Research on image content description in Chinese based on fusion of image global and local features

Dongyi Kong, Hong Zhao, Xiangyan Zeng

1 School of Computer and Communication, Lanzhou University of Technology, Lanzhou, China, 2 Department of Mathematics and Computer Science, Fort Valley State University, Fort Valley, GA, United States of America

* zhaoh@lut.edu.cn

Abstract

Most image content modelling methods are designed for English description which is different form Chinese in syntax structure. The few existing Chinese image description models do not fully integrate the global features and the local features of an image, limiting the capability of the models to represent the details of the image. In this paper, an encoder-decoder architecture based on the fusion of global and local features is used to describe the Chinese image content. In the encoding stage, the global and local features of the image are extracted by the Convolutional Neural Network (CNN) and the target detection network, and fed to the feature fusion module. In the decoding stage, an image feature attention mechanism is used to calculate the weights of word vectors, and a new gating mechanism is added to the traditional Long Short-Term Memory (LSTM) network to emphasize the fused image features, and the corresponding word vectors. In the description generation stage, the beam search algorithm is used to optimize the word vector generation process. The integration of global and local features of the image is strengthened to allow the model to fully understand the details of the image through the above three stages. The experimental results show that the model improves the quality of Chinese description of image content. Compared with the baseline model, the score of CIDEr evaluation index improves by 20.07%, and other evaluation indices also improve significantly.

1. Introduction

Image content description, also known as image semantic understanding, uses computer vision and deep learning technology to extract the semantics contained in the image, and use natural language processing technology to generate a reasonable text description [1, 2]. Image content description belongs to the task of cross modal transformation, which is different from image classification and object detection task of computer vision. Image semantic understanding includes extracting the semantic information in an image and converting it into fluent text descriptions. Image content description breaks through the barriers between computer vision
and natural language processing, and organically integrates the two directions in the field of artificial intelligence. The challenge is to accurately transform the image features obtained by computer vision into the text information generated by natural language processing.

With the wide application of deep learning, convolutional neural network has been used in the field of image content description. Through convolutional neural network, the model can extract image features independently to understand the semantic information contained in the image, and get better results in the task of image content description. However, there are some problems in the current work. First, most of the existing researches either use convolution neural network to extract global image features [3–8] or use Fast R-CNN [9] and other target detection networks to recognize targeted local objects in the image [10–12]. They do not integrate the global features and the local features, which limits the improvement of the model description performance. Secondly, in the descriptive statement generation stage of the existing models [7, 13–15], the word vector with the highest probability among the candidate words is chosen by the model as the final word and output it directly as a sentence. However, the sentence obtained in this way is not necessarily the best descriptive result. Thirdly, most of the current research work focus on English image description [5–8, 10], There are only a few researches on image content description in Chinese [4, 11, 16].

We propose a Chinese image content description method based on the fusion of global and local features. The main contributions are as follows: (1) In the encoding stage, the global and local features of the image are first extracted respectively. Then, the feature fusion module is constructed by using convolution kernel, Multilayer Perceptron (MLP) and other technologies to fuse the global and local features. (2) In the decoding stage, attention weights are calculated for image features and a new gate unit is added to the LSTM network to adjust image fusion features with attention information, where the model can better focus on the global and local features in the image fusion features to improve the decoding ability of the model. (3) In the description statement generation stage, beam search [17] is used to search and combine the candidate word vectors generated by the model to optimize the description generation results and further improve the model description generation effect.

2. Related work

2.1 Encoder-decoder structure

Mao et al. [18] proposed a multimodal Recurrent Neural Networks (m-RNN) model. In this model, the encoder decoder structure is used in the field of image content description for the first time. A convolutional neural network is used to encode the image, and a Recurrent Neural Network (RNN) [19] is used to decode the extracted features, realizing cross modality fusion between image feature information and text description information. Vinyals et al. [6] proposed the Neural Image Caption (NIC) model, which replaces the decoder in the m-RNN model with LSTM [20], which makes the model have strong long-term memory ability and improves the description performance of the model. Wu et al. [21] constructed a large-scale ICC dataset with the most comprehensive scenes and the richest language description. The dataset contains 300000 images and 1.5 million Chinese description sentences. The NIC model is used to verify the dataset, and the results show that the dataset effectively improves the performance of the existing models.

