# CIC Chinese Image Captioning Based on Image Label Information

Xindong You[1], Likun Lu[2](✉), Hang Zhou[1], and Xueqiang Lv[1]

[1] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing, China
{youxindong,lxq}@bistu.edu.cn
[2] Beijing Institute of Graphic Communication, Beijing, China
lklu@bigc.edu.cn

**Abstract.** Although image captioning technology has made great progress in recent years, the quality of Chinese image description is far from enough. In this paper, we focus on the problem of Chinese image captioning with the aim to improve the quality of Chinese image description. A novel framework for Chinese image captioning based on image label information (CIC) is proposed in this paper. Firstly, image label information is extracted by a multi-layer model with shortcut connections. Then the label information is input into the neural network with an extension of LSTM, which we coin L-LSTM for short, to generate the Chinese image descriptions. Extensive experiments are conducted on various image caption datasets such as Flickr8k-cn, Flickr30 k-cn. The experimental results verify the effectiveness of the proposed framework (CIC). It obtains 27.1% and 21.2% BLEU4 average values of Flickr8k-cn and Flickr30k-cn, respectively, which outperforms the state-of-art model in Chinese image captioning domain.

**Keywords:** Chinese image caption · Convolution neural network · Recurrent neural network · Deep learning · Chinese image tagging

## 1 Introduction

Image caption is an important work in the field of computer vision and natural language processing. It can generate descriptive sentences according to the image content, which is an effective method to narrow the "semantic gap" [1] between low-level visual feature and semantic information of the image [2–4]. Image caption is useful and practical for many application scenarios, including helping children or impaired people understand images, visual intelligent chat robot and image retrieval, which has great commercial value and attracted the great interest of researchers [5, 6]. However, image caption is a challenging task, which not only needs to recognize the objects in the image, but also needs to use the natural language to represent the attributes of objects, and then organized the relationship between them by natural language manner.

Up to now, there are three kinds of methods to generate image captions, they are template-based methods, retrieval-based methods, and deep learning-based methods,

Y. Weng et al. (Eds.): TridentCom 2020, LNICST 380, pp. 108–119, 2021.
https://doi.org/10.1007/978-3-030-77428-8_9

among which the deep learning-based methods are most advanced and achieve the best performance. Deep learning models, usually use the "encoder-decoder" framework. An image is encoded as a vector using a convolution neural network (CNN) and captions are decoded from the vector using a recurrent neural network (RNN) for end-to-end training of the entire system [13–17]. Currently, much of the image captioning work is concentrating on generating English sentences. Few image captioning work conducted on Chinese sentence generated due to the lack of datasets on image description labeled by Chinese. Rich meanings of Chinese words and the complex sentence structure make the Chinese image description be more challenging work. In this paper, we focus on Chinese image caption issues. We propose a label-based image caption model (CIC). Firstly, we design an image label prediction network with higher accuracy for image label feature generation. Secondly, the L-LSTM proposed in this paper which uses the image label features obtained from the label prediction network to generate descriptive sentences. Finally, we optimize the loss function by using label features, which further improves the quality of descriptive sentences. Extensive experiments conducted on various benchmark datasets such as Flickr8k-cn, Flickr30k-cn show that CIC achieves state-of-the-art performance.

The rest of this article is arranged as follows. We review the related work about image caption in Sect. 2, then introduce the proposed model CIC in Sect. 3, and show the experimental results in Sect. 4, finally give the conclusion in Section.

## 2 Related Work

Traditional image captioning methods can be divided into two categories: template-based methods and retrieval-based methods.

Template-based methods first use an object detector to detect the objects in the image, predict the object attributes and the relationship between objects, then fill the pre-designed template and finally form descriptive sentences [7–9]. Template-based methods are heavily dependent on the quality of object detection and limited by the pre-designed template, the generated descriptive sentences are single and lack of diversity. Retrieval-based methods first retrieve a similar subset of the images to be described in the training set based on the image visual feature, then generate candidate descriptive sentences by reasonably organizing the corresponding descriptive sentences of the similar subset images, and finally sort the candidate descriptions and select the optimal results [10–12]. The retrieval-based methods rely excessively on the training dataset, and the resulting descriptions are limited to the descriptions of the training set. With the rise of deep learning, these two kinds of methods are no longer favored.

