Chinese description generation of dual attention images based on multi-modal fusion

Xiyun Lv
College of Electrical Information, Southwest Minzu University, Chengdu, Sichuan, China
E-mail:1045106108@qq.com

Abstract: In view of the lack of attention to image features in most of the current research on image Chinese description generation, the low quality of image details, and the low accuracy of generated language description, a multi-modal fusion dual attention image description generation model is proposed. The model uses the pre-trained VGG19 network to extract image features, LSTM extracts key information features of text, and combines image features and text features through attention stitching to obtain the attention weight of each spatial position of the image, and then input it into the LSTM language model to generate descriptive sentences. In the AI Challenger 2017 test set, the BLEU-1 and CIDEr scores of this algorithm can reach 71.2 and 106.1 respectively, which are better than baseline models based on a single attention structure, such as NIC, and perform manual observation and comparison of the effects of several models. The model can generate natural language descriptions that can better represent the details of the picture.

Keywords. Multi-modal fusion; Long and short-term memory network; Image description; Attention mechanism

1. Introduction
The task of image text description technology is to describe and understand the language of a given image. The research content spans the fields of natural language processing and computer vision. It is a key technology for the multi-modal conversion of image generation text. This technology was first proposed by Farha-di et al.[1]. Assuming that I represents a picture and S represents its corresponding descriptive sentence, there are two-tuples(I, S). This technology needs to complete the multiplication from I to S. The goal of modal conversion is to make the machine describe the subject in the image, the scene in the image, the connection between the objects in the image, the attributes between them, and the activities they participate in with a Chinese sentence. In recent years, the multi-modal processing problem of fusing image information and text information has aroused the interest of researchers. Inspired by the seq2seq[2] model in the field of machine translation, the encoder-decoder framework based on deep neural networks has brought fresh blood to the field of image description generation.
The m-RNN (multimodal Recurrent Neural Network) model proposed by Mao et al.[3] abroad pioneered the use of deep learning methods to solve the problem of image description generation. The NIC (Neural Image Caption) model proposed by Vinyals et al.[4] uses an encoder-decoder framework. Compared with the m-RNN model, the image encoder of NIC uses a better CNN (specifically, GoogleNet[5]) to extract the global features of the image. Xu et al.[6] introduced the attention mechanism in the human visual system into the image description generation algorithm for the first time. In addition to introducing the attention mechanism to improve the effect of image description, Jia et al.[7] Semantic information is used as an additional input of LSTM to further guide the model to generate a more appropriate description of the image content. Ren et al. in 2017[8] introduced a new image description method based on reinforcement learning.

Most of the current research on image description generation technology is mainly focused on the English description of the image. Due to the complex structure and rich meaning of the Chinese sentence, it is difficult to be understood by the machine. Secondly, the Chinese data set of image description is relatively rare, which makes the Chinese image of the image. Descriptive research is more difficult. At present, most of the researches on Chinese image description have problems such as errors in the description of the image content in the generated Chinese description sentences and poor coherence and readability of the sentences. Based on the existing research, this paper proposes a multi-modal fusion dual attention image Chinese description automatic generation model. In the encoding stage, VGG19 and LSTM pre-trained on ImageNet are used to extract image and text features, and perform splicing and fusion to obtain feature weights. In the decoding stage, the attention LSTM module is used to generate descriptive sentences. The verification experiment was carried out on the AI Challenger2017 image Chinese description data set, which has the largest scale of Chinese image description data.

2. Image description generation model

Most of the existing image description generation models are based on the Encoder-decoder architecture. The Encoder part uses CNN for effective feature extraction and encoding of images, and the Decoder part uses RNN to generate description sentences to complete the multimodal conversion from image to text.

2.1. NIC model

The NIC model is a typical Encoder-decoder architecture model proposed for automatically generating semantic descriptions. The model proposed in this paper is improved based on the NIC[4] model as a framework. It is abstracted as a mathematical optimization problem, and the training process of the model is based on formula (1) to maximize the mapping between image features and sentences.

$$\theta^* = \arg \max \sum_{I, S} \log p(S | I, \theta)$$  \hspace{1cm} (1)

$\theta$ represents model parameters, which are learned by the network itself. I represents an image, and S represents a sentence described by the text. The model first uses Inceptionv3 to encode the image as a static representation, and then uses LSTM to decode this representation into a meaningful sentence to describe the content of the image.

