official approach to constructing training sets, validation sets, and test sets. Among them, there are 210,000 pictures in the training set, a total of 1.05 million sentence descriptions, 30,000 pictures in the verification set and the test set, and a total of 150,000 sentences. The data format contains images and corresponding Chinese descriptions in five sentences. The following Table 1 is an example.

Table 1. Examples of image and summary in Flickr8kCN.

		
Description 1: In the gym, there is a man with sleeves up and a man in a black jacket looking at the same mobile screen. Description 2: There is a man in a black coat and a man in a white coat in the room looking at the same cell phone. Description 3: There are two men in different clothes in the room looking at the same cell phone. Description 4: In the bright room, a man in short sleeves and a man in white clothes were looking at the same cell phone. Description 5: A man in a short-sleeved jacket and a man in a white jacket stood in the room and looked at the cell phone together.	Description 1: A medical worker and a woman in black were talking at a table in the house. Description 2: There is a man in white and a woman in black in the house. Description 3: In the bright room, a doctor in a white coat was sitting in front of a woman with her legs crossed. Description 4: There is a woman in a black jacket in the office talking to a doctor in a white coat. Description 5: In the bright room, a medical worker and a woman with her right hand on her lap were sitting in a chair.	Description 1: In front of three people on the playground, a man with a pen in his right hand signed a child's clothes. Description 2: In front of three people on the playground, a man in short sleeves signed a child's clothes. Description 3: In front of three people on the court, a man with a pen in his right hand was writing on a child's clothes. Description 4: In front of three people on the playground, a right-handed pen holder signed a child's clothes. Description 5: In front of three people on the playground, a man with a pen in his right hand was signing a boy's name.

4.2. Setup

4.2.1. Chinese Participle

Chinese natural language processing NLP is different from English natural language processing. For example, in the NLP field, the word is the smallest language unit that can be used independently. Due to the particularity of Chinese, Chinese needs to be able to perform subsequent NLP tasks better. Text is used for word segmentation and English does not need word segmentation. Chinese word segmentation is also a basic difference between English and Chinese NLP. In Chinese word segmentation task, there are two kinds of ambiguity phenomena: cross ambiguity and combination ambiguity. There are three major technical methods for solving segmentation ambiguity. Classes are rule-based methods, statistical-based methods, and a combination of rules and statistics. Therefore, we first construct the word coding matrix of the Flickr8k-CN image annotation set in the Chinese description dataset of Flickr8k-CN image, including text preprocessing and word segmentation, establish a dictionary, and establish a word index in Chinese sentences [14].

Step 1: Preprocess the Flickr8k-CN image annotation set, that is, the caption data set in the Flickr8k-CN image Chinese description data set;

Step 2: Using jieba to segment Chinese caption, and filtering out words with a frequency greater than 4 to form a word frequency dictionary;

Step 3: Setting < S> and < /S> as the start and end identifier of the text occupies a word position, and the generated dictionary size is 8560;

Step 4: Each word in the dictionary is uniquely heat-coded, the word vector dimension is equal to the dictionary size of 8560, and the value of the word vector is 1 to represent the index value of the word in the dictionary.

4.2.2. Model Settings

In this paper, a pre-trained convolutional neural network ResNet-152 is used to extract image features. It is a network structure that has obtained the best results in the image classification and detection contest of ImageNet and MSCOCO. Residual structure in residual network can learn residual function according to the input of each layer of the network. This residual function effectively solves the problem that the deep network model is difficult to train. It can improve the depth of the network and obtain high accuracy at the same time. Specifically, the pool5 layer of ResNet-152 is used to obtain a 2048-dimensional image feature.

The model is trained by using RMSP optimization algorithm with attenuation rate of 0.9. The initial learning rate is 0.001 and the batch size is 256.

