A lightweight convolutional neural network for large-scale Chinese image caption*

ZHAO Dexin (赵德新)**, YANG Ruixue** (杨瑞雪)**, and GUO Shutao** (郭淑涛)
Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384, China

(Received 14 June 2020; Revised 1 September 2020)

Image caption is a high-level task in the area of image understanding, in which most of the models adopt a convolutional neural network (CNN) to extract image features assigning a recurrent neural network (RNN) to generate sentences. Researchers tend to design complex networks with deeper layers to improve the performance of feature extraction in recent years. Increasing the size of the network could obtain features of high quality, but it is not an efficient way in terms of computational cost. A large number of parameters brought by CNN makes the research difficult to apply in human daily life. In order to reduce the information loss of the convolutional process with less cost, we propose a lightweight convolutional neural network, named as Bifurcate-CNN (B-CNN). Furthermore, recent works are devoted to generating captions in English, in this paper, we develop an image caption model that generates descriptions in Chinese. Compared with Inception-v3, the depth of our model is shallower with fewer parameters, and the computational cost is lower. Evaluated on the AI CHALLENGER dataset, we prove that our model can enhance the performance, improving BLEU-4 from 46.1 to 49.9 and CIDEr from 142.5 to 156.6 respectively.

Document code: A **Article ID:** 1673-1905(2021)06-0361-6

DOI https://doi.org/10.1007/s11801-021-0100-z

Image caption is a challenging task in artificial intelligence that relates to computer vision and natural language processing. As a cross-disciplinary research issue, the essence of the image caption is to automatically generate a descriptive sentence for a given image. In order to generate human-readable sentences with a clear expression of the content, the feature extracted by convolutional neural network (CNN) is of great importance. The establishment of a connection between image and semantic interpretation has been considered an effective way to bridge the semantic gap[1], and the image caption attempt to establish such a rational connection. Different from other image understanding tasks, such as image classification, object detection and semantic segmentation, image caption is a high-level research in this area, on which researchers tend to capture detailed entities rather than just lining out simple labels. Significantly, it has a wide range of applications such as helping the blind and visually impaired people interpret visual information into detailed textual descriptions automatically. The current researches of the image caption are mainly about how to generate image caption in English. Obviously, this research should not be restricted by the language. Due to the rich meanings of Chinese words and the complex sentence structure, the image caption in Chinese is more challenging and essential. In this paper, we propose a lightweight CNN named as Bifurcate-CNN (B-CNN) for Chinese image caption.

According to the different ways of generation, there are three ways for image captions: template-based method, retrieval-based method and neural network-based method. With the development of deep learning, the neural network-based method has become the mainstream approach to solve image caption tasks in recent years[2-6]. Most of the models utilize encoder-decoder structures, combining a CNN with an RNN. The NIC model[2] proposed by Vinyals et al is the foundation of the image caption model based on the encoder-decoder framework. This model reduces the complexity of the image caption greatly, thus methods based on deep neural networks always take the NIC model as the initial pipeline. Liu et al. take Inception-v3[7] as the backbone structure of the feature extractor to generate captions in Chinese with multi-label keywords for an image[8]. Another Chinese image caption model with tag enhancement proposed by Lan et al. is designed to improve the quality of sentence generation[9]. In our previous work, we proposed a multimodal neural network model[10], in which Inception-v3 is used to extract the visual features of the image, ATTssd is proposed to supply the image attribute information, and CNNm is designed to reinforce the important message in sequence generation.

As the neural networks have been designed deeper and

* This work has been supported by the National Natural Foundation of China (No.61571328).
** E-mail: yangruixue1995@outlook.com

deeper, the computing cost is much heavier. It implies that the computer source even could not reach the level to put these researches based on a huge amount of data into use in daily life such as mobile vision analyzation and big-data scenes. Moreover, the deeper the network is designed, the easier gradient vanishing would happen. The highway networks[11] inspired by the Long and Short Time Memory (LSTM) network take gating functions into the model in order to train directly through simple gradient descent. However, because the sigmoid functions are widely used in the network, the continuous values between 0 and 1 make it difficult to get accurate results, and it would bring more parameters at the same time. Besides, while the dataset is not big enough, during the construction of the network, the over-fitting phenomenon is more likely to appear. Most of the models are trained on small datasets, such as flickr8k-cn and flickr30k-cn, we consider to train a convincing model for large-scale Chinese image caption. Flickr8k-cn is a dataset containing artificial annotations and the machine translations, while overly simplistic, may nevertheless be insufficient in large-scale captioning. As is emerged in the paper[12], We found that the translation of "bar" is wrong, and some sentences are not translated properly. Furthermore, flickr30k-cn only gives the machine-translated Chinese dataset, and there are no artificially labeled captions. In this paper, AI CHALLENGER dataset[13], a large-scale Chinese dataset with human annotations proposed in 2017, is used to train the model. More importantly, the quality and quantity of this dataset are completely higher than the other ones. Tab.1 manifests the comparisons between these datasets. We use AI_CH to refer to AI CHALLENGER.

