Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/32/20230201

VGG16 based on dilated convolution for face recognition

Fanxing Meng

Electronic and Information Engineering, Beijing University Of Technology, Beijing, 100000, China

Mengfanxing@emails.bjut.edu.cn

Abstract. Face recognition has wide applications in fields such as security systems, biometrics, and human-computer interaction. However, traditional face recognition methods face challenges in capturing details and reducing model complexity. To address these issues, this paper proposes a new method based on VGG16, which improves recognition accuracy and reduces parameter quantity by introducing dilated convolution and parameter pruning. First, the hole convolution is introduced to expand the Receptive field and capture more details to enhance the ability of the model in distinguishing facial features. Next, parameter pruning is applied to reduce redundant parameters, optimize model structure, and improve computational efficiency. This article conducted experimental evaluation on the classic face recognition dataset CK+ dataset. The results show that the proposed method is significantly superior to the traditional VGG16 model in terms of recognition accuracy. At the same time, the use of pruning technology significantly reduces the number of parameters in the model and improves computational efficiency. The experimental outcomes conclusively validate the effectiveness and feasibility of the proposed method.

Keywords: VGG16, dilated convolution, face recognition.

1. Introduction

Facial expressions are one of the important nonverbal communication methods between people and the most direct manifestation of emotional transmission. By recognizing and understanding facial expressions, we can obtain information about a person's emotional state, emotions, and intentions. In the contemporary era, with the improvement of computer hardware, facial expression recognition technology has found application in numerous domains, including human-computer interaction, affective computing, intelligent security, and various other fields [1].

In traditional machine learning, features are extracted manually. Once the data volume is too large, feature extraction will be a very complex process. In deep learning, neural networks are mainly used for feature extraction, avoiding the tedious process of manual extraction, and the feature extraction effect is better [2]. This implies that it offers an effective and automated approach for analyzing facial expressions, thereby enhancing the performance and expanding the potential applications of facial expression recognition systems. This has great potential in fields such as social media analysis, personalized recommendations, facial animation, and emotional assistance therapy.

At present, deep learning based facial expression recognition has made significant progress. Researchers have proposed a series of deep neural network models, for example, prominent techniques

DOI: 10.54254/2755-2721/32/20230201

employed for extracting and learning expression features from face images or videos include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and attention mechanisms. These advanced methodologies have demonstrated their effectiveness in capturing meaningful spatial information and modeling temporal dependencies, thereby facilitating accurate and robust facial expression recognition. At the same time, there are also a large number of publicly available datasets, such as FER2013, CK+, and FERG, which provide annotated data for facial expression recognition, promoting the development and evaluation of research.

In this paper, we use deep learning to recognize facial expressions in video. We mainly use Convolutional neural network to extract facial expression features from a single image. We use hole convolution to solve the problem of internal data structure loss and spatial hierarchical information loss caused by standard convolution, and on this basis, we conduct appropriate channel pruning, which has played a role in promoting facial expression feature extraction.

2. Method

2.1. Introduction to VGG16

To enhance the precision of facial expression recognition, this study employs VGG16 as the foundational neural network model [3]. VGG16, a deep Convolutional Neural Network (CNN), has been specifically chosen for its compatibility with facial expression recognition tasks. Its Deep structure and surface structure can extract rich and abstract features, and is sensitive to subtle expression differences. The VGG16 model has a smaller convolutional kernel size, which helps to extract more detailed information from images. VGG16 combines convolution and fully connected layers, possessing excellent feature representation and classification capabilities, comprehensively perceiving images and learning expression differences.

VGG16 was proposed by a research team at the University of Oxford in 2014. Its model has a 16-layer deep network structure, which includes 13 convolutional layers and 3 fully connected layers. Within the convolutional layer segment, there exist a total of 13 convolutional layers. The initial convolutional layers employ smaller 3×3 convolutional kernels, while the subsequent convolutional layers utilize larger kernels such as 3×3, 4×4, and 5×5. Following each convolutional layer, a ReLU activation function is applied to incorporate nonlinearity. To down sample the feature map and decrease the size and parameter count, a pooling layer is inserted between two adjacent convolutional layers. The most commonly employed pooling methods include 2×2 maximum pooling and mean pooling (Figure 1).