2.2 Attention mechanism

Xu et al. [7] proposed an image description method with a feature attention mechanism inspired by work in machine translation. It calculates the weights of words in the description text and generates an image feature vector with the weight information. The decoding network
can adjust image features with different weights at different times, which can enhance image features and alleviate overfitting. Lu et al. [8] proposed an image description model based on an adaptive attention mechanism, which calculates the importance of image description vocabulary in images by introducing visual sentinels. The visual sentinels decide whether the final predicted vocabulary is generated directly using a language model or by using the attention mechanism to calculate the attention weights of the word vectors. This assigns greater weight to the more important features in the image. Liu et al. [4] proposed a Chinese image content description model based on via visual attention [22] and topic modeling, which reduces the bias between image semantics and description statements by adding visual attention mechanism, and improves the accuracy of description statement generation by extracting the theme features of images through theme modeling. Zhao et al. [23] proposed a Chinese description method of image content based on image feature attention and adaptive attention fusion. The two attention mechanisms of literature [7] and literature [8] are fused in depth to extract more accurate information about the main features in the image, which effectively improves the image understanding and description capability of the model.

2.3 Local image features

Anderson et al. [10] proposed a Bottom-Up and Top-Down (BUTD) attention model, in which Faster R-CNN was used to extract local features in the image and a bottom-up attention mechanism was used to identify the image feature areas, and then a top-down attention mechanism was used to determine the weight values of image features. Ma et al. [11] proposed an improved Chinese image description model based on a global attention mechanism. This model adds global image features to the BUTD, which effectively overcomes the semantic loss caused by the loss of global features. However, this model does not deeply fuse global and local features. Li and Chen [14] proposed an image description model based on the fusion of image local features and label attributes. The model uses the target detection method and attribute trainer to extract the local features of the image and the Attributes as high-level semantic of the image features, and decodes the two features after fusion. Zhang et al. [12] used Faster R-CNN to extract local image features, used visual semantic attention model to generate visual keywords, and added optimized pointer network to the model, so that the model can receive variable length input sequence.

The above methods [3–8, 10–12, 23] obtain either global or local features, which will lead to incomplete image semantic features. Only [11, 14] combined global and local features, but did not deeply integrate the two image features, resulting in poor image content description effect.

3. Chinese image description model

3.1 Model design

Based on the encoder decoder framework, this research constructs a Chinese image description model with global and local image feature fusion. The encoder decoder structure was first proposed by Cho et al. [24], also known as sequence to sequence (seq2seq) structure [25], which is a model structure in deep learning, as shown in Fig 1.

Mao et al. [18] introduced this structure into the field of image content description. The encoder encodes the input information into an intermediate semantic vector, and the decoder decodes the semantic vector to get the output result. The two parts are independent but closely related to each other, which is conducive to the conversion between different modes. The encoding and decoding process is as follows:

\[ V_C = \text{Encoder}(x_1, x_2, \ldots, x_m) \]  

(1)
\[ y_t = \text{Decoder}(V_e; y_1, y_2, \ldots, y_{t-1}) \]  
(2)

Where \( x_1, x_2, \ldots, x_m \) is the input sequence of the encoder decoder structure, \( V_e \) is the semantic vector generated by the encoder, and \( y_t \) is the output value of the decoder at \( t \) time.

### 3.2 Encoding stage

The model’s extraction of image features is divided into two modules: global feature extraction and local feature extraction. The global image feature extraction module uses CNN to extract the global features of the image. The local image feature extraction module uses the backbone network to extract the underlying features of the image, and then uses the image Region Proposal Network (RPN) to search the local objects and get their coordinate information. Finally, the Region of Interest (RoI) pooling network is used to extract the local features that represent the image details from the underlying features through feature mapping. The fusion of global features and local features generates the image content, as shown in Fig 2.

#### 3.2.1 Global image feature extraction

The global image feature extraction module is a pre-trained ResNet152 [26], where the average pooling layer and the fully connected layer of \( 1 \times 1 \) are replaced by the average pooling layer of \( 14 \times 14 \). The global image feature vector \( V_g \) is extracted by the network:

\[ V_g = \{ v_1, v_2, \ldots, v_i, \ldots, v_n \} \]  
(3)

Where \( v_i \in \mathbb{R}^M \) is the image feature at any position in the image feature vector, and \( M \) is the dimension size of the image feature; \( n \) is the number of image features.

#### 3.2.2 Local image feature extraction

The common features extracted by the model through ResNet50 are shared by the subsequent local candidate region generation network and RoI pooling network to form the underlying image feature \( V_p \) as follows:

\[ V_p = \{ v'_1, v'_2, \ldots, v'_i, \ldots, v'_k \} \]  
(4)

Where \( v'_i \in \mathbb{R}^N \) is the image feature at any position in the common feature graph, and \( N \) is the dimension size of the image feature; \( k \) is the number of image features.

