English Character Recognition based on Deep Learning

Yuanhong Su

Beijing 101 Middle school, Beijing 100190, China
s20213040655@cau.edu.cn

Abstract. Since the traditional method of English character recognition had already reach its limit, there's a need of a new pattern which can deal with various problems such as noises, distorted words, obstacles and more. Deep learning is a way out. The unique neural network structure of deep learning determines its advantages in the field of recognition, but at the same time, it will also face the interference of the noise of input data. Therefore, using the preprocessing algorithm for input images can effectively improve the recognition accuracy. In this paper, an English character recognition framework based on Convolutional Neural Network is proposed. The framework is based on VGG-19 model and was trained by datasets from ImageNet, a pretreatment based on principal component analysis is used to reduce noise in the image, and a untreated version of the same test set was entered into the model as a control group for comparison. In the experiment, the framework work successfully and shows a high accuracy in the self-collected test set. And with pretreatment dealt by PCA algorithm, a higher accuracy is shown.

Keywords: Deep learning, English characters, Image enhancement, PCA

1 Introduction

Computers obtain data from input information, therefore, it is necessary to develop machine vision, in order to let the machine has the ability to get information from images, it is necessary to be able to identify texts [1]. Text is a high-level information carrier, and it contains more information than other carriers. With the correct recognition, it can be more effective to gather information from the picture. Through the recognition of the label, a better understanding of the commodity could be made, through the recognition of the words on the financial statements, a company's financial situation could be analyzed, a English recognition model can be used in many area, such as real-time translation, industrial automation, intelligent robot producing and more. Therefore, it is important to build a high accuracy model. Deep learning model is a strong and effective classifier, and it can be applied to text recognition [2], its unique recognition pattern make it widely used in character recognition. With the ongoing development of deep learning these days, Region-CNN and other efficient model was developed, and the current text recognition model based on machine learning, has reached its accuracy over 95%.

B. H. Ahmad (ed.), *Proceedings of the 2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*, Advances in Intelligent Systems Research 180,
https://doi.org/10.2991/978-94-6463-370-2_39

However, the current recognition model still has its defects. The traditional optical recognition is mainly for the recognition of the printing body, which assumes that there is a clean background and high definition in the image [3]. Therefore, its recognition accuracy in general scenarios is not satisfactory. Therefore, a general text recognition model with high accuracy is very needed.

At present, the widely used character recognition model is VGG-19 model, the purpose of this paper is on the basis of VGG-19 model and improve the original model by using of image enhancement to pretreat the input image, improve text contrast and reduce environmental interference, and further reduce the error in the output, to improve recognition accuracy. In this paper, author proposed a deep learning model based on PCA pretreatment and VGG-19 model. This model may be applied in real-time translation, intelligent robot construction and other fields in the future.

2 Related Works

2.1 Recognition Process

The current text recognition process can be divided into several steps:

1) Detect text area: it's used to divide the recognition area from the whole image and conduct text recognition in the divided area.

2) Image pretreatment: usually the divided area is not expected to be perfect, so the image needs to be pre-processed to enhance the recognition accuracy. Common treatments include removing noise, emphasizing the text, and weakening the impact of the background, and pressing the image to a size suitable for recognition.

3) Extract features: features detected by shape and color is not universal, in order to identify the text, there's a need to extract the characteristics of each text, then compare with the detection data, and get the final identification result. Now, there are many characteristics are studied and defined, including edge features, structure characteristics, etc.

4) Use algorithm to distinguish text: the features could be seen as a set of multi-dimensional vector, and using the algorithm ot identify is equivalent to find one or more functions, which shows the character's type. By Taylor expansion and Fourier transform, it is proved that any function can be stimulated by continue changing the arguments, which constitutes the hidden layer in machine learning. Commonly used identifiers include support vector machine, logistic regression, naive Bayes, random forest, etc.

5) Screening the possible results: filter the final text, correct the errors by using the character confidence, context and other methods, and output the most likely correct results.

2.2 Research Defects

The tradition Optical Character Recognition technology has many problems, and it is basically only suitable for clear images with text written in printing form. For universal text recognition, there are the following problems:

- The complex color of the text background. The recognition of the basic is used for pure color background, or a small amount of noise, which can be solved by image processing, but generally, text images often have background stripe interference and incomplete text, block, or have the prospect of text color, therefore, using image enhancement technology to screen key area is very necessary.
- Text font difference. There are many commonly used fonts, whose characteristics are often quite different, and some are even changed by adding shadows, missing and blocking, also, different fonts are mixed occasionally, which brings more difficulties for text recognition and leads to identification errors.
- Distortion of the text. In the image, text does not always have vertical or horizontal arrangement, but often appears oblique and curved distribution, and the text interval is relatively random, which is difficult to be used as a reference for dividing words.
- Character adhesion. Most English fonts use conjoined characters, resulting in multiple characters connected together, it is difficult to divide a single character, and will lead to possible identification errors, and image quality problems are more likely to lead to incorrect division of conjoined characters.

