Proceedings of the 2nd International Conference on Computing Innovation and Applied Physics (CONF-CIAP 2023)
DOI: 10.54254/2753-8818/5/20230457

Towards global recognition: Chinese recognition with non-local attention

Yifan Guo

Dalian University of Technology, Wanli Subdistrict, Jinzhou District, Dalian City, Liaoning Province

qiaoan412@mail.dlut.edu.cn

Abstract. As a part of image recognition, Chinese character recognition has a great application market in China, such as license plate recognition, logistics information recognition and so on. In recent years, the application of CNN has set off a frenzy of computer vision. Especially in the task of image recognition, CNN is widely used because of its high accuracy and few calculation parameters. However, CNN's local operation of using repeated filters to process images also has its shortcomings, for example, it can't pay attention to the relationship between distant pixels in the image. To solve this problem, we tried to add non-local operation to CNN to improve its performance. We chose the classic model of ResNetV2-50 as the foundation, and added non-local blocks to it. We compared the results of the two models and found that the accuracy increased by 4%.

Keywords: convolutional neural network, non-local block, Chinese character recognition.

1. Introduction

In the field of artificial intelligence, computer vision is a very important component. The main task of computer vision is to extract information by processing the collected pictures or videos. Computer vision has also been used in many ways in recent years [1]. Thanks to the rise of social media, which has led to a huge increase in image data, computer vision has also been used in many ways in recent years. Well-known areas include self-driving cars, medical image processing, face recognition and more.

Image recognition is one of the most important tasks in computer vision processing. Every image has its different features, so the basic task of image processing lies in how to exclude redundant information and extract the key main features. In recent years, many kinds of neural networks with different structures have been applied to the field of image recognition. At present, the most effective tool is convolutional neural network (CNN), and its application in the task of image recognition reaches a high accuracy [2]. With the introduction of AlexNet model in 2012, there is also a frenzy of deep learning computer vision [3].

In recent years, license plate recognition has been widely used in various aspects [4], such as intelligent community by identifying the license plate to determine whether the vehicle belongs to the community, the traffic management system also added license plate recognition to record illegal vehicles. Chinese license plates are made up of Chinese characters, letters and numbers. The Chinese characters are abbreviations of 31 provinces, and different provinces have different characters on the

DOI: 10.54254/2753-8818/5/20230457

front of their license plates. If the computer can recognize these 31 kinds of Chinese characters, it will be of great help to license plate recognition. Some people have studied Chinese character recognition before, but CNN was not used [5].And there also has a great application market in the logistics field. Because the logistics number also contains the abbreviation of the Chinese provinces to judge the delivery province of this express.

The dataset composes of 1400 labeled chinese character images. They were divided into 31 categories based on the province, each with an image of a Chinese character representing the province's abbreviation. Label values are assigned ranging from 0 to 30 to the images of Chinese characters in these 31 provinces, which was convenient for viewing the results of neural network recognition later.

In 2015, ResNet neural network was proposed. The core of its work was that it proposed framework of depth residual [6]. This framework is a good solution to the degradation problem when the neural network has too many layers, the accuracy will decrease. By adding a jump link, it links the output of the previous layer directly to the output of the next layer. After the ResNet50 model, a bottleneck framework with a small amount of parameters is added to reduce the computational power consumption [7].

In this paper, ResNetV2-50 model is set as the baseline to train the Chinese character image dataset. But when some images were fuzzy and the features were not obvious, the recognition effect is not good. We think this may be because it only considers the local operation, and may need to introduce more global information in this task. Previous studies also have related parts such as attention [8]. To solve this obstacle, the non-local block is employed. The traditional CNN uses a filter to filter a small region in an image so as to obtain the feature values of this small region. Therefore, it is a typical local operation, which does not pay much attention to the relationship between distant pixels. Non-local neural networks, on the other hand, consider all possible locations associated with a location when calculating its output [9]. As a plug and play module, the non-local block can be combined with many neural networks and achieves high accuracy in the field of facial identification [10] and so on.

In summary, ResNetV2-50 is utilized to train the dataset, and added non-Local blocks on this basis to build a new model to test the performance of non-local blocks. We also compared the results of these two models and gave conclusions for further discussion.

2. Method

In this paper, two models are established, namely ResNetV2-50 and non-local ResNetV2-50.

