IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 13, 2019, accepted July 1, 2019, date of publication July 8, 2019, date of current version September 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927169

A Dilated CNN Model for Image Classification

XINYU LEI, HONGGUANG PAN, AND XIANGDONG HUANG
College of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an 710049, China

Corresponding author: Hongguang Pan (hongguangpan@163.com)

This work was supported in part by the National Science Foundation of China under Grant 61603295, in part by the Outstanding Youth Science Fund of the Xian University of Science and Technology under Grant 2018YQ2-07, in part by the China Postdoctoral Science Foundation under Grant 2017M623207, and in part by the Natural Science Basic Research Plan in Shaanxi Province under Grant 2018JM6003.

ABSTRACT The dilated convolution algorithm, which is widely used for image segmentation, is applied in the image classification field in this paper. In many traditional image classification algorithms, convolution neural network (CNN) plays an important role. However, the classical CNN has the problem of consuming too much computing resources. To solve this problem, first, this paper proposed a dilated CNN model which is built through replacing the convolution kernels of traditional CNN by the dilated convolution kernels, and then, the dilated CNN model is tested on the Mnist handwritten digital recognition data set. Second, to solve the detail loss problem in the dilated CNN model, the hybrid dilated CNN (HDC) is built by stacking dilated convolution kernels with different dilation rates, and then the HDC model is tested on the wide-band remote sensing image data set of earth's terrain. The results show that under the same environment, compared with the traditional CNN model, the dilated CNN model reduces the training time by 12.99% and improves the training accuracy by 2.86% averagely, compared with the dilated CNN model, the HDC model reduces the training time by 2.02% and improves the training and testing accuracy by 14.15% and 15.35% averagely. Therefore, the dilated CNN and HDC model proposed in this paper can significantly improve the image classification performance.

INDEX TERMS Image classification, CNN, dilated convolution, hybrid dilated CNN.

I. INTRODUCTION

Image classification is one of the most basically and widely used field in computer vision [1]. In recent years, convolution neural network (CNN) has achieved great success in the field of image classification due to the fast and accurate feature extraction function and end-to-end trainable network framework [2]–[4].

Generally, the image classification models include the following three types: 1) full convolution neural network (FCN); 2) condition random field (CRF); 3) dilated CNN. The FCN is first proposed in [5], which replaces the last softmax layer in classic CNN with the convolution layer, outputs a feature map, converts the feature map into the size of original input image through up-sampling so as to realize pixel-level classification, and it can precess input images of any size. The CRF is a conditional probability distribution model that outputs random variables with the given input variables, and it can obtain local and large-scale dependencies in the image to refine the feature map [6]. The dilated CNN is a special CNN, whose convolution kernels are formed by inserting holes into traditional convolution kernels, and it can help reduce the consumption of computational resources when extracting image features from the network and expand the receptive field without increasing the number of parameters [11].

The development of CNN can be traced back to the 1990s. Lecun et al. designed LeNet-5, which laid the foundation for modern CNN [7], [8]. Krizhevsky et al. proposed a deeper and wider LeNet-5 named AlexNet, which won the champion in the ImageNet Competition in 2012 and set off a research upsurge of CNN [3]. The VGGNet, proposed by Oxford University in 2014, adopted smaller $3 * 3$ size convolution kernels and stacked them to replaced larger size convolution kernels, which can obtain better network performance with fewer training parameters. The VGGNet obtained the first and second place in ILSVRC localization and classification, respectively. In 2015, He et al. proposed the ResNet, which solved the problem of gradient disappearance in deep CNN by adding shortcut to the VGGNet, and further improved the accuracy of image classification [2].

In the field of image classification, many scholars have proposed their methods. Pang et al. proposed a novel fused

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Shen.

CNN that fused the features extracted in the shallow and deep layers for the biomedical image classification [9]; Ivan et al. studied an image classification method on a tight representation budget, it focused on very short image descriptor which might be lost during the training process [10]; Zhou et al. proposed a novel data augmentation strategy for the Siamese model and introduced a joint decision mechanism into the model which can better improve the classification performance [11]. All the above methods are well-behaved in the image classification filed and widely applied in many different fields. However, the structures of these CNN models are very complex, and the consumption of computing resources and the reduction of efficiency cannot be overlooked. Therefore, it is necessary to propose a more efficient image classification method.

Dilated convolution, which is generally used in image semantic segmentation, has attracted more and more attention in recent years, and many excellent models have been proposed based on it. Zhang et al. proposed a dilated convolution model for single image super-resolution, which mixed the traditional convolution kernels with dilated convolution kernels to formed a mixed convolution. This model can achieve good generalization ability by capturing the correlation between low-resolution and high-resolution images [12]. Li et al. proposed a CSRNet for understanding the highly congested scenes, which was composed of the front-end CNN for feature extraction and the back-end dilated CNN for enlarging receptive fields and replacing pooling operations. It obtained the good effect on the ShanghaiTech Part A dataset [13]. Wei et al. proposed an augmented classification network based on multi-dilated convolutional blocks which can produce dense object localization, and thus realizing weakly- and semi- supervised semantic segmentation of images [14]. In the above literatures, the dilated convolution is used as an optimization method to expand the receptive field and obtain more information, it is not a main method to extract image features and other algorithms are needed to improve the model performance.

