Chinese image caption of Inceptionv4 and double-layer GRUs based on attention mechanism

Yongbin Pan\textsuperscript{1,3,4,5}, Lidan Wang\textsuperscript{1,3,4,5*}, Shukai Duan\textsuperscript{2,3,4,5}, Xiuling Gan\textsuperscript{1,3,4,5}, Liangyi Hong\textsuperscript{1,3,4,5}

\textsuperscript{1}College of Electronic and Information Engineering, Southwest University, Chongqing 400715, China; \\
\textsuperscript{2}College of Artificial Intelligence, Southwest University, Chongqing 400715, China; \\
\textsuperscript{3}Brain-inspired Computing \& Intelligent Control of Chongqing Key Lab, Chongqing 400715, China; \\
\textsuperscript{4}National \& Local Joint Engineering Laboratory of Intelligent Transmission and Control Technology, Chongqing 400715, China; \\
\textsuperscript{5}Chongqing Brain Science Collaborative Innovation Center, Chongqing 400715, China

*Corresponding author’s e-mail: ldwang@swu.edu.cn

Abstract. In recent years, there has been a wave of research on English image caption at home and abroad. However, due to the particularity of Chinese image caption task, the research on Chinese image caption has not made good progress. In order to solve this problem, a new Chinese image caption model is implemented. Firstly, the AI challenge dataset is enhanced, and then the Chinese text data of the dataset is preprocessed by Chinese word segmentation tool word2vec. Secondly, based on the encoder-decoder framework, the image visual features are extracted by Inceptionv4 network, the attention mechanism is incorporated in the process of feature extraction and the Chinese sentences are generated by double-layer GRUs network. In the process of training, Adam is used to optimize the algorithm. Finally, A GUI interface is designed to better show the experimental effect. Experiments show that the new Chinese image caption model can automatically generate more fluent and more accurate Chinese caption sentences, and the trained model has excellent performance in many evaluation indexes.

1. Introduction

Image caption can be regarded as the process of transforming an input image into text information data through neural network algorithm. It is the combination of computer vision and natural language processing[1]. The ultimate goal is to enable the computer to recognize the scene content of the image and automatically generate natural language text to describe the image [2]. At present, in the research of image caption, vinyals[2] proposed NIC (Neural Image Capture) network for image caption task. The network can be trained in the way of end-to-end(Encoder-Decoder), and it shows great potential in experiments. Donahue [3] et al used VGG net as CNN to extract image information, and then input it into LSTM to output image description. Jia [4] proposed an extended long short memory network (glstm) for image caption task. Fang [5] et al took the lead in using multi instance learning mechanism to generate language text description of image information. Wu [6] et al provide another way to improve
the original model, which is using high-level semantic features. Chen [7] et al modified part of the recurrent neural network (RNN) structure of the decoder itself in the encoder-decoder structure, so that the RNN network can not only translate images into text description data, but also apply the network model in reverse. It means that the image characteristics are obtained from the text description data. Li [8] et al created the Chinese image caption dataset flickr8kcn, and then designed the Chinese image caption model cs-nic. In the encoder part of this model, the Google net network structure is used to extract the image feature information, and the long-term and short-term memory network (LSTM) is used as the decoder to generate the image text information description.

In summary, this article uses the AI challenge dataset[16] after data enhancement, based on the encoder-decoder framework, and designed an Chinese image caption network that integrates attention mechanism. Based on the four evaluation indicators of BLEU, METEOR, CIDEr, and Perplexity[9], this paper makes a detailed comparison and evaluation of common image description models. Finally, this paper designs a GUI interface for the designed network, which is convenient for the operator to operate intuitively.

2. The architecture of I-GRUs model

2.1. Inceptionv4 model

In 2015, Inceptionv3 was proposed by Szegedy et al[10]. Inceptionv3 model is the third generation of perception network structure of Googlenet.Inceptionv4 is developed on the basis of Inceptionv3. Inceptionv4 improves the model structure of Inceptionv3 and is the further development of concept model. The concept structure is highly adjustable. Szegedy et al[11] carefully adjusted the size of each layer of the model to balance the amount of computation among the subnets of each model by adjusting the size of each layer. The encoder uses CNN to extract the visual features of the image, and the CNN algorithm used in this paper is Inceptionv4 [11].

Inceptionv4 is trained on Imagenet [12]. In this paper, the trained network model is used to extract visual features, and Inceptionv4 is used as the encoder network. Finally, the image feature information collected by the network is input into the decoder network.

2.2. GRU model

In the original RNN neural network, the neural network can't memorize information, and it doesn't refer to the information that has appeared in the past when it processes the events that appear at every moment.

In order to solve the above problems, LSTM and GRU are used as the extension models of traditional RNN neural network. The core of LSTM model is a memory cell, which is controlled by three kinds of gates: forgetting gate, input gate and output gate. Compared with LSTM network, GRU model has a simpler structure, which is only controlled by two gates, reset gate and update gate. This model reduces a lot of network training parameters and improves the accuracy and computational efficiency of network model[13]. The network structure of GRU is shown in Fig. 1.