By using RPN and Non-Maximum Suppression (NMS) [27] algorithm, the local target object is screened out from the image bottom feature \( V_p \) and its coordinate information \( G \) is predicted. Then, according to the coordinate information \( G \) of the candidate box, the features are extracted from the bottom feature \( V_p \) through the mapping relationship, and the image content description in Chinese

![Fig 1. Encoding and decoding structure of Chinese description of image content.](https://doi.org/10.1371/journal.pone.0271322.g001)
feature $V_p$ of the ROI region is obtained as follows:

$$V_R = \text{RoI}(V_p, G)$$

Finally, the RoI pooling network is used to extract image feature $V_R$ of the region of interest (ROI) corresponding to the candidate box, and local feature vector $V_l$ with fixed size is obtained as follows:

$$V_l = \text{RoI Pooling}(V_R)$$

### 3.2.3 Global and local image feature fusion.

The global and local features of the image are sent to the image feature fusion module, as shown in Fig 3. The feature fusion module consists of three components: global image feature information processing, local image feature information processing and image feature fusion.

In the first component, the global image feature $V_g$ is fed into the convolution layer with convolution kernel size of $2 \times 2$ to extract the global feature, which is then sent to a Multilayer Perceptron network. The main functions include: (1) dimensionality reduction on the characteristic vector to reduce the complexity of the model and prevent over fitting; (2) weighting image features to facilitate the subsequent image feature fusion. The calculation process is as follows:

$$V_a = \text{MLP}(\text{Conv}^2d(V_g))$$
Where \( \text{Conv}^{2d}_{2 \times 2}(\cdot) \) represents 2D convolution operation with convolution kernel size of 2×2, and \( \text{MLP}(\cdot) \) represents Multilayer Perceptron network.

In the second component, the local image feature \( V_l \) is passed through a MLP network and two convolution layers with kernel size of 2×2 and 5×5. Compared with the global image feature \( V_g \), the local feature \( V_l \) has more quantity and more detailed information. Therefore, when processing the local feature \( V_l \), the convolution layer with the convolution kernel size of 5×5 is used to extract the feature at a deeper level. Then sent it to the Multilayer Perceptron network for processing. The calculation process is as follows:

\[
V'_l = \text{MLP}(V_l)
\]  

\[
V_b = \text{MLP}(\text{Conv}^{2d}_{5 \times 5}(\text{Conv}^{2d}_{2 \times 2}(V'_l)))
\]

Where \( \text{Conv}^{2d}_{5 \times 5}(\cdot) \) means to perform 2D convolution operation with convolution kernel size of 5×5.

In the third component, the global image feature \( V_g \) and local image feature \( V_b \) are fused. Firstly, the matrix addition operation is used to fuse the two features. In the model training, the proportion between the two features can be dynamically adjusted to achieve the best fusion effect. Then the fused features are sent to the convolution layer with convolution kernel of 1×1 for further fusion, and finally the fused image feature \( V_f \) is obtained. The calculation process is as follows:

\[
V'_f = \text{MLP}(V'_f)
\]
as follows:

\[
V_f = \text{Conv}2d_{1 \times 1}(V_a \oplus V_b)
\]  

(10)

Where, \(\oplus\) represents matrix addition operation, and \(\text{Conv}2d_{1 \times 1}()\) represents 2D convolution operation with convolution kernel size of \(1 \times 1\).

3.3 Decoding phase

In the decoding stage, the model maps the lexical information described in the text of the training set to the corresponding image feature area through the image feature attention mechanism. The calculation process is as follows:

(1) As shown in Fig 4, the attention weight of each region of the image feature at \(t\) time is calculated. Firstly, MLP is used to couple the image feature \(V_f\) with the hidden information \(h_{t-1}\) output by the decoder at the last time. Then, the above calculation results are sent into SoftMax function to calculate the weight value \(\phi_i\) of the \(i\)-th image feature region at time \(t\), and the weight distribution \(\phi\) of each region of the image can be obtained. The sum of the weight distribution is 1, that is \(\sum \phi_i = 1\). These weight distributions represent the attention degree of the word vector information at time \(t\) to each region of the image, as follows:

\[
d_i = W_{f\_att}(\text{ReLU}(W_e \cdot [h_{t-1}, V_f] + b_e) + b_{f\_att})
\]

\[
\phi_i = \frac{\exp(d_i)}{\sum_{i=1}^m \exp(d_i)}
\]

(12)

Where \(W_{f\_att}, W_e, b_{f\_att}\) and \(b_e\) are the weight parameters and bias parameters that the

![Fig 4. Attention weight of image feature.](https://doi.org/10.1371/journal.pone.0271322.g004)
Multilayer Perceptron needs to learn, ReLU [28] represents the Rectified Linear Unit, and $m$ represents the number of image features.