With the development of deep learning, researchers have proposed image captioning methods based on deep learning. In 2015, Vinyals et al. proposed the image captioning model (Neural Image Caption, NIC) based on convolution neural network and Long Short-Term Memory (LSTM) [13]. The NIC model uses CNN to extract the visual feature of the image and takes the feature as the input of LSTM. The LSTM outputs the predicted words in turn to describe the contents of the image. The model is relatively simple, but the quality of generated sentences needs to be improved. Subsequently, the researchers improved on the NIC model. In 2015, Xu et al. introduced the Attention

Mechanism (AM) into the model for the first time, which enables the model to capture the local information of the image [14]. In 2016, Jia et al. used semantic information to guide LSTM in generating descriptions [15]. In 2017, Rennie et al. applied reinforcement learning to model training of image caption [16]. The objective of the existing model training is to maximize the probability of generating the correct descriptive sentence, which is inconsistent with the evaluation criteria in the inference. The method based on reinforcement learning can directly optimize the evaluation index in training, and obtain better results. In 2017, Dai et al. optimized the NIC model using Conditional GAN, making the generated sentences more natural and diverse [17]. In 2018, Anderson et al. proposed a combined bottom-up and top-down attribute mechanism [18], which enables the model to obtain image regions with significant semantics when predicting descriptions, and the performance is further improved.

The above-related work is all about the research on English image caption, the research on Chinese image caption is relatively few, and the quality of descriptions is needed to be improved. In 2017, Ze-Yu Liu et al. fused visual features and label features with LSTM in four ways to improve the effectiveness of the NIC model [19]. In 2018, Wei-Yu Lan et al. used the top10 image prediction labels to sort and select the generated sentences of NIC model, which makes the selected sentences closer to the label information and improves the image descriptive ability of NIC model [20].

## 3 CIC: Label-Based Chinese Image Captioning Model

We first describe the image label prediction network FC-PT in Sect. 3.1, then introduce our proposed image caption generator CIC in Sect. 3.2, the training process of the model is explained finally.

### 3.1 Image Label Prediction Network FC-PT

In the dataset of image caption, each picture described by five descriptive sentences, each sentence can vividly describe the content of the image. We choose the nouns, verbs, and adjectives in the sentence as the image labels. More specifically, Firstly, we use Chinese word segmentation tool Boson to segment the words in the descriptive sentences and selectively retain the nouns, verbs, and adjectives according to the word frequency. Then we construct a label vocabulary of the remaining words and tag the label information for each picture, therefore we can obtain the training data of label prediction network. Figure 1 shows the part-of-speech distribution of Flickr8k-cn and Flickr30k-cn dataset image labels.

The image label prediction network proposed in this paper consists of two parts, one is a feature extraction network based on CNN, the other is a feature classification network. as shown in Fig. 2