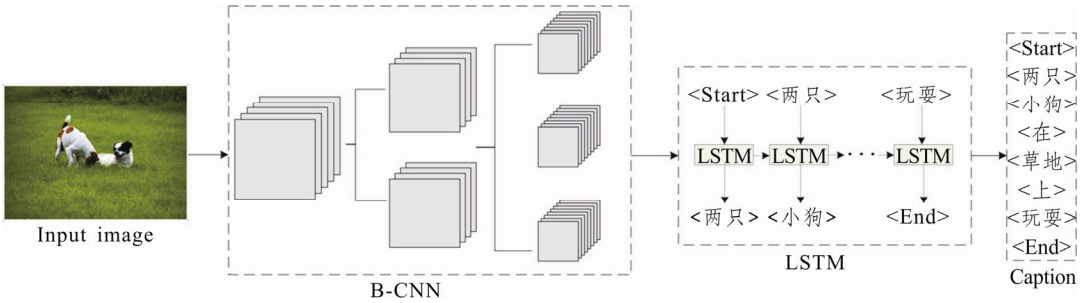

Fig.1 Overview of the image caption model

Tab.1 Dataset

Name	Train	Validation	Test	Language
Flickr8k_cn	6k	1k	1k	Chinese
Flickr30k_cn	29 783	1k	1k	Chinese
AI-CH	210k	30k	30k	Chinese

In this paper, a lightweight model is proposed to address the problems of the traditional models demanded large amounts of computer resources. Most competitive models have an encoder-decoder framework. Here, in our model, the encoder B-CNN maps an input image to a continuous feature vector. Given the feature vector, the language decoder LSTM then generates an output sequence step by step (as shown in Fig.1). The generation at time step t is

$$S_t=LSTM(x,\ S_{t-1}). \tag{1}$$

B-CNN has a total of 36 layers, which is shallower than the other ones used most frequently in recent years such as 101-layer ResNet[14], 152-layer ResNet[14], Inception-v3[7], Inception-v4[15] and Inception-ResNet-v2[15]. Taking inspiration from the ideas of ResNet and bifurcate path, we focus our attention on designing a network wider and shallower, instead of stacking with deeper and narrower modules as the traditional way does. This substitution extends the range of the receptive field, reduces the number of parameters, and it does not cause loss of expression. At the same time, nonlinear activation functions are widely used in the model to improve performance.

B-CNN is composed of two parts. We push seven stacked residual blocks adding up to 21 layers in total at the beginning, adopting four bifurcate modules as the backbone of the network. At the end of the model followed a full connectional layer and a pooling layer. As shown in Fig.2, the residual block is consisted of three convolutional layers piling up with a skip connection which explicitly increases the "depth" without adding extra layers and computational cost. The output of the block is

$$y=H(x,\ W_i)+xW_s, \tag{2}$$

where $H(x,\ W_i)$ represents residual mapping and W_s is only the expression of dimensions. In order to get features extracted by convolutional layers rather than the original features, we increase the step size of the skip connection. And the initial convolution channel sized 64 rather than 32 aims to offset the decrease of the depth. Merging together shallow features and deep features by adding the former output to the end of a block is helpful to extract more image content features. Moreover, the skip connections make our model converge quickly and easier to optimize. The details of the 7 residual blocks are illustrated as Tab.2 described as Conv2_X, Conv3_X and Conv4_X.

Replacing the large convolution kernels with the small ones has been proved to be an efficient method to reduce

the parameters and space complexity. The substitutions could get the same final feature maps as the large one extracted, as shown in Fig.3. To the extreme, we state several filters sized 1×d and d×1 to replace the large ones sized d×d. However, simply stacking convolutional layers with small kernels cannot solve the problem in a permanent cure. Applying the bifurcate modules could not only reduce the dimension of the channels greatly but also could extract more information than the simply stacked ones. It mainly because more nonlinear activation functions are used in our model combining with different receptive fields to enhance the representation. Fig.4 shows the modified module composed of small kernels with channels in low-dimension.

