Deep Neural Network for Handwritten Digital Recognition Based on Attention Mechanism

Yibo Hao[1], Jianbo Chen[2]

[1] Shenzhen Foreign Language School, Shenzhen 518000, China

[2]University of Chinese Academy of Sciences, Beijing 100049, China

Keywords: Convolutional neural network, Attention mechanism, Handwritten digital recognition

Abstract: Handwritten digital recognition is a hot spot in the field of artificial intelligence, which has already played a very important role in the society. A lot of handwritten digital recognition algorithms have been developed in the last few decades. However, most of previous algorithms failed to better extract the semantic and useful local parts from the original image, which is important for the enhancement of signal noise ratio. In this paper, an attention mechanism based neural network is proposed to better extract the salient local information automatically, and simultaneously filter out background or noise in the handwritten digital images. By this way, our method is able to only focus on the important information of the input image, and consequently better understanding the semantic information of the image. Experiments has been conducted on a widely used dataset MNIST, and the results show that our method is able to achieve great performance on the handwritten digital recognition task.

1. Introduction

Handwriting digital recognition is to automatically recognize the digit from the handwriting digital images with computer. In recent years, the application of handwriting recognition is more and more extensive and important. Due to the great potential of this task, there have been a huge amount of algorithms developed to solve the handwritten numeral recognition problems [2,3,4,5,6,9,22]. However, those algorithms and are flawed. In the recent years, the computer community has witnessed three main kinds of approaches for handwriting digital recognition and all the object recognition tasks.

The first kind of approaches is to use handcrafted feature to solve the recognition problems [2,3,4,5]. For example, Biglari et al. [2] used Local Binary Patterns (LBP) for handwritten digital recognition. The LBP feature is shown in Fig. 1. A simple LBP records the contrast, or difference, between a pixel and its surroundings. The image on the far left is the original. The aim is to detect some information of a pixel point. A threshold processing is performed for the middle grid of nine squares (the number in the grid is the gray value of pixel point).Pixels greater than or equal to the center point are marked as 1 and pixels less than are marked as 0.Finally, the 11110001 binary number around the central pixel point is converted into decimal number to obtain the LBP value. At first the original function was to assist image local contrast and was not a complete feature description, as a result, if the identification problem is more difficult, the effect it can get will not be well.

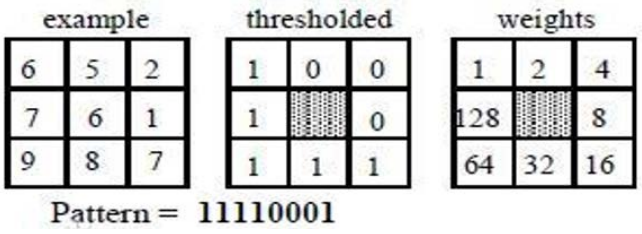

Fig.1 Local Binary Pattern

The second algorithm is to use sparse representation based methods [6,7,8]. For example, Qu et

  DOI: 10.25236/mmmce.2020.022

al.[6] proposed a locality-sensitive sparse representation method toward optimized prototype classifier. The learned dictionary can not only preserve local data structures, but also require the reconstruction of a pattern to get as close as possible to the prototype optimized by the minimum classification error (MCE) approach. So this method is able to help improve the classification accuracy effectively. However, the obtained sparse encoding may be highly redundant.

In recent years, deep learning has made great progress in many artificial intelligence areas, such as computer vision[11,12,17,20]. For the recognition tasks, deep learning based methods are also becoming very dominating. For example, Wen et al. [9] analyzed the difference between convolution neural network (CNN) network and traditional neural network (NN) in the recognition of handwritten digital datasets. They further designed a deep convolutional neural network (DCNN) based on Alex network (AlexNet). However, large size of convolutional neural network always leads to an excessive number of parameters and an over-fitting problem.

Through the analysis, comparison and research of these three algorithms, their advantages and disadvantages are summarized. To avoid the above mentioned problems from happening, an attention based model is developed in this paper. which can effectively solve the problems cause by the first three methods. The proposed method is elaborated in section 2.

2. Methodology

In this section, we discuss the process methods for digital recognition algorithm. First, a backbone neural network is used as a basic architecture. Second, from the verification and modification of the attention module, we can get a new architecture of the algorithm. Finally, to properly train the algorithm, a loss function is implemented on the output of the network.

