Image recognition algorithm based on spatial transform convolutional neural network

XING Zhen-zhen
Department of Computer Engineering, Taiyuan Institute of Technology, Taiyuan 030008, China
tg667788@xzcstudio.com

Abstract: Aiming at the problem of classification and recognition of noisy handwritten digits, a connection method is proposed to add a spatial transformation network to a convolutional neural network. The spatial transformation network can not only obtain the output results, but also understand the parts of the input data that have the greatest influence on the results and perform feature extraction, which can improve the performance of the model and enhance the interpretability of the model. The experiment was performed on the Cluttered MNIST dataset, and the effects of the conventional convolutional neural network and the improved neural network with spatial transformation network were compared. The experimental results show that the prediction accuracy of convolutional neural networks based on spatial transformation can reach 97.65%, and the results obtained by this method are better than those of conventional convolutional neural networks.

Chinese library classification number: TP391

1. Introduction
Convolutional Neural Networks (CNN) have been widely used in the field of image recognition since they were first proposed\(^1\). Convolutional neural network is multi-layer with an important algorithm in deep learning. The main structure includes convolutional layer, pooling layer, fully connected layer, etc.\(^2\). Convolutional neural network can train and predict images. The training process is mainly to learn the parameters of the convolution kernel on the convolutional and the weight coefficient between the layers. The prediction is calculated through the input image and network correlation coefficients, and the category label is obtained\(^3\).

This paper discusses the recognition method based on spatial transformation convolutional neural network in handwritten digital images with disturbing information. The traditional convolutional neural network has a high accuracy rate in the recognition of handwritten digital MNIST dataset.\(^4\) In this experiment, the Cluttered MNIST dataset is used, and some shape interference such as strokes are added. The classifier needs to find the position of the number in the image and exclude the interference of the newly added shape. The Spatial Transformer network (STN) can extract the digits to be identified to understand which areas of the input image have the greatest influence \(^5\), thus adding the spatial transform network to improve the accuracy.

2. Algorithm principle

2.1 Principle of convolutional neural network
A typical convolutional neural network mainly includes an input layer, a convolutional layer, a pooling
layer, a fully connected layer, and an output layer\[^6\]. As shown in figure 1.

![Figure. 1 Structure diagram of convolutional neural network](image)

The input layer is usually the original image, which is generally represented by a two-dimensional matrix. According to the needs of the experiment, the convolutional and the pooling layer usually appear multiple times in combination. The fully connected layer can perform weighted summation on the input and output through the activation function\[^7\]. The fully connected layer is a deep neural network structure. The experiment uses convolutional neural network for handwritten digit recognition, which mainly includes three parts.

(1) Data sets and preprocessing

The experiment uses the Cluttered MNIST data set. Each image is 60*60 pixels composed of pictures and labels. The data set consists of the training data set (picture + label) and the test data set (picture + label). Train with the training data set first, and test with the test data set. The data set has 60,000 handwritten digital samples, of which 50,000 are training and 10,000 are test. Some samples of the Cluttered MNIST data set are shown in figure 2.

![Figure. 2 Cluttered MNIST partial sample instance](image)

Before training the model, the data set needs to be preprocessed. The first is to convert the image information into a one-dimensional array, and then normalize, so that the value of each pixel of all images falls from (0, 255) to (0, 1). Finally, the training data label and the test data label are subjected to an effective coding conversion, which is convenient for the prediction of the results in the later period.

(2) Build the network model

Building a network model mainly includes the depth of the network, functions and parameters of each layer. The design of these parameters needs to be determined according to the characteristics and the amount of data for specific applications\[^8\]. The construction of convolutional neural network mainly includes the creation of convolutional layers, pooling layers, fully connected and output layers\[^9\]. Specific steps are as follows:

Step 1 Lead in the required modules and libraries. Using Keras to build a deep network, and the core data structure is models. Neural networks are organized by the simplest linear stacked model sequential, which imports it into the network model.

Step 2 Create a convolutional layer. Set the number and size of convolution kernels, input the channel of the picture, and use Relu function as the activation function, which can perform non-linear
conversion on the extracted features of the convolution.