4.2.3. Evaluation

In this paper, the three methods of BLEU, METEOR and CIDEr [15] are used to evaluate the generated sentences. The BLEU method first calculates the number of matches between the reference sentence and the generated n-gram in the sentence, and then calculates the ratio of the number of n-grams in the generated sentence as the evaluation index. It mainly focuses on the accuracy of the word

or phrase in the generated sentence. The METEOR method first uses the arbitrary matching method to search for the maximum value of the matching sentence in the reference sentence and the generated sentence according to the exact matching, synonym matching and prefix matching. When the maximum values of the three matchings are the same, the selection button is selected. The matching with the least number of intersections in the order two-two matching is used as the alignment; by continuously iterating, the alignment set is generated, and then the ratio of the number of elements in the collection to the total number of words in the reference sentence is used as the recall rate, and the ratio of the total number of words in the generated sentence. As an accuracy rate, the final value is then calculated using the harmonic mean. In addition to BLEU and METEOR, there is the latest CIDEr evaluation method, which uses the concept of "consensus" to calculate the degree of matching between candidate sentences and reference sentence sets, and more to reflect the semantic quality of the generated sentences.

4.3. Results

In this section, NIC model, CNIC model based on attention mechanism and dual attention model are compared and analyzed from three aspects: confusion, model results and model scores. In order to verify the improvement and promotion of the model on the basis of this theory, in order to improve the persuasiveness of the model in generation, this paper also compares the performance of the current authoritative mainstream description model in this field, so as to further verify the effectiveness of the model in image description.

1) First of all, this paper compares the performance of NIC model, attention model and this model by using the method of reference [11] and the confusion of the model in the verification set. As shown in **Figure 5**, the NIC model of Google is the best in 100,000 iterations, the CNIC model based on attention is the best in 130,000 iterations, and the da-nic model of this paper is the best in 110,000 iterations. Compared with the former two models, the model has better convergence in the training stage.

2) This paper uses the dual attention mechanism to improve the correlation between the generated description sentences and images, and learn the visual and text attention models together to explore the subtle interaction between the visual and language. In order to verify whether visual attention pays attention to the information of image area and whether text attention plays a guiding role in the generated words. We try to select two representative pictures in the test set and put them into three models to generate descriptions. Through further analysis of the description effect, we verify the richness and relevance of the model in generating description statements. **Figure 6** shows the two images selected from the validation set and the generated Chinese description.

The first NIC model in **Figure 6** is described in Chinese as "standing on the beach by a person". The description is too vague and vague. After joining the

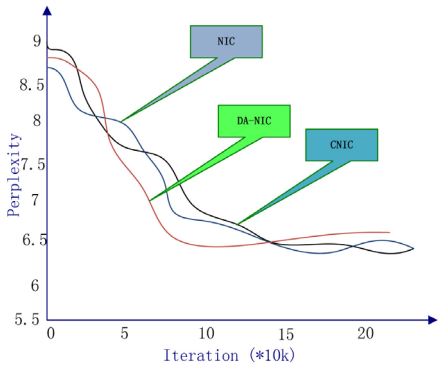

Figure 5. Perplexity curves of three models.

(Image)	(Caption)
	NIC: A man is picking up things at the seaside ATT: In the distance, a man stooped to pick up things at the seaside D-ATT: A little girl stooped to pick up shells by the sea at sunset
	NIC: A group of people are standing on the playground ATT: A group of people are playing football on the grass D-ATT: A group of little boys in white and red are playing football on the playground

Figure 6. Example of model results.

visual attention, the model notices people's movements and behaviors, and successfully recognizes that the person's movements are bent and the act is picking up things. The appearance of actions and behaviors will affect the accuracy of scene recognition. Therefore, the results generated by NIC model only show the person in the picture, and the sentences with visual attention show that the person is bending to pick up things. Compared with the results of the above two models, the double attention model in this paper, when describing, judges the image background as the setting sun, and judges it as a little girl according to the rough outline of the characters, and judges it as a shell picking according to the text attention creativity. This model improves the quality and richness of description generation.

The Chinese description of NIC model of the second image in **Figure 6** is "a group of people standing on the playground". The model grasps the main contents of the image "a group of people" and "the playground", and does not capture the smaller contents such as "football" and "goal", so the generated description is inconsistent with the meaning of the image. The model with attention mechanism focuses on the small object "football", and judges that the group is playing football according to the characteristics learned before. Compared with the results of the above two models, the double attention model in this paper is better in detail description. According to the figure's outline, we can catch that this is a group of little boys, and according to the color characteristics of the figure, we can divide the little boys into "red dressed" and "White dressed". This model improves the level of detail description.