Specifically, suppose the image is represented as $I$, and the corresponding descriptive sentence is marked as $(S_0, ..., S_N)$, and then the predicted words are generated according to formulas (2)-(4)[4].

$$x_{-1} = \text{CNN}(I)$$ \hspace{1cm} (2)

$$x_t = W_s S_t, t \in \{0, 1, ..., N-1\}$$ \hspace{1cm} (3)

$$P_{st} = \text{LSTM}(x_t), t \in \{0, 1, ..., N-1\}$$ \hspace{1cm} (4)

Among them, the input image feature $x_{-1}$ is the network data of the LSTM network at $t=-1$. $S0$ and $SN$ represent the beginning and ending symbols of the sentence, respectively. We is a two-dimensional matrix containing the length of the word vector.
2.2. Multimodal cyclic attention image Chinese description generation model framework

The framework of the Chinese description generation model for multimodal cyclic attention images proposed in this paper is shown in Figure 1 below. This paper adopts a dual-circular attention mechanism. At the encoder stage, the attention mechanism is adopted for image features and text features, and the attention mechanism is adopted when the decoder generates description sentences. The encoder is composed of an image feature extraction part and a text feature extraction part, and features modeling based on convolutional neural networks. The decoder is composed of a multi-modal sentence generation network. The decoder models the sequence based on a long short-term memory network (LSTM), and the output is a Chinese sentence description of the image. In the training process, the softmax cross-entropy loss function is used to update the model parameters.

![Figure 1. Multi-modal fusion dual attention image description production model framework](image)

2.2.1. Feature Fusion Encoder

The model uses the VGG19 model trained on ImageNet for image feature extraction, uses the output feature map of its last convolutional layer as the image feature, and the dimension is W*H*D, and then it is converted to L*D. L is the number of image feature regions, and the feature of each region is represented by a vector of dimension D. The keyword information is extracted from the text descriptive sentence. The keyword information contains the key information in the image description sentence. Each word is encoded into a D-dimensional vector, and then the characteristic representation of the keyword information is obtained through the LSTM network. After the image and text features, we use the method of splicing and fusion, and then calculate the attention weight of each spatial position of the image, and then use this weight and the hidden layer output state of the LSTM as the input of the language model LSTM.

2.2.2. Language model decoder

The decoder can decode the feature vector obtained by the encoder into natural language sentences conforming to human thinking. The decoder in this paper consists of an LSTM network and an attention mechanism module.

For the LSTM unit, it consists of three control units and two inputs. The two input terminals are gate $x_t$ and hidden layer state $h_{t-1}$. A core cell Cell is introduced. The state of Cell is controlled by three gates, including input gate $i_t$, forget gate $f_t$, and output gate $o_t$. The input gate $i_t$ is used to control the current sequence unit input $x_t$ and the previous sequence unit output $h_{t-1}$ to obtain the current LSTM information. The forget gate $f_t$ is used to maintain and clear the previous unit state signal of the LSTM. The output gate $o_t$ is used to calculate the output of the current unit of the LSTM. In forward propagation, the calculation of LSTM is based on formula (5), a-g[9].
\[
\begin{align*}
(a) \quad & i_t = \sigma(W_x x_t + W_{h_{t-1}} b_{i}) \\
(b) \quad & f_t = \sigma(W_x x_t + W_{h_{t-1}} b_{f}) \\
(c) \quad & o_t = \sigma(W_x x_t + W_{h_{t-1}} b_{c}) \\
(d) \quad & \tilde{c}_t = \phi(W_x x_t + W_{h_{t-1}} b_{\tilde{c}}) \\
(e) \quad & c_t = f_t \odot \tilde{c}_t + i_t \odot c_t \\
(f) \quad & h_t = o_t \odot \phi(c_t) \\
(g) \quad & P(S_{t+1} | I, S_t, S_{t-1}, ..., S_0) = \text{Soft} \max(W_f h_t)
\end{align*}
\]

\(\sigma\) is the sigmoid function, \(\phi\) is the hyperbolic tangent function, \(\odot\) is the multiplication operation, \(\tilde{c}_t\) is the candidate for updating the previous state in the memory unit, \(c_t\) is the state of the memory unit after the update, and \(W^*\) and \(b^*\) are the parameters of the model learning. Each word is represented by a one-hot encoding vector \(S_i\) equal to the size of the vocabulary.

We introduce an attention mechanism when generating descriptions, and different context vectors are generated at each time step of LSTM. Given an image and its corresponding descriptive sentence, the image is extracted by VGG19 features, denoted as \(I[I_1, ..., I_L], I \in \mathbb{R}^L\). The text feature is extracted by LSTM, denoted as \(T\). We splice the text keyword information behind each regional feature of the image, and then input it to the decoding end, as shown in Figure 2 below.