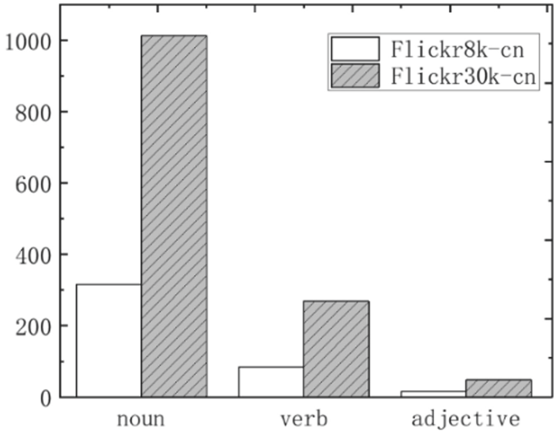


**Fig. 1.** Part-of-speech distribution of Flickr8k-cn and Flickr30k-cn

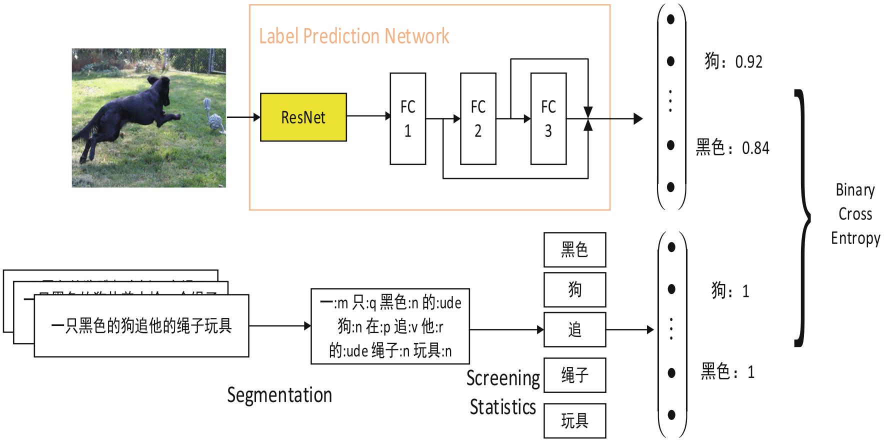


**Fig. 2.** Image label prediction classification network

### 3.2 CIC: Label-Based Chinese Image Captioning Model

The Chinese image captioning model CIC proposed in this paper is composed of CNN and L-LSTM. CNN is used as the encoder to extract image convolution features, and L-LSTM is used as the decoder to decode the image convolution features to the target descriptive sentences. Specifically, L-LSTM first accepts the image convolution feature and ignores the output at this time. Then, after inputting a start symbol <Start> and a predicted label feature, L-LSTM outputs a vector composed of the predicted probability of the words in the vocabulary and selects the word with the highest probability as the output. Then the word and the prediction label feature are used as the input of the next time, and the prediction is continued until the end symbol <End> is predicted, and the overall structure is shown in Fig. 3.

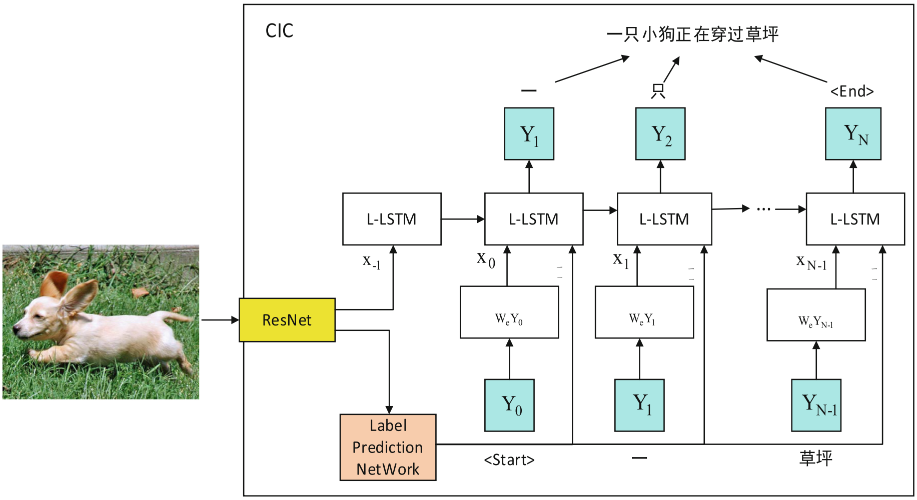


**Fig. 3.** Overall CIC architecture

The encoder CNN in the CIC model is a neural network used to process grid data. CNN model consists of a series of linear or non-linear transformation modules such as convolution layer, pooling layer, and activation layer. Deep CNN model is generally used to extract image features. After multiple times of convolution, pooling and activation of image data, the extracted features are more abstract and more expressive, which make breakthroughs in image classification, object detection, and other visual tasks. The CNN used in this model is ResNet-152. ResNet-152 is the best performance of ImageNet 2015 Image Classification Competition.

# 4 Experiment

## 4.1 Datasets

The datasets used in this paper are Flickr8k-cn and Flickr30k-cn. Flickr8k-cn and Flickr30-cn have been translated from English image caption datasets Flickr8k and Flickr30k into Chinese by machine translation method in Reference [21]. The Flickr8k-cn dataset contains 8,000 annotated images and 40,000 Chinese description sentences, as shown in Fig. 5. The Flickr30k-cn dataset contains 30,000 annotated images and 150,000 Chinese description sentences as shown in Fig. 6. In this paper, Flickr8k-cn and Flickr30k-cn are segmented according to the data segmentation method used in Kapathy et al. [22]. Flickr8k-cn includes 6000 training data, 1000 verification data, 1000 test data, Flickr30k-cn includes 28000 training data, 1000 verification data, 1000 test data (Fig. 4).

