Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/36/20230426

Research on image recognition in human-computer interaction based on convolutional neural networks

Kairan Yang

School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, China

yangk@stmail.ujs.edu.cn

Abstract. Image recognition is an important research direction in human-computer interaction, which has broad development prospects, aiming at enabling computers to understand and interpret image content. Early image recognition methods mainly rely on hand-designed feature extraction algorithms for image analysis and classification. However, this method has significant limitations, which may not provide accurate recognition results for complex images and is hard to adapt to different application scenarios. With the advancement of deep learning and increased processing power in recent years, Convolutional neural network (CNN) -based image recognition methods have achieved remarkable achievements in human-computer interaction. Therefore, based on the image recognition of CNN in human-computer interaction, this paper first studies the CNN model in depth then describes the application of CNN in human-computer interaction, then enumerates different design methods for comparative analysis, and finally summarizes the advantages and disadvantages of CNN in application and proposes improvements to existing problems. The research shows that image recognition based on CNN is better than the traditional network model and has higher accuracy, but it still has some disadvantages. To address these issues and produce more effective and precise image understanding and interaction, it is required to research and enhance the model's structure and algorithm. The appearance and development of CNN have greatly promoted the development of image recognition technology, which has been widely used in human-computer interaction and has made great breakthroughs.

Keywords: image recognition, convolutional neural networks, human-computer interaction.

1. Introduction

With the development of human-computer interaction, image recognition based on CNN has using widely. For example, it can verify or identify people's identities in contactless palm print recognition [1]. The application in behavior recognition can detect whether the driver is distracted while driving the car, thus reducing traffic accidents [2]. It can be used in the medical field to rapidly diagnose diseases, such as the diagnosis of breast cancer [3] and the analysis of clinical characteristics of diabetes [4]. In addition, it can also apply to map navigation, virtual reality, and other aspects.

The neural network model is essential for image recognition. One of the most basic neural network models, the single-layer perceptron can only resolve linear issues. Then, Liu et al. introduced a deep multi-layer perceptron, and Xie et al. suggested a multi-layer perceptron using a fractional gradient descent method, both of which can handle issues that single-layer perceptrons cannot solve [5, 6]. Both single-layer and multi-layer perceptrons require manual feature extraction, which leads to the poor performance of traditional image recognition models. CNN has had remarkable success as deep learning has progressed [7]. It can automatically extract main features from image data by learning feature expression and can process large-scale complex image data. Battleday et al. also mentioned that CNN could process complex images, greatly expanding the scope of traditional cognitive models, and increasingly explored as a structural model of the human visual system, which can also provide a basis for higher-level models [8]. Now there are also many new CNN models. A differential convolutional neural network was created by Sargül et al. [9], while a recurrent convolutional neural network with gating was created by Wang et al. [10], all of which can improve the performance of CNN. It can see from the literature that CNN has many advantages, but it needs to be improved in terms of model parameters and complexity [11]. In addition, it is also challenging for image recognition in complex scenes, such as occlusion and angle change.

This paper aims to understand the role of image recognition based on CNN in human-computer interaction. For this purpose, then this paper narrates the CNN model, and the application of image recognition based on CNN in the field of human-computer interaction will be described in detail through literature and then list two different design methods of CNN: the STN-CNN composite model and the modular CNN model. Finally, based on the application of CNN in human-computer interaction, the advantages and disadvantages are summarized, and the shortcomings are improved.

2. Research works

The CNN model's structure and principle are first introduced in this section. After having a general understanding of it, it will explain the application of CNN in human-computer interaction and its different design methods.

2.1. CNN model

The convolutional layer, pooling layer, fully connected layer, and SoftMax layer are the components of CNN. The structure shows in Figure 1. The input layer receives images, the convolutional layer extracts features, and the pooling layer reduces dimensions, and then feature images enter the fully connected layer and proceed with data integration and classification, finally obtaining classification results in the SoftMax layer.