The attention mechanism is used to calculate the attention degree of different regions of the input feature in combination with the last hidden state of the LSTM at each time step. The formulas are shown in (6)-(8).

\[
\begin{align*}
\alpha_i &= f_{\text{att}}([I, T], h_{t-1}), \\
\alpha_i &= \text{soft} \max(e_i), \\
z_t &= \sum_{i=1}^{L} \alpha_i [I, T], \\
\end{align*}
\]

Where \(e_i\) is the output of the attention mechanism at time \(t\), \(\alpha_i\) is the weight of the region \(i\) of the input feature at time \(t\), and \(Z_t\) is the context vector input to the LSTM model at time \(t\). The context vector will be combined with the word embedding vector at the current moment into the LSTM model. At each time step, the hidden state \(h_t\) of the LSTM model is input into a softmax function, and then a probability distribution of all words is generated. The word with the highest probability is the word generated at the current time step.

![Figure 2. Dual attention LSTM model](image)

---

\(\text{Figure 2. Dual attention LSTM model}\)
2.2.3. Loss function

In this paper, the commonly used image description generation task cross entropy function is used as the loss function to train the model, and the formula is shown in (9). When the LSTM generates the next word, the model will make the probability distribution \( p(a_t | x, s_{t-1}) \) of action \( a \) based on the current input \( a_t \) and the historical state \( S_{t-1} \).

\[
\log p(S | x) = \sum_{t=0}^{h} \log p(S_t | x, S_{0..t-1}; \theta).
\]  

(9)

Where \( x \) is the input of the model, \( S \) is the image description sentence; \( \theta \) is the parameter of the model. During the training process, the model parameters will be updated by minimizing the cross-entropy function.

3. Experimental results and analysis

In this chapter, we will introduce our experiment and result analysis. Based on the CNN-RNN model framework, a multi-modal fusion dual attention image description generation model is realized. Through several commonly used evaluation indicators: BLEU (bilingual evaluation understudy)[10], ROUGE_L (recall-oriented understudy for gisting evaluation)[11], CIDEr (consensus-based image description evaluation)[12]. Comparison of experimental data verifies the effectiveness and accuracy of the proposed model. The models compared in the experiment include Google NIC[4], CS-NIC[14], CIC[15], RAL[16] and our proposed multimodal cyclic attention image Chinese description generation model. These two models are the latest research results in this field in the past two years.

3.1. AI Challenger 2017 image Chinese description data set

Deep learning requires a lot of data to drive. Nowadays, most of the image description generation data sets are based on English, such as Microsoft COCO, Flickr8k and Flickr30k, etc., which is different from the current English-based image description generation. Chinese description of the image. Compared with the English description of the image, the sentence structure of the Chinese description is more flexible. We use AI Challenger 2017 image Chinese description data set[17] for training. This data set is currently the largest image Chinese description data set created for image description problems. It contains 300,000 pictures and 1.5 million Chinese descriptions, including 210,000 training sets, 30,000 verification sets, and 30,000 test sets A. The test set B is 30,000 pieces, and each picture has five Chinese description sentences that are manually annotated and conform to Chinese natural language habits and elaborate on the main visual information of the image (such as main characters, scenes, actions, etc.). Specific examples are shown in Table 1 below.

| Table 1. Examples of images in the AI Challenger dataset and their description sentences |
| --- |
| ![Example Image 1](image1.png) |
| ![Example Image 2](image2.png) |
| ![Example Image 3](image3.png) |
3.2. Experiment and result analysis

The experimental platform is Ubuntu 14.04, the configured GPU is NVIDIA 1080ti, tensorflow 1.11.0 and python 3.6 are selected, and the number of model training iterations is 131,250. The decoder uses a two-layer LSTM structure, the number of hidden units is 512, and the word embedding size is set to 512. Unlike English sentences that use spaces as clear word boundary markers, Chinese sentences do not have such markers. Therefore, we use jieba Chinese word segmentation tool, jieba is an open source software for Chinese word segmentation, which can mark Chinese sentences as meaningful word lists. Use Adam as the optimizer for model training, the batch size is 8, and the learning rate is 0.001.

Regarding the evaluation indicators generated by image descriptions, BLEU[10], ROUGE_L[11], CIDEr[12], METEOR (metric for evaluation of translation with explicit ordering)[13] are commonly used. BLEU is currently the most popular evaluation index based on the n-gram mechanism. In this experiment, we choose BLEU-1, BLEU-2, BLEU-3, BLEU-4, CIDEr, ROUGE-L to verify the model performance. The above indicators are used to evaluate the quality of the generated language. The higher the value, the better the quality of the generated language.