![Fig.1 GRU Network](image-url)
In Figure 1: r represents reset gate; z represents update gate; Ht-1 represents output state of previous time; h0 represents hidden state information of current time; Ht represents output value of current time. The figure is expressed as follows:

\begin{align}
z_t &= \sigma(W_z \cdot [h_{t-1}, x_t]) \\
r_t &= \sigma(W_r \cdot [h_{t-1}, x_t]) \\
h_u &= \tanh(W_r \cdot [r_t \cdot h_{t-1}, x_t]) \\
h_t &= (1 - z_t) \cdot h_{t-1} + z_t \cdot h_u
\end{align}

In the formula, \(\sigma\) represents the sigmoid function, \(\ast\) represents the operation of multiplying the corresponding elements of the vector; W represents the weight matrix.

2.3. I-GRUS model

The Chinese image caption model in this article is based on the encoder-decoder structure of the NIC network model, where I stands for Inceptionv4 and GRUs stands for double-layer gated recurrent unit.

The I-GRUs model proposed in this paper is based on the encoder and decoder architecture, and consists of Inceptionv4 and a double-layer gated recurrent unit GRUs. As shown in Fig. 2, in the encoder and image feature extraction stage, this paper chooses the large-scale data set ImageNet. The trained Inceptionv4 is used as a pre-training model to extract image features, and as an encoder to extract visual feature information of the image. The attention mechanism[14] is incorporated in the process of feature extraction, and the extracted image visual feature information is sent to the GRUs network unit as the decoder. The encoder uses a double-layer gated recurrent unit GRUs network structure to construct the language model, and the first layer of GRUs is used to receive the feature information of the image and the word embedding vector processed by Word2vec are used for preliminary modeling and transformation of features. The second layer of GRUs is used to receive the output of the first layer of GRUs and the output of the previous time step to form multi-modal features, and the image The feature information and the text feature are mapped to the same feature space.

The entire model of I-GRUs is expressed as:

\begin{align}
x_{t-1} &= \text{Inceptionv4}(I) \\
x_t &= m_w w_t, t \in \{0, 1, \cdots, N-1\} \\
p_{t+1} &= \text{GRUs}(x_t), t \in \{0, 1, \cdots, N - 1\}
\end{align}

Where I represents the input image, \(x_{t-1}\) represents the acquired image feature, \(x_t\) is the input of each time step of GRUs, \(N\) is the maximum time step of GRUs, \(m_w\) is the mapping matrix of the vocabulary, \(w_t\) is the Word2vec[15] code of the vocabulary entered in the GRUs network at each time step, \(w_0\) represents the character at the beginning of the sentence, \(w_{N+1}\) represents the character at the end of the
sentence, \( m_w \) represents the word embedding vector corresponding to the vocabulary, \( p_{t+1} \) is the vocabulary generation probability of the next time step in the sentence.

3. Test Results and Discussions

3.1. Dataset

The AI Challenge dataset [16] used in this article is a large-scale artificially labeled Chinese dataset proposed in the 2017 AI Challenge competition. In the AI Challenge dataset, there are 210,000 images in the training set and 30,000 images in the verification set. There are 30,000 pictures in the test set.

Based on the AI Challenge dataset, this paper enhances the AI Challenge dataset in the form of image geometric transformation (flip, rotate, crop, deform, zoom), making the experimental dataset have 1.05 million images in the training set and the verification set. There are 150,000 pictures and 150,000 pictures in the test set.

3.2. The Experimental Results

During the experiment, the deep learning model training framework used in this article is Tensorflow 1.8, the image input size of the Inceptionv4 network is 299*299*3, and the language model is constructed with double-layer gated recursive units GRUs, and the cross-entropy loss function is used to do For the loss function, the algorithm optimizer used is based on the original optimizer SGD [19], and Adam[18] is used as the optimizer.

The experimental results in this paper are evaluated by four indicators including BLEU, METEOR, CIDEr, and Perplexity [9]. BLEU is the current evaluation standard for machine translation, but BLEU is an evaluation index that only focuses on accuracy. The METEOR evaluation index not only considers the semantic accuracy of the generated sentence, but also considers the recall rate of the generated sentence. The CIDEr evaluation index can reflect the generated The semantic similarity between the sentence and the reference sentence. Perplexity is also called perplexity. It is usually used in natural language processing problems to evaluate the quality of language training models. The greater the sentence probability, the smaller the Perplexity, and the better the language model. In this article Tab.1 compares the Inceptionv3-LSTM[17] model, the Inceptionv3-Attention-LSTM[14] model with the attention mechanism, Inceptionv3-LSTMs model in which a single-layer LSTM is changed to a double-layer LSTM in the decoder, I-GRU model with only single-layer GRU in the decoder and the I-GRUs model of a double-layer GRUs in the decoder. It can be seen from Tab.1 that the B-1, B-2, B-3, B-4, METEOR and CIDEr indicators of the I-GRUs model significantly exceed those of other models. Tab.2 compares the changes in the Perplexity evaluation index value of each model as the training step length increases. From Tab.2, it can be seen that as the training step length increases, the Perplexity value of the I-GRUs model changes more than other models. The value of I-GRUs is the smallest under the training step.