Step 3 Create a pooling layer. The type of pooling is Max Pooling with a size of 2*2, which aggregates different features in a local area of the image.

Step 4 Recreate the convolutional layer and pooling layer.

Step 5 Create a fully connected layer. In the fully connected layer, the two-dimensional feature map output by the previous layer is first flattened into one dimension, that is, a flat layer is created, and then a dense hidden layer is created. The activation function still uses ReLU. In order to prevent overfitting, the dropout function is added, and the ratio parameter of disconnected neurons is set to 0.5.

Step 6 Create an output layer. Set parameters such as output category and activation function. The experiment uses the Softmax activation function.

(3) Experimental training and prediction

Set the parameters when training the data set. The loss function uses the cross-entropy (categorical_crossentropy) method and the adam optimizer. The standard for evaluating the model is the judgment of accuracy. Finally, it is needed to create an evaluation model for training and set the batch size batch, that is, how many samples are processed for gradient update, and set the number of training cycles epochs and the display process.

By training the input data, output feature maps at each layer, and based on the input data, use the fully connected network to output the conditional probability distribution. Use the created evaluation model to test and verify the accuracy of the model. Modify the training model many times and train repeatedly until the accuracy is satisfactory.

2.2 Principles of spatial transformation network

The spatial transformation network allows spatial operations on the data in the network. After the target image has operations such as zooming, translation, and rotation, the model can still be correctly classified\[10\]. This module can be inserted into the existing convolutional neural network structure. The new network structure can actively perform spatial transformation on the feature map without any additional supervision or modification of the process\[11\]. The spatial transformation network consists of the input layer U, transformation layer and output layer V. The transformation layer is composed of three parts: localisation net, grid generator and sampler\[12\]. The spatial transformation is as follows.

Step 1 Use convolutional neural network to automatically learn advanced features and capture contextual information in images.

Step 2 In order to make the convolutional neural network better adapt to salient and abnormal tasks, the feature sub-network structure is redesigned to output the transformation matrix based on the space transformer network. Extracting a few local features can effectively capture the edge pixels in the salient area, while embedding in the above model, reducing the influence of the prominent background area \[13\].

Step 3 Detect the region of interest through the linear combination of global and local feature information.

2.3 Convolutional neural network model based on spatial transformation

Convolutional neural network is a powerful model, but after a certain change or interference of the target image, the accuracy of its classification and prediction will be reduced \[14\]. According to the idea of model optimization, this paper considers that the data sets are all disturbed handwritten digital images. Based on the traditional convolutional neural network model, a spatial transformation network is introduced to form a new network structure (ST-CNN). The spatial transformation network can detect salient and abnormal areas in the image and align and straighten them \[15\]. The specific steps of the network design are as follows:

Step 1 Define the positioning network. The feature map U is used as an input. A fully connected or convolutional structure can be used. The goal is to output the spatial transformation parameter $\theta$, record $\theta=floc(U)$.

Step 2 Use the spatial transformation parameter $\theta$ output by the positioning network to transform
the input feature map. Taking affine transformation as an example to output a certain position on 
\((x_i^f, y_i^f)\), and map to a location on the input map with the parameter \((x_i^e, y_i^e)\) and the calculation formula is as follows:

\[
(1) \left(\begin{array}{c}
 x_i^e \\
 y_i^e \\
\end{array}\right) = \theta \left[\begin{array}{ccc}
 x_i^f & y_i^f & 1 \\
\end{array}\right] = \left[\begin{array}{ccc}
 \theta_{11} & \theta_{12} & \theta_{13} \\
 \theta_{21} & \theta_{22} & \theta_{23} \\
\end{array}\right] \left(\begin{array}{c}
 x_i^f \\
 y_i^f \\
 1 \\
\end{array}\right)
\]

(1) Affine transformation calculation formula

Step 3 Fill according to the filling rules defined in the previous period (usually using bilinear interpolation). The pixel value corresponding to the coordinate point in U is obtained according to the coordinate point of V for filling, without the need for matrix operations. It should be noted that the filling is not performed directly. If the calculated coordinates are decimals, certain processing needs to be performed, and then the other pixel values around the filling should be considered. The filling formula is as follows.