2.3 Technical Background

In 2006, the Deep Neural Network algorithm was proposed by Hinton et al. It can simulate the deep transmission of multiple layers of neurons in the human brain, which creates its advantage in complex cognition [4]. The defect lies in the large amount of computation, but with the development of hardware, this problem has been solved, and DNN is now widely used in speech recognition devices and other recognitions.

In 2011, Microsoft reported in a paper that by using DNN, their large-scale speech recognition accuracy has increased by 30%.

In 2012, Baidu Voice Assistant was launched, which applied DNN technology to software services. The data showed that DNN significantly improved the accuracy of speech recognition.

In 1989, Lecun first proposed the concept of approximating the current CNN model [5]. He soon released the LeNet5 model, which uses the convolution layer composed of convolution layer, pooling layer and nonlinear activation function, and uses the convolution to get spatial features from the image, also, a special activation function, called Tanh, was used.

In 2012, Geoffrey Hinton won the first place in the ImageNet competition using CNN, which greatly promoted the development of image recognition technology.

2.4 Defects of Existing Technology

Despite the continuous improvement and innovation, there are still many problems in the existing recognition mode, which can be mainly divided into the following categories.

1) Rely on manually defined features

Although many people have done a lot of research, there is still no general feature definition at present, and the recognition of different characters still depends on manual identification features.

2) Random coding hinders classifier-based identification

In daily life there will sometimes have meaningless random coding, such as pouch production serial number, these no semantic string will make recognition based on the semantics and dictionary set lead to wrong answer, at the same time, the long (more than 20 characters) words will lead to classifier identification efficiency exponentially reduced, and lead to most of the decline of accuracy on the overall recognition model.

3) Lack of a character location function

When the overall identification method is adopted, it is often difficult to determine the text and the location of each character in the original image. The character-based identification method will not make such errors, but it will lead to errors in the text.

4) Training sample production is tedious

Almost all deep learning models rely on detailed annotated training samples, whose annotation information includes the specific location and text content of each character, and even a single character sample needs to be divided. Moreover, the training model needs to use variety of samples, so the annotation workload of such samples is very large.

3 Research Methods

In this paper, author established a neural network based on the VGG-19 model, and test the same set of input images separately with using the PCA algorithm as pretreatment and not. Then compare the recognition accuracy between them.

3.1 PCA

PCA algorithm is based on linear transformation, to change the original data to a dataset that reflect the original data characteristics its set of dimensions have no correlation between the feature vector set. Rearrange the dataset and let the new data set can reflect the original data, which is sorted by those can mostly reflect original data. Intercept the top selection, and abandon other data, then translate them back to the original data so the new data have lower noise. PCA is mainly used to extract the main characteristic components of the data, and can reduce the high dimensional data, which can effectively reduce the data complexity [6].

When applying PCA to images, it is necessary to express the image as a set of feature vectors, and the image in the computer is stored as one or more matrices. The singular values of these matrices can be converted into the required data form of PCA.The main process of PCA is shown in Figure 1.

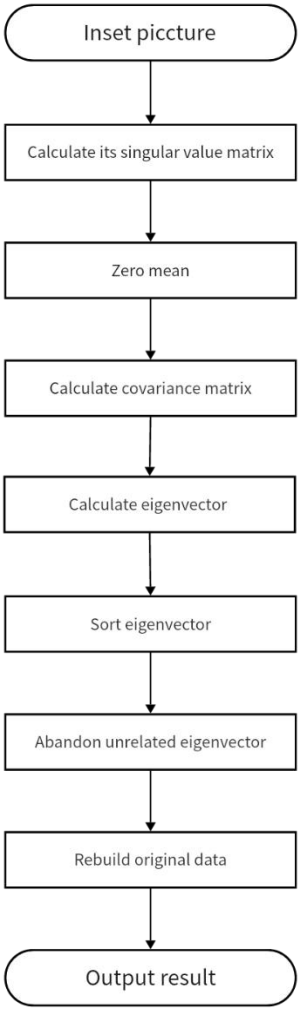

Fig. 1. The main process of PCA (Photo/Picture credit: Original)

Singular value. The singular values formulate the maximum action direction vector of the corresponding matrix, which is defined as

- If there is a vector v that makes the product of the matrix A and $Av = \lambda v$, which λ and v are the eigenvalues and eigenvectors of A.
- Let two matrices U and V, which are both orthogonal matrix, and Σ is a diagonal matrix, could let matrix $A = U\Sigma V^T$, the values on the diagonal are called singular values, and the columns in U and V is the left and right singular vector of A [7].