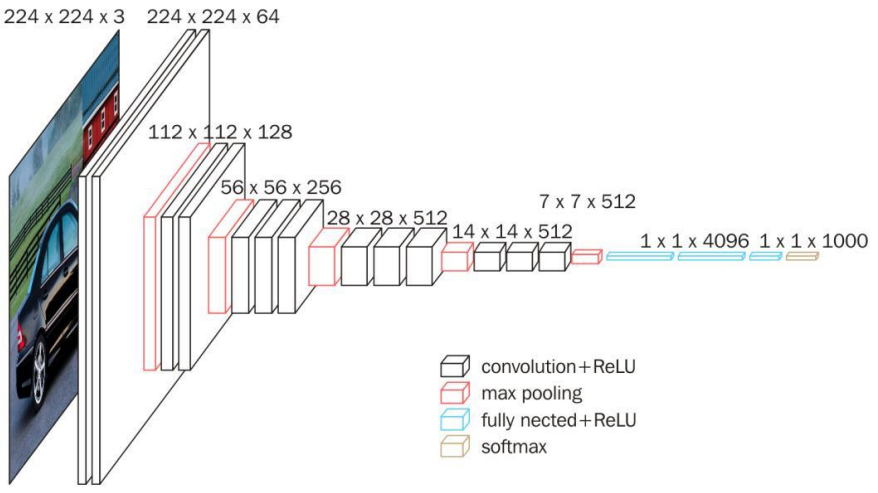

Figure 1. Structure diagram of vgg16 [3].

DOI: 10.54254/2755-2721/32/20230201

In the full connection layer, there are three hidden layers with 4096 neurons, and each hidden layer is followed by ReLU Activation function. The final output layer is a softmax layer with a number of output categories for multi category classification [4].

The VGG16 model extracts high-level features of images by stacking multiple convolutional layers and fully connected layers, thereby achieving the tasks of image classification and target recognition. The traditional VGG16 model has some shortcomings in facial expression recognition. Firstly, the VGG16 model is constructed based on traditional convolutional and pooling layers, which use relatively large convolutional kernels and pooling windows when processing inputs, resulting in the loss of some detailed information [5]. This may affect the model's ability to perceive subtle changes in facial features.

In addition, the convolution operation of the VGG16 model is continuous and does not consider the spatial relationships and contextual information between different regions. This is not ideal in facial expression recognition, as expressions often involve specific regions and local features, requiring the model to accurately capture this information. For these problems, introducing Dilated Convolution can solve them to some extent.

2.2. Implementation of dilated convolution

Hole convolution, also known as dilation convolution or hole convolution, is a special convolution operation used in Convolutional neural network. Compared to traditional convolution operations, hollow convolution has an additional hyperparameter called inflation rate, which defines the spacing between adjacent convolutional kernels. Traditional convolution operations scan the input feature map with a fixed filter size, while dilated convolution introduces gaps within the filter based on the expansion rate, resulting in voids in the convolution operation [6].

In this article, adding empty convolutions can help improve the accuracy of facial expression recognition. First, it expands the Receptive field by increasing the inflation rate, so that it can capture broader context information, which is particularly important for face recognition. In addition, cavity convolution can also by incorporating pooling layers between the convolutional layers, the number of parameters is reduced, thereby decreasing the model's overall complexity and computational requirements. This reduction is achieved while ensuring that the receptive field remains unchanged. Consequently, the model's efficiency is enhanced as it continues to capture the necessary information effectively [7]. Most importantly, hollow convolution avoids the use of pooling operations for down sampling, thereby preserving more detailed information and meeting the requirements of face recognition for high-resolution, detail rich images.