\[
(2) V_i = \sum_n \sum_m U_{nm} \ast k(x_i^f - m; \varnothing_x) \ast k(y_i^f - n; \varnothing_y)
\]

(2) Filling calculation formula

From above, \(\varnothing_x\) and \(\varnothing_y\) are sampling kernels, \(k\) is the parameter, which defines the form of image interpolation (such as bilinear). \(U_{nm}\) is the input feature map. \(U\) position is \((n, m)\). \(V_i\) is the output feature map, \(V\) position is \((x_i^f, y_i^f)\). The formula for using bilinear interpolation is as follows:

\[
(3) V_i = \sum_n \sum_m U_{nm} \ast \max(0,1 - |x_i^f - m|) \ast \max(0,1 - |y_i^f - n|)
\]

(3) Bilinear interpolation calculation formula

3. Experimental analysis

3.1 On CNN in Cluttered MNIST classification tests

The experiment uses the CNN model to train and recognize the Cluttered MNIST data set. The model uses multiple convolutional layers and pooling layers. The size of the convolutional kernel in the convolutional layer is 3*3, the Relu activation function is used, and the pooling layer uses the 2*2 maximum pooling method. To prevent over-fitting, a dropout layer is added, and finally an output layer is added, and the softmax activation function is used. The network structure is shown in table 1.

| Structure       | Parameter |
|-----------------|-----------|
| Convolution layer 1 | 320       |
| Convolution layer 2 | 18496     |
| Convolution layer 3 | 36928     |
| Convolution layer 4 | 36928     |
| The connection layer 1 | 991360    |
| The connection layer 2 | 1290      |

Table 1 shows the amount of parameters of the CNN network input as a 60*60 grayscale image. The total number of parameters participating in the training is 1,085,322. After the network is built, the training method is defined. The loss function adopts the cross entropy method. The standard for evaluating the model is accuracy. Taking 128 data each time during training and train for 30 cycles, the accuracy of the final prediction is 96.76%, and the loss value is 12.49%. The prediction results are shown in figure 3.
Among them, the abscissa represents the number of training cycles, and the blue and orange curves in the ordinate of figure 3(a) represent the accuracy of the training and prediction results respectively. The blue and orange curves on the ordinate represent the loss values of the training and prediction results respectively. The accuracy of the final prediction is 96.76%, and the loss value is 12.49%.

3.2 ST-CNN in Cluttered MNIST on the classification of the test

The ST-CNN model is used to train and recognize the Cluttered MNIST data set. When constructing the ST-CNN network, first construct the positioning network structure. The network structure is shown in table 2.

| Structure              | Parameter |
|------------------------|-----------|
| Convolution layer 1    | 832       |
| Convolution layer 2    | 51264     |
| Convolution layer 3    | 36928     |
| The connection layer1  | 387250    |
| The connection layer2  | 306       |

The total number of parameters participating in the training is 476,580. Using the spatial transformation parameters output by the positioning network, a spatial transformation network is constructed. The spatial transformation and the convolutional neural network are combined to form ST-CNN. Its network structure is shown in table 3.

| Structure              | Parameter |
|------------------------|-----------|
| Convolution layer 1    | 476580    |
| Convolution layer 2    | 320       |
| Convolution layer 3    | 18496     |
The connection layer1 36928
The connection layer2 409856
Structure 2570

The total number of parameters participating in the training is 944,750. To define training methods, the loss function also uses the cross entropy method. The standard for evaluating the model is accuracy. During the training, 128 data were taken each time, and the prediction results of 30 training cycles were shown in figure 4.