2.1 Network Backbone

When the system receive the image, the neural network will use a convolution kernel to process the input image and extract a corresponding feature of the image. Then a down-sampling operation will be performed. This process will be repeat again to hierarchically extract more abstract and semantic feature. In the end, the network uses two full connections and get the final high level information. For example, in the picture, an image which has a 32x32 matrix, with a convolution, it will divide the image into six 28x28 images. Then, the algorithm will perform the down-sampling, this process can decrease the size of image and the size of the matrix will be a half, the size will be 14x14 and number of matrix will increase. Then the process will be repeat again. After that, it will use two full connections, then we can get the information we want.

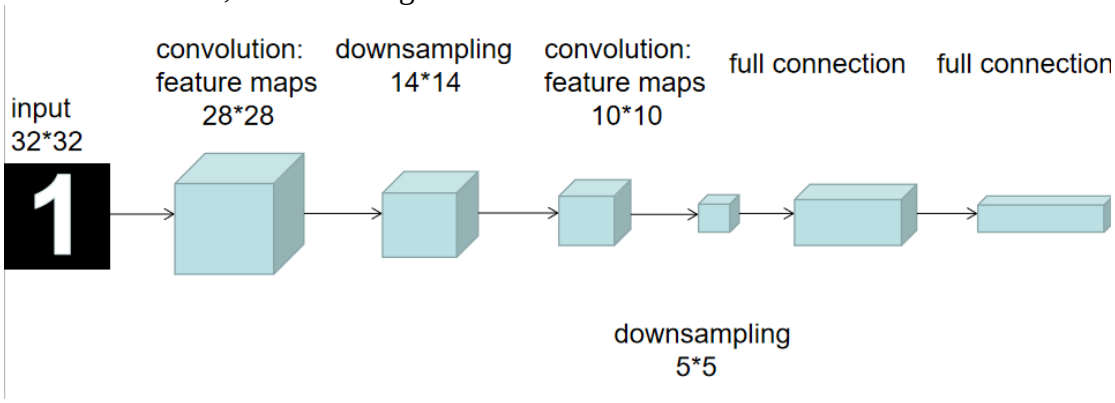

Fig.2 The Overall Framework of the Proposed Model

2.2 Attention Module

In addition to the basic neural network backbone, the algorithm exploits the attention mechanism to automatically focus on the important local regions of interest. The function of attention mechanism is to filter out excess information, because like human beings, when it observes images, it will focus on some particular regions and ignore other useless information, so the attention

mechanism can help it concentrate on useful information of the image.

The detail of the implemented attention module is illustrated in Fig 3. In this module, the attention mask is first extracted from original feature map with 1x1 convolution filters. Then the attention mask is elementary produced with the origin map and produces the attention map.

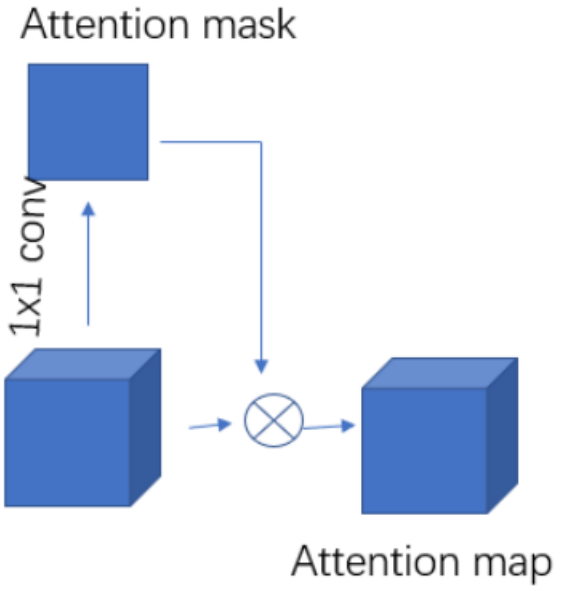

Figure 3The implemented attention module. In this module, the attention mask is first extracted from original feature map with 1x1 convolutional filters. This mask is then elementwise product with origin feature map.

2.3 Loss Function

Loss function is also important to the learning of neural network [13,14,15]. In the training procedure the cross entropy function. It is a function that maps a random event or its value of related random variables to a non-negative real number to represent the risk or loss of the random event. In applications, loss functions are often associated with optimization problems are learning criteria, that is, loss minimization functions are used to solve and evaluate models. The loss function is formulated as equation 1:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} \hat{y}_i \log(y_i) \quad (1)$$

In machine learning, given independent identically distributed learning samples $(X, y) \in \mathcal{X} \times \mathcal{Y}$ and models $\hat{y} = f(X, w)$, the loss function is the quantification of the difference in