A Depthwise Separable Network for Action Recognition

Shengwei Zhou, Liang Bai, Yang Yang, Haoran Wang and Kesong Fu

ABSTRACT

Action recognition is one important but challenging tasks in computer vision. The 3D convolutional neural network is one mainstream method for action recognition because it can extract temporal and spatial information in the video simultaneously. However, 3D convolutional neural network has a serious drawback which is that its parameter quantity is too large. Depthwise convolution is a form of group convolution, which can effectively reduce the parameter of convolution kernel, and has been widely applied in 2D convolutional neural network. Therefore, we propose to introduce depthwise convolution into the 3D convolutional neural network. We choose 3D resnet as our basic model, and construct our model by replacing the 3D convolution kernel in the baseline with the depthwise convolution, we named our proposed model as depthwise separable network (DSN). We conducted experiments on UCF101 and HMDB51 dataset. The experimental results show that by introducing the depthwise convolution, our DSN network can not only reduce the parameters of the baseline, but also can moderately improve the accuracy.

KEYWORDS

Action recognition, 3D convolution, Depthwise convolution.

INTRODUCTION

Action recognition has attracted much attention in recently because of wide range of application such as dangerous video recognition and human-computer interaction. Different from the task in the image field, the video contains both spatial information and temporal information. The key problem in action recognition is how to effectively extract the spatiotemporal information in the video. In recent years, 3D convolutional neural networks have been recognized as a mainstream method for action recognition because 3D convolution is capable of simultaneously extracting spatial and temporal information[1-6]. Recently, the emergence of large-scale action recognition datasets and the development of computer performance have also contributed to the development of 3D convolutional neural networks.

Although 3D convolutional neural network has been extensively studied, they still have some serious drawbacks. One of the main problems is that the parameter size of the 3D convolutional neural network is too large, which limits the further development of them. In the field of images, depthwise...
convolution has been widely used in 2D convolutional neural networks to reduce parameters[7-9]. By introducing depthwise convolution, the 2D convolutional neural network can separately model the channel information and spatial information of the image, and can greatly reduce the amount of network parameters. However, to date, depthwise convolution has never been introduced into 3D convolutional neural networks. Therefore, we propose to introduce depthwise convolution into the 3D convolutional neural network to reduce the parameters.

The main contributions of this article include: 1) By introducing depthwise convolution into the 3D convolutional neural network, we built a lightweight 3D convolutional neural network named as depthwise separable network (DSN); 2) We conducted experiments on two classical datasets UCF101[10] and HMDB51[11], and found that our model not only reduces the amount of parameters in the baseline, but also increases its accuracy. 3) We analyzed the training process of baseline and our Channel-Separated Convolutional Network, and found that depthwise convolution can provide a regularization function, which can prevent model overfitting and promote early convergence. It is worth mentioning that our proposed method is universal and can be easily implemented in any other 3D convolutional neural networks.

RELATED WORK

METHOD FOR ACTION RECOGNITION

Since the neural network achieved amazing results in the ImageNet competition in 2012[12], the neural network has developed rapidly, and a series of neural network structures have emerged for action recognition. So far, the action recognition method mainly includes two mainstream methods. The first method is the two-stream convolution neural network. K. Simonyan et al. [13] first proposed the two-stream neural network framework, which consists of two branches networks of identical structure, one for processing images in the video and the other for processing the optical flow of the video. L. Wang et al. [14] introduced a sparse sampling method into the two-stream neural network to better extract long-range information. After that, a series of methods based on two-stream neural networks have been appeared[15-17]. However, there is an unavoidable problem in the two-stream networks, that is, them needs to pre-process the video to extract the optical flow, which not only requires additional computational cost but also limits the application of video real-time processing.

Another mainstream method for action recognition is the 3D convolutional neural network. S.Ji et al.[1] first proposed using 3D convolutional neural networks to process action recognition tasks, because the 3D convolution kernel can extract spatial information and temporal information in the video simultaneously. D. Tran et al.[2] proposed a standard 3D convolutional neural network architecture named as C3D. Since then, the action recognition framework based on 3D convolution has attracted wide attention[20-22]. Recently, the emergence of large video dataset has further promoted the development of 3D convolutional neural networks. However, there is still a problem with the 3D convolutional neural networks, which is that their parameter quantity is still too large, and it is prone to over-fitting, especially on some smaller datasets. Therefore, we mainly focus on how to reduce the parameter of the 3D convolutional neural networks without dropping the accuracy.
DEPTHWISE CONVOLUTION

The idea of group convolution is to divide the input feature maps into different groups and then convolve them separately. The group convolution was first used in AlexNet[12] in the 2012 Imagenet competition, it was used to segment the network to solve the problem of GPU memory limitations. Since then, group convolution has been introduced into many neural network to reduce the amount of parameters, such as ShufflfeNet[7], ResNeXt[18]. Depthwise convolution is one extreme form of group convolution, where the feature map grouped by each channel. Depthwise convolution can greatly reduce the amount of parameters, making the network model calculations faster. F. Chollet. et al. [8] introduced the depthwise convolution into the Inception model and built an extreme version of Inception named Xception. Howard et al.[9] built a lightweight neural network for mobile applications by introducing depthwise convolution. However, all of these network structures are based on 2D convolutional neural networks to solve problems in the image domain, and we are mainly focus on introducing depthwise convolution into 3D convolutional neural networks for action recognition.