The dilated convolution is also used in some other fields. Zhang et al. proposed a dilated CNN to classify environmental sounds, introduced LeakyReLU activation function to weigh the network sparsity and input information, and achieved superior performance on UrbanSound8K data set [15]. Kudo et al. proposed a dilated CNN model for image classification and object localization, which replaced the traditional convolutional kernels in ResNet with dilated convolutional kernels and verified on the ImageNet50 data set. Its parameters were reduced by 94% and its convergence rate was faster compared with the traditional ResNet [16]. In the above literatures, in order to improve the performance of the CNN, the traditional convolution layers in the model are completely replaced by the dilated convolution layers which can improve the efficiency. However, the gridding effect caused by the stack of the dilated convolution kernels is not paid much attention, which may cause the loss of important continuity details in the image. To solve these problems, Yu et al. proposed a dilated residual networks by replacing the convolution kernels in original ResNet with the dilated convolution kernels, removing the max-pooling in the model to reduce the high-amplitude and high-frequency activations which can be propagated to later the layers and ultimately exacerbate gridding effect, meanwhile using the HDC to further reduce the gridding effect [17]. Liu et al. proposed a multi-scale residual CNN based on dilated convolution for image denoising, which can extract more information from original images by using the dilated ResNet and avoid gridding effect by using the HDC [18]. In the above literatures, the HDC is introduced into the dilated ResNet to optimize the model performance, and the traditional max-pooling is removed to further improve the accuracy.

In this paper, the dilated CNN model is proposed through replacing the convolution layers in traditional CNN by the dilated convolution layers, and expands the receptive field without increasing parameters, so that it can improve the network performance without increasing the network complexity. Compared with our work, the existed structures are more complex, the training time required and difficulty are greatly improved while improving the accuracy. To verify the performance of dilated CNN, a simple dilated CNN model is built and tested on the Mnist handwritten digital recognition data set, and the HDC model is built by stacking the dilated convolution kernels with different rate and tested on the wide-band remote sensing image data set of earth's terrain. The experiments show that, the dilated CNN model consumes less training time than traditional CNN, and the HDC model has higher accuracy and less time consumption than the dilated CNN model. The HDC model proposed in this paper is not only simple in structure and easy to be repeated, but also retains pooling operations to further enhance the performance, therefore the reliability of the proposed model is higher and the accuracy is kept at a high level.

II. CNN AND DILATED CNN

In this section, we mainly focus on the introduction of the CNN and dilated CNN, including the principle, the calculation process and the characteristic of them.

A. THE BRIEF INTRODUCTION OF CNN

Like the traditional neural network, the CNN is composed of input layer, hidden layer and output layer. The difference is that the input of CNN is image (the pixel matrix), and the output is the image feature obtained by the convolution calculation [19].

Fig.1 is a classical LeNet-5 structure, which consists of two convolution-pooling layers and three full connected layers. It is a simple and efficient nonlinear multi-layer learning model, which has achieved good performance in Mnist data sets and been widely applied in various fields.

In the CNN, the convolution kernel is the most important part, which is also the origin of the name ''convolution neural network''. The convolution kernel is a two-dimensional matrix with size $n * n$, in which each point has a

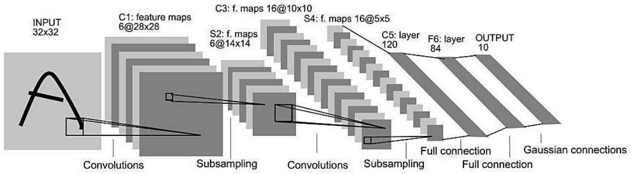
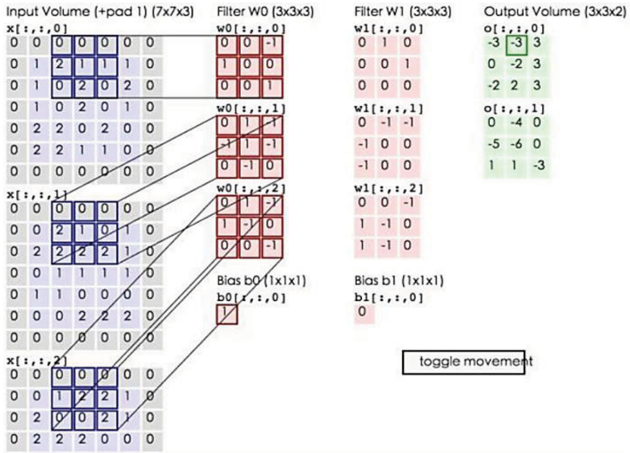


FIGURE 1. The structure of LeNet-5 [7].