In the given scenario, the convolutional kernel size is 3 × 3. The figure showcases the empty convolution with an expansion rate of d. Figure 2 highlights that ordinary convolution is a particular case of empty convolution, precisely an empty convolution with an expansion rate of 1. Taking Figure 3 as an example, with an expansion rate of 2, the original 3 ×3 The convolution kernel of 3 increases the Receptive field to 5 × 5.

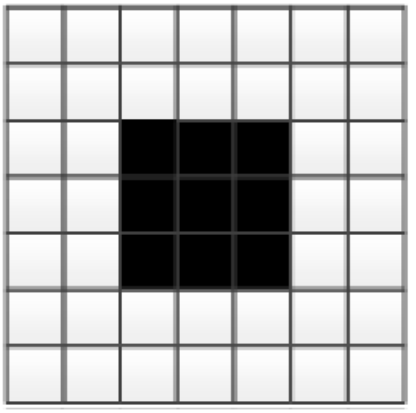

Figure 2. The ordinary convolution.

DOI: 10.54254/2755-2721/32/20230201

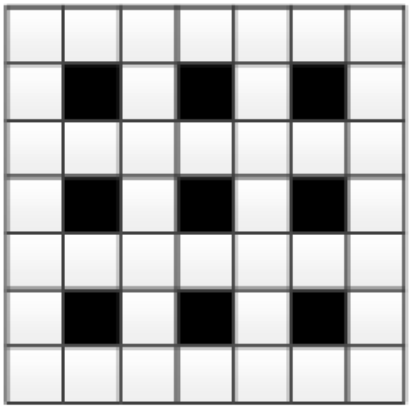

Figure 3. An example with an expansion rate of 2.

This article explores different scenarios for the application location and expansion rate of dilated convolution. After extensive experimentation, it was determined that the most optimal recognition rate was achieved by incorporating dilated convolution with an expansion rate of 2 in the final layer of the VGG-16 network.

2.3. Pruning

Neural network pruning is a technique that optimizes the structure of a neural network model by reducing redundant parameters. Pruning can reduce the size and computational complexity of the model by removing unimportant connections, neurons, or layers in the network, while maintaining model performance.

The VGG16 model consists of 13 convolutional layers and 3 fully connected layers, resulting in a total of approximately 140.6 million parameters. In this article, channel pruning is used to remove redundant and unimportant channels in the model to reduce the number of parameters. Firstly, we choose L1 regularization as the pruning algorithm [8]. L1 regularization punishes the channel weights of the convolutional layer and obtains the corresponding channel importance score. A smaller channel weight means that the contribution of the channel to the model is smaller, so it can be considered to delete these channels. Subsequently, we sort the channels based on the L1 norm score. After sorting, delete the channels with lower scores to achieve the desired pruning ratio. In this example, we choose a 30% pruning ratio [9]. After deleting the channel, fine tune or retrain the pruned model to recover the performance degradation caused by pruning.

Through pruning operations, this article achieves model compression and improves computational efficiency. Reduce the storage requirements of the model. This makes the deployment of the model on devices with limited memory more feasible and enables faster loading and execution.

3. Experiments:

3.1. Dataset introduction + Experimental details

The data set in the real scene generally has Confounding such as lighting, occlusion, low pixels, etc. It is relatively difficult to extract features. The CK+ data set used in this paper is the data set in this real scene. The CK+ dataset is a dynamic video dataset [10-11]. At the same time, it is a competition level dataset. Compared with the dataset recorded in the laboratory, it has added some Confounding, among which the Confounding mainly include occlusion, too low pixels, background changes, etc. Because of the existence of these Confounding, it is more realistic.

The CK+ (Cohn Kanade) dataset is a commonly used facial expression recognition dataset for training and evaluating facial expression recognition algorithms. This dataset was jointly created by researchers from Stanford University and the University of California, San Diego. The CK+ dataset contains facial expression sequences performed by a group of volunteers in a laboratory environment.