PROPOSED METHOD

DEPTHWISE CONVOLUTION

In the conventional convolution process, each convolution kernel receives data from all channels in the previous layer as input, meaning that the connection between the input channel and the output channel is fully connected, as in Figure 1(a). But this kind of dense connections in conventional convolution leads to a large amount of parameters, especially for 3D convolutional neural networks.

Group convolution is a classic method for reducing the amount of parameters in convolution. Group convolution divides the input channel into several groups, and then each convolution kernel accepts only the channels in the same group as input, as in Figure 1(b). Depthwise convolution is an extreme form of group convolution, where each group has only one channel, indicating that each convolution kernel has only one channel as input, as in Figure 1(c).

Figure 1. Several different kinds of convolution for the situation of 4 input channels and 4 output channels. The rectangular boxes in the figure represent a group. a) conventional conv: all channels in a group. b) group conv: channels are divided into 2 groups. c) depthwise conv: each channel is a separate group.
PROPOSED ARCHITECTURE

The 3D convolutional neural networks is one of perfect methods for action recognition because it can extract spatial information and temporal information simultaneously. But one serious drawbacks of the 3D convolutional neural networks is that the parameter quantity is too large. So we are mainly focus on how to reduce the parameters of the 3D convolutional neural network by introducing depthwise convolution.

We chose 3D resnet[6] as our basic model, because it is one of the most representative network structures in computer vision. After that, we built our model by introducing depthwise convolution into the 3D resnet. The specific implementation details of introducing depthwise convolution into 3D resnet is shown in Figure 2, where dw represents the depthwise convolution. For simplicity, we skip the skip connections in the figure. The key to implementation is that the depthwise convolution requires the number of input channels be consistent with the number of output channels. For the basic block, we first add one 1x1x1 convolution before the 3x3x3 convolution to keep the dimensions consistent, then replace the 3x3x3 convolution with the depthwise convolution, as in Figure 2 (a). For the bottleneck block, since the input and output dimensions of the original 3x3x3 convolution are consistent, so we just replace it with the corresponding depthwise convolution, as in Figure 2(b).

![Figure 2. The specific implementation details of introducing depthwise into 3D resnet. The implementations for basic blocks and bottleneck blocks are different. The dw in the figure represents depthwise convolution.](image)

During the experiment, we chose 3D resnet 50 as our basic model, the structure of which is almost the same as 2D resnet 50 [19], except that its convolution and pooling kernel are extended from 2D to 3D. After that, we replaced the 3x3x3 convolution in the basic 3D resnet with 3x3x3 depthwise convolution to build our model, which is named as depthwise separable network(DSN).

In conventional 3D convolutional neural networks, channel interaction and spatiotemporal interaction are modeled simultaneously. However, in our depthwise separable network, the modeling process of channel interaction and spatiotemporal interaction is separated, where 1x1x1 convolution is used to model channel interactions, and 3x3x3 depthwise convolution is used to model spatiotemporal interaction. The architecture of our model is described in TABLE I, where dw represents depthwise convolution.
TABLE I. THE ARCHITECTURE OF OUR DEPTHWISE SEPARABLE NETWORK.

| Layer Name | Architecture | Output size |
|------------|--------------|-------------|
| conv1      | 7 x 7 x 7, stride 1,2,2 | 64 x 16 x 56 x 56 |
| conv2      | 3 x 3 x 3 max pool, stride 2 | 64 x 8 x 28 x 28 |
| conv3      | 1 x 1 x 1 | 128 x 4 x 14 x 14 |
| conv4      | 1 x 1 x 1 | 256 x 2 x 7 x 7 |
| conv5      | 1 x 1 x 1 | 256 x 1 x 4 x 4 |
| avg        | average pool | 256 x 1 x 1 x 1 |
| fc         | fully connected layer | 101-d / 51-d |

EXPERIMENT

DATASET

We conducted experiments on two classic datasets of the action recognition: UCF101[16] and HMDB51[17]. The UCF101 dataset is collected from YouTube. It contains 101 types of actions categories and a total of 13320 videos, the average length of which is about 7 seconds. They are mainly belong to five categories: human-object interaction, human movement, instrumental performance, human-to-human interaction and sports. The HMDB51 dataset is collected from YouTube and Google. It contains 51 categories of actions, 6849 videos, which are about 3 seconds long in average. The videos in HMDB51 dataset are mainly divided into four categories: general facial actions, object operations, general body movements, and interaction with objects.

RESULTS

To test the validity of our proposed model, we first train the basic model and our depthwise separable network from scratch on both the UCF101 and HMDB51 datasets. However, A K. Hara et al.[6] confirmed that these two datasets are still too small for training 3D convolutional neural networks, while the kinetics[3] dataset has enough data to train 3D convolutional neural networks. So in order to further verify the validity of our proposed model, we also pre-train these two models on the large dataset kinetics[3] and then fine-tune them on the UCF101 and HMDB51 dataset. The input to all models is 3 channels x 16 frames x 112 pixels x 112 pixels. During the test, the model outputs the clip-level prediction of 16 frames, after which these clip-level prediction are integrated into the video-level prediction.