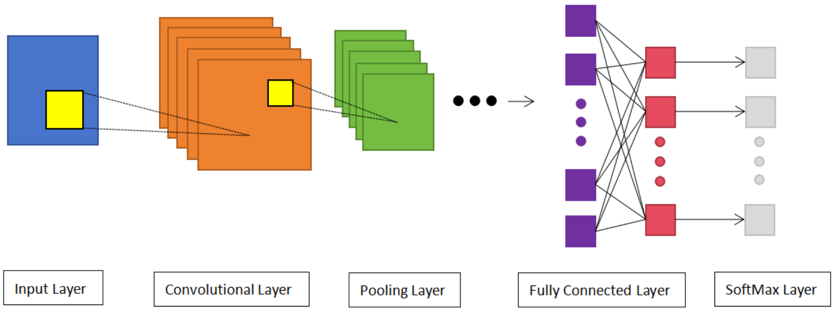

Figure 1. CNN model structure.

2.1.1. Convolution layer. The main process of extracting image features occurs in the convolution layer. Images are revolutionized by the convolution kernel. Specific operations are as follows: Assume that a 3x3 image picture and a 2x2 convolution kernel carry out a convolution operation with step size 1. The convolution kernel covers the upper left position of the image, and the values on the corresponding pixel are multiplied and then summed. After the calculation, the convolution kernel will shift one pixel to the right and repeat the above operations, as shown in Figure 2.

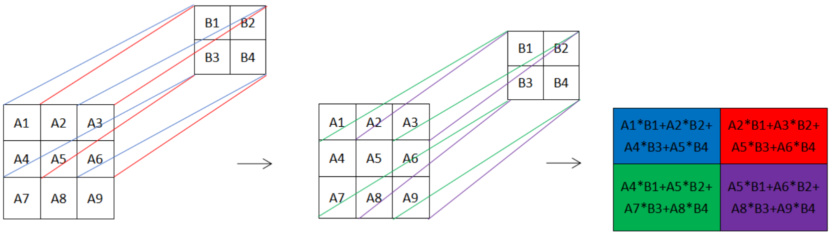

Figure 2. Convolution operation procedure.

2.1.2. Pooling layer-ReLU activation function. To decrease the computation amount, the dimensionality of the extracted features needs to be reduced after the convolution operation. The common methods include max pooling, average pooling, etc. Because max pooling has the advantage of preserving the original image features to the greatest extent, this paper uses max pooling. As shown in Figure 3, assume that there is a 4x4 feature image and a 2x2 filter for max pooling with a step size of 2. The filter covers the upper left position on the feature map and takes the maximum value. Then the filter will shift two pixels to the right and repeat the above operations. After that, pass into the ReLU activation function, as shown in Figure 4. It will change the negative value in the obtained feature image to 0, and the positive value will remain unchanged. The purpose of this activation function is to specify the image features within a fixed range and increase the nonlinear features of the model.

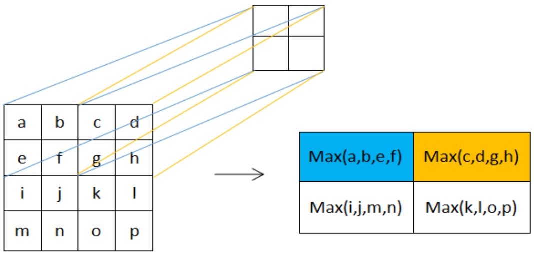

Figure 3. Max pooling operation procedure.

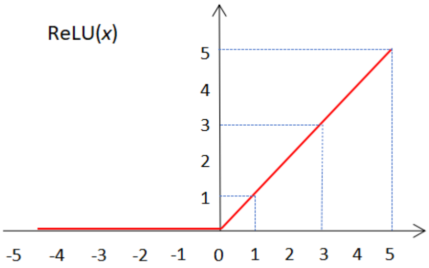

Figure 4. ReLU activation function.