(2) The attention weight is mapped to the image features. Firstly, the attention model of image features focuses on the target in the image features by threshold $\lambda_t$. Then, the weight distribution $\phi_t$ calculated above is applied to the corresponding image region, and finally the image feature vector $q_t$ with weight information at time $t$ is obtained, as follows:

$$\hat{\lambda}_t = \sigma(W_\beta h_{t-1})$$

(13)

$$q_t = \hat{\lambda}_t \sum_{i=1}^{L} \phi_t v_i$$

(14)

Where $L$ is the number of image feature regions and $W_\beta$ is the weight parameter that threshold $\lambda_t$ needs to learn.

(3) The generated image feature attention vector is adjusted dynamically. As shown in Fig 5, A new gate unit $r_t$ is added to the traditional LSTM network as follows:

$$r_t = \sigma(W_r \cdot h_{t-1} + b_r)$$

(15)

The gate unit dynamically adjusts the image feature $q_t$ with attention information, so that the attention mechanism of image feature can fully pay attention to the global and local features in the image fusion features. In this way, attention can be paid to the information in the...
global and local image fusion features more accurately. The calculation process is as follows:

\[ v_t = r_t \otimes q_t \]  \hspace{1cm} (16)

Where \( \otimes \) is the matrix multiplication operation.

Through the above calculation process, the dynamically adjusted image fusion feature \( v_t \) with attention weight information is obtained. The detailed calculation process of the LSTM network input and output is as follows:

1. Given the image feature vector \( v_t \) and the word vector \( w_t \) in the training dataset, the LSTM network input \( x_t \) is obtained as follows:

\[ x_t = \{ v_t; w_t \} \]  \hspace{1cm} (17)

Where, {;} represents the splicing of two vectors.

2. The semantic hiding state \( h_t \) of LSTM at the current time is obtained from the previous value and the fused vector \( x_t \), as follows:

\[ h_t = LSTM(x_t, h_{t-1}) \]  \hspace{1cm} (18)

3. Finally, a SoftMax function is applied to the output of the fully connected layer as follows:

\[ p_t = \tanh(W_p h_t + b_p) \]  \hspace{1cm} (19)

\[ y_t = \text{softmax}(W_y p_t + b_y) \]  \hspace{1cm} (20)

Where \( W_p, W_y \) and \( b_p, b_y \) are the weight parameters and bias parameters that the Multilayer Perceptron needs to learn.

3.4 Description statement generation phase

In the model reasoning and testing stage of image description generation, a probability vector is generated where each element represents the probability value of each word vector in the dictionary at the current time. In the description generation stage, a greedy search algorithm is mostly used to find the word vector with the highest probability as the predicted word vector at the current time. Although this algorithm can ensure that each word is optimal by itself, they may be less desired when combined into a sentence.

To improve the efficiency of the search, we use the beam search algorithm based on breadth-first search. It generates all successors of the states at the current level, sorting them in increasing order of heuristic cost. However, it only keeps a predetermined number of optimal nodes while pruning the other nodes. This algorithm can reduce the computational cost and yield sentences that are more fluent.

4. Experimental design and result analysis

4.1 Dataset and evaluation index

The ICC Chinese image description dataset [21] was used for the experiment. The data set contains 210000 training pictures, 30000 verification pictures and 30000 test pictures. Each picture has five Chinese image description sentences corresponding to it.

In order to conduct a fair comparison between different models, BLEU (1–4) [29], METEOR [30], ROUGEL [31] and CIDEr [32], which are widely used in the field of image
content description, are used as evaluation indexes. The CIDEr evaluation index is specially designed for image description task, which can objectively evaluate the performance of image description model.

4.2 Data preprocessing and experimental parameter setting

Before the model training, the images of the original dataset are uniformly scaled to 256*256 pixels. In order to increase the generalization ability of the model, the scaled images are randomly cropped to 224*224 pixels and randomly rotated by 15°. Using the "jieba" word segmentation tool, the description text labels in the dataset are segmented, and the words with frequency greater than 5 are reserved. Each word is represented by a unique number to form a dictionary of the dataset. The final size of the dictionary is 7768.