The experimental results are shown in TABLE II. It can be seen that by introducing depthwise convolution, we greatly reduced the parameter of the basic network directly from 45.3M to 12.3M. What’s more, whether train from scratch or trained with pre-trained model, our depthwise separable network improve the accuracy of the basic model significantly, which proves the effectiveness of depthwise convolution. The results show that the amount of parameters...
contained in the 3D convolutional neural network may be too large, which leads to the phenomenon that the model is prone to over-fitting and the effect is poor.

**TABLE II. THE ACCURACIES COMPARISON ON UCF101 AND HMDB51 DATASET.**

| Model                        | Parameters | UCF101 | HMDB51 |
|------------------------------|------------|--------|--------|
| 3D ResNet50 (baseline)       | 45.3M      | 44.3   | 19.7   |
| Depthwise separable network  | 12.3M      | 45.9   | 21.8   |
| 3D ResNet50 (pretrained)     | 45.3M      | 89.1   | 60.8   |
| Depthwise separable network (pretrained) | 12.3M | 89.4 | 61.5 |

To further explore the effect of depthwise in 3D convolutional neural networks, we analyzed the accuracy curve of the basic model and our depthwise separable network during the training process. Fig.3 is training accuracy curve of the baseline and DSN when training on the UCF101 from scratch. It can be seen that, compared with the basic model, our DSN improves both the training accuracy and the validation accuracy, and converges in advance. Therefore, we believe that depthwise convolution can provide a regularization effect on 3D convolutional neural networks, promoting model convergence in advance and improving accuracy.

![Figure 3](image)

*Figure 3. Training accuracy curve of the baseline and DSN when training on the UCF101 from scratch. Compared to the basic model, the DSN not only converges faster, but also have higher accuracy.*

**CONCLUSION**

For the problem that the parameters of the 3D convolutional neural network are too large, we propose using the depthwise convolution to replace the original 3D convolution. We chose 3D resnet50 as the basic model, and constructed a lightweight 3D convolutional neural network named as depthwise separable network (DSN). To test the validity of our proposed model, we selected 3D resnet50 as our basic model, and carried out experiments on two classic action recognition datasets UCF101 and HMDB51 from both scratch and pre-trained...
model. The experimental results show that our model can reduce the parameters of the benchmark while improving model accuracy.

We specifically analyzed the training process of the basic model and our depthwise separable network, and concluded that the depthwise convolution can provide a regularization effect to the 3D convolutional neural network, thus promoting the early convergence of the model and improving the accuracy. We believe that depthwise convolution is a much important for 3D convolutional neural networks and it can be easily introduced into any 3D network structure. In the following work, we will further introduce the depthwise convolution into other 3D convolutional neural networks to further verify its practicability and effectiveness.

REFERENCES

1. S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. IEEE TPAMI, 35(1):221–231, 2013. 2.
2. D. Tran, L. Bourdev, R. Fergus, L. T Orresani and M. Paluri. Learning spatio-temporal features with 3D convolutional networks. In ICCV, 2015.
3. J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In CVPR, 2017.
4. D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri. A closer look at spatio-temporal convolutions for action recognition. In CVPR, 2018.
5. X. Wang, R. Girshick, A. Gupta and K. He. Non-local neural networks. In CVPR, 2018.
6. K. Hara, H. Kataoka and Y. Satoh. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? In CVPR, 2018.
7. X. Zhang, X. Zhou, M. Lin, and J. Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. CoRR, abs/1707.01083, 2017. 1, 2, 4
8. F. Chollet. Xception: Deep learning with depthwise separable convolutions. In CVPR, 2017. 1, 2, 3.
9. A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. CoRR, abs/1704.04861, 2017. 1, 2, 3.
10. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. CoRR abs/1212.0402 (2012).
11. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T.A., Serre, T.: HMDB: A large video database for human motion recognition. In: ICCV. (2011) 2556–2563.
12. A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In NIPS, 2012. 2.
13. K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In NIPS, 2014.
14. L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In ECCV, 2016.
15. Z. Lan, Y. Zhu and A. G. Hauptmann. Deep local video feature for action recognition. In CVPR, 2017.
16. R. Girdhar, D. Ramanan, A. Gupta, J. Sivic and B. Russell. ActionVLAD: Learning spatio-temporal aggregation for action recognition. In CVPR, 2017.
17. G. A. Sigurdsson, S. Divvala, A. Farhadi and A. Gupta. Asynchronous temporal fields for action recognition. In CVPR, 2017.
18. S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In CVPR, 2017. 2.
19. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
20. C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In Proceedings of the Advances in Neural Information Processing Systems (NIPS) 2016. 3.

21. K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3D residual networks for action recognition. In Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition, 2017.

22. Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In Proceedings of the International Conference on Computer Vision (ICCV), 2017. 8.