PAPER

Image Captioning Algorithm Based on Multi-Branch CNN and Bi-LSTM

Shan HE†, Nonmember, Yuanyao LU†(a), Member, and Shengnan CHEN†, Nonmember

SUMMARY The development of deep learning and neural networks has brought broad prospects to computer vision and natural language processing. The image captioning task combines cutting-edge methods in two fields. By building an end-to-end encoder-decoder model, its description performance can be greatly improved. In this paper, the multi-branch deep convolutional neural network is used as the encoder to extract image features, and the recurrent neural network is used to generate descriptive text that matches the input image. We conducted experiments on Flickr8k, Flickr30k, and MSCOCO datasets. According to the analysis of the experimental results on evaluation metrics, the model proposed in this paper can effectively achieve image caption, and its performance is better than classic image captioning models such as neural image annotation models.

key words: image captioning, multi-branch CNN, Bi-LSTM, encoder-decoder

1. Introduction

Image Captioning is an important issue in the field of artificial intelligence. It involves multiple ways to make the computer “view” the images better, that is, algorithms can give natural language sentences that describe the content of the image. The early research of image captioning mainly focused on extracting low-level features, such as image edges, corners and optical flow [1]. With the development of Computer Vision (CV) [2], the understanding of intermediate features such as image segmentation [3], target detection and recognition [4] has become the focus of many studies. Recently, with the emergence of large-scale image datasets and hardware innovation (GPUs), deep neural networks [5] have re-entered public vision and brought tremendous innovation in many research fields. In particular, Convolutional Neural Network (CNN) has stimulated tremendous progress in computer vision. Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) are widely used in sequence learning, such as machine translation [6] and image captioning [7].

Since the task of image captioning involves image processing and generating text-type sentences, it is a comprehensive problem of computer vision and natural language processing (NLP) [8]. Computer vision involves processing and analyzing digital data including images and videos. With a deep structure, deep neural networks can extract complex features from images and videos to solve various problems such as image classification and target recognition. It can understand the details of digital content from the bottom to the top. Natural language processing also plays a vital role in generating descriptive sentences from image features. The new algorithms used to generate words and sentences in natural language processing also provide a broad development space for the realization of image captioning tasks. The recurrent neural network is helpful to process sequence data, which can be used in this subject to generate the words in the description in order.

Methods for image captioning fields are generally divided into three categories. The first type is the retrieval-based method [9], [10]. Given the image to be queried, the system will search the database for visually similar images, retrieve them from the nearest neighbor annotations and convert them to the best description. The second type is the template-based method [11], [12] that generates a description that conforms to predefined grammar rules according to a given template. However, both the two methods have certain limitations that cannot be applied to new scenes and are less relevant to human descriptions. The third category is based on the encoder-decoder model. The encoder-decoder model was originally used for machine translation [13]. The input of the encoder is the source language and the output of the decoder is the translated target language. Inspired by the latest progress of machine translation in deep learning, an image captioning method based on the encoder-decoder framework is automatically generated. In this framework, the source language of the input can be regarded as an image and the target language of the output is the corresponding descriptive sentence.

Kiros et al. [14] introduced the encoder-decoder framework into the field of image captioning. They proposed a joint image-text embedding model and multimodal neural language model. Inspired by the field of machine translation, Vinyals et al. [15] proposed an image captioning model called neural image annotation (NIC). They replaced the encoder recurrent neural network in the machine translation model with a deep convolutional neural network and used the image as input. The model proposed by Donahue et al. [16] also uses a deep convolutional neural network for encoding and a long-short-term memory network for decoding to generate a textual description of the input image. Karpathy et al. [17] proposed a multimodal embedding model combining visual and linguistic modalities for generating image descriptions. The Stanford University team cre-
ated an image semantic description system Neuraltalk [18] similar to NIC, which uses other models to map image regions to sentence segments. Xu et al. [19] introduced the attention mechanism of the human visual system to the image description generation algorithm. Unlike NIC, this model uses the feature map of the last convolutional layer of the convolutional neural network as the image feature. At the decoding stage, the attention mechanism helps the model to dynamically select the characteristics of specific areas that require attention.