2.1. ResNetV2-50

This model is based on Keras and paper [7]. ResNetV2 offers two improvements over its predecessor. On the one hand, compared with ResNetV1-50, which puts ReLU layer behind convolutional layer, ResNetV2 adopts pre-activation structure, which can also improve model regularization in Batch Normalization (BN) layer. In addition, compared with ResNet18 and ResNet34, ResNet50 itself has new bottlenecks in structure.

2.1.1. Bottleneck. As shown in Figure 1, each Bottleneck contains three convolution blocks, which are 1×1,3×3 and 1×1. The inputs go through the BN layer and ReLU layer before each convolution block. The results after three levels of convolution will add to the input to form new outputs. This module is called Bottleneck. And the residual unit is formed by a jump link in the bottleneck.

DOI: 10.54254/2753-8818/5/20230457

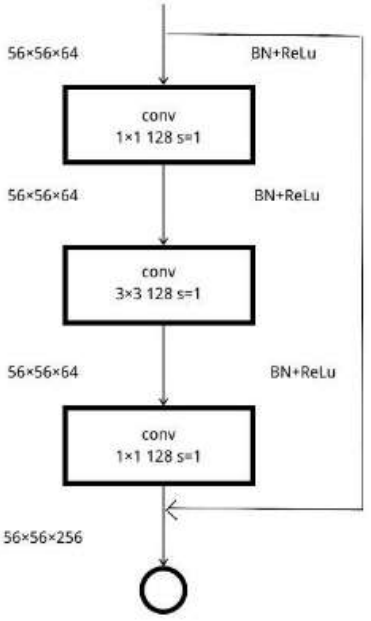

Figure 1. This figure shows the framework of bottleneck.

2.1.2. The structure of ResNetV2-50. As shown in Figure 2, ResNetV2-50 consists of five phases. The purpose of Stage 1 is to process the dataset first. It uses 7×7 convolution layer with st size 2, BN layer, ReLU layer and 3×3 Maxpool layer with stride of 2 to process the original 224×224×3 image into 64×64×56 image. as shown in Figure 2, and Stages 2 to 5 use 3, 4, 6, and 3 Bottleneck blocks to stack, and jump links are added to each block to form residual units. Finally, an output image of 7×7×2048 is obtained (as shown in Figure 5). At last the outputs go through the Average pooling layer, Flatten layer and full link layer to get the final results.

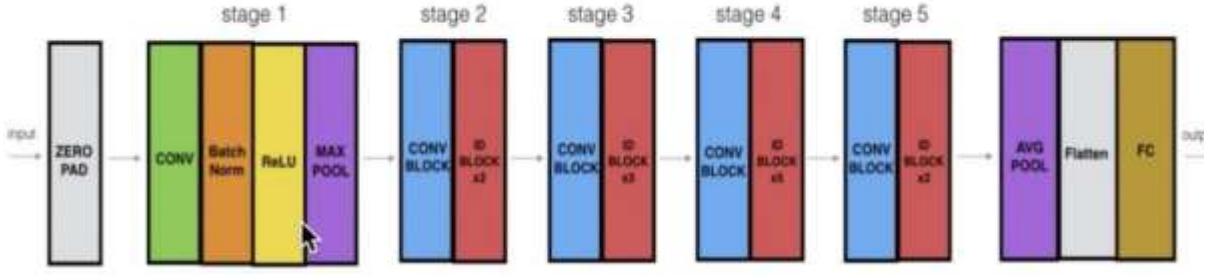

Figure 2. This figure shows the structure of ResNetV2-50. The CONV block and ID block which are introduced below are two types of bottleneck.

2.1.3. CONV block and ID block. These two types of modules both belong to the bottleneck part, and the difference lies in whether it is necessary to add a 1×1 convolution kernel and a BN layer to change the dimension of the input image when connecting the input to the output. According to Figure 3, the dimensions of input and output of CONV block are different, while those of ID block are the same.

DOI: 10.54254/2753-8818/5/20230457

So, a 1×1 convolution kernel and a BN layer are added into the CONV block, the ID block does not need to add them.