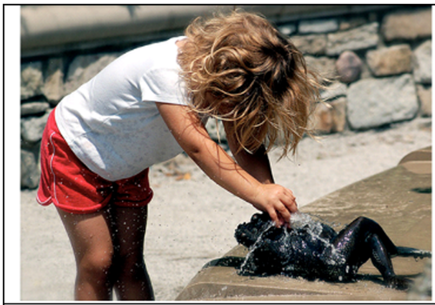
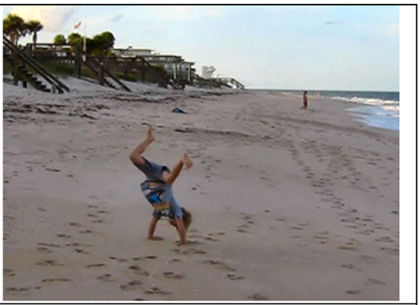
|  |  |
|---|---|
| （1） 一个孩子在海边玩喷泉。 | （1）一个男孩在海滩上倒立。 |
| （2） 一个孩子玩一个喷泉。 | （2）一个孩子在沙滩上做倒立。 |
| （3） 一个女孩看了一个喷水池。 | （3）一个孩子在海滩上做倒立。 |
| （4） 一个小女孩在水上玩耍在水上雕塑。 | （4）一个小男孩在海滩上做侧手翻。 |
| （5） 小女孩在一个公共喷泉里玩玩具。 | （5）一个小男孩在沙滩上做了一个倒立。 |

**Fig. 4.** Examples of Flickr8k-cn dataset

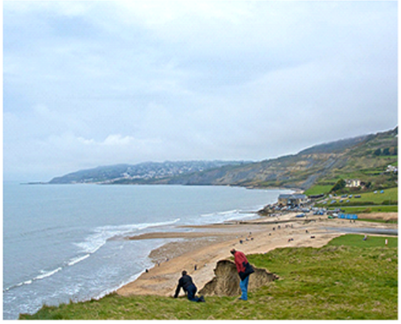
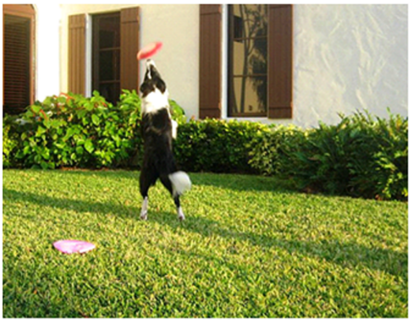
|  |  |
|---|---|
| （1） 男人看悬崖附近的海滩。 | （1）一只狗跳到院子里抓一个红色的飞盘。 |
| （2） 两只男人站在悬崖顶上俯瞰沙滩。 | （2）黑色和白色的狗追逐草地上红色的飞盘。 |
| （3） 这两只人从悬崖顶上看了海岸线。 | （3）黑色和白色的狗试图在空中抓住飞盘。 |
| （4） 两只人站在悬崖俯瞰大海。 | （4）狗跳上一个绿色的草坪接飞盘。 |
| （5） 两只人在海滩边的悬崖边。 | （5）一只黑白相间的狗在院子里追飞盘。 |

**Fig. 5.** Examples of Flickr30k-cn dataset

## 4.2 Experiment Details

The lab environment is configured as follows: Intel Xeon E5-2603 v4 processor, 64G RAM, Nvidia Tesla k80 graphics card, operating system Ubuntu 16.03.1, development language python 2.7, and deep learning framework tensorflow 1.6.

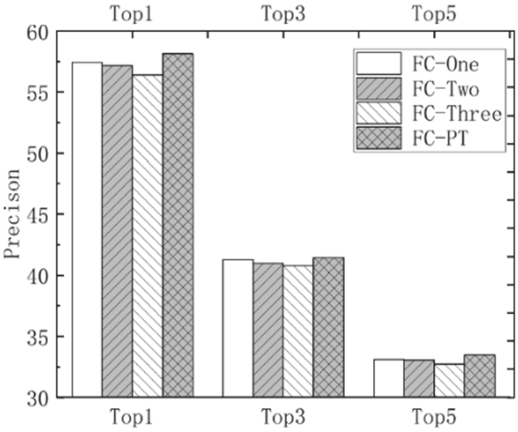


**Fig. 6.** Comparison of Flickr8k-cn accuracy

The Flickr8k-cn training set includes 6000 images, 30000 Chinese descriptive sentences, and 7784 words. The Flickr 30k-cn training set includes 28,000 images, 140,000 Chinese descriptive sentences, and 19735 words. In order to eliminate the interference of low-frequency words, we empirically preserve the nouns, verbs, and adjectives that appear at least twice in the five Chinese descriptive sentences of the same picture in Ref. [20], and the words whose overall word frequency is more than 20 times are used as the label vocabulary. Finally, the Flickr8k-cn label vocabulary contains 416 words, and the Flickr30k-cn label vocabulary contains 1330 words.