3) The evaluation scores of da-nic, NIC and CNIC are compared, and the comparison results are shown in Table 2 and Table 3. Through comparison, it can be found that da-nic model effectively improves the model of using object category information or scene category information as prior knowledge alone; on flickr8k CN data set, its bleu4 index reflecting sentence consistency and accuracy is 4.4 higher than NIC model, 3.5 higher than CNIC model; on the rouge-l index reflecting the correlation between image and generation description, It is 3.9 higher than CNIC model and 5.3 higher than NIC model. In terms of cider index reflecting semantic richness, it is 2.1 higher than CNIC model, 7.5 higher than NIC model, and in the same way, Bleu and cider index are higher than NIC and CNIC model in flickr30k CN data set. This fully demonstrates the effectiveness of the proposed model.

4) In addition, the performance of the proposed model is compared with other mainstream models. As shown in Table 4, in the flickr8k CN data set, the performance of the model proposed in this paper exceeds the f-nic model [16] based on fluency guidance and the m-nic model based on multimodality in the reference [17] and the other two models in the cider index, the rouge-l index and the meter index in the bleu4 index and the rouge-l index in the flickr30k CN dataset. On the other hand, it performs well, surpassing the other two models, especially on the rouge-l, which improves the performance compared with the attention based model, as shown in Table 5. This shows that the method used in this paper pays more attention to the correlation between the image and the generated description, and the quality of the generated sentences is also higher.

Table 2. Comparison of experimental results of dual attention model and benchmark model on flickr8k CN dataset.

Model	BLEU-4	Rouge	CIDEr
NIC	32.7	61.1	108.1
CNIC	33.5	62.5	113.5
DA-NIC	37.1	66.4	115.6

Table 3. Comparison of experimental results of dual attention model and benchmark model on flickr30k CN dataset.

Model	BLEU-4	Rouge	CIDEr
NIC	26.5	52.3	99.8
CNIC	29.4	54.6	107.5
DA-NIC	33.5	63.2	110.6

Table 4. Performance comparison between our model and other models on flickr8k CN dataset.

Model	BLEU-4	Rouge	CIDEr	METEOR
F-NIC	34.5	63.2	110.6	31.4
M-NIC	33.6	65.3	109.5	32.1
DA-NIC	37.1	66.4	115.6	32.6

Table 5. Performance comparison between our model and other models on flickr30k CN dataset.

Model	BLEU-4	Rouge	CIDEr	METEOR
F-NIC	31.5	60.8	109.5	23.7
M-NIC	29.1	58.1	107.6	22.5
DA-NIC	33.5	63.2	110.6	25.3

Table 6. Visualizations of dual attention layers.

A			
	(Four boys play volleyball on the beach at sunset)	(Boy) (Volleyball)	(Four)
	Original image	Visual attention	Text attention
B			
	(A little girl in blue belt trousers is walking on the grass with a dog)	(Little girl) (Dog)	(Blue belt trousers)
	Origin image	Visual attention	Text attention

4.4. Visualizations of Dual Attention Layers

In this paper, we use the dual attention mechanism to guide the correlation between the generated description sentences and images. In order to verify whether the attention mechanism pays attention to the image region information, and whether the region information plays a guiding role in the generated words, we try to use the hot spot map to visualize and view the image region of interest. Table 6 shows the two images selected from the validation set and their generated focus visualization images.

In the above-mentioned attention visualization hot spot graph, the bright color indicates the area to be concerned, and the dark color indicates the area to be ignored. The scene of figure A is very dark and the characters are very small, which is difficult to describe accurately. However, through the attention mechanism, we can successfully focus on the four boys, volleyball and other entities, figure B shows the area of little girl, dog, blue belt pants and so on.