DOI: 10.54254/2755-2721/32/20230201

Each volunteer will exhibit a series of facial expressions, including seven basic expressions of happiness, sadness, anger, surprise, disgust, and fear, as well as a neutral expression. Each facial expression sequence consists of a series of consecutive facial images.

The CK+ dataset comprises 327 video sequences obtained from 123 subjects. The length of each sequence ranges from 10 to 60 frames. The dataset provides grayscale images, and each image has a resolution of either 640×480 or 640×490 pixels. It is a dynamic video dataset that forms a video frame dataset by performing frame truncation operations on the video, and each video frame has been pre labelled with emoticons.

3.2. Experimental details

In this article, it is required that the CNN network output multiple continuous facial features simultaneously. Therefore, the model inputs n facial images each time, and each facial image can share the CNN network weight for feature extraction. After setting the final n value to 8, 8 consecutive facial images are passed in at once. VGG16 adopts the VGG-16-FACE model with pre training weight, and the initial Learning rate is 0.01. The optimization algorithm is the random Gradient descent, the momentum parameter value is set to 0.9, and the training times are 200. For input, the image set included in the dataset was used, with each image size of 224 × 224. After the network training is completed, the test set is fed into the model and the test results are obtained. The experimental code was completed using PyCharm on the Ubuntu 20.04 system, with a virtual machine memory of 4GB and a host CPU model of R9-7945HX.

The F1 score, also referred to as the F-Measure, is a common metric employed in statistics and machine learning to assess the classification model's accuracy. It is calculated as the harmonic mean of Precision and Recall. By incorporating both Precision and Recall, the F1 score provides a comprehensive evaluation of the model's accuracy and ability to recall positive instances. This score is valuable in measuring the model's performance stability and generalization capabilities when making predictions.

Formula for calculating F1 score:

$$F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP+EP}} \tag{2}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

Among these, True Positive (TP) denotes the number of accurate predictions made by the model. False Positive (FP) refers to the number of instances that were incorrectly predicted as belonging to a particular class. On the other hand, False Negative (FN) represents the number of instances that should have been predicted correctly but were mistakenly predicted as negative.

The calculation formula for accuracy:

$$\text{accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{4}$$

Among them, TN (True Negative) is the number that predicts other classes as correct.

3.3. Feature visualization

3.3.1. Confusion matrix. The Confusion matrix presents the relationship between the prediction outcomes of a classification model on the test dataset and the actual labels, represented in matrix form. It provides a visual representation of how well the model performs in classifying the data. Through the Confusion matrix, we can clearly understand the prediction of the classification model for each category, which is helpful to identify which categories the classification model predicts more accurately (Figure 4).

DOI: 10.54254/2755-2721/32/20230201

	Anger	Contempt	Fear	Happiness	Neutral	Sadness	Surprise
Anger	18	0	1	0	0	0	0
Contempt	0	10	0	0	1	0	0
Fear	1	0	12	2	0	0	0
Happiness	0	0	0	27	0	0	0
Neutral	0	1	2	0	12	2	1
Sadness	1	0	0	1	0	23	1
Surprise	0	0	0	0	0	0	23

Figure 4. CK+ dataset traditional VGG16 test results.

	Anger	Contempt	Fear	Happiness	Neutral	Sadness	Surprise
Anger	18	0	1	0	0	0	0
Contempt	0	10	0	0	0	0	0
Fear	1	0	12	1	0	0	0
Happiness	0	0	0	28	0	0	0
Neutral	0	1	2	0	13	1	0
Sadness	1	0	0	1	0	24	1
Surprise	0	0	0	0	0	0	24

Figure 5. CK+ dataset traditional VGG16+dilated Convolutional test results.

After adding hollow convolution, some test videos that were originally incorrectly recognized are now recognized correctly, and the original correct recognition remains correct. It can be seen that hollow convolution has an improvement in the accuracy of the original model's recognition (Figure 5).