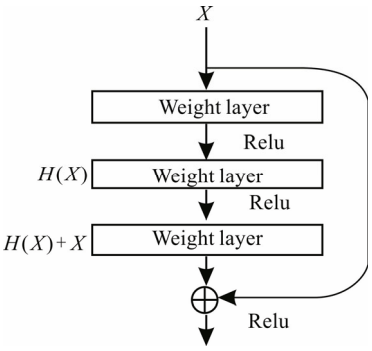

Fig.2 Residual block

Tab.2 The first part with 7 residual blocks

Module name	Inner structure	
Conv1	3×3, 64, stride 2	
	3×3, 64	
Max pooling	3×3, stride 2	
Conv2_X	3×3, 64 3×3, 64 3×3, 64	×2
Conv3_X	3×3, 128, stride 2 3×3, 128 3×3, 128	×3
Conv4_X	3×3, 256 3×3, 256 3×3, 256	×2

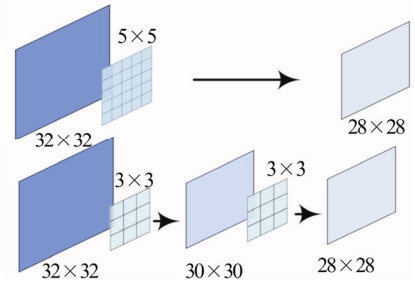

Fig.3 Convolutional process with different filters

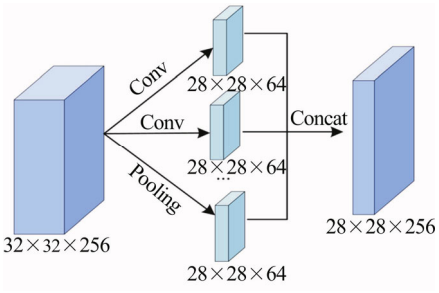

Fig.4 Modified convolutional block with bifurcations

The detailed structures of these four modules with bifurcations are illustrated in Figs.5—8. The map sizes of each convolutional layer are set as below. Every module has multiple bifurcation paths, and the branch convolutional features obtained from each path are finally combined to the final representation of the image. Adopting the small kennels with lower dimension channels, the parameters is effectively reduced. We fix the image, randomly crop, and flip horizontally to get an input size of 299×299×3. After processing with a series of stacked residual blocks, we get a map size of 37×37×256. Model_1 is mainly used to reduce the number of pooling channels, which has the advantage of saving computation time with fewer parameters, and the size of 35×35×256 is obtained. Assigning bifurcate paths composed of several stacked convolutional layers avoids the appearance of large channels between input channels and output channels.