Input Volume (+pad 1) (7x7x3)

x[:,:,0]

0 0 0 0 0 0 0
0 1 2 1 1 1 0
0 1 0 2 0 2 0
0 1 0 2 0 1 0
0 2 2 0 2 0 0
0 2 2 1 1 0 0
0 0 0 0 0 0 0

x[:,:,1]

0 0 0 0 0 0 0
0 0 2 1 0 1 0
0 2 2 2 2 1 0
0 0 1 1 1 1 0
0 1 1 0 0 0 0
0 0 0 2 2 2 0
0 0 0 0 0 0 0

x[:,:,2]

0 0 0 0 0 0 0
0 0 1 2 2 1 0
0 2 0 0 2 1 0
0 2 2 2 0 0 0

Filter W0 (3x3x3)

w0[:,:,0]

0 0 -1
1 0 0
0 0 1

w0[:,:,1]

0 1 -1
-1 1 -1
0 -1 0

w0[:,:,2]

0 1 -1
1 -1 0
0 0 -1

Bias b0 (1x1x1)

b0[:,:,0]

1

Filter W1 (3x3x3)

w1[:,:,0]

0 1 0
0 0 1
0 0 0

w1[:,:,1]

0 -1 -1
-1 0 0
-1 0 0

w1[:,:,2]

0 0 -1
1 -1 0
1 -1 0

Bias b1 (1x1x1)

b1[:,:,0]

0

Output Volume (3x3x2)

o[:,:,0]

-3 -3 3
0 -2 3
-2 2 3

o[:,:,1]

0 -4 0
-5 -6 0
1 1 -3

FIGURE 2. The convolution process.

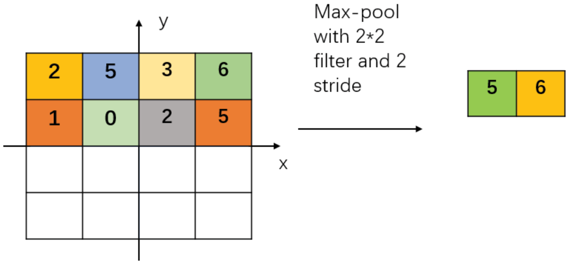
corresponding weight. A convolution kernel corresponds to a neuron, and the size of convolution kernel is called the receptive field of the neuron. When an image is input into the CNN, one calculation process starts: k convolution kernels in the system carry out convolution calculation on the image, that is, the weight values in the convolution kernel and the pixel values at the corresponding position of the image are summed within the receptive field. Then, the convolution kernels slide to the next position of the image according to the step size and repeat the above process until all the pixels in the image are counted. At this point, the output pixel matrix is the feature map of the original image, and k convolution kernels output k feature maps. The process is shown in Fig.2 and expressed as:

$$o_w = [\frac{i_w - n + 2p}{s}] + 1 \tag{1}$$

$$o_h = [\frac{i_h - n + 2p}{s}] + 1 \tag{2}$$

where, o_w and o_h are the size of the output feature graph, i_w and i_h are the size of the input image, s is the sliding step of the convolution kernel, and p is the number of pixels filled.

The pooling layer can further reduce the size of image and retains important information. The pooling methods mainly include maximum pooling, average pooling, etc. Here, we select the maximum pooling method, and its calculation process is shown in Fig.3. A filter with size $2 * 2$ selects the pixel point with the largest value in the range of $2 * 2$ on the image, which is retained as the feature point of this region. Then, the filter slides down to the next range and repeats this process, a feature map is formed.



2	5	3	6
1	0	2	5

Max-pool with 2*2 filter and 2 stride

5	6

FIGURE 3. The pooling process.

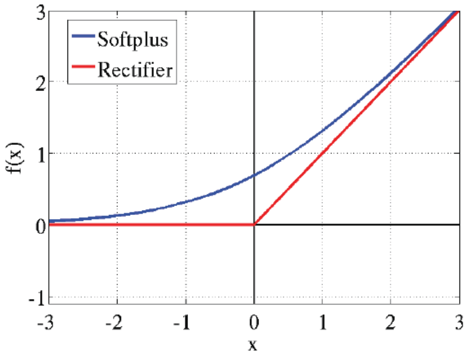


FIGURE 4. The ReLU activation function with softplus function.

To improve the fitting ability of CNN, the nonlinear factors need to be added. Therefore, the nonlinear activation function is adopted to map the output characteristic graph of CNN. The commonly used nonlinear activation functions include tanh, sigmiod, ReLU, etc. Among them, the ReLU function (shown in Fig.4) is the most common one in the CNN because it can better simulate the brain environment, the Softplus function ($Softplus(x) = \log(1 + e^x)$) is a smooth ReLU function. The ReLU activation function is expressed as:

$$f(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{3}$$

In the traditional CNN, on the one hand, due to the existence of pooling layer, some small pixel points on the image will be missed in the calculation process, resulting in the loss of network accuracy; on the other hand, when the layers of CNN are deepened, the network needs more parameters, and leads to more consumption of computing resources. To solve the above problems, Yu et al. proposed the dilated convolution model, which inserted a set of holes (with 0 weights) into the original convolution kernel and formed the dilated convolution kernel to replaced the original one [20]. The above process is shown in the Fig.5.

B. THE BRIEF INTRODUCTION OF DILATED CNN

To cope with more complex situations and achieve higher network accuracy, the depth of CNN is increased by stacking layers in traditional methods. However, the disadvantages are also very prominent. The back propagation of gradient may lead to the vanishing gradient with the increasing of network layers, then the network performance tends to be

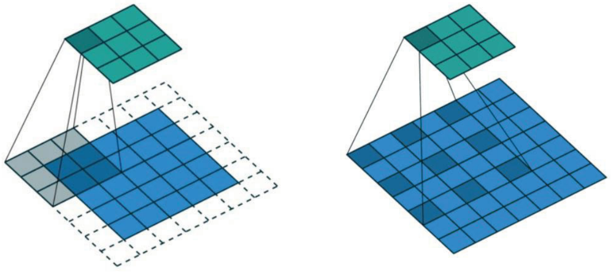


FIGURE 5. The calculation process of dilated convolution.