	Anger	Contempt	Fear	Happiness	Neutral	Sadness	Surprise
Anger	18	0	1	0	0	0	0
Contempt	0	10	0	0	1	0	0
Fear	1	0	12	1	0	0	0
Happiness	0	0	0	28	0	0	0
Neutral	0	1	2	0	12	2	0
Sadness	1	0	0	1	0	23	1
Surprise	0	0	0	0	0	0	24

Figure 6. CK+ dataset traditional VGG16+dilated convolutional + pruning test results.

After channel pruning, as the number of parameters decreases, the recognition accuracy also decreases slightly, but the overall recognition rate remains at a high level (Figure 6).

3.4. Performance comparison

The table presents the test outcomes obtained from the CK+ dataset. When applied to video data, the standard VGG16 model attains an accuracy of 90.58%, precision of 83.33%, recall rate of 88.89%, and an F1 score of 0.86 in facial expression recognition. These findings suggest that the traditional VGG16 model has accomplished favorable results in recognizing facial expressions (Table1).

DOI: 10.54254/2755-2721/32/20230201

Table 1. The test outcomes obtained from the CK+ dataset.

	Accuracy	Precision	Recall	F1
VGG16	90.58%	83.33%	88.89%	0.86
VGG16+Dilated Convolutional	95.65%	93.33%	93.33%	0.93
VGG16+Dilated Convolutional+ Pruning	94.20%	89.36%	93.33%	0.91

To enhance the accuracy and performance of facial expression recognition even further, this research introduces dilated convolution and parameter pruning techniques to the VGG16 model. Through experimental testing, notable advancements in performance were observed. Upon the incorporation of dilated convolution and subsequent testing, the revised VGG16 model demonstrated an enhanced accuracy of 95.65% in facial expression recognition on the CK+ dataset. The model achieved an accuracy rate of 93.33%, a recall rate of 93.33%, and an F1 score of 0.93. These results exhibit the positive impact of integrating dilated convolution on improving the model's ability to recognize expressions. This is due to the fact that the hole convolution expands the Receptive field of the network, enabling it to better capture the details of facial expressions.

Subsequently, by applying parameter pruning techniques, this article successfully reduced the parameter count of the VGG16 model. As a result, the model's operations have become more efficient, leading to favourable performance on the CK+ dataset. Following the application of parameter pruning, the VGG16 model achieved an accuracy of 94.20% in facial expression recognition. The accuracy rate stood at 89.36%, the recall rate at 93.33%, and the F1 score at 0.91. Although the accuracy and F1 score have decreased due to the reduction of some parameters by pruning, the magnitude of the decrease is within an acceptable range, and the accuracy is still higher than traditional VGG16.

Table 2. Parameter quantity of VGG16.

Name	Parameter quantity before pruning	Parameter quantity after pruning
Convolutional Layer1	1,792	1,792
Convolutional Layer2	73,728	51,710
Convolutional Layer3	1,048,576	734,003
Convolutional Layer4	1,179,648	825,754
Convolutional Layer5	2,359,296	1,652,349
Convolutional Layer6	2,359,296	1,652,349
Convolutional Layer7	2,359,296	1,652,349
Convolutional Layer8	4,718,592	3,892,008
Convolutional Layer9	4,718,592	3,892,008
Convolutional Layer10	4,718,592	3,892,008
Convolutional Layer11	4,718,592	3,892,008
Convolutional Layer12	4,718,592	3,892,008
Convolutional Layer13	4,718,592	3,892,008
Fully connected layer1	102,760,448	62,949,474
Fully connected layer2	16,777,216	16,777,216
Fully connected layer3	4,194,304	4,194,304
Output layer	4,096	4,096
total	140,630,976	111,922,730

Through pruning, we observed a significant reduction in the parameter count of the VGG16 model. After 30% channel pruning, the parameter quantity of the model is approximately 111.9M, which is nearly 20% less than the original model's 140.6M. This accomplishment is obtained by applying pruning algorithms to remove a specific portion of channels, effectively reducing the number of parameters within the Convolutional Layer. Channel pruning technology plays a crucial role in decreasing model parameters, subsequently reducing the computational complexity of the model. Through the elimination

DOI: 10.54254/2755-2721/32/20230201

of certain channels, the computational complexity of multiplication and addition operations within the Convolutional Layer is reduced. This outcome translates to enhanced inference speed and higher computational efficiency (Table 2).