This paper uses deep learning to design an image captioning model based on the encoder-decoder structure. The extended deep convolutional neural network is used as an encoder to extract image features, and the Bi-LSTM network is used to generate descriptive sentences. This paper focuses on the end-to-end automatically generated image captioning model, and the main contributions are as follows:

(1) Improve the deep convolutional neural network to improve its accuracy while keeping the number of hyperparameters unchanged. Designed a highly modular convolutional neural network based on Resnet. It uses the method of stacking modules with the same topology structure to simplify the complexity of the network, making the multi-branch convolutional neural network easier to optimize and generalize.

(2) A method of extracting image features using a convolutional neural network with extended branches is proposed to further optimize the automatically generated image captioning model based on the encoder-decoder architecture. Bi-LSTM is adopted to predict each word that makes up a sentence.

(3) Calculated the scores of the model proposed in this paper on the automatic evaluation metrics BLEU [20], METEOR [21] and CIDEr [22], and compared with other image captioning algorithms. The experimental results on the large-scale datasets Flickr8k [23], Flickr30k [24] and MSCOCO [25] show that the method proposed in this paper can effectively improve the performance of the image captioning algorithm.

2. Method

2.1 Residual Learning

In deep learning, the network performance is usually improved by increasing the number of network layers. However, deeper neural networks are difficult to train. Beyond a certain depth, traditional deep networks will suffer from under-fitting problems caused by optimization difficulties. Residual learning makes the training of deep-level networks easier and helps to deal with more complex deep learning tasks and models. The residual network is easier to converge and its accuracy can still be obtained when the network is deep.

Deep residual learning can use residual mapping to solve the problem of network degradation. As shown in Fig. 1, the required bottom layer mapping is $H(x)$ and $x$ represents the input of the first layer. The stacked nonlinear layers are used to fit a new mapping $F(x) = H(x) - x$. Then the original required mapping can be transformed into $F(x) + x$. The purpose of residual learning is to make it easier to optimize the residual mapping $F(x)$ than to optimize the original mapping. If the added layer is constructed as an identity map, the error rate of the deeper model will not be greater than its shallow structure, that is, $F(x)$ tends to 0.

The direct connection in Fig. 1 represents a shortcut connection, that is, skip one or more layers to perform identity mapping and directly add the output of the shortcut connection to the output of the stacked layer. This kind of identity fast connection neither increases the parameters of the network nor increases the computational complexity. The entire network can still be trained end-to-end through stochastic gradient descent and back propagation algorithms.

2.2 Sparse Connection

During the gradual evolution of deep learning, some model building bottlenecks have gradually emerged. The large width of the convolutional layer and the huge number of parameters will also increase the computing resources, which is a great pleasure for the memory and time required by the training environment. Therefore, it is inefficient to build better neural network models simply by increasing the number of layers and neurons. To solve this problem, we use a sparsely connected network architecture instead of a fully connected network architecture to construct a convolutional neural network. The dense connection architecture and sparse connection architecture of the network are shown in Fig. 2.

When using random sparse connections, a large number of non-uniform sparse matrices will increase the difficulty of calculation. Therefore, in order to ensure the effec-
tiveness of sparse connections, the architecture of the Inception module implements dense components to approximate sparse convolutional neural networks. As shown in Fig. 3, multiple convolution and pooling operations are performed on the input. Since only a small number of neurons are effective, the size of the convolution filter is set to be small, and different sizes of convolution are used (5×5, 3×3, 1×1) to obtain different image details. Then the results are output in parallel to the next layer.

2.3 Multi-Branch CNN

Commonly used convolutional neural networks for image captioning tasks are VGG and Resnet, both of which use stacked modules to deepen the depth of the network. In this paper, the convolutional neural network architecture can reduce the selection of hyperparameters by repeatedly stacking modules with the same topology structure, and reduce the risk of the hyperparameters selected being over-adapted to certain datasets.