2.1.3. Fully connected layer, SoftMax layer. After maximum pooling, the data is flattened into a column of data bars, as shown in Figure 5, which enter the fully connected layer as in Figure 6 x1, x2, x3 ...,xn, calculates the corresponding y value by the formula (1) (the weight refers to w and the bias term refers to b),

$$y_i = \sum_{n=1}^{N} (x_n \times w_{nj}) + b \tag{1}$$

and then passes it into the SoftMax layer, which converts the values into probabilities by formula (2),

$$\frac{e^{y_j}}{\sum_{k=1}^{K} e^{y_k}} \tag{2}$$

while ensuring that their sum is 1, and selects the category corresponding to the maximum value as this result.

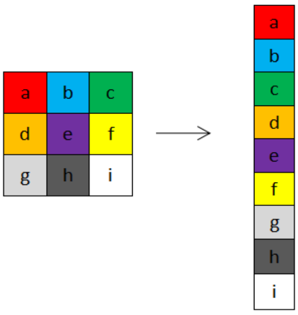

Figure 5. Flattening.

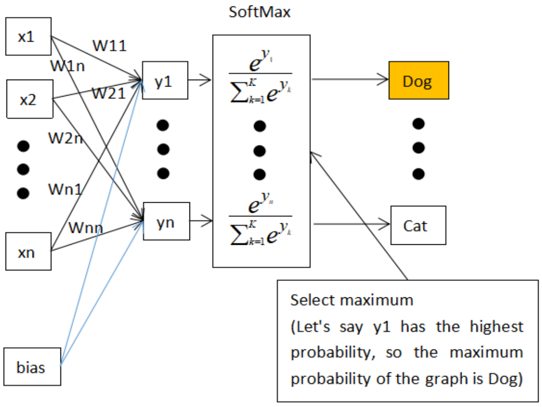
SoftMax
x1
x2
xn
bias
W11
W1n
W21
W2n
Wn1
Wnn
y1
yn
$\frac{e^{y_1}}{\sum_{k=1}^{K} e^{y_k}}$
$\frac{e^{y_n}}{\sum_{k=1}^{K} e^{y_k}}$
Dog
Cat
Select maximum
(Let's say y1 has the highest probability, so the maximum probability of the graph is Dog)

Figure 6. SoftMax layer operation.

2.2. *Application of CNN model in human-computer interaction*

The CNN model has made significant achievements in human-computer interaction and has a broad spectrum of applications due to its significant advantages. The following introduces the application of CNN in intelligent driving.

2.2.1. Traffic sign recognition. In literature [12], CNN is used to recognize traffic signs. When a car recognizes a traffic sign, it can make good action feedback. This literature found a total of 43 types of traffic sign images, as shown in Figure 7, and unified these road sign images into the same size.

Figure 7. Types of traffic signs [12].

Then, divide the image into three color components of R, G, and B, count each pixel's value in the image and perform the histogram. The X-axis of the histogram is the pixel value size, namely brightness, and the Y-axis is the number of pixel values, as shown in Figure 8. Then the image is histogram equalization. Histogram equalization is a method of gray-scale transformation. It commonly uses for image processing. The pixel distribution of the image is adjusted using the image histogram, stretch the image contrast and balance the histogram, consequently widening the difference between the background and the target image. In processing traffic sign images, histogram equalization can better obtain the distribution of road signs in images and make road sign images clearer and of higher quality. After histogram equalization, it becomes Figure 9.

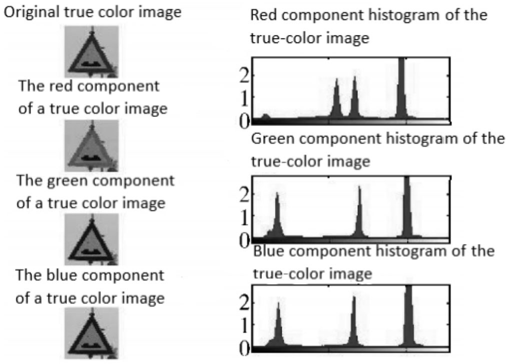

Figure 8. Color image histogram [12].