In the encoder, the IoU threshold of NMS algorithm is set to 0.5, and the top 100 candidate frames with higher prediction probability are selected from the filtered candidate frame set. In the decoder, the word vector input dimension and network output dimension of LSTM network are set to 512, and the hidden layer dimension of image feature attention is set to 512. In the model training phase, the Batch Size of batch training is set to 128, and the initial learning rate of encoder and decoder is 0.0001. The model uses Adam [33] to optimize the parameters. In the back-propagation, the gradient of each round of training is trimmed to prevent the gradient explosion of the model. When the word vector is generated by the model, dropout [34] technology is used to prevent the model from over fitting. The parameter value of dropout is 0.5.

4.3 Experimental results and analysis

4.3.1 Model training and performance comparison. In the model encoder network, the global feature extraction module ResNet152 and the local feature extraction module ResNet50 and RPN are initialized with the parameters of the pre-training model; the decoder network parameters are initialized with random parameters. In the initial stage of model training, the decoder network does not have the ability to decode. In order to prevent the large error produced by the decoder, the network parameters of the encoder are fixed in the initial stage. When the evaluation index score of the model in the verification set converges, the network parameters of the encoder are fixed. The encoder network and decoder network are trained jointly.

The evaluation index scores of the validation set in each epoch of the model training are shown in Fig 6. In order to prevent the random parameters of the decoder network from affecting the pre-training parameters of the encoder, the parameters of the encoder network were frozen in the first 20 epochs. At the 21st epoch, the score of the evaluation index increases significantly because the encoder parameter freeze was removed. Henceforward the coder and decoder of the model were trained jointly, which broke through the bottleneck of the model decoder network, and the evaluation index score of the model has been significantly improved.

In the decoder of Chinese image description method in reference [23], image feature attention mechanism and adaptive attention mechanism are added, and the two attention mechanisms are fused. This method can effectively improve the image understanding ability of the model. In order to verify the influence of the attention fusion mechanism on the global and local image feature fusion mechanism proposed in this research, a comparative experiment as shown in Table 1 is designed with the same dataset.
1. GLF: the model is global and local image feature fusion (GLF) model without attention mechanism.

2. GLF-ATTF: this model is a global and local image feature fusion model with the attention fusion mechanism proposed in reference [23].

3. GLF-ADPATT: this model is a global and local image feature fusion model with a single adaptive attention mechanism.

4. GLF-IFATT: this model is a global and local image feature fusion model with a single image feature attention mechanism.

According to the experimental results in Table 1, the attention fusion mechanism proposed in reference [23] only improve the performance by a small margin. The score of CIDEr evaluation index only increases by 1.401%, and other evaluation index values are almost the same. In addition, a single adaptive attention mechanism does not improve the performance either. When the single attention mechanism of image features is combined with the global and local
image feature fusion method, the scores of all evaluation indices improve compared with those without any attention mechanism, and the score of CIDEr evaluation index increases by 3.439%. Therefore, it can be verified that in the fusion method of global and local image features proposed in this research, the number of local image features is far more than that of global image features. If the adaptive attention mechanism is applied to this model, due to the addition of local detail features, the performance of the model does not improve significantly. The attention mechanism of image feature can calculate the attention information of any image feature according to the word vector, so it has better performance when combined with global and local feature fusion methods.

In order to verify the effectiveness of this model, comparative experiments with other models are carried out using the same dataset, as shown in Table 2.

1. Baseline NIC [21]: this model is the experimental result of NIC model on ICC dataset, and NIC model is often used as the baseline model in this field.
2. BUTD [10]: this model is the Bottom-Up and Top-Down attention model proposed in reference [10], and its experimental results are the reproduction results of reference [11] on ICC dataset.
3. GATT [11]: this model is a global attention model proposed in reference [11], which adds global image features to the BUTD model and uses attention mechanism to improve the performance of the model.
4. G-IFATT [7]: this model uses a single global image feature, and adds the image feature attention [7] to the image Chinese description model, which can reproduce the results on the ICC dataset.

According to the experimental results in Table 2, compared with the Baseline NIC model, the GLF-IFATT model has greatly improved the evaluation indices, and the score of CIDEr evaluation index has increased by 13.96%. Compared with the BUTD model that only uses partial image features, the GLF-IFATT model has a 15.26% improvement in the score of the CIDEr evaluation index. Compared with the GATT model, the GLF-IFATT model yields a 13.33% improvement in the score of the CIDEr evaluation index. Compared with the G-IFATT model, the GLF-IFATT model increases the CIDEr evaluation score by 10.78%.