The deepening of the layers of VGG and Resnet is limited by various complex factors. Through carefully designed topology structure, the Inception module could obtain accuracy close to a large dense network in the case of lower computational complexity. However, since the number and size of the filters of the Inception module must be gradually customized for each individual input conversion, the complexity of the structure increases, making it difficult to modify and difficult to apply to other datasets and tasks.

In order to solve the above problems, this paper proposes a convolutional neural network based on Resnet, and uses the multi-branch idea of Inception network to improve Resnet. Figure 4 is a unit after modular expansion, where m is the number of convolutions and n is the number of branches expanded. Since the topology of each sub-module in the network is the same, the network structure is more concise and modular while maintaining the network capacity, while keeping the parameter scale unchanged. In each sub-module with the same structure, the weight layer is the convolution layer and the ReLU layer serves as activation function. A batch normalization (BN) is added between each convolution and activation function to ensure that the variance of the forward propagation signal is not zero, which plays a role in regularization.

To ensure that the learning effect of the deeper network is not lower than that of the shallow network, the shortcut connection is retained in the network structure, allowing the model to learn the identity function and reducing the problem of gradient disappearance. The multi-branch network structure in Fig. 4 can increase the receptive field, so that the convolutional neural network can separately encode information from multiple modalities, and get more feature information when used for feature extraction. Meanwhile, the number of hyperparameters in the model is consistent with the single-branch model. This structure makes it easier to adjust parameters during the model training process, while the simple modular structure improves the generalization performance of the model.

3. Model

This paper follows the encoder-decoder structure of the NIC model to construct an image captioning model. As shown in Fig. 5, the variable-length input is encoded into a vector of fixed dimensions, and then decoded into the required output sentence. The realization of the image captioning model in this paper is divided into two stages: 1. The feature extraction stage, which extracts each level of features contained in the image and expresses it in the form of a feature vector; 2. The sentence generation stage, which uses the feature vector as input, the word sequence will be generated one by one and combined into a meaningful description of the image.

We implement multi-branch expansion based on the Resnet152 network to obtain the expanded convolutional neural network structure as shown in Fig. 6. In the process...
of extracting image features via convolution operation, the value in the output matrix only depends on the corresponding area in the original image matrix, that is, receptive field. For images with multiple channels (such as RGB channels), the depth of the convolution kernel is the same as the depth of the input image. The matrix multiplication results of each channel are superimposed and an offset is added to obtain a compressed single-depth channel convolution feature output. The first convolution layer is mainly for capturing low-level features, such as edges, colors, gradient direction, etc. By increasing the number of layers, the deep convolutional neural network can capture high-level image features and can understand the images in the dataset more comprehensively like humans.

The task of the decoder part is to train the language model, that is, to predict each word that makes up the sentence based on the image and all previous words. As shown in Fig. 7, in the expanded form of the decoder model, the image and each word are connected to the Bi-LSTM memory unit with the same parameters. In the expanded structure, all recurrent connections are converted to feed-forward connections. If \( S = (S_0, \ldots, S_N) \) represents a correct descriptive sentence for image \( I \), the expanded form of the decoder can be interpreted as:

\[
\begin{align*}
x_{-1} &= \text{CNN}(I) \\
x_t &= W_e S_t, t \in \{0, \ldots, N-1\} \\
p_{t+1} &= \text{BiLSTM}(x_t), t \in \{0, \ldots, N-1\}
\end{align*}
\]

whereas \( S_0 \) represents the specific start word, \( S_N \) represents the specific end word, respectively marking the beginning and end of the description sentence. Bi-LSTM gives a signal that a complete sentence will be generated by the end word. In the training process of the encoder-decoder model, words are mapped to the same space where the image is mapped by a visual convolution neural network through the word embedding \( w_e \).

The loss function is expressed as follows, which is the sum of the negative log-likelihood of the correct word at each moment. The loss function is minimized during training:

\[
L(I, S) = \sum_{i=1}^{N} \log p_t(S_i)
\]

4. Experiment and Analysis

4.1 Datasets

In order to accelerate the convergence of the model during the training process and reduce overfitting, we implement the pre-train of the encoder part, that is, the multi-branch convolutional neural network, which is pre-trained on large-scale dataset ImageNet. The experiment of image captioning model is based on three datasets including images and English annotations, Flickr8k, Flickr30k and MSCOCO, which are also the most commonly used benchmark datasets for image captioning tasks. Table 1 is the image captioning evaluation datasets in the experiment of this paper.