Simultaneously, reducing the parameter count has the added benefit of decreasing the storage space required by the model. Consequently, this simplifies the deployment and utilization of the model in resource-constrained environments.

4. Conclusion

This paper presents an optimized approach to enhance the accuracy of video facial expression recognition by building upon the traditional VGG-16 network framework. The optimization strategy entails two key techniques: hollow convolution (or dilated convolution) and parameter pruning. To begin with, the introduction of hollow convolution allows for the capturing of spatial information from facial features at various scales, without incurring additional computational costs. By adjusting the expansion rate of the convolution kernel, this method effectively expands the receptive field and enriches the model's ability to perceive intricate facial details. Furthermore, parameter pruning is employed to reduce the complexity of the model while maintaining its performance. Through precise analysis of network weights, unessential connections are identified and eliminated, resulting in a more compact and efficient model. By incorporating both hollow convolution and parameter pruning techniques, the proposed approach aims to enhance the accuracy of video facial expression recognition while optimizing computational costs and model efficiency. This article conducted extensive experiments on the CK+ dataset to evaluate the effectiveness of the proposed method. The results indicate that our method significantly improves accuracy compared to the benchmark VGG16 model without cavity convolution and pruning. The introduction of hollow convolution enables the model to capture fine-grained facial features, enhancing the ability to distinguish between individuals. In addition, parameter pruning effectively reduces the number of parameters in the network, improves computational efficiency without affecting performance.

This study contributes to the advancement of facial recognition technology by addressing the challenges of capturing fine-grained facial features and reducing model complexity. This method has broad application potential in fields such as biometric recognition systems, monitoring technology, and facial based authentication systems.

[1] Aderhold J, Davydov V Yu, Fedler F, Klausing H, Mistele D, Rotter T, Semchinova O, Stemmer J and Graul J 2001 J. Cryst. Growth 222 701

References

[1] Peng Z, Weiwei K, and Jinbao T. Face expression recognition based on multi-scale feature attention mechanism. *Comput. Eng. Appl.*, 2022,**58 (01)**: 182-189

[2] Xuan H. Research on face recognition methods in unrestricted scenes. *Sichuan Univ.*, 2021.

[3] Jinxiang L. Image recognition and target detection based on VGGNet. *Yanshan Univ.*, 2021.

[4] Jie Z. Research on the Application of Deep Learning Based Face Recognition Algorithms in Video Surveillance. *Electric. Des. Eng.*, 2023,**31 (13)**: 182-186.

[5] Jiehao W. Research on 3D facial expression recognition method based on deep learning. *Xi'an Univ. Tech.*, 2023.

[6] Qianqian L, Weixing W and Qin Y, etc. Research on audio-visual multimodal emotion recognition based on deep learning. *Comp. Dig. Eng.*, 2023,**51 (03)**: 695-699

[7] Shuyu D. Occlusive facial expression recognition based on deep learning. *Shandong Univ.*, 2022.

[8] Junling X. Research on Natural Scene Facial Expression Recognition Method Based on Deep Learning. *Chongqing Univ. Posts Telecomm.*, 2022.

[9] Huihua X, Ming L and Yan W, etc. Facial expression recognition based on DE Gabor features. *J. Nanchang Hangkong Univ. (Natural Science Edition)*, 2021,**35 (02)**: 82-91+124

[10] Ting Z, Research on facial expression recognition based on deep learning. *South China Univ. Techn.*, 2022.

DOI: 10.54254/2755-2721/32/20230201

[11] Dongdong Q, Lile H, Lin H. Improved Lightweight Face Recognition Algorithm. *J. Intel. Sys.*, 2023,**18 (03)**: 544-551