Table 1. Evaluation scores of attention mechanism comparison experiment.

| Model     | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGEL | CIDEr |
|-----------|-------|-------|-------|-------|--------|--------|-------|
| GLF       | 0.790 | 0.677 | 0.579 | 0.494 | 0.392  | 0.657  | 1.570 |
| GLF-ATTF  | 0.788 | 0.674 | 0.576 | 0.493 | 0.394  | 0.658  | 1.592 |
| GLF-ADPATT | 0.788 | 0.674 | 0.575 | 0.491 | 0.394  | 0.658  | 1.578 |
| GLF-IFATT | 0.798 | 0.687 | 0.591 | 0.507 | 0.397  | 0.665  | 1.624 |

Table 2. Evaluation scores of comparison experiment with other models.

| model                  | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGEL | CIDEr |
|------------------------|-------|-------|-------|-------|--------|--------|-------|
| Baseline-NIC [21]      | 0.765 | 0.648 | 0.547 | 0.461 | 0.370  | 0.633  | 1.425 |
| BUTD [10]              | 0.780 | 0.662 | 0.561 | 0.476 | 0.379  | 0.637  | 1.409 |
| GATT [11]              | 0.782 | 0.665 | 0.564 | 0.477 | 0.380  | 0.638  | 1.433 |
| G-IFATT [7]            | 0.792 | 0.670 | 0.566 | 0.479 | 0.382  | 0.653  | 1.466 |
| GLF-IFATT              | 0.798 | 0.687 | 0.591 | 0.507 | 0.397  | 0.665  | 1.624 |
Therefore, the global and local images along with the feature fusion mechanism can effectively improve the image description performance of the model.

4.3.2 Description generation experiment based on beam search. The beam search algorithm is used to optimize the description generation stage of the model. In order to verify the influence of different beam widths on the model evaluation index score, we use the optimal model obtained above and use beam search algorithms with 10 different beam widths to optimize the model description generation stage, and finally use the test set to calculate the evaluation index score of the model.

The evaluation scores of different beam widths are shown in Table 3. It can be seen from the table that when the beam width is 2 and 3, the beam search algorithm has the most obvious optimization effect on the description generation of the model. When the beam width is greater than 3, the evaluation index score of the model begins to increase slowly. When the beam width is 7, the evaluation index score reaches the maximum, and the model achieves the optimal performance.

The results with the beam width value of 7 are compared with the Baseline-NIC model and the greedy search algorithm as shown in Table 4.

1) GLF-IFATT-GS: GLF-IFATT model uses greedy search for description generation.
2) GLF-IFATT-BS: GLF-IFATT model uses beam search for description generation.

It can be seen that the beam search algorithm, compared with the greedy search algorithm, improves the scores of all evaluation indicators. The BLEU4 evaluation score is increased by 6.90%, and the CIDEr evaluation score is increased by 5.36%. Compared with the Baseline-NIC model, the model optimized by beam search algorithm significantly improved the evaluation index scores. The BLEU4 evaluation score increase by 17.57% and the CIDEr evaluation score increase by 20.07%. Therefore, the beam search algorithm can significantly improve the image description performance of the model.

Table 3. Impact of different beam widths on model evaluation scores.

| Beam Size | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGEL | CIDEr |
|-----------|-------|-------|-------|-------|--------|--------|-------|
| 1         | 0.799 | 0.688 | 0.592 | 0.509 | 0.395  | 0.664  | 1.619 |
| 2         | 0.810 | 0.705 | 0.610 | 0.528 | 0.398  | 0.669  | 1.677 |
| 3         | 0.816 | 0.712 | 0.619 | 0.538 | 0.400  | 0.671  | 1.701 |
| 4         | 0.815 | 0.713 | 0.620 | 0.540 | 0.399  | 0.671  | 1.702 |
| 5         | 0.813 | 0.711 | 0.620 | 0.540 | 0.399  | 0.671  | 1.704 |
| 6         | 0.813 | 0.711 | 0.620 | 0.541 | 0.398  | 0.670  | 1.707 |
| 7         | 0.813 | 0.712 | 0.621 | 0.542 | 0.399  | 0.671  | 1.711 |
| 8         | 0.811 | 0.709 | 0.618 | 0.539 | 0.397  | 0.669  | 1.700 |
| 9         | 0.811 | 0.710 | 0.619 | 0.540 | 0.397  | 0.669  | 1.704 |
| 10        | 0.810 | 0.710 | 0.619 | 0.540 | 0.397  | 0.669  | 1.705 |