![Figure 4. ST-CNN value test results](image)

The prediction accuracy of the ST-CNN method is 97.65%, and the predicted loss value is 10.66%. In order to compare the classification and prediction of disturbing images by the two algorithms, CNN and ST-CNN-based neural network structures were used for training on the Cluttered MNIST data set. The results are shown in table 4.

|                  | CNN         | ST-CNN      |
|------------------|-------------|-------------|
| Training accuracy| 98.60%      | 97.66%      |
| Training accuracy| 96.76%      | 97.65%      |
| Training loss    | 7.96%       | 7.54%       |
| Forecast loss    | 12.49%      | 10.66%      |

It can be seen from the results that although the training accuracy of the conventional CNN method is high, the actual prediction accuracy is relatively low. However, the prediction accuracy of the ST-CNN method is almost the same as the training accuracy. And the prediction accuracy of ST-CNN method is higher than that of CNN method, and the predicted loss value is lower than that of CNN method.
4. Conclusion
This paper proposes a method of convolutional neural network based on spatial transformation for the classification and prediction of handwritten digital images with interference information such as shapes and strokes. Due to the unique feature of spatial transformation network that can detect the salient and abnormal areas in the image and align and straighten it. Combined with the convolutional neural network, it can solve the problem of disturbing classification and recognition in the image. Through the analysis and testing of the Cluttered MNIST data set, the prediction accuracy of the ST-CNN method is 97.65%, the prediction accuracy of the conventional CNN method is 96.76%, and the accuracy is improved by 0.89%. The predicted loss value of the ST-CNN method is 10.66%, and the predicted loss value of the CNN method is 12.49%. The loss value is reduced by 1.83%. It achieves the desired effect. Experiments show that ST-CNN method has higher accuracy and lower loss value in the recognition of handwritten digital images with interference information, which is better than conventional CNN methods.

References
[1] Chang Liang, Deng Xiaoming, Zhou Mingquan, et al. 2016. Convolutional Neural Network in Image Understanding[J]. Journal of Automation, 42(9): 1300-1312.
[2] Li Yandong, Hao Zongbo, Lei Hang. 2016. Summary of Research on Convolutional Neural Network [J]. Computer Applications, 36(9):2508-2515, 2565.
[3] Xi Y, Liu X. 2017. Multi-task convolutional neural network for pose-invariant face recognition[J]. IEEE Transactions on Image Processing, PP(99):1.
[4] Du Kuo, Li Ya. 2019. Research and optimization of digital classifier based on convolutional neural network [J]. Modern Electronic Technology.
[5] Zhang Tao, Yang Jian, Song Wenai, et al. Improved design method of convolutional neural network model[J]. Computer Engineering and Design, 40(7): 1885-1890.
[6] Zhang Xinsheng, Gao Teng, Gao Dongdong. 2018. A new deep spatial transformer convolutional neural network for image saliency detection[J]. Design Automation for Embedded Systems, 22 (3), 243-256.
[7] Szegedy C, Liu W, Jia Y, et al. 2015. Going deeper with convolutions [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE.
[8] Xia Wuji, Color Se Chajia, Zha Xi Ji, et al. 2019. Tibetan handwritten digit recognition based on convolutional neural network [J]. Modern Electronic Technology, 42(5): 79-86.
[9] Zhang Qianyu, Guan Shu, Xie Hongwei, et al. 2018. Image adaptive target recognition algorithm based on deep feature learning [J]. Journal of Taiyuan University of Technology, 49(4): 592-598.
[10] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// International Conference on Machine Learning. JMLR. org. 2015: 448-456.
[11] Han J, Zhang Z, Cummins N, et al. 2017. Strength modelling for real-world automatic continuous affect recognition from audiovisual signals[J]. Image and Vision Computing, 65: 6-86.
[12] Yang Jiashu, Mei Tiancan, Zhong Sidong. 2018. Application of CNN taking into account local characteristics in classification of remote sensing images[J]. Computer Engineering And Applications, 54 (7): 188-195.
[13] Yuan Jiajie, Zhang Ling, Chen Yunhua. 2019. Deep neural network image recognition based on attention convolution module [J]. Computer Engineering and Applications, 55(8): 9-16.
[14] Feng Jianzhou, Song Shasha, Wang Yuanzhao et al. 2019. Entity relation extraction method based on improved attention mechanism[J]. Acta Electronic Sinica, 47(8): 1692-1700.
[15] Wang Tao, Wang Hongzhang, Xia Yi, et al. 2019. Human gait recognition based on convolutional neural network and attention model[J]. Journal of Sensor Technology, 32(7): 1027-1032.