Label prediction network parameter configurations, as shown in Table 1.

**Table 1.** Label prediction network configuration parameters

| Parameter name | Flickr8k-cn | Flickr30k-cn |
|---|---|---|
| Batch size | 256 | 256 |
| Learning rate | 0.001 | 0.001 |
| Number of epochs | 50 | 50 |
| ResNet-152 image feature dimension | 2048 | 2048 |
| Number of hidden layer units | 512 | 1024 |
| Dropout | 0.5 | 0.5 |

The Chinese Image Captioning Model CIC parameter configurations, as shown in Table 2.

**Table 2.** The Chinese image captioning model CIC configuration parameters

| Parameter name | Flickr8k-cn | Flickr30k-cn |
|---|---|---|
| Batch size | 100 | 100 |
| Learning rate | 0.001 | 0.001 |
| Number of epochs | 50 | 50 |
| Number of L-LSTM units | 512 | 512 |
| ResNet-152 image feature dimension | 2048 | 2048 |
| Word embedding dimension | 512 | 512 |
| Label feature dimension | 416 | 1330 |
| Loss optimization coefficient $\alpha$ | 0.2 | 0.2 |

## 4.3 Evaluating the Proposed Label Prediction Network

Tables 3, 4, and 5 show the results of micro_Precision@k, micro_Recall@k, and micro_F1@k for different label prediction networks. FC-One represents the single-layer fully connected network. FC-Two represents the two-layer fully connected network. FC-Three represents the three-layer fully connected network. FC-PT represents our proposed label prediction network. We take Flickr8k-cn label prediction network results in Table 5 as an example to compare the proposed FC-PT network with FC-One, FC-Two, FC-Three. Experiments show that with the increase of network depth, network degradation occurs, that is, the accuracy and recall rate of the net decrease. The accuracy between FC-One and FC-Three decreases by 0.5% and the recall rate decreases by 0.57%. The FC-PT proposed in this paper is based on FC-Three network with residual structure, which improves the accuracy and recall rate to 33.49% and 39.54%, respectively. It shows that the proposed method can solve the problem of multi-layer label prediction network degradation. But as a whole, the accuracy and recall rate of label prediction network still have much room for improvement, which needs to be further studied in the future.

**Table 3.** Label prediction network top1 results comparison

| | Flickr8k-cn | | | Flickr30k-cn | | |
|---|---|---|---|---|---|---|
| Network structure | micro_Precision@1 | micro_Recall@1 | micro_F1@1 | micro_Precision@1 | micro_Recall@1 | micro_F1@1 |
| FC-One | 57.43% | 13.56% | 21.94% | 46.10% | 8.44% | 14.27% |
| FC-Two | 57.17% | 13.50% | 21.84% | 45.73% | 8.37% | 14.15% |
| FC-Three | 56.40% | 13.32% | 21.55% | 44.89% | 8.22% | 13.89% |
| FC-PT | 58.16% | 13.73% | 22.22% | 46.39% | 8.49% | 13.36% |

**Table 4.** Label prediction network top3 results comparison

| | Flickr8k-cn | | | Flickr30k-cn | | |
|---|---|---|---|---|---|---|
| Network structure | micro_Precision@3 | micro_Recall@3 | micro_F1@3 | micro_Precision@3 | micro_Recall@3 | micro_F1@3 |
| FC-One | 41.29% | 29.25% | 34.24% | 35.39% | 19.43% | 25.09% |
| FC-Two | 40.99% | 29.04% | 34.00% | 34.65% | 19.03% | 24.57% |
| FC-Three | 40.78% | 28.89% | 33.82% | 34.60% | 19.00% | 24.53% |
| FC-PT | 41.47% | 29.38% | 34.39% | 35.45% | 19.47% | 25.13% |