Table 4. Comparison of different search methods in this model.

| Model      | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGEL | CIDEr |
|------------|-------|-------|-------|-------|--------|--------|-------|
| Baseline-NIC | 0.765 | 0.648 | 0.547 | 0.461 | 0.370  | 0.633  | 1.425 |
| GLF-IFATT-GS | 0.798 | 0.687 | 0.591 | 0.507 | 0.397  | 0.665  | 1.624 |
| GLF-IFATT-BS | 0.813 | 0.712 | 0.621 | 0.542 | 0.399  | 0.671  | 1.711 |

https://doi.org/10.1371/journal.pone.0271322.t003

https://doi.org/10.1371/journal.pone.0271322.t004
In addition, the beam search algorithm is applied to the GLF model without the attention mechanism to verify its effectiveness, as shown in Table 5.

Among them:

1) GLF-GS: GLF model uses greedy search for description generation.
2) GLF-BS: GLF model uses beam search for description generation.

The experimental results show that the beam search algorithm also improves the scores of various evaluation indicators in the GLF model that does not use the attention mechanism. For example, the BLEU4 score increase by 6.88% and the CIDEr score increase by 6.69%.

In addition, a qualitative comparison of the image description sentences generated by the G-IFATT model with only global image features and the GLF-IFATT model with the global and local image feature fusion is shown in Table 6.

In the first example of Table 6, the description generated by the G-IFATT model appears "草莓(strawberries)", but there is no strawberries in the image, while the GLF-IFATT model accurately describes the details of "右手拿着桶(with a bucket in his right hand)" and correctly describe the number of people in the image. In the second example, the GLF-IFATT model successfully describes the details of "helmet"; In the third example, the G-IFATT model does not correctly describe the number of people in the image, while the GLF-IFATT model accurately describes the details.

5. Conclusion

This paper proposes a Chinese description model of image content based on the fusion of global and local image features. Based on the encoder decoder network structure, the model is improved in the encoding phase, the decoding phase and the description generation phase. In the encoding stage, the global and local detail information of the image are extracted respectively, and the extracted features are sent to the feature fusion module to obtain fusion features. In the decoding stage, the attention mechanism of image features is added to pay attention to more important image features, which is specifically effective for the fused image features. In the description generation phase, the beam search algorithm is used to optimize the generation.

### Table 5. Comparison of different search methods of GLF model.

| Model | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGE-L | CIDEr |
|-------|-------|-------|-------|-------|--------|---------|-------|
| GLF-GS | 0.790 | 0.677 | 0.579 | 0.494 | 0.392 | 0.657 | 1.570 |
| GLF-BS | 0.807 | 0.703 | 0.609 | 0.528 | 0.395 | 0.665 | 1.675 |

### Table 6. Subjective comparison of description effects.

| Image Name | Image ID | Image | Description |
|------------|----------|-------|-------------|
| ae6247fa311e8ca5eb727e7ea1b77733ae882ae7.jpg | 7750225178903012713 | G-IFATT | 大棚里有两个弯着腰的人在采摘草莓 (Two hunched over people picking strawberries in a greenhouse) |
| 006440a6c45bd1c1d0cdd15c1fbad853b99e9d93b7e.jpg | 675623727970939378 | GLF-IFATT | 大棚里站着一个右手拿着桶的男人 (A man with a bucket in his right hand is standing in the greenhouse) |

* Images can be found and viewed in the test dataset "AIC-ICC Dataset Test A" by image name and image ID.
process of Chinese description of the model image. Finally, the effectiveness of the proposed model was verified by comparative experiments.

Author Contributions
Conceptualization: Dongyi Kong.
Formal analysis: Dongyi Kong.
Investigation: Dongyi Kong.
Methodology: Dongyi Kong.
Resources: Hong Zhao.
Software: Dongyi Kong.
Validation: Dongyi Kong.
Visualization: Dongyi Kong.
Writing – original draft: Dongyi Kong.
Writing – review & editing: Hong Zhao, Xiangyan Zeng.