4.2 Experimental Results

We compare the proposed model (denoted as Ours) with the following state-of-the-arts image captioning models on Flickr8k, Flickr30k and MSCOCO datasets: (1) Deep VS [17], NIC [15] and m-RNN [26] models are end-to-end Multi-modal networks. They use pre-trained convolutional neural networks (such as VGG or ResNet) as encoders, and use recurrent neural networks as language models. (2) Soft-attention and hard attention methods [19] introduce two alternative attention mechanisms for image captioning. The method based on soft-attention is to train the model through the standard back propagation method, while the method based on hard-attention is to train the method by maximizing the variational lower bound. (3) The Spatial model is a spatial attention mechanism model capable of extracting spatial image features. The Adaptive attention mechanism model can use visual markers instead of a single hidden

| Dataset   | Train | Val  | Test  |
|-----------|-------|------|-------|
| Flickr8k  | 60000 | 1000 | 1000  |
| Flickr30k | 29,783| 1000 | 1000  |
| MSCOCO    | 82,783| 5000 | 5000  |
Table 2 Performances compared with the state-of-the-art on Flickr8k dataset

| model         | B@1 | B@2 | B@3 | B@4 | METEOR |
|---------------|-----|-----|-----|-----|--------|
| Deep VS       | 57.9| 38.3| 24.5| 16.0| -      |
| NIC           | 63.0| 41.0| 27.0|     |        |
| Soft-Attention| 67.0| 44.8| 29.9| 19.5| 18.9   |
| Hard-Attention| 67.0| 44.8| 29.9| 19.5| 18.9   |
| SCA-CNN-VGG   | 65.5| 46.6| 32.6| 22.8| 21.6   |
| SCA-CNN-ResNet| 68.2| 49.6| 35.9| 25.8| 22.4   |
| Ours          | 70.3| 50.9| 36.9| 26.7| 23.2   |

Table 3 Performances compared with the state-of-the-art on Flickr30k dataset

| model         | B@1  | B@2  | B@3  | B@4  | METEOR | CIDEr |
|---------------|------|------|------|------|--------|-------|
| Deep VS       | 57.3 | 36.9 | 24.0 | 15.7 | -      | 24.7  |
| NIC           | 66.3 | 42.3 | 27.7 | 17.3 | -      | -     |
| m-RNN         | 60.0 | 41.0 | 28.0 | 19.0 | -      | -     |
| Soft-Attention| 66.7 | 43.3 | 28.8 | 19.1 | 18.5   | -     |
| Hard-Attention| 66.9 | 43.9 | 29.6 | 19.9 | 18.5   | -     |
| SCA-CNN-VGG   | 64.6 | 45.3 | 31.7 | 21.8 | 18.8   | -     |
| SCA-CNN-ResNet| 66.2 | 46.8 | 32.5 | 22.3 | 19.5   | -     |
| Spatial       | 64.4 | 46.2 | 32.7 | 23.1 | 20.2   | 49.3  |
| Adaptive      | 67.6 | 49.4 | 35.4 | 25.1 | 20.4   | 55.1  |
| Ours          | 69.6 | 50.9 | 36.6 | 26.1 | 21.2   | 55.3  |

Table 4 Performances compared with the state-of-the-art on MSCOCO dataset

| model         | B@1  | B@2  | B@3  | B@4  | METEOR | CIDEr |
|---------------|------|------|------|------|--------|-------|
| Deep VS       | 62.5 | 45.0 | 32.1 | 23.0 | -      | 66.0  |
| NIC           | 66.6 | 46.1 | 32.9 | 24.6 | -      | -     |
| m-RNN         | 67.0 | 49.0 | 35.0 | 25.0 | -      | -     |
| Soft-Attention| 70.7 | 49.2 | 34.4 | 24.3 | 23.9   | -     |
| Hard-Attention| 71.8 | 50.4 | 35.7 | 25.0 | 23.0   | -     |
| SCA-CNN-VGG   | 70.5 | 53.3 | 39.7 | 29.8 | 24.2   | -     |
| SCA-CNN-ResNet| 71.9 | 54.8 | 41.1 | 31.1 | 25.0   | -     |
| Spatial       | 73.4 | 56.6 | 41.8 | 30.4 | 25.7   | 102.9 |
| Adaptive      | 74.2 | 58.0 | 43.9 | 33.2 | 26.6   | 108.5 |
| Ours          | 77.4 | 60.9 | 45.8 | 35.3 | 27.9   | 112.6 |