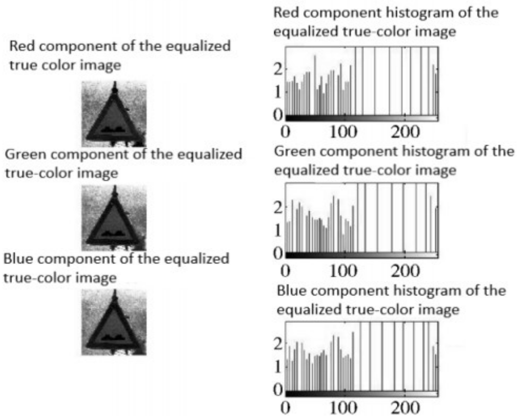

Figure 9. Histogram after equalization [12].

After image preprocessing, establish a LeNet5 structured CNN model, which is a two-layer convolutional neural network, that is, containing two hidden layers. The first hidden layer has a convolutional layer that is 28x28x6, and a subsampling layer that is 14x14x6. The convolutional layer in the second hidden layer is 10x10x16, while the subsampling layer is 5x5x16. The layer that is fully connected has a size of 120x1, and the structure shows in Figure 10.

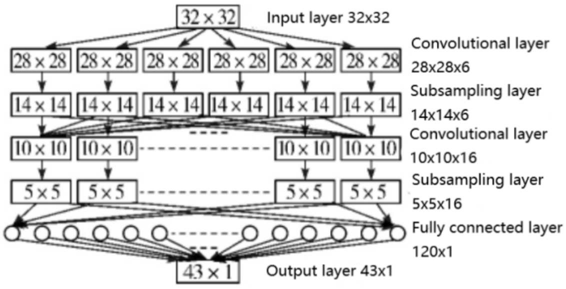
32 × 32
Input layer 32x32
28 × 28
Convolutional layer 28x28x6
14 × 14
Subsampling layer 14x14x6
10 × 10
Convolutional layer 10x10x16
5 × 5
Subsampling layer 5x5x16
Fully connected layer 120x1
43 × 1
Output layer 43x1

Figure 10. CNN model of LeNet5 structure [12].

After that, carry out feature extraction, dimensionality reduction, and other operations for image recognition. However, experiments showed that the model's accuracy was not high, So the model has been improved. Reduce the convolution kernel to maintain the original features of the image as much as possible and increase the batch size and training times. The final experiment shows that the accuracy of the improved model reaches 93.2%.

2.2.2. Fatigue-driving recognition. In literature [13], CNN is used to identify the fatigue expression, determine whether the crew is tired and distracted when flying the aircraft, and give fatigue warning tips to reduce the accident casualty rate and other losses. This paper collects the fatigue state, normal state, and speaking state expressions by the LFW face data set. Clip the face area of the image and expand the data set by three data enhancement methods, mirroring, random clipping, and noise. The picture with pre-processing and data augmentation shows in Figure 11. The improved CNN model used in this paper shows in Table 1, in which add a 1x1 convolution kernel, which can not only extend the model's depth but also does not increase the additional calculation amount. Then use the above expression data for training. Afterward, facial expressions can be captured in real-time using an image acquisition device. These facial expressions are then input into a trained model for several convolutional and pooling operations, followed by calculations in fully connected layers. Finally, the results are passed through an output layer to recognize the outcome. The model presents the expression confusion condition, but the mean accuracy of fatigue, normal, and vocal expression is 88.3%.

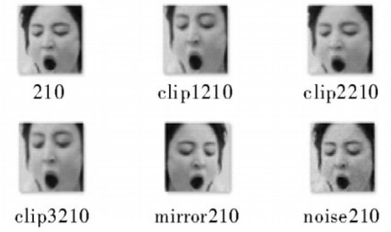

Figure 11. Fatigue expression image [13].

Table 1. Improved CNN model structure [13].