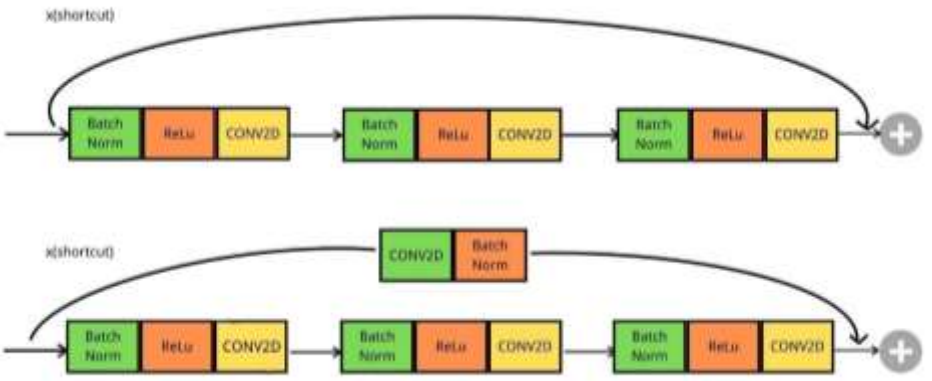

Figure 3. The upper part of this image shows the structure of ID block, and the lower part shows the structure of CONV block. The difference between them lies in the shortcut part.

2.2. Non-local ResNetV2-50

The non-local block can be inserted into the convolution layer in the model of ResNetV2-50 as a single block. However, if the number of image features is too small, it is not meaningful to use non-local blocks. Therefore, in the non-local block, it should first judge the filters of this input image. If filters>=256, the non-local block can be inserted into the convolution layer.

2.2.1. The principle of non-local blocks. As a local operation, convolution is a building block that only deals with one local neighbour at a time, which makes it unable to capture the relationship between distant pixels. However, the nonlocal operation calculates the weighted sum of the features of all the locations on the image when calculating the response of a certain location, which enables it to capture the remote dependencies well. The formula for non-local operations is

$$y_i = \frac{1}{C(x)} \sum_{\forall x} f(x_i, x_j) g(x_j) \tag{1}$$

Here x_i is the position of the output, x_j is all the positions that may be related to x_i,y_i is the output signal of the same size as x_i, f(x_i,x_j) function is used to calculate all the scalars between x_i and x_j, g(x_j) is used to calculate the input signal representation of x_j, which is realized by a 1×1 convolution kernel in practice. C(x) is the normalization parameter, which is set to the number of x positions The function f has 4 types gaussian, embedded gaussian, dot product and concatenation.

$$f(x_i, x_j) = e^{(x_i^T x_j)} \tag{2}$$

$$f(x_i, x_j) = e^{\theta(x_i)^T \phi(x_j)} \tag{3}$$

$$f(x_i, x_j) = \theta(x_i)^T \phi(x_j) \tag{4}$$

$$f(x_i, x_j) = ReLU(w_f^T [\theta(x_i), \phi(x_j)]) \tag{5}$$

The embedded Gaussian is used in this paper.

2.2.2. The construction of non-local block. In order to insert the non-local block into the ResNetV2-50 model, It is wrapped as a block, which is similar to the block in the ResNet network. The definition is given below: Here y_i is the output of the above formula, and w_z is actually a 1×1 convolution operation.

DOI: 10.54254/2753-8818/5/20230457

The $+x_i$ part adds the original input to the end, allowing the non-local block to be combined with other convolution block without changing its previous structure. And the figure 4 shows the process of the whole non-local block.

$$z_i = W_z y_i + x_i \tag{6}$$

F function is actually a matrix operation. The whole operation of non-local block can also be considered as matrix multiplication plus convolution operation, just as × in the figure 4 represents matrix multiplication and + in the figure 4 represents Element-wise add. And C represents the output channel, that is, the number of output filleters.

3. Data pre-processing

The original dataset contains 1400 Chinese characters images, all of which are 28×28 grey images. The 20 percent of training set are chosen as the test set. But the original image was too small for training, so the size of the original image are changed to 160×160×1. And the dataset is trained on both models for comparison.

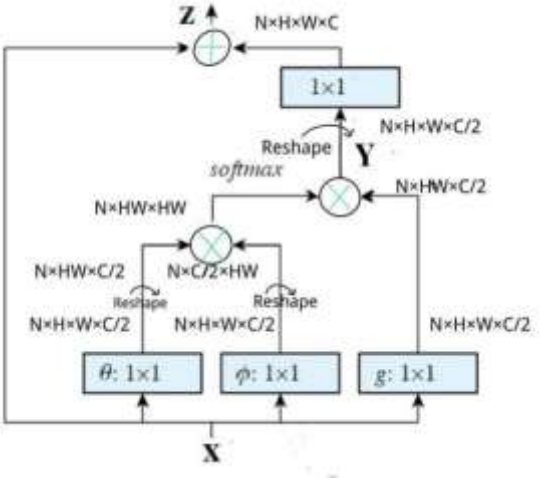

Figure 4. This figure presents the structure of non-local blocks.

4. Results

After 200 batches of training, the model of ResNetV2-50 achieved an accuracy of 89.6%.Under the same epochs and batch, the accuracy of the non-local ResNetV2-50 model increased by nearly 4% to 93.3%. This shows that adding non-local blocks is indeed very helpful for Chinese character image recognition.

Table 1. Comparison of the two different models.

Models	Accuracy
ResNetV2-50	89.6%
Non-Local ResNetV2-50	93.3%

As shown in figure 5, the recognition results of nine pictures are displayed. Most of the results are correct. But when the image is too fuzzy such as the picture right in the middle of the second line, the recognition result of ResNetV2-50 is wrong. However, the recognition result of non-local ResNetV2-50 is correct, which shows the performance of non-local ResNetV2-50 is better than ResNetV2-50.

DOI: 10.54254/2753-8818/5/20230457

Figure 5. This figure presents the nine examples of the results of Chinese characters image recognition. The GT represents ground truth. The Res represents the recognition result of ResNetV2-50 model. The Non represents the recognition result of non-local ResNetV2-50 model.

5. Discussions

For convolution operation, its receptive field is only the size of convolution kernel, which is called local. But according to Figure 6, Chinese characters are composed of strokes, just as English words are composed of letters. Different English words may contain the same letters, and different Chinese characters may also have the same strokes. Therefore, there will be some limitations if only partial consideration is given. In this task, the global information is needed to be considered to better recognize different Chinese characters. And non-local blocks can increase the receptive field to the global size, and this part is realized by calculating the global response value weighted by a certain position. In this way, the later layers can get more global information, so the recognition accuracy is greatly improved.

Figure 6. This figure presents the examples of Chinese characters image, which are used in dataset.

DOI: 10.54254/2753-8818/5/20230457

According to the results, the accuracy is improved greatly. although there are some limitations in this paper.

Firstly, only one form of embedded Gaussian function was used in the model in this paper, and there are three other forms of functions that have not been tested. In the future, non-local blocks in other functions can be further investigated.

Secondly, the insertion of non-local blocks will increase the number of parameters to be trained, which will lead to the increase of computational complexity. If it is applied to the neural networks with more layers, it may consume too much computational power. Therefore, how to balance computational complexity and expected performance is also a research direction.

6. Conclusions

Since only convolution neural network is used to identify Chinese character images, it can only extract local information, which is not enough to identify different Chinese characters. And the results of image recognition are not very good if the filters of image are too small. To solve this obstacle, this work proposes to introduce non-local module into traditional convolutional neural network. In addition, we conduct extensive experiments to validate the effectiveness of non-local module. The results show that the non-local blocks can greatly improve the performance of ResNetV2-50 model in image recognition. The accuracy also is improved by 4%.

References

[1] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. (2015) ”Deep learning.” nature 521.7553: 436444.

[2] Y. Lecun; L. Bottou; Y. Bengio and P. Haffner. (1998) ”Gradient-based learning applied to document recognition”. IEEE

[3] Krizhevsky, A , Sutskever, I. and Hinton, G.E. (2012) ”ImageNet Classification with Deep Convolutional Neural Networks”.

[4] Shyang-Lih Chang; Li-Shien Chen; Yun-Chung Chung and Sei-Wan Chen. (2004) ”Automatic license plate recognition”. IEEE

[5] C.-L. Liu; S. Jaeger and M. Nakagawa. (2004) ”Online recognition of Chinese characters: the state-of-the-art”. IEEE

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.(2015) ”Deep residual learning for image recognition”.IEEE.doi:CVPR.2016.90

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.(2016)”Identity mappings in deep residual networks”.

[8] Vaswani, AshishShazeer, NoamParmar, NikiUszkoreit, JakobJones, LlionGomez, Aidan NKaiser, LukaszPolosukhin, and Illia. (2017) ”Attention Is All You Need”. arXiv.1706.03762

[9] X. Wang, Ross B. Girshick, A. Gupta,and Kaiming He. (2018) “Non-local Neural Networks,”. IEEE. DOI:CVPR.2018.00813.

[10] B Li,D Lima. (2021) ”Facial expression recognition via ResNet-50”. j.ijcce.2021.02.002.