state to provide backup options for the decoder [27]. (4) The SCA-CNN model [28] combines spatial attention mechanism and channel attention mechanism in the convolutional neural network to identify each feature entry in the multi-layer feature map. The comparison results on BLEU, METEOR and CIDEr evaluation metrics are shown in Table 2, Table 3 and Table 4.

B@1, B@2, B@3, and B@4 represent the scores on BLEU evaluation metrics when the n-gram is 1 to 4, respectively. According to the analysis in Table 2, the image captioning model proposed in this paper is significantly better than the NIC model on the BLEU evaluation metric, which fully shows the effect of different feature extractors on the performance of the image captioning model. At the same time, the performance of the model proposed in this paper is also better than other image captioning models, which provides different ideas for improving the image captioning algorithm based on the encoder-decoder structure.

Figure 8 is the score of our proposed model and Google’s NIC model on BLEU. The comparison between different data sets shows that the improvement of evaluation metric score on the MSCOCO dataset is more obvious.
For example, on the Flickr8k dataset, the model proposed in this paper has a B@1 score increase of 11.59% compared with the NIC model. On the MSCOCO dataset, the score by the model proposed is increased by 16.22% over the NIC model. This result shows that when large-scale datasets are used to train the model, the method of increasing the receptive field can obtain more image information in a larger dataset, thereby effectively improving the model training effect.

Figure 9 shows two examples of automatic generation of image description model. In the test, the maximum length of the generated description is set to 20 words. The inaccuracy of model generation results covers a large range, from negligible detail errors to inconsistent image descriptions. In this paper, the multi-branch deep convolutional neural network is used to detect the objects in the image. The receptive field corresponding to the convolution characteristics of multi-branch can cover more details of the image, which can reduce the inaccuracy of the generated description.

5. Conclusion

In this paper, the multi-branch deep convolutional neural network is used as the encoder to extract image features, and Bi-LSTM is used as the language model to generate descriptive text for image captioning task. The experimental results show that the method proposed in this paper is superior to other classic image captioning methods in three types of evaluation metrics. It is verified that the improved image feature extraction part can effectively improve the performance of the image captioning algorithm. Among them, the SCA-CNN model improves the convolutional neural network by introducing an attention mechanism, but the change of the neural network structure by this method greatly increases the complexity of the model. The model in this paper can achieve more accurate recognition of the objects and details in the image by improving the feature extraction, and more effectively extract the complete visual semantic information, so that the generated description is closer to the human description. In addition, the topology of the multi-branch convolutional neural network proposed in this paper is simple and easy to generalize.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (61971007 & 61571013).