Through 131250 iterations, 5000 verification images in the AI Challenger image Chinese description dataset are used to compare and verify the baseline NIC[4], CS-NIC[14], CIC[15], RAL[16] and our model. In the iterative process of model training, the optimized parameters of the model can be viewed through TensorBoard. Figure 3 below is the loss function graph of the proposed model. It can be seen from the figure that during the training process, the loss function continues to decay, and the performance of the model continues to improve.

| 1. In the kitchen there is a woman with a bowl in her left hand and a girl watching a man with a knife in his right hand cutting vegetables | 1. There is a sitting man and a woman kneeling on a mat playing chess on the board by the lake |
| 2. A woman with long hair and a shawl has a smiling girl in front of her, watching a man chopping vegetables in the kitchen | 2. Two casually dressed people are playing chess on the table by the water |
| 3. In front of a woman in the kitchen, there is a girl holding a bell pepper in her left hand, watching a smiling man cutting vegetables | 3. There are two smiling people playing chess on the board by the lake |
| 4. There is a man bending over in the kitchen cutting vegetables next to two people | 4. Two men in dark shirts are playing chess outdoors |
| 5. A man with a knife in his right hand is cutting vegetables beside two smiling people in a clean room | 5. There are two people in different clothes playing chess on the table by the sea |

| 1. A woman in a long-haired shawl and a man in a sleeveless top perform in the hall | 1. There is a sitting man and a woman kneeling on a mat playing chess on the board by the lake |
| 2. A man in a sleeveless shirt is playing guitar next to a woman with a microphone in her right hand | 2. Two casually dressed people are playing chess on the table by the water |
| 3. A man in a vest is playing guitar in the room next to a woman | 3. There are two smiling people playing chess on the board by the lake |
| 4. There is a man playing the guitar next to a smiling woman in the room | 4. Two men in dark shirts are playing chess outdoors |
| 5. A woman with a microphone in his hand and a man with a guitar in his hand play musical instruments in the room | 5. There are two people in different clothes playing chess on the table by the sea |

![Figure 3. Model loss function](image-url)
Figure 4 is a diagram of the improvement of the accuracy of each evaluation index of the model. The final data is in Table 2, which mainly verifies the quantitative validity of the three models on BLEU-1, 2, 3, 4, Rouge-L, and CIDEr. Table 2 shows the performance comparison between my model and the CS-NIC model, CIC model, NIC model, and RAL model.

As can be seen from the table 2, on the AI Challenger2017 image Chinese description test set, this model has improved all indicators compared to the NIC, CS-NIC, CIC, and RAL models, especially in BLEU-1 by 3.5% and 8.4% respectively, 4.9% and 2%, respectively, increased by 13.4%, 31.5%, 12.8% and 7.2% on CIDEr. The results show that our model has better performance.

Although these existing automatic evaluation indicators can provide positive reference value for algorithm evaluation, they cannot fully and objectively evaluate the content richness of each description sentence and the relevance of images. We do not believe that the true level of algorithm evaluation has surpassed that of humans. Therefore, the six most representative images in the data set and the corresponding description sentences generated during the five model testing process are selected to intuitively display the actual effect of the model. The specific effects are shown in Table 3 and Table 4.

| Model  | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | Rouge-L | CIDEr |
|--------|--------|--------|--------|--------|---------|-------|
| NIC[4] | 67.7   | 53.7   | 42.8   | 34.3   | 56.7    | 92.7  |
| CS-NIC[14] | 62.8 | 48.0   | 37.0   | 28.7   | 51.8    | 74.6  |
| CIC[15] | 66.3   | 52.5   | 41.7   | 33.2   | 56.5    | 93.3  |
| RAL[16] | 69.2   | 55.4   | 44.6   | 36.1   | 57.5    | 98.9  |
| Ours   | 71.2   | 56.8   | 45.8   | 37.2   | 60.9    | 106.1 |

As can be seen from the table 2, on the AI Challenger2017 image Chinese description test set, this model has improved all indicators compared to the NIC, CS-NIC, CIC, and RAL models, especially in BLEU-1 by 3.5% and 8.4% respectively, 4.9% and 2%, respectively, increased by 13.4%, 31.5%, 12.8% and 7.2% on CIDEr. The results show that our model has better performance.