**Table 5.** Label prediction network top5 results comparison

| | Flickr8k-cn | | | Flickr30k-cn | | |
|---|---|---|---|---|---|---|
| Network structure | micro_Precision@5 | micro_Recall@5 | micro_F1@5 | micro_Precision@5 | micro_Recall@5 | micro_F1@5 |
| FC-One | 33.12% | 39.11% | 35.87% | 29.92% | 27.39% | 28.60% |
| FC-Two | 33.08% | 39.06% | 35.82% | 29.56% | 27.06% | 28.25% |
| FC-Three | 32.72% | 38.63% | 35.43% | 29.01% | 26.56% | 27.73% |
| FC-PT | 33.49% | 39.54% | 36.26% | 30.34% | 27.77% | 29.00% |

Figure 7 specifically illustrate that portion of the Precision in the table. As shown in Fig. 7, the prediction accuracy of the network decreases with the increase of network depth, and the network FC-PT proposed in this paper adds a residual structure on the basis of FC-Three, which improves the prediction accuracy and verifies the effectiveness of the method proposed in this paper.

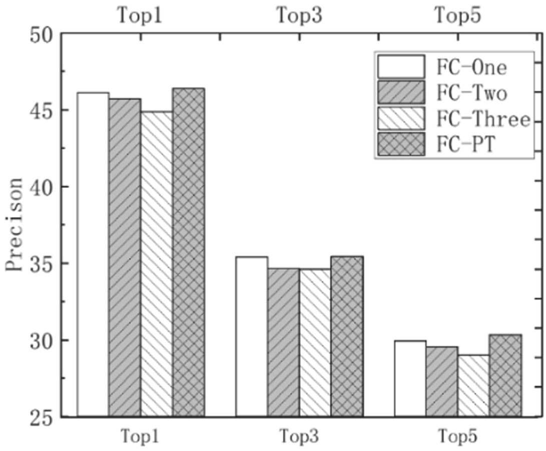


**Fig. 7.** Comparison of Flickr30k-cn accuracy

### 4.4 Evaluating the Proposed Chinese Image Captioning Model (CIC)

The proposed Chinese image caption network CIC is compared with CNIC-Ensemble in Ref. [17] and sentence rearrangement-MLP in Ref. [18]. As shown in Table 6, BLEU4, ROUGE-L, and CIDEr are greatly improved. Compared with CNIC-Ensemble, the experimental results are improved by 5.9%%, 2.1%, and 22.6%, and compared with sentence rearrangement-MLP, the experimental results are improved by 1.7%, 1.7%, and 2.4%. First, the improvement of Bleu indicates that the generated sentence has more co-occurrence words with the reference sentence, and the description is more accurate. Second, the improvement of ROUGE-L indicates that the generated description has higher accuracy and recall rate. Again, a significant increase in the CIDEr indicates that the generated sentence is more similar to the reference sentence. As shown in Table 7, CIC also performs well on Flcikr30k data. The CIC proposed in this paper has achieved the best results on Flickr8k-cn and Flickr30k-cn due to the existing Chinese image captioning model.

**Table 6.** Comparison of Flickr8k-cn Chinese image caption networks

| Method | BLEU4 | ROUGE-L | CIDEr |
|---|---|---|---|
| NIC19 | 18.7 | 44.2 | 27.9 |
| CNIC-Ensemble 19 | 20.5 | 45.4 | 30.1 |
| Google Model 20 | 23.6 | 44.8 | 47.4 |
| Word rearrangement-MLP 20 | 23.4 | 44.9 | 47.2 |
| Sentence rearrangement-MLP 20 | 24.7 | 45.8 | 50.3 |
| CIC | 26.4 | 47.5 | 52.7 |

**Table 7.** Comparison of Flickr30k-cn Chinese image caption networks

| Method | BLEU4 | ROUGE-L | CIDEr |
|---|---|---|---|
| NIC19 | – | – | – |
| CNIC-Ensemble 19 | – | – | – |
| Google Model 20 | 18.2 | 40.0 | 32.5 |
| Word rearrangement-MLP 20 | 18.0 | 39.9 | 32.5 |
| Sentence rearrangement-MLP 20 | 20.0 | 41.9 | 35.6 |
| CIC | 19.9 | 42.0 | 38.1 |

## 5 Conclusion and Future Work

We propose a label-based image captioning model CIC in this paper. Firstly, a label prediction network is proposed to capture the image label features, which effectively solves the problem of network degradation with the deepening of the number of label prediction network layers. Secondly, we present an L-LSTM architecture, in which image label features are input into the image caption network. However, the model proposed in this paper still has room for improvements. In the aspect of label prediction network, it will continue to improve the accuracy of label prediction network and try to use high-quality label features to generate descriptive sentences. In image caption network, more network architectures can be constructed, such as introducing image local features through the attention mechanism to further improve all aspects of the generation network indicators.