Kind	Kernel	Output	Drop
Input layer		48x48x1	0
Conv layer 1	1x1	48x48x32	0
Conv layer 2	3x3	48x48x32	0
Pool layer 1	2x2	24x24x32	0
Conv layer 3	3x3	24x24x32	0
Pool layer 2	2x2	12x12x32	0
Conv layer 4	5x5	12x12x64	0
Pool layer 3	2x2	6x6x64	0
FC layer 1		1x1x2048	50%
FC layer 2		1x1x1024	50%
Output layer		1x1x3	0

2.2.3. Sidewalk recognition. In literature [14], CNN is used to identify sidewalks, and the algorithm in this literature can be applied to blind-guide systems or intelligent driving technology. The algorithm flow chart shows in Figure 12.

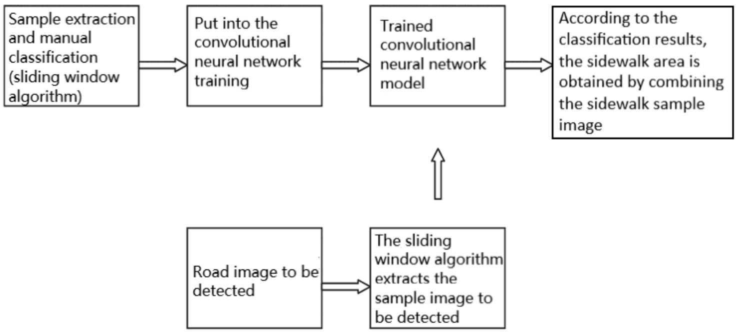
Sample extraction and manual classification (sliding window algorithm)
Put into the convolutional neural network training
Trained convolutional neural network model
According to the classification results, the sidewalk area is obtained by combining the sidewalk sample image
Road image to be detected
The sliding window algorithm extracts the sample image to be detected

Figure 12. Flow chart of sidewalk recognition algorithm [14].

Firstly, set the size and moving length of the sliding window and calculate the image by the sliding window algorithm to obtain multiple overlapping rectangular blocks, as shown in Figures 13 and 14. Then, the sidewalk and road are manually classified, then trained in the CNN model. Next, extract the sample images to be tested by the slider algorithm and send them to the CNN model formed for recognition. CNN's score indicates that to obtain the sidewalk area, as shown in Figure 15. At this time, the obtained area is a lot of rectangular boxes. Finally, merge the identified adjacent sidewalk areas, and take the largest area as the final recognition result.

Figure 13. Road sample [14].

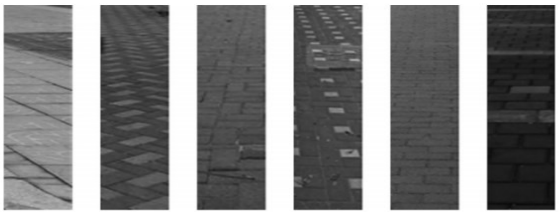

Figure 14. Sidewalk sample [14].

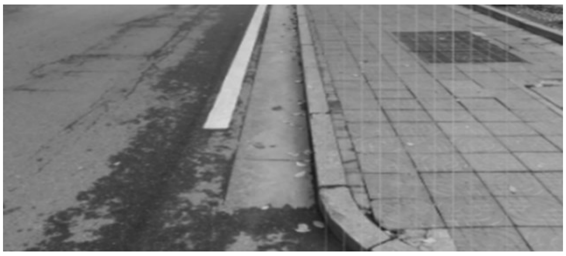

Figure 15. Slide window diagram of the sidewalk area [14].

The experiment shows that this algorithm suits all kinds of sidewalk recognition. As long as increasing the types of sidewalk samples fed into the CNN can identify more sidewalks. Due to certain limitations in the sliding window design, if the road is deviated in the image sample to be tested and does not extend from the bottom to the top, in that way, the detected sidewalk area may be inaccurate.

2.3. Different design methods

CNN is an important breakthrough in human-computer interaction. But it also has some shortcomings. Many researchers hope to make up for and solve these shortcomings, so they want to design more perfect models based on CNN. This article also lists two different design methods.