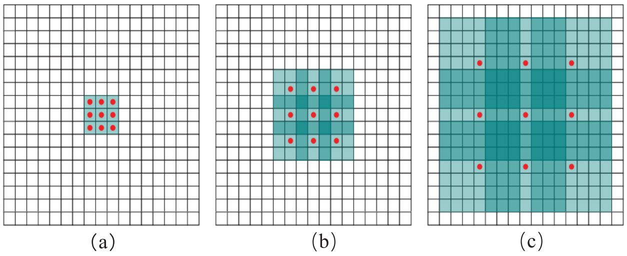


FIGURE 6. The stacking effect of dilated convolution kerne [20].

saturated or even drops sharply. The dilated CNN model is a solution to the above problems.

The dilated CNN model is formed by using the dilated convolution kernels to instead the traditional convolution kernels. The dilated convolution kernel is shown in the Fig.6, where (a) is a traditional convolution kernel of size $3 * 3$, a hole (a point with weight of 0) is inserted around each point in the matrix in (a) and transform into (b), similarly, (c) is a 3-hole dilated convolution kernel. As Fig.6 shows, the receptive field of the convolution kernel in (a) is $3 * 3$, in (b) is $7 * 7$, and in (c) is $15 * 15$. The size of the receptive field increases with the increase of the number of inserted hole, however, the number of parameters in (a), (b) and (c) are still the same. Therefore, using such a dilated convolution kernel to process images can make the convolution kernel obtain more information without increasing the computation.

III. THE DESIGN AND TESTING FOR DILATED CNN MODEL

In this section, we mainly focus on the structure design of dilated CNN model, and then, to verify the effectiveness, the performance of dilated CNN and the traditional CNN is compared under the same circumstances.

A. THE DILATED CNN MODEL DESIGN

Fig.7 is a traditional CNN model with two convolution-pooling layers and two full connection layers, which is improved based on the classical Lenet-5. The softmax function is used to classify the feature map output of full connection layers, and a vector representing probability classification is the output with value in $(0 - 1)$. The softmax function is:

$$\sigma(z_j) = \frac{e^{z_j}}{\sum_{j=1}^{J} e^{z_j}} \tag{4}$$

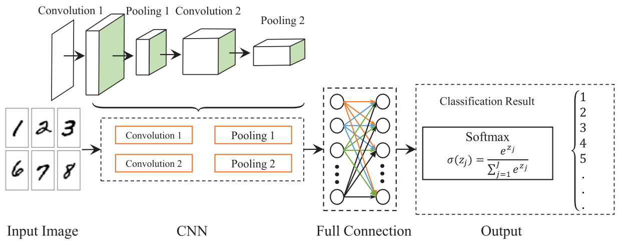


FIGURE 7. The structure of traditional CNN model.

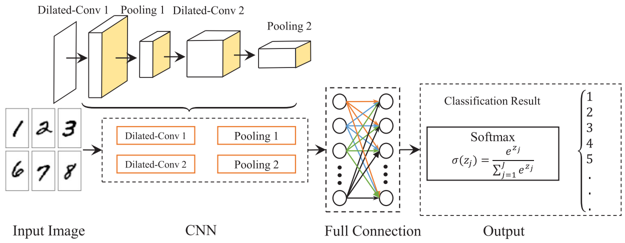


FIGURE 8. The structure of dilated CNN model.

TABLE 1. The hardware environment configuration.

Item	Details
CPU	Intel(R) Core(TM) i7-8550U
CPU frequency	1.99GHz
Memory	8GB

where, z_j is the j-th element in the array z, and J is total number of elements in the array z.

To highlight the performance improvement of the traditional CNN by dilated convolution, we replace the traditional convolution kernels in Fig.7 with dilated convolution kernels, and design the dilated CNN model, see Fig.8.

B. THE PERFORMANCE TESTING OF DILATED CNN MODEL

The experimental hardware environment configuration is shown in Table.1. The performance indexes tested in this section include training accuracy, testing accuracy and time consuming. The common neural network frameworks include TensorFlow, Caffe, Keras, etc. In this paper, TensorFlow is selected to build the dilated and traditional CNN models.

The Mnist data set is the handwritten digital recognition data from the National Institute of Standards and Technology, which contains the training set of 60,000 samples and the testing set of 10,000 samples. Its structure is shown in Fig.9. In this section, the Mnist data set is used to train and test the dilated CNN model and the traditional CNN model. Since the Mnist handwritten digital recognition data set contains ten kinds numbers $(0 - 9)$, J is set to 10 in the softmax function.

On this data set, 1000, 2000, and 5000 times of training are carried out, and the performance indexes of the traditional CNN and dilated CNN are compared respectively. The results

TABLE 2. The parameters of the dilated CNN and traditional CNN.

Training rounds	Time required (s)		Training accuracy		Testing accuracy	
	Traditional CNN	Dilated CNN	Traditional CNN	Dilated CNN	Traditional CNN	Dilated CNN
1000	423.072	382.716	0.9448	0.9680	0.9841	0.9835
2000	445.788	386.928	0.9611	0.9829	0.9877	0.9874
5000	465.984	390.348	0.9628	1	0.9879	0.9878

TABLE 3. The performance improvement of the dilated CNN model.