References

[1] C.G. Harris and M. Stephens, “A combined corner and edge detector,” Proc. Alvey Vision Conference 1988, pp.23.1–23.6, 1988.
[2] N.M. Oliver, B. Rosario, and A.P. Pentland, “A Bayesian computer vision system for modeling human interactions,” IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.831–843, 2000.
[3] J. Dai, K. He, and J. Sun, “Instance-aware semantic segmentation via multi-task network cascades,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3150–3158, 2016.
[4] R. Fergus, P. Perona, and A. Zisserman, “Weakly supervised scale-invariant learning of models for visual recognition,” Int. J. Comput. Vision., vol.71, no.3, pp.273–303, 2007.
[5] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature., vol.521, no.7553, pp.436–444, 2015.
[6] I. Sutskever, O. Vinyals, and Q.V. Le, “Sequence to sequence learning with neural networks,” Advances in neural information processing systems, pp.3104–3112, 2014.
[7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3156–3164, 2015.
[8] E. Cambria and B. White, “Jumping NLP curves: A review of natural language processing research,” IEEE Comput. Intell. Mag., vol.9, no.2, pp.48–57, 2014.
[9] A. Farhadi, M. Hejrati, M.A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, “Every Picture Tells a Story: Generating Sentences from Images,” Computer Vision – ECCV 2010, Lecture Notes in Computer Science, vol.6314, pp.15–29, Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.
[10] V. Ordonez, G. Kulkarni, and T.L. Berg, “Im2text: Describing images using 1 million captioned photographs,” Advances in neural information processing systems, pp.1143–1151, 2011.
[11] M. Mitchell, X. Han, J. Dodge, et al., “Midge: Generating image descriptions from computer vision detections,” Proc. 15th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pp.747–756, 2012.
[12] Y. Yang, C.L. Teo, H. Daumé III, et al., “Corpus-guided sentence generation of natural images,” Proc. Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp.444–454, 2011.
[13] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1724–1734, 2014.
[14] R. Kiros, R. Salakhutdinov, and R.S. Zemel, “Unifying visual-semantic embeddings with multimodal natural language models,” NIPS 2014, arXiv:1411.2539, 2014.
[15] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge,” IEEE Trans. Pattern Anal. Mach. Intell., vol.39, no.4, pp.652–663, 2017.
[16] J. Donahue, L.A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, “Long-term recurrent convolutional networks for visual recognition and description,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.2625–2634, 2015.
[17] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3128–3137, 2015.
[18] A. Karpathy, A. Joulin, and L.F. Fei-Fei, “Deep fragment embeddings for bidirectional image sentence mapping,” Advances in neural information processing systems, pp.1889–1897, 2014.
[19] K. Xu, J. Ba, R. Kiros, et al., “Show, attend and tell: Neural image caption generation with visual attention,” International conference on machine learning, pp.2048–2057, 2015.
[20] K. Papineni, S. Roukos, T. Ward, and W-J. Zhu, “BLEU: a method for automatic evaluation of machine translation,” Proc. 40th Annual Meeting on Association for Computational Linguistics - ACL ’02, pp.311–318, 2001.
[21] S. Banerjee and A. Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” Proc. ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp.65–72, 2005.
[22] R. Vedantam, C.L. Zitnick, and D. Parikh, “CIDEr: Consen-
sus-based image description evaluation,” 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.4566–4575, 2015.

[23] C. Rashtchian, P. Young, M. Hodosh, et al., “Collecting image annotations using Amazon’s Mechanical Turk,” Proc. NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk, 2010.

[24] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” Transactions of the Association for Computational Linguistics, vol.2, pp.67–78, 2014.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C.L. Zitnick, “Microsoft COCO: Common objects in context,” Computer Vision – ECCV 2014, Lecture Notes in Computer Science, vol.8693, pp.740–755, Springer International Publishing, Cham, 2014.

[26] J. Mao, W. Xu, Y. Yang, et al., “Deep captioning with multimodal recurrent neural networks (m-rnn),” ICLR 2015, arXiv:1412.6632, 2015.

[27] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.3242–3250, 2017.

[28] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.6298–6306, 2017.

Shan He received the B.S. degree from the Faculty of Electronic Engineering and Computer Science, Ningbo University, China in 2015 and received the M.S. degree from the School of Electronic and Information Engineering, North China University of Technology, Beijing, China in 2020. Her research interests include artificial intelligence and image processing.

Yuanyao Lu received his Ph.D. from the Chinese Academy of Sciences, Beijing, China, in 2006. He is currently a professor with the Department of Electronic Engineering, North China University of Technology, Beijing, China. His current research is mainly on image processing and visual perception, pattern recognition, and artificial intelligence. He has published more than 20 papers in international journals and conferences.

Shengnan Chen received the B.S. degree from Xi’an University of Posts & Telecommunications, China in 2020 and is studying for the M.S. degree from the School of Information Science and Technology, North China University of Technology, Beijing, China now. Her research interests include artificial intelligence and behavior recognition.