2.3.1. Modular CNN model. When an image is rotated and scaled, features will be lost in single-module CNN processing, and recognition accuracy will decline. Therefore, a modular CNN model is proposed in the literature [15]. It consists of multiple modules. These modules consist of the convolutional layer, pooling layer, and ReLU activation function, which are stacked in parallel. Pass normal image, rotating image, and scaled image into several parallel main modules for pre-processing and divided into several parts. After each main module, connect several parallel sub-modules, and these pre-processed images are transferred into sub-modules respectively for feature extraction. Finally, the features extracted from all sub-modules are aggregated and sent to the SoftMax layer, as shown in Figure 16. Design a gate unit to improve the model structure as not all sub-modules can extract a large number of characteristics useful for classification, as shown in Figure 17. It comprises the Sigmoid activation function and the linear layer. The features extracted by submodules will enter this structure and compare the value to the cutoff. Whenever the feature value falls below the threshold value, it means that this feature is useless. In this way, those useless modules can be deleted, which can minimize the model's parameters and decide how many modules to use. Comparing it to the standard CNN, this model has fewer model parameters, higher feature extraction efficiency, and faster model convergence. However, this approach also has a significant flaw in that as the number of modules increases, so will the amount of processing.

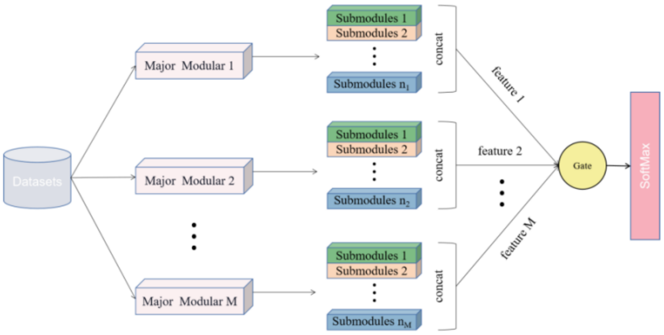

Figure 16. The structure of the modular CNN model [15].

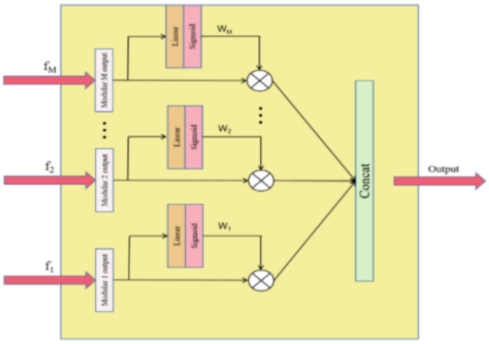

Figure 17. Gate unit structure [15].

2.3.2. STN-CNN composite model. When an image is small, with few main features, or when the object in the image is tilted, ordinary CNN may not be able to identify it accurately. Therefore, literature [16] proposes a new CNN object detection model, which improves the algorithm of the CNN model and combines it with the STN model. STN network model is an attention mechanism used to enhance the CNN model. Its basic idea is to swap out the spatial information between the input picture and another space before compressing the feature data using maximum or average pooling. Retain the important details in this way to cut down on calculations and increase recognition accuracy. The picture may be rotated and scaled using the space transformer module in the STN network architecture, and via these changes, significant local information about the images can be extracted, as shown in Figure 18.