Singular values are the absolute values of the eigenvalues of the orthogonal matrix, which is shown below

$$\sqrt{Q * Q} = \sqrt{U\Lambda^2 U^T} = U|\Lambda|U^T \tag{1}$$

PCA. From the mathematical definition, PCA is to find another set of orthogonal bases that can represent the original data to the greatest extent, and is a linear combination of orthonormal bases, and the data can be expressed by the linear combination of each basis.

Let X be the original data set, the converted new data set is Y, and P is the linear conversion relationship, then the conversion relationship is $Y = PX$.

And let the column vector of X be X_i and Y be Y_i, and the row vector of P be P_i

Therefore, the Y can be expressed as

$$Y = \begin{bmatrix} p_1 \cdot x_1 & \cdots & p_1 \cdot x_n \\ \vdots & \ddots & \vdots \\ p_m \cdot x_1 & \cdots & p_m \cdot x_n \end{bmatrix} \tag{2}$$

$$y_i = \begin{bmatrix} p_1 \cdot x_i \\ \vdots \\ p_m \cdot x_i \end{bmatrix} \tag{3}$$

The p vector becomes the new base of the x vector. At this time, the problem changes to find the optimal P and optimize the data expression of X.

In statistics, the covariance is used to judge the optimal P, and the formula is

$$\sigma_{AB}^2 = \frac{\sum_{i-1}^{n}(a_i - a)(b_i - b)}{n-1} \tag{4}$$

Since $\sigma_{AB}^2 \geq 0$ and because 0 is equivalent. We have A and B are independent of each other, if A and B are equivalent, the covariance can be expressed as

$$\sigma_{AB}^2 = \frac{1}{n-1} AB^T \tag{5}$$

If both orders A and B are X, the covariance matrix of X can be obtained.

The data size on the diagonal of the matrix reflects the importance of the data. In order to minimize the redundancy as possible, the final target matrix should meet the value of 0 except for the diagonal. It can be completed by finding the matrix P, so that the covariance matrix of Y is all 0 except for the diagonal. At this time, the dimension reduction of the data has been completed, and each dimension is independent of each other, so the whitening can be used.

Divide each bit of Y by the standard deviation of each dimension, and then the variance of each dimension is equal, and multiply it with the feature vector matrix Z. Then the lower of noise is finished.

3.2 CNN

CNN is a basic architecture of deep learning network, and it is inspired by the cognitive mechanism which is from natural biological vision. Because its network model is often composed of multiple layers, it is also known as DNN, which is one of the typical models in the field of deep learning [8]. It is consist of pooling layer,

convolutional layer and fully connected layer. The input data of each convolution layer is locally connected to the output data of other convolution layer, and the output value is calculated by the parameter weighting [9]. CNNs differs to traditional ANNs since they are comprised of neurons that self-optimize through learning. The only notable difference between CNNs and traditional ANNs is that CNNs are primarily used in the field of pattern recognition within images. This allows it to get the result quicker and more accurate [10].

3.3 Model Training

The general CNN training method is about two steps. The first step is forward propagation, by collecting features from input data in the process of network. The second step is back propagation. This process compares the results obtained through forward propagation with the labels of the training set and finds the error between the two through a specific valuation function. Using gradient descent method, the weight parameters are updated from deep to shallow.

When selecting the deep learning model, because of the simplicity and efficiency of the VGG-19 model, it was selected for the base of the model. The existing OpenCV and PyTorch auxiliary programs were used to help finish the program.

The structure of the model is shown in Figure 2.

Data insert		
Convolutional layer 1	3*3*64	
	3*3*64	
Maxpool 1		
Convolutional layer 2	3*3*128	
	3*3*128	
Maxpool 2		
Convolutional layer 3	3*3*256	3*3*256
	3*3*256	3*3*256
Maxpool 3		
Convolutional layer 4	3*3*512	3*3*512
	3*3*512	3*3*512
Maxpool 4		
Convolutional layer 5	3*3*512	3*3*512
	3*3*512	3*3*512
Maxpool 5		
FC 1		
FC 2		
FC 3		

Fig. 2. The structure of the model (Photo/Picture credit: Original)

The training process is shown in Figure 3.