5. Conclusion

In this paper, an image Chinese description model is built using large-scale Chinese datasets obtained from machine translation. Aiming at the problem that the image description does not match the image style, a dual attention

mechanism is proposed to improve the quality of sentence generation. Experiments on Flickr8k-cn data set show that the dual attention mechanism of image and text can effectively improve the quality of final sentence prediction. Compared with the Google NIC model guided by attention mechanism, the CIDEr index on the Flickr8k-cn was increased from 108.1 to 115.6 [10], and on the manual translation test set, the CIDEr index on the Flickr8k-cn was increased from 27.9 to 31.5. At the same time, experiments show that dual attention mechanism can further improve the quality of final prediction sentences, which indicates that image description model still has a lot of room to improve. Image description should not be limited to one sentence, but should be combined with image time and scene to enhance the richness and comprehensiveness of image description text. This is also the future research work. In addition, the following work applies the model to the current popular AI Challenger Chinese image data set.

Acknowledgements

We would like to acknowledge all respondents, and colleagues those worked with us, I will listen carefully to your opinions and correct them.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Chen, X., Ma, L., Jiang, W., Yao, J. and Liu, W. (2018) Regularizing RNNS for Caption Generation by Reconstructing the Past with the Present. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 18-23 June 2018, 1-9. https://doi.org/10.1109/CVPR.2018.00834

[2] Mathews, A., Xie, L. and He, X. (2018) SemStyle: Learning to Generate Stylised Image Captions Using Unaligned Text. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 18-23 June 2018, 8591-8600. https://doi.org/10.1109/CVPR.2018.00896

[3] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., *et al.* (2017) Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, 18-23 June 2018, 6077-6086. https://doi.org/10.1109/CVPR.2018.00636

[4] Bernardi, R., Cakici, R., Elliošš, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A. and Plank, B. (2016) Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. *Journal of Artificial Intelligence Research*, **55**, 409-442. https://doi.org/10.1613/jair.4900

[5] Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T. and Rohrbach, M. (2016) Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. *Proceedings of the* 2016 *Conference on Empirical Methods in Natural Language Processing*, Austin, TX, November 2016, 457-468. https://doi.org/10.18653/v1/D16-1044

[6] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 27-30 June 2016, 770-778. https://doi.org/10.1109/CVPR.2016.90

[7] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D. (2015) Show and Tell: A Neural Image Caption Generator. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 3156-3164. https://doi.org/10.1109/CVPR.2015.7298935

[8] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. (2015) Learning Deep Features for Discriminative Localization. 2016 *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 27-30 June 2016, 2921-2929. https://doi.org/10.1109/CVPR.2016.319

[9] Ellioš, D., Frank, S. and Hasler, E. (2015) Multilingual Image Description with Neural Sequence Models. arXiv preprint arXiv:1510.04709.

[10] Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L. and Xu, W. (2015) Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering. *Computer Science*, 1-10.

[11] Rabuñal Dopico, J.R., Dopico, J. and Pazos, A. (2008) Encyclopedia of Artificial Intelligence: Volume 3. Encyclopedia of Artificial Intelligence. Information Science Reference. IGI Publishing, Hershey, PA. https://doi.org/10.4018/978-1-59904-849-9

[12] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. (2002) BLEU: A Method for Automatic Evaluation of Machine Translation. In: *Proceedings of* 40*th Annual Meeting of the Association for Computational Linguistics*, 311-318. https://doi.org/10.3115/1073083.1073135

[13] Michael Denkowski, A.L. (2010) METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages. *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR*, Uppsala, Sweden, July 2010, 339-342.

[14] Hori, C. (2003) Evaluation Methods for Automatic Speech Summarization. 8*th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 1-4 September 2003.

[15] Vedantam, R., Zitnick, C.L. and Parikh, D. (2014) CIDEr: Consensus-Based Image Description Evaluation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 4566-4575. https://doi.org/10.1109/CVPR.2015.7299087

[16] Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollar, P., Gao, J., He, X., Mitchell, M., Plaš, J., Zitnick, L. and Zweig, G. (2015) From Captions to Visual Concepts and Back. 2015 *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, 7-12 June 2015, 1473-1482. https://doi.org/10.1109/CVPR.2015.7298754

[17] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L. and Parikh, D. (2015) VQA: Visual Question Answering. 2015 *IEEE International Conference on Computer Vision*, Santiago, Chile, 7-13 December 2015, 2425-2433. https://doi.org/10.1109/ICCV.2015.279