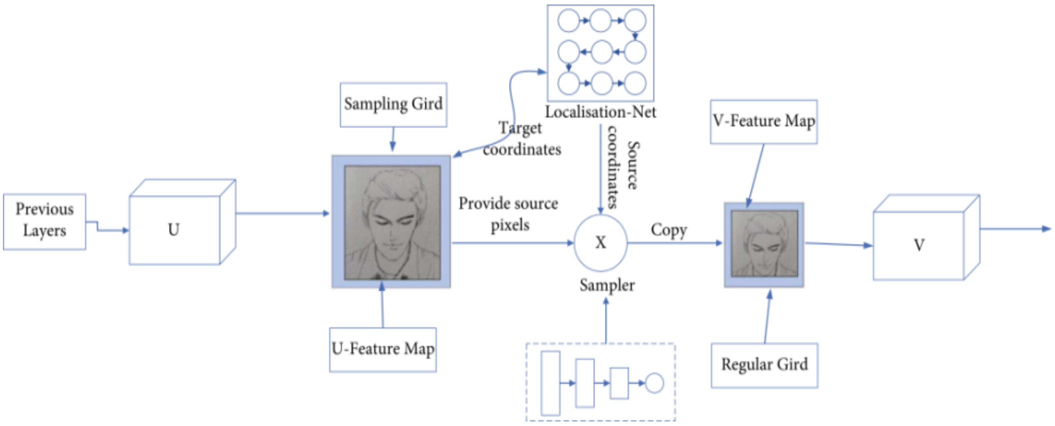

Figure 18. STN model structure [16].

Images are processed through the STN model and entered into the improved CNN model for operation, as shown in Figure 19. The convolution layer and pooling layer produce feature maps, which are combined in the Max-Avg Features to obtain richer feature data. The classification result is then converted into probability in the Softmax layer and then sent to GAP together with the feature map to gain a fixed-length feature vector, attempting to cut back on the parameters. Then, using the Loss function, determine where the categorization result and the actual diverge, calculate the loss value, and modify the network model parameters according to this value for optimization. Finally, evaluate train accuracy and test accuracy. This model can increase generalization capacity and recognition rate while lowering interference brought on by changing lighting and viewing angles. Because it is a complex model, the model will bring high computing costs.

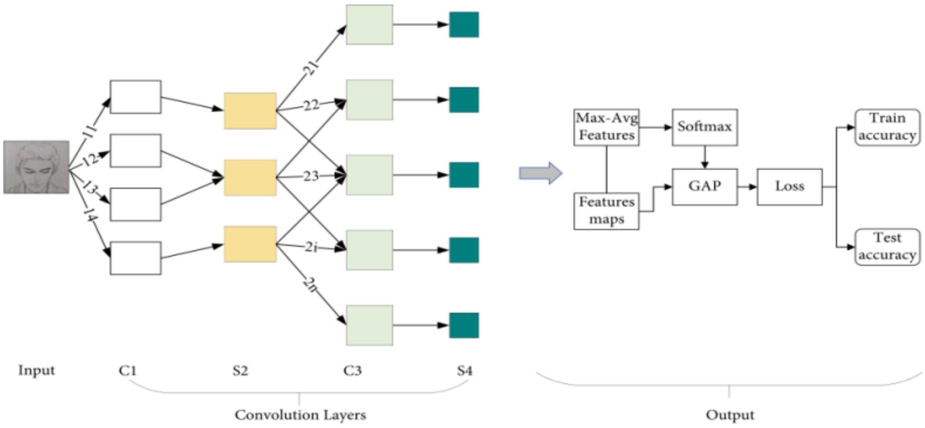

Figure 19. Improved CNN model structure [16].

3. Discussion

Through the in-depth study of the CNN model, we have a good understanding of its operation process and principle. This section will discuss the advantages and disadvantages of CNN in human-computer interaction based on the second section and give improvement methods based on these disadvantages.

3.1. Advantages

CNN has many advantages in human-computer interaction. It can learn high-level abstract features, adapt to complex tasks and data, and deal with high-dimensional data without pressure. In the application introduced in the second section, CNN also has many advantages. It has a high recognition accuracy and thus can reduce the incidence of accidents. In traffic sign recognition, the model has a low error rate and strong noise resistance performance. In fatigue driving recognition, the model has the advantages of strong feature distinguishing ability, simple process, and good real-time performance. In sidewalk recognition, the model has the advantages of high recognition efficiency, good generalization ability, and better robustness under different backgrounds, illumination, or occluder.

3.2. Disadvantages

CNN also has drawbacks in these applications, which require lots of data for training and testing. In fatigue-driving recognition, there will be confusion in facial expression recognition, resulting in inaccurate recognition or errors. In sidewalk recognition, due to the limitations of sliding window design, only recognize the sidewalk image extending from the bottom to the forward. If the road in this image is deviated and inclined, the recognition rate is not high.