Although these existing automatic evaluation indicators can provide positive reference value for algorithm evaluation, they cannot fully and objectively evaluate the content richness of each description sentence and the relevance of images. We do not believe that the true level of algorithm evaluation has surpassed that of humans. Therefore, the six most representative images in the data set and the corresponding description sentences generated during the five model testing process are selected to intuitively display the actual effect of the model. The specific effects are shown in Table 3 and Table 4.

| Image | NIC | RAL | Ours |
|-------|-----|-----|------|
| A man in a hat plays golf on the grass | Two men in jerseys compete for football on the field | On an outdoor open space, a man in a hat is handing a man in short sleeves |
| A man in a hat plays golf on the grass | Two men in jerseys playing football on the field | Two men in short sleeves are working on the open space outdoors |
| A man in a hat plays golf on the grass | Two men in football uniforms playing football on the field | Four men in long sleeves squatted on the clearing |
As shown in Table 3 and Table 4, use the NIC, CS-NIC, CIC, RAL model and our model to generate the description sentence of the same image. $S_{\text{NIC}}$, $S_{\text{CS-NIC}}$, $S_{\text{CIC}}$ and $S_{\text{RAL}}$ are image description sentences generated by their respective models and image description sentences of our model.

It can be seen that the complexity of these six pictures increases, and the difficulty of image description also increases. It can be seen from the table that the people and objects in the first two pictures of the two tables are relatively simple and easy to identify. All five models accurately describe the content of the pictures. The third picture in the two tables is the most complex picture, not only contains more characters, the background and content are also more complex, which increases the difficulty of generating descriptions for recognition. It can be seen from the table that the description of the secondary image by the three models is not very accurate. Among them, the NIC baseline model and the RAL model only recognize two characters, and the costume information of the characters is not accurate. Our model can not only identify more characters come out and can more accurately describe the actions of the main characters in the picture. It can be seen that the description sentences generated by our model are more accurate.

### 4. Conclusion

This paper proposes a new method of inputting image features and key information features into dual attention LSTM module to generate image description sentences. According to the experimental results, the performance of this model is better than baseline models such as NIC. It can generate smooth and smooth image description sentences, and the focus of the same image description can be controlled by keyword information, which increases the diversity of image descriptions to a certain extent. Although we have made some progress, there are still some problems. Our model cannot accurately identify the clothing information and more specific activities of the characters similar to the third picture in the two tables. In the future, we will consider adding more additional knowledge to optimize the model to obtain better performance to meet actual application requirements.

### Acknowledgments

The authors gratefully acknowledge the support of the innovative research projects for graduate students of Southwest Minzu University (No. CX2019SZ20).

### References

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, et al. Every picture tells a story:
generating sentences from images[J]. Lecture Notes in Computer Science, 2010, 21(10):15-29.
[2] Sutskever I., Vinyals O., Le Q V. Sequence to Sequence Learning with Neural Networks[J]. Advances in neural information processing systems, 2014.
[3] MAO J, XU W, YANG Y, et al. Explain images with multimodal recurrent neural networks[J]. arXiv preprint arXiv:1410.1090, 2014.
[4] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3156-3164.
[5] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions[J]. 2014.
[6] Xu K., Ba J., Kiros R., et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[J]. Computer ence, 2015:2048-2057.
[7] LU J, XIONG C, PARikh D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, 6.
[8] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep Reinforcement Learning-Based Image Captioning with Embedding Reward. In Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). 1151–1159.
[9] Xingjian Shi, Zhourong Chen, et al. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting[J]. arXiv:1506.04214v2.
[10] Wu J, Zheng H, Zhao B, et al. Large-Scale Datasets for Going Deeper in Image Understanding[C]// 2019 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2019.
[11] Lin C. Rouge : a package for automatic evaluation of summaries[C].the Workshop on Text Summarization Branches Out, Barcelona, Spain, 2004.
[12] Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based Image Description Evaluation[J]. 2015.
[13] Satanjeev B. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments[J]. ACL-2005, 2005:228-231.
[14] Li, Xirong, Lan, Weiyu, Dong, Jianfeng, et al. Adding Chinese Captions to Images.[J]. 2016.
[15] Aneja J, Deshpande A, Schwing A. Convolutional Image Captioning[J]. 2017.
[16] Zhang C, Dai Y, Cheng Y, et al. Recurrent Attention LSTM Model for Image Chinese Caption Generation[C]// 2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS). 2018.
[17] Wu J, Zheng H, Zhao B, et al. AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding[J]. 2017.