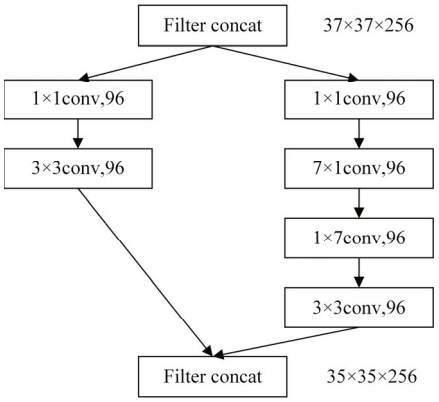

Fig.5 Model_1 module

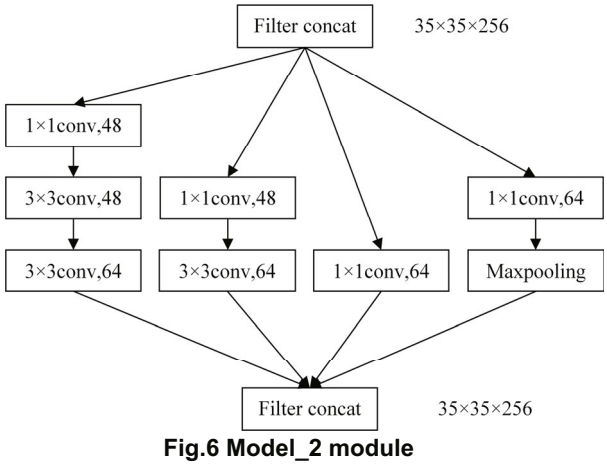

Fig.6 Model_2 module

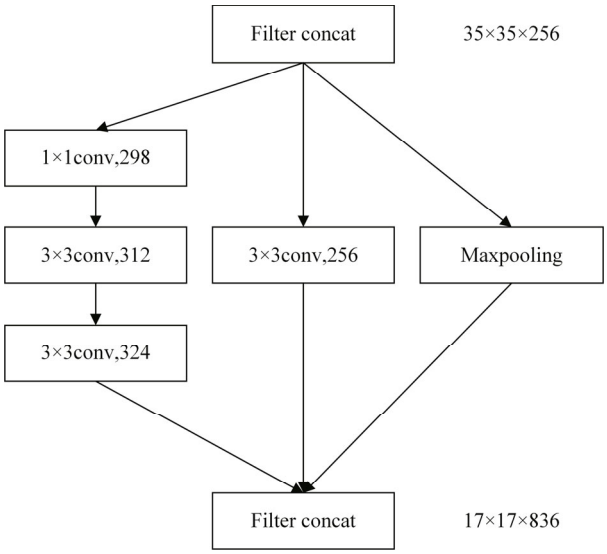

Fig.7 Model_3 module

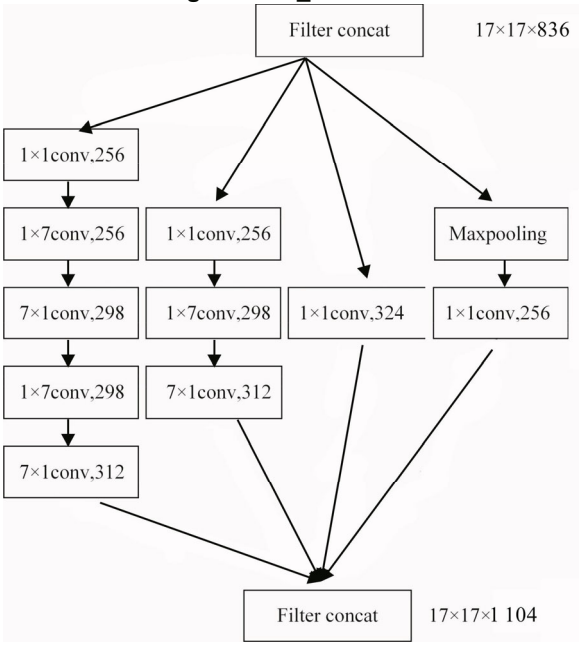

Fig.8 Model_4 module

The parameter count of each convolutional process is computed as

$$para=\alpha N(dm)^2+bias, \quad (3)$$

where N is the number of the convolution kernels, d is the size of the kernel, m is the input channels, and we presuppose output channels $n=\alpha m$. Obviously, as the network scales up to deeper with large spatial filters with a growing of channels, the growth rate of the computational cost is proportional to the filter size and the dimension of channels.

As shown in Tab.3, comparing with Incepton-v3, the total parameters are significantly disparate which means that more layers usually signify more parameters. To be precise, we compute the parameters of every portion in B-CNN as Tab.4. It is safe to draw the conclusion that our lightweight network with bifurcations could greatly decrease the parameters and improve the computational efficiency.

Tab.3 Comparisons between Inception-v3 and B-CNN

Model	Conv layers	Total para
Inception-v3	47	24 734 048
B-CNN	36	22 015 628

Tab.4 The parameters in B-CNN

Module name	Para
Seven residual blocks	4 759 234
Model_1	400 000
Model_2	223 632
Model_3	2 413 878
Model_4	4 281 884
FC	9 937 000

Our model is developed in Pytorch using Ubuntu 18.04 as the operating system. We run experiments on NVIDIA GPUs with 1080Ti, and the Chinese word segmentation tool is jieba. The vocabulary size of the AI CHALLENGER dataset is 8564 adding with <Start> and <End>. <Start> is the symbol of the beginning and <End> is used as the end of a sentence. We assign the cross-entropy loss function with Adam optimizer. We also employ the migration learning method to load the decoder parameters which was trained in Inception-v3 for 20 rounds. We adjust the learning rate to 0.0001 to train the B-CNN for 10 epochs, with that, setting the learning rate to 0.00001 to train end-to-end for another 5 epochs.We evaluated our model by BLEU [16], METEOR [17], ROUGE-L [18] and CIDEr [19] which are the acknowledged metrics in the image captioning task. We use B@1, 2BLEU-4. As presented in Ref.[12], Baseline is the officially published indicator in the AI CHALLENGER dataset obtained by Inception-v3 combined with LSTM, using TensorFlow as the deep learning framework. Tab.5 shows the results in these metrics on the AI CHALLENGER. Assembling B-CNN with LSTM, the results are obviously better than the traditional methods especially BLEU-4 and CIDEr. Among these methods, our B-CNN significantly improves the performance of the image caption by keeping consistent in the rest sections of the caption model.