**Acknowledgments.** This work is supported by National Natural Science Foundation of China under Grants No. 61671070, Defense-related Science and Technology Key Lab Fund project 61420061903011, National Language Committee of China under Grants ZDI135-53, and Project of Developing University Intension for Improving the Level of Scientific Research–No. 2019KYNH226, Qin Xin Talents Cultivation Program, Beijing Information Science & Technology University No. QXTCP B201908.

## References

1. Cui, P., Zhu, W., Chua, T.S., et al.: Social-sensed multimedia computing. IEEE Multimedia **23**(1), 92–96 (2016)
2. Li, X., Uricchio, T., Ballan, L., et al.: Socializing the semantic gap: a comparative survey on image tag assignment, refinement, and retrieval. ACM Comput. Surv. **49**(1), 14:1–14:39 (2016)
3. Ariji, Y., Yanashita, Y., Kutsuna, S., et al.: Automatic detection and classification of radiolucent lesions in the mandible on panoramic radiographs using a deep learning object detection technique. Oral Surg. Oral Med. Oral Pathol. Oral Radiol. **128**(4), 424–430 (2019)
4. Zhuge, Y., Zeng, Y., Lu, H.: Deep embedding features for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 9340–9347 (2019)
5. Agrawal, A., Lu, J., Antol, S., et al.: VQA: visual question answering. Int. J. Comput. Vis. **123**(1), 4–31 (2015)
6. Das, A., Kottur, S., Gupta, K., et al.: Visual dialog. In: IEEE Conference on Computer Vision & Pattern Recognition. IEEE Computer Society (2017)
7. Farhadi, A., Hejrati, M., Sadeghi,, M.A., et al.: Every Picture Tells a Story: Generating Sentences from Images. Computer Vision – ECCV 2010, pp. 15–29. Springer, Berlin (2010)
8. Yang, Y., Teo, C.L., Hal Daumé, III., et al.: Corpus-guided sentence generation of natural images. In: Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2011)
9. Kulkarni, G., Premraj, V., Ordonez, V., et al.: Babytalk: understanding and generating simple image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **35**(12), 2891–2903 (2013)
10. Kuznetsova, P., Ordonez, V., Berg, A.C., et al.: Collective generation of natural image descriptions. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers, vol. 1. Association for Computational Linguistics (2012)

11. Verma, Y., Gupta, A., Mannem, P., et al.: Generating Image Descriptions Using Semantic Similarities in the Output Space (2013)
12. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2Text: describing images using 1 million captioned photographs. In: International Conference on Neural Information Processing Systems, pp. 1143–1151. Curran Associates Inc. (2011)
13. Vinyals, O., Toshev, A., Bengio, S., et al.: Show and tell: a neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164. IEEE Computer Society (2015)
14. Xu, K., Ba, J., Kiros, R., et al.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 2048–2057. JMLR. org, Lille (2015)
15. Jia, X., Gavves, E., Fernando, B., et al.: Guiding the long-short term memory model for image caption generation. In: IEEE International Conference on Computer Vision, pp. 2407–2415. IEEE (2016)
16. Rennie, S., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7008–7024. Hawaii (2017)
17. Dai, B., Fidler, S., Urtasun, R., et al.: Towards Diverse and Natural Image Descriptions via a Conditional GAN (2017). arXiv:1703.06029
18. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering (2018). arXiv:1707.07998
19. Liu, Z.Y., Ma, L.L., Wu, J., Sun, L.: Chinese image captioning method based on multimodal neural network. J. Chin. Inform. Process. **31**(06), 162–171 (2017)
20. Lan, W.Y., Wang, X.X., Yang, G., Li, X.R.: Improving chinese image captioning by tag prediction. Chin. J. Comput. 1–14 (2018)
21. Li, X., Lan, W., Dong, J., et al.: Adding Chinese captions to images. In: Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval, pp. 271–275. ACM, New York (2016)
22. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions. Computer Vision and Pattern Recognition, pp. 3128–3137. IEEE (2015)