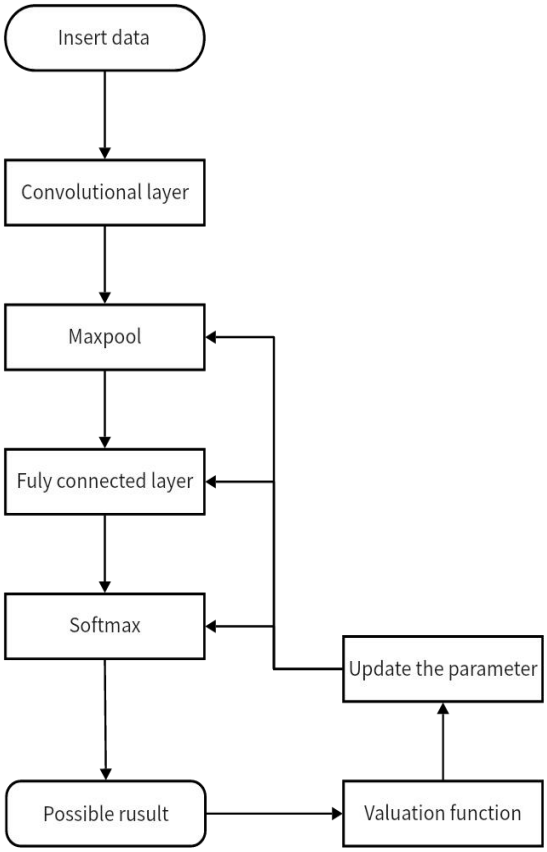

Fig. 3. The training process of CNN (Photo/Picture credit: Original)

The training set used was obtained from the handwritten English dataset from ImageNet, and the common printed dataset intercepted by the author.

The test set included 150 clear images collected from Baidu images and Google images, including 50 single letter images, 100 images containing short English text or a single word, and 50% of the images were manually added noise.

4 Experimental Results

In the test set's result, the test set used the PCA algorithm shows that among 50 single letter images, 50 were identified correctly, with an accuracy of 100%, and in 100 text images, there were a total of 1634 characters, among them, 159 characters were wrong or failed to identify, and the final identification accuracy was 90.2%.

In the test set without using the PCA algorithm, 49 of 50 single letter images were identified correctly, with 98% accuracy, and in 100 text images, there were a total of 1634 characters, among them, 253 characters were wrong or failed to recognize, and

the final recognition accuracy was 84. 5%.

After noise reduction through the principal component analysis, the final accuracy of single character recognition was improved by 2%, and the accuracy of short text or word recognition was increased by 5.7%, proving that image noise reduction through PCA can improve the recognition accuracy based on VGG-19 model.

5 Conclusion

This paper introduces the structure and working principle of convolutional neural network and principal component analysis algorithm through the research of English character recognition using CNN. Through the research, this paper found that using PCA algorithm for input data noise reduction can improve CNN's recognition accuracy. Although the model has shown a good result in the test data set, the test data in specific or extreme environment isn't enough, so the accuracy of the model can only reflect the PCA algorithm's use on conventional image optimization, the model still has large improvement space, especially in the text positioning and character segmentation. The model could probably improve the existing text recognition and natural scene character recognition technology, and might be applied in real-time translation, industrial automation and other fields.

There is still a large development space for text recognition based on deep learning, and there are two main optimization methods.

- Reduce the algorithm space-time complexity

At present, the recognition algorithm has high spatial and temporal complexity, which leads to excessive resource occupation in use. In the future, we can try to use more streamlined and optimized algorithms to reduce the computational amount to increase the practicability of the model

- Enhance the generalization performance of the model

Deep learning-based algorithms require large numbers of training samples, but manually collected and annotated samples can lead to huge manpower and time consumption, the use of adversarial neural networks could be able to reduce labor costs.

References

1. P. Huang : "Deep learning based scene text recognition." Diss. Zhejiang U (2016).
2. Y. Liu, et al. : Overview of the aplication of deep learning in scene text recognition technology. Computer Engineering and Application 58 (4), 52-63 (2022).
3. R. Ma : "Design and implementation of a natural scene text recognition system based on deep learning algorithm." Diss. Jilin U (2015).
4. A. Krizhevsky, I. Sutskever, and G.E. Hinton : ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. (2017).
5. Y. Lecun, et al. : Backpropagation applied to handwritten zip code recognition. Neural Computation, 1(4), 541-551. (1989).

6. A. Maćkiewicz, W. Ratajczak, : Principal components analysis (PCA), Computers & Geosciences, 19(3), 303-342 (1993).
7. F. Smithies : The eigen-values and singular values of integral equations. Proceedings of the London Mathematical Society, s2-43(1), 255-279. (2006).
8. Y. Zhang : “Research on traffic sign detection algorithm based on artificial neural network.” Diss. Northwest Normal U. (2021).
9. F. Zhou, L. Jin, and J. Dong: Review of Convolutional neural network research. Journal of Computer Science, 40 (6): 1229-1251. (2017).
10. O'Shea, Keiron , and R. Nash : An Introduction to Convolutional Neural Networks. Computer Science (2015).