Tab.5 Performance on AI CHALLENGER dataset compared with other methods

Method	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr
Baseline [12]	76.5	64.8	54.7	46.1	37.0	63.3	142.5
Inception-v3 + LSTM	77.1	65.1	54.8	46.9	38.1	64.2	143.0
BCNN + LSTM	**77.5**	**65.7**	**55.5**	**49.9**	**39.2**	**65.9**	**156.6**

Fig.9 shows some captions generated by different models, all of these images have rich content with complex semantics. It is easy to see that all the methods could generate fluent sentences, but there are some differences in details. Take the first image as an example, our model with B-CNN could recognize three people rather than two. If look more closely, we will find there are three people in the image. In the third image, the object we detected is "helmet" rather than the "hat". Moreover, the background of the image depicted is "woods" which is more suitable than the "road" generated by the Inception-v3 combined with LSTM model. The caption generated by our B-CNN combined with LSTM is more comprehensive and pertinent.

Fig.9 Qualitative results for images on AI CHALLENGER dataset

In this work, a lightweight neural network model named B-CNN for automatically generating image captions in Chinese is proposed. Compared with other deep neural networks, our lightweight model could get better performance with the idea of bifurcation paths. It was verified that our approach could get significant improvements in captioning quality with our simplified architecture. As a next step, we consider to introduce the reinforcement learning mechanism to train a better generator, and take the optimization of the decoder into account.

References

[1] Li X, Uricchio T, Ballan L, Bertini M, Snoek C.G and Bimbo A.D, ACM Computing Surveys **49**, 1 (2016).

[2] Vinyals O, Toshev A, Bengio S and Erhan D, Show and Tell: A Neural Image Caption Generator, IEEE Conference on Computer Vision and Pattern Recognition, 3156 (2015).

[3] Jia X, Gavves E, Fernando B and Tuytelaars T, Guiding the Long-Short Term Memory Model for Image Caption Generation, IEEE International Conference on Computer Vision IEEE Computer Society, 2407 (2015).

[4] Lu J, Yang J, Batra D and Parikh D, Neural Baby Talk, Conference on Computer Vision and Pattern Recognition, 7219 (2018).

[5] Rennie S J, Marcheret E, Mroueh Y, Ross J and Goel V, Self-Critical Sequence Training for Image Captioning, IEEE Conference on Computer Vision and Pattern Recognition, 7008 (2017).

[6] Yang J, Sun Y, Liang J, Ren B and Lai S, Neurocomputing **328**, 56 (2019).

[7] Szegedy C, Vanhoucke V, Ioffe S, Shlens J and Wojna Z, Rethinking the Inception Architecture for Computer Vision, IEEE Conference on Computer Vision and Pattern Recognition, 2818 (2016).

[8] Liu Z, Ma L, Wu J and Sun L, Journal of Chinese Information Processing **31**, 162 (2017). (in Chinese)

[9] Lan W, Wang X, Yang G and LI X, Chinese Journal of Computers **42**, 136 (2019). (in Chinese)

[10] Zhao D, Chang Z and Guo S, Neurocomputing **329**, 476 (2019).

[11] Srivastava R, Greff K and Schmidhuber J, Training Very Deep Networks, Advances in Neural Information Processing Systems, 2368 (2015).

[12] Kulkarni G, Premraj V, Dhar S, Li S, Choi Y, Berg A and Berg T, Baby Talk: Understanding and Generating Simple Image Descriptions, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2891 (2014).

[13] Wu J, Zheng H, Zhao B, Li Y, Yan B, Liang R, Wang W, Zhou S, Lin G, Fu Y, Wang Y and Wang Y, Large-Scale Datasets for Going Deeper in Image Understanding, IEEE International Conference on Multimedia and Expo (ICME), 1480 (2019).

[14] He K, Zhang X, Ren S and Sun Y, Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition, 770 (2014).

[15] Szegedy C, Ioffe S, Vanhoucke V and Alemi A A, Inception-v4, inception-resnet and the impact of residual connections on learning, AAAI Conference on Artificial Intelligence, 4278 (2017).

[16] Papineni K, Roukos S, Ward T and Zhu W, Bleu: A Method for Automatic Evaluation of Machine Translation, 40th Annual Meeting of the Association for Computational Linguistics, 311 (2002).

[17] Banerjee S and Lavie A, METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, Meeting of the association for computational linguistics, 65 (2005).

[18] Lin C, ROUGE: A Package for Automatic Evaluation of Summaries, Meeting of the Association for Computational Linguistics, 74 (2004).

[19] Vedantam R, Zitnick C L and Parikh D, CIDEr: Consensus-Based Image Description Evaluation, Computer Vision and Pattern Recognition, 4566 (2015).