3.3. Improvement

A large amount of training data can be obtained by expanding the data set through data-enhancing processing methods such as rotation and scaling. Use the attention mechanism to zoom in on important or desired areas to highlight features. Through the method of image enhancement, reduce the image noise to make it clear, which can reduce the computational amount of subsequent model recognition and improve accuracy.

4. Conclusion

Through the study of the CNN model, we can understand that it has superior performance and powerful functions, so it can process images of many complex scenes, solve the limitations and problems of traditional neural network models, and image recognition based on the CNN model is widely used in human-computer interaction. The primary goal of this article is to quickly present CNN's guiding concept, each component's structure, and the use of human-computer interaction, such as traffic sign recognition, fatigue-driving recognition, and sidewalk recognition. After that, two improved CNN models are described: the modular CNN model and STN-CNN composite model. Both models solve many problems that ordinary CNNs cannot handle. These applications and improved models pave the way for further study of CNN and provide some references for other researchers. Although the CNN model has many advantages and has achieved remarkable results in human-computer interaction, it still has some defects, which will be the focus of future research. In the future, through continuous research and exploration of a more efficient and accurate CNN model and improve the model according to shortcomings, to further improve the image recognition performance of human-computer interaction and promote the development of the human-computer interaction field.

References

[1] Stoimchev M, Ivanovska M and Štruc V 2021 Learning to combine local and global image information for contactless palmprint recognition *Sensors (Basel)* 22 73

[2] Yan X, He J, Wu G, Zhang C and Wang C 2022 A proactive recognition system for detecting commercial vehicle driver's distracted behavior *Sensors (Basel)* 22 2373

[3] Kwak D, Choi J and Lee S 2023 Rethinking breast cancer diagnosis through deep learning based image recognition *Sensors (Basel)* 23 2307

[4] Wang R, Li P and Yang Z 2022 Analysis and recognition of clinical features of diabetes based on convolutional neural network *Comput Math Methods Med* 2022 7902786

[5] Liu R, Li Y, Tao L, Liang D and Zheng HT 2022 Are we ready for a new paradigm shift? A survey on visual deep MLP *Patterns (N Y)* 3 100520

[6] Xie X, Pu YF and Wang J 2023 A fractional gradient descent algorithm robust to the initial weights of multilayer perceptron *Neural Netw* 158 154-170

[7] Serre T 2019 Deep learning: the good, the bad, and the ugly *Annu Rev Vis Sci* 5 399-426

[8] Battleday RM, Peterson JC and Griffiths TL 2021 From convolutional neural networks to models of higher-level cognition (and back again) *Ann N Y Acad Sci* 1505 55-78

[9] Sarıgül M, Ozyildirim BM and Avci M 2019 Differential convolutional neural network *Neural Netw* 116 279-287

[10] Wang J, Hu X 2022 Convolutional neural networks with gated recurrent connections *IEEE Trans Pattern Anal Mach Intell* 44 3421-3435

[11] Li Z, Liu F, Yang W, Peng S and Zhou J 2022 A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE Trans Neural Netw LearnSyst* 3 6999-7019

[12] Chen B, Lin N 2018 Traffic sign recognition based on convolutional neural network *Computer And Modernization* 30 103-107,113

[13] Wang J, Li Y, Cao Y and lü S 2022 Fatigue driving recognition based on convolutional neural network *Aeronautical Computing Technology* 52 60-63,68

[14] Hu C, Xiong P and Zhou X 2017 Research on sidewalk recognition algorithm based on convolutional neural network *Information and Computers (Theory)* 26 53-55

[15] Wu W, Pan Y 2022 Adaptive modular convolutional neural network for image recognition *Sensors (Basel)* 22 5488

[16] Wang P, Qiao J and Liu N 2022 An improved convolutional neural network-based scene image recognition method *Comput Intell Neurosci* 2022 3464984.