| Model                  | Evaluation Index |
|------------------------|------------------|
|                        | B-1 | B-2 | B-3 | B-4 | M   | C   |
| Inceptionv3-LSTM       | 61.1 | 40.8 | 28.9 | 20.1 | 17.1 | 61.2 |
| Inceptionv3-Attention-LSTM | 66.6 | 44.1 | 31.3 | 21.2 | 20.3 | 80.2 |
| Inceptionv3-LSTMs      | 67.1 | 45.3 | 34.5 | 23.6 | 22.9 | 80.3 |
| I-GRU                  | 67.2 | 43.4 | 29.1 | 20.2 | 20.8 | 66.1 |
| I-GRUs                 | 68.8 | 48.7 | 34.7 | 26.8 | 23.9 | 85.6 |

| Model                  | Num-Step |
|------------------------|----------|
|                        | 10000 | 20000 | 30000 | 40000 | 50000 |
| Inceptionv3-LSTM       | 7.83  | 7.23  | 6.89  | 6.34  | 5.85  |
### 3.3. The Visual Results

We use the trained Inceptionv3-LSTM, Inceptionv3-Attention-LSTM and I-GRUs networks to make beautiful, concise, and friendly user interaction software interfaces in the PyQt5 framework of Python. At the same time, the program is based on Qt's signal-slot function program mechanism, which has good robustness during program operation. The Qt can map and connect the signal function with multiple corresponding slot functions. When the program is running, all the slot functions bound by the signal function will be executed. The operation buttons of the entire software interface are simple in design and have a friendly graphical interface, which can intuitively display the detection content and meet the needs of users.

It only takes 2-3 seconds from input to detection for a picture, which is robust and can adapt to the needs of different environments. As shown in Fig. 3.

![Fig. 3 The Visual Result of GUI](image_url)

### 4. Conclusion

Based on the structural characteristics of the encoder-decoder of NIC model, this paper designs a new Chinese image caption model based on the original NIC network model, using AI challenge which is expanded by us, using the Incepionv4 network which integrates attention mechanism, combined with the double-layer gating recursive unit GRU network. Our model is better than other network models in BLEU, METEOR, CIDEr, Perlexity, and the generated image caption sentences can be more accurate. Finally, we use PyQt to make a graphical user interface, which can clearly convey the generated image Chinese caption sentences to users. This research is helpful to reduce visual impairment for people with visual impairment. It is of great significance to the early education of infants, image retrieval, and has great reference value for the future research of neural network model of Chinese image caption.

### Acknowledgments

Our project was supported by the National Key R&D Program of China(Grant No.2018YFB1306600); the National Natural Science Foundation of China (Grant Nos. 62076207, 62076208, U20A20227).
References

[1] Huang Tian. Research on Image Description Algorithm Based on Deep Learning[D].Beijing: Beijing University of Posts and Telecommunications,2019.(in Chinese).

[2] Kulkarni G, Premraj V, Ordonez V, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.

[3] Donahue J, Hendricks L A, Rohrbach M, et al. Long-term Recurrent Convolutional Networks for Visual Recognition and Description[C].2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017:677-691.

[4] Jia X, Gavves E, Fernando B, et al. Guiding the long-short term memory model for image caption generation[C].Proceedings of the 2015 International Conference on Computer Vision. Santiago: IEEE Computer Society, 2015:2407-2415.

[5] Fang H, Gupta S, Iandola F, et al. From Captions to Visual Concepts and Back[C].2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015.

[6] Wu Q, Shen C, Liu L, et al. What value do explicit high level concepts have in vision to language problems[J]. Computer science, 2015:203-212.

[7] Chen X, Zitnick C L. Mind's eye: A recurrent visual representation for image caption generation[C].IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2015:2422-2431.

[8] Li X R, Lan W Y, Dong J F, et al. Adding Chinese Captions to Images[C].Proceedings of the 2016 Association for Computing Machinery(ACM) on International Conference on Multimedia Retrieval.New York, USA: ACM, 2016: 271-275.

[9] Neubig G. Neural Machine Translation and Sequence-to-sequence Models: A Tutorial[J]. Computer Science-Learning.2017.

[10] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the Inception Architecture for Computer Vision[J]. 2015, 28:18-26.

[11] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C].Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[12] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-252.

[13] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation , 1997, 9(8):1735-1780.

[14] Kelvin X, Jimmy L B, Ryan K, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[C].International Conference on Machine Learning.2015.

[15] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer science, 2013.

[16] 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).Shanghai:ICME,2019:1480-1485.

[17] Kulkarni G, Premraj V, Ordonez V, et al. Babytalk: Understanding and generating simple image descriptions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(12): 2891-2903.

[18] Kingma, Diederik P, Ba, Jimmy, Adam: A Method for Stochastic Optimization[C].International Conference on Learning Representations. 2015.

[19] Ye Junxiang. Pathological image region classifier based on improved convolutional neural network[J]. Digital Technology and Application ,2019,37(07):56-57.(in Chinese).