References
1. Hossain M D Z, Sohel F, Shiratuddin M F, et al. A comprehensive survey of deep learning for image captioning[J]. ACM Computing Surveys (CoUR), 2019, 51(6): 1–36.
2. Liu X, Xu Q, Wang N. A survey on deep neural network-based image captioning[J]. The Visual Computer, 2019, 35(3): 445–470.
3. Zhao W, Wu X, Luo J. Cross-Domain Image Captioning via Cross-Modal Retrieval and Model Adaptation[J]. IEEE Transactions on Image Processing, 2020, 30: 1180–1192.
4. Liu M, Hu H, Li L, et al. Chinese Image Caption Generation via Visual Attention and Topic Modeling[J]. IEEE Transactions on Cybernetics, 2020.
5. Guo L, Liu J, Lu S, et al. Show, Tell, and Polish: Ruminant Decoding for Image Captioning[J]. IEEE Transactions on Multimedia, 2019, 22(8): 2149–2162.
6. Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 3156–3164.
7. Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention [C]. In: Proceedings of the 2015 International Conference on Machine Learning. New York: ACM, 2015: 2048–2057.
8. Lu J S, Xiong C M, Devi P, et al. Knowing when to look: adaptive attention via a visual sentinel for image captioning[C]. In: Proceedings of the 2017 International Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2017: 3242–3250.
9. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 39(6): 1137–1149.
10. Anderson P, He X D, Buehler C, et al. Bottom-up and top-down attention for image captioning and visual question answering[C]. In: Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2018: 6077–6086.
11. Ma S L, Zhang G B, Jiao Y, et al. Improved method for image caption with global attention mechanism [J]. Journal of Xidian University, 2019, 46(02): 17–22.
12. Zhang S, Zhang Y, Chen Z, et al. VSAM-Based Visual Keyword Generation for Image Caption[J]. IEEE Access, 2021, 9: 27638–27649.
13. Liu Y, Peng D, Zhang M. Learn and Tell: Learning Priors for Image Caption Generation[J]. Applied Sciences, 2020, 10(19): 6942.
14. Li N. and Chen Z., “Image Cationing with Visual-Semantic LSTM,” in Proc. IJCAI, 2018, pp. 793–799.
15. Chu Y, Yue X, Yu L, et al. Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention[J]. Wireless Communications and Mobile Computing, 2020, 2020.
16. Lan W., Wang X., Yang G., and Li X.-R., “Improving Chinese image captioning by tag prediction,” Chin. J. Comput., vol. 42, no. 1, pp. 136–147, Jan. 2019. https://doi.org/10.11897/SP.J.1016.2019.00136
17. Freitag M, Al-Onaizan Y. Beam search strategies for neural machine translation[J]. arXiv preprint arXiv: 1702.01806, 2017.
18. Mao J H, Xu W, Yang Y, et al. Explain images with multimodal recurrent neural networks[J]. arXiv preprint arXiv: 1410.1090, 2014.
19. Goodfellow I, Bengio Y, Courville A. Deep learning[M]. Cambridge: MIT Press, 2016: 367–415.
20. Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.
21. Wu J, Zheng H, Zhao B, et al. Large-scale datasets for going deeper in image understanding[C]. In: Proceedings of the 2019 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2019: 1480–1485.
22. Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[C]/Proceedings of the 27th International Conference on Neural Information Processing Systems. MIT Press, 2014(2): 2204–2212.
23. Zhao H, Kong D Y. Chinese description of image content based on fusion of image feature attention and adaptive attention[J]. Journal of Computer Applications, 2021, 41(09): 2496–2503.
24. Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv: 1406.1078, 2014.
25. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[J]. arXiv preprint arXiv: 1409.3215, 2014.
26. He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]. In: Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2016: 770–778.
27. Wang D, Chen X, Yi H, et al. Improvement of Non-Maximum Suppression in RGB-D Object Detection [J]. IEEE Access, 2019, 7: 144134–144143. https://doi.org/10.1109/ACCESS.2019.2945834
28. Rossig A, Petkovic M. Advances in verification of ReLU neural networks[J]. Journal of Global Optimization, 2020: 1–44.
29. Papineni K, Roukos S, Ward T, et al. BLEU: A method for automatic evaluation of machine translation[C]. In: Proceedings of the 2002 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2002: 311–318.
30. Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language[C]. In: Proceedings of the 2014 Ninth Workshop on Statistical Machine Translation. Stroudsburg, PA: Association for Computational Linguistics, 2014: 376–380.
31. Lin C Y. ROUGE: A Package for Automatic Evaluation of Summaries[C]. In: Proceedings of the 2004 Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2004: 74–81.
32. Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based image description evaluation[C]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2015: 4566–4575.
33. Kingma D, Ba J. Adam: a method for stochastic optimization[J]. arXiv preprint arXiv: 1412.6980, 2017.
34. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014; 15(1):1929–1958.