Training rounds	Training duration improvement	Training accuracy improvement	Testing accuracy improvement
1000	9.54%	2.46%	-0.06%
2000	13.20%	2.27%	-0.03%
5000	16.23%	3.86%	-0.01%
Average	12.99%	2.86%	-0.03%

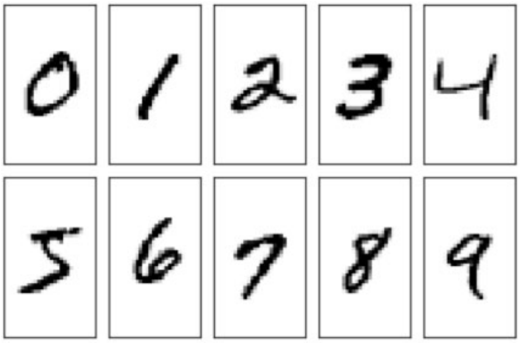


FIGURE 9. The examples of mnist handwritten digital recognition data set.

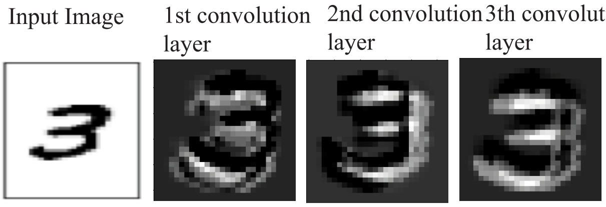


FIGURE 10. The feature maps in the training process.

are shown in Table.2 and Table.3. It can be seen that, under the same experimental environment, the training time of dilated CNN model is reduced by 12.99% averagely. With the training rounds increasing, the training accuracy of the dilated CNN model and the traditional CNN model both increase, and the training accuracy of the dilated CNN model is always higher than that of the traditional CNN model, which is 2.86% averagely. When the training rounds are 5000, the network performance is saturated, and the training accuracy can not be improved furtherly. With the training rounds increasing, the testing accuracy of dilated CNN decreases slightly than that of traditional CNN. As an example, the feature maps (the first, second and third convolution layer) of handwritten 3 generated in the training process of dilated CNN are shown in Fig.10.

C. FURTHER ANALYSIS

It can be concluded from section III-B, simply stacking the dilated convolution kernels can shorten the training time and increase the training accuracy in some extent, but can not effectively improve the testing accuracy. The reasons are as follows:

- the discontinuity between the dilated convolution kernel leads to the omission of some pixels, which may lead to the neglect of the continuity information on the image;
- when extracting the image feature map, if the size rate is fixed, the large and small size information cannot be taken into account simultaneously.

Similar to the image segmentation filed, these problems will affect the training and testing accuracy of the dilated CNN model. To solve the problems, Wang et al. proposed a special dilated CNN model named HDC model. The dilated convolution kernels of HDC model have different dilation rates in the different layers and can make a series of convolution operations to cover the area of a square completely without holes or missing information. The above measures can solve the information loss and precision reduction problem caused by the holes in dilated convolution kernels [21]. Therefore, the HDC model can be established for image classification.

IV. THE DESIGN AND TESTING FOR HDC MODEL

In this section, we mainly focus on the design and performance testing of HDC model, and to highlight the effectiveness of the HDC model, the traditional CNN model and dilated CNN model are compared with the HDC model.

A. THE HDC MODEL DESIGN

To solve the problems appeared in the previous section and further improve the performance of the dilated CNN model, we build the HDC model whose structure is shown in Fig.12. In Fig.12, the dilation rates are set as 1, 2, 5 and 1, 2, 5 in the dilated convolution kernels, respectively, so they can cover every single pixel point on the image and ensure that the key information won't be lost to the greatest extent when extracting the feature map. Then the HDC model is built with 6 dilated convolution-pooling modules, 2 full connection layers and a softmax function. To avoid overfitting, the two-layer Dropout function is adopted to optimize the model. The designed HDC model can not only retain the original advantages of the dilated CNN, but also make up for the deficiency that the dilated CNN model will lose the image details.

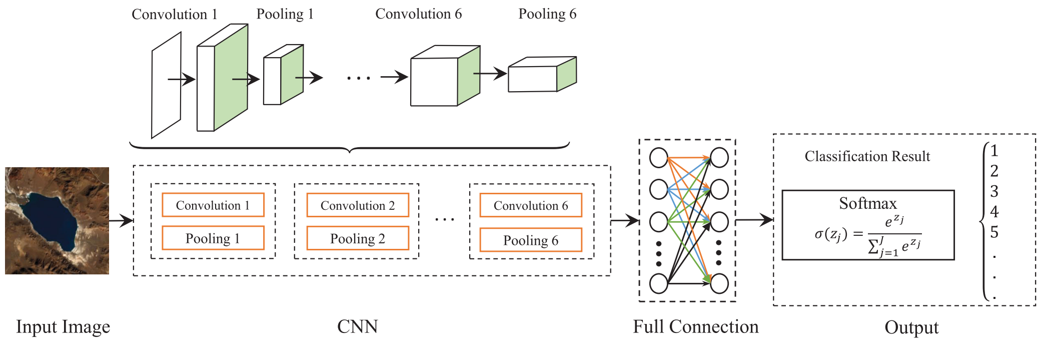


FIGURE 11. The structure of traditional CNN model.

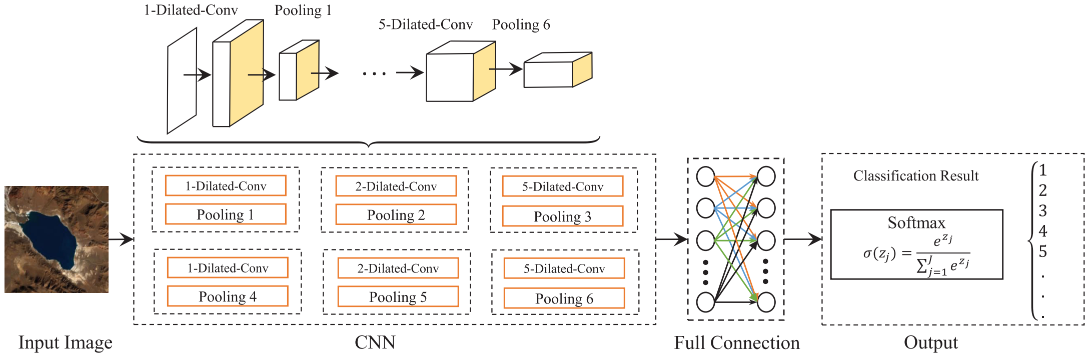


FIGURE 12. The structure of HDC model.

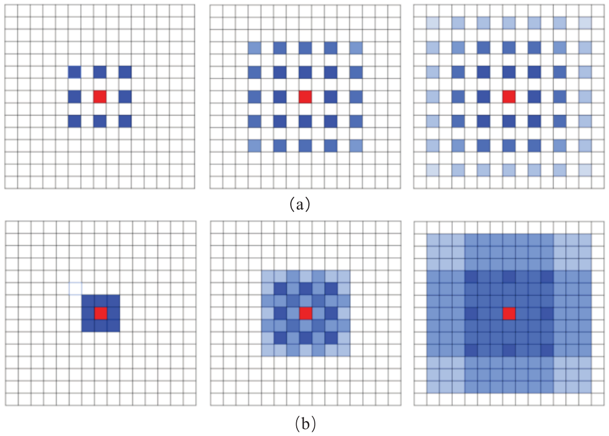


FIGURE 13. The stacking effect of dilated convolution kernels [21].

The choice of dilation rate is very important when designing the structure of HDC model, and it should meet:

$$M_i = \max[M_{i+1} - 2r_i, M_{i+1} - 2(M_{i+1} - r_i), r_i] \qquad (5)$$

where $i = 1, 2, \cdots, n$; r_i is the dilation rate in the i-th layer; M_i is the largest r_i in the i-th layer [21]. The stacking of the dilated convolution kernels is shown in Fig.13, where (a) is the stacking effect of three dilated convolution kernels with size $3 * 3$ and $r_i = 2$; (b) is the stacking effect of three dilated convolution kernels with $r_i = 1$, 2 and 3. It can be seen that (a) always leave some holes between pixels, and these holes are the places on the image where small information exist; in (b), every single step in the HDC model can fill the holes remained by last convolution computation and finally get the $13 * 13$ receptive field without any holes.

B. THE PERFORMANCE TESTING AND ANALYSIS OF HDC MODEL

The wide-band remote sensing image set of the earth's terrain is selected as the data set of the HDC model, it has 10,000 pictures divided into 6 types including deserts, oceans, farmland, mountains, lakes and cities (some examples are shown in Fig.14). The data set is taken by Tiangong-2 which is China's first space laboratory, mainly responsible for earth observation and space earth system science, new space application technology, space technology and space medicine applications and experiments. The wide-band remote sensing image is a kind of image with more information than the ordinary RGB image, which can obtain the image, polarization and spectrum of the object at the same time.

The HDC model is compared with the dilated CNN model and traditional CNN model under the same condition and parameters. The traditional CNN model is shown in Fig.11. Firstly, during the training process, 50 pictures are randomly selected in the training set as a batch, and the process is repeated for 100, 200, 350 and 500 times, respectively, and the Adam algorithm is used to improve the generalization ability of the model. Then, after the model converges,

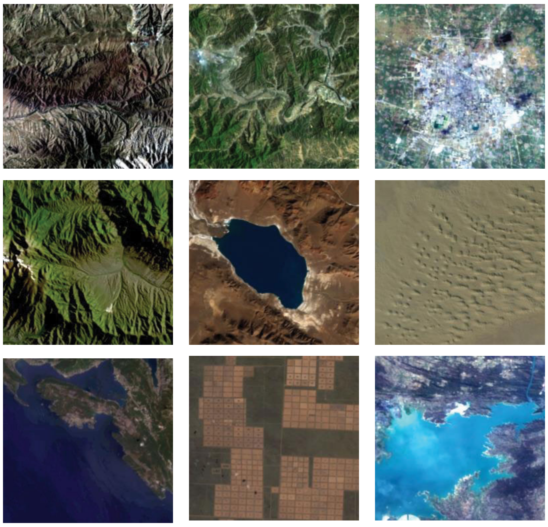


FIGURE 14. **The examples of wide-band remote sensing images of the earth's terrain.**

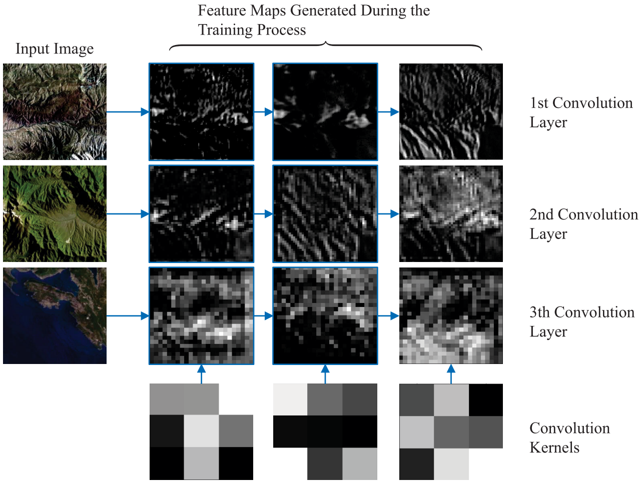


FIGURE 15. **The examples of feature maps and convolution kernels in HDC model.**

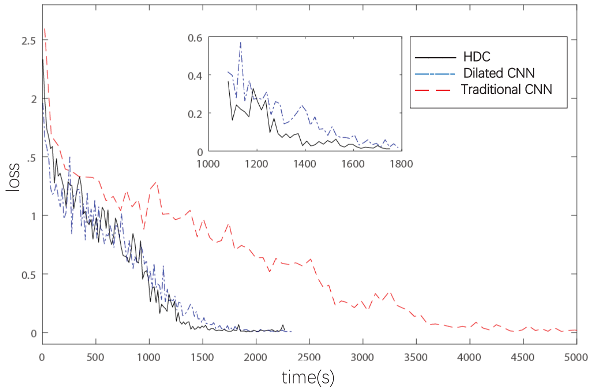


FIGURE 16. **The cost function comparison.**

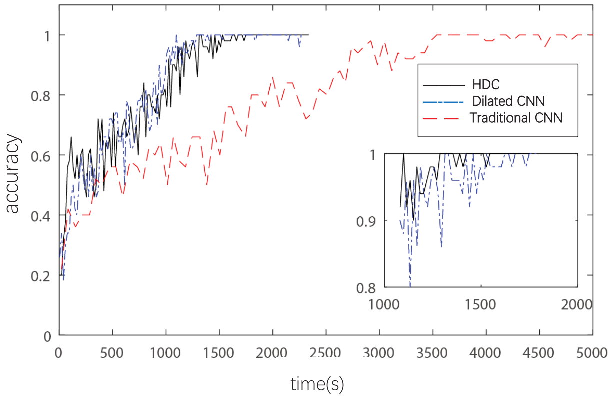


FIGURE 17. **The training accuracy comparison.**

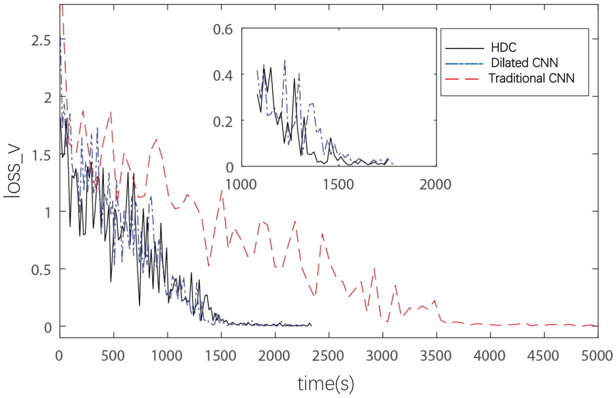


FIGURE 18. **The cost function comparison of cross validation.**

the random images are selected from the testing set whose capacity is 1,000 to test the model performance. Finally, the cross-validation set is used to verify the model performance and select the optimal result.

Fig.15 shows the feature maps and convolution kernels when the input images are processed in the HDC model. It can be seen that the weights of dilated convolution kernel are updated constantly with the iteration of the training, the output feature maps are also changed with different weights. Meanwhile, the Fig.15 also obviously shows that, in the HDC model, the feature maps output by the deep convolution layers have a higher convolution level than the shallow layers.

Fig.16, Fig.17 and Fig.18 show the performance indexes in the training process of HDC model, which include the loss function, training accuracy and loss function of cross validation set. It can be seen that, the dilated CNN model and HDC model have higher convergence speed than traditional CNN model in terms of all the performance indexes, meanwhile, the insets of Fig.16, Fig.17 and Fig.18 show that the HDC model converges faster than the dilated CNN model and has a better performance.

The precise data of performance indexes are shown in Table.4 and Table.5. As can be seen, the training time required by the HDC model is always less than that of the dilated CNN model. When the training rounds are 100, 200, 350, and 500, the training time of the HDC model is 0.73%, 2.11%, 2.44% and 2.78% less than that of the dilated CNN model, respectively, and 2.02% averagely.

TABLE 4. The experimental result comparison of traditional CNN, dilated CNN and HDC model.

Training rounds	Time required (s)			Training accuracy			Testing accuracy		
	Traditional CNN	Dilated CNN	HDC	Traditional CNN	Dilated CNN	HDC	Traditional CNN	Dilated CNN	HDC
100	1245.08	364.74	362.09	0.6602	0.4837	0.6458	0.6042	0.6040	0.8032
200	2510.51	730.12	714.69	0.8160	0.6839	0.8245	0.8031	0.7893	0.9468
350	4375.28	1279.44	1248.30	0.8835	0.9208	0.9440	0.9364	0.9218	1
500	6155.13	1830.95	1780.01	1	1	1	1	1	1

TABLE 5. The performance improvement of HDC (compared with dilated CNN).

Training rounds	Time required improvement	Training accuracy improvement	Testing accuracy improvement
100	0.73%	33.51%	32.98%
200	2.11%	20.56%	19.95%
350	2.44%	2.52%	8.48%
500	2.78%	0	0
Average	2.02%	14.15%	15.35%

In terms of the training accuracy, the HDC model is 14.15% better than the dilated CNN model averagely. When the training rounds are 100, 200 and 350, the training accuracy of HDC model has considerable improvements, which are 33.51%, 20.56% and 2.52% respectively; when 500, there is no improvement, because the performance of the HDC and dilated CNN model tends to be saturated, i.e., simply increasing the training rounds can not provide a significant performance improvement to the model, therefore, when training rounds reach a certain level, the training accuracy of the HDC model will not continue improving.

In terms of the testing accuracy, the HDC model is 15.35% better than the dilated CNN model averagely. When the training rounds are 100, 200 and 350, the testing accuracy of HDC is improved by 32.98%, 19.95% and 8.48%, respectively; when 500, the improvement is none, still, as the network performance tends to be saturated, the increase of training rounds cannot continue to significantly improve the model performance, resulting in the stability of the testing accuracy.

V. CONCLUSIONS

In this paper, the dilated CNN model and the HDC model is proposed for image classification. The dilated convolution kernels are used to replace the traditional convolution kernels, and the verification is carried out on the Mnist handwritten digital recognition data set and the wide-band remote sensing image data set of the earth terrain. Experiments show that, the dilated CNN model has less time consuming and higher training accuracy on the Mnist data set compared with traditional CNN model; the training and testing accuracy of HDC model are both higher, and the time consuming is less than those of dilated CNN model on the remote sensing image data set.

REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 1, Dec. 2015, pp. 1026–1034.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2016, pp. 770–778.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Computer Science*, 2014.

[5] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[6] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn.*, 2001, pp. 282–289.

[7] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[8] Y. LeCun, B. Boser, and J. S. Denker, "Handwritten digit recognition with a back-propagation network," in *Proc. Neural Inf. Process. Syst.*, 1990, pp. 396–404.

[9] S. Pang, A. Du, M. A. Orgun, and Z. Yu, "A novel fused convolutional neural network for biomedical image classification," *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 107–121, 2018.

[10] I. Sikirić, K. Brkić, P. Bevandić, I. Krešo, J. Krapac, and S. Šegvić, "Traffic scene classification on a representation budget," *IEEE Trans. Intell. Transp. Syst.*, to be published.

[11] Y. Zhou, X. Liu, J. Zhao, D. Ma, R. Yao, B. Liu, and Y. Zheng, "Remote sensing scene classification based on rotation-invariant feature learning and joint decision making," *EURASIP J. Image Video Process.*, p. 3, 2019.

[12] Z. Zhang, X. Wang, and C. Jung, "DCSR: Dilated convolutions for single image super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1625–1635, Apr. 2019.

[13] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[14] Y. Wei, H. Xiao, H. Shi, Z. Jie, J. Feng, and T. S. Huang, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7268–7277.

[15] X. Zhang, Y. Zhou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proc. 22nd Int. Conf. Digit. Signal Process. (DSP)*, Aug. 2017, pp. 1–5.

[16] Y. Kudo and Y. Aoki, "Dilated convolutions for image classification and object localization," in *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, 2017, pp. 452-455.

[17] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 472–480.

[18] C. Liu, Z. Shang, and A. Qin, *A Multiscale Image Denoising Algorithm Based on Dilated Residual Convolution Network*, 2018.

[19] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.

[20] F. Yu and V. Koltun, *Multi-Scale Context Aggregation by Dilated Convolutions*, 2015.

[21] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, "Understanding Convolution for Semantic Segmentation," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, vol. 1, Mar. 2018, pp. 1451–1460.

XINYU LEI received the bachelor's degree from the Xi'an University of Science and Technology, in 2017, where she is currently pursuing the master's degree. Her research interests include image classification, image segmentation, machine learning, and deep learning and their applications.

HONGGUANG PAN received the bachelor's degree from the Xi'an University of Science and Technology, in 2007, and the Ph.D. degree from Xi'an Jiaotong University, in 2003, Xi'an, China, respectively. He is currently a Lecturer with the Xi'an University of Science and Technology. From September 2013 to March 2015, he studied with Lehigh University, USA, as a Visiting Ph.D. Student. His research interest covers artificial intelligence, model predictive control, and brain-machine interface and their applications.

XIANGDONG HUANG received the bachelor's and master's degrees from the Xi'an University of Science and Technology, in 1982 and 1987, respectively, where he is currently an Associate Professor. His research interests include intelligent detection and control technology, fieldbus control system (FCS), and pattern recognition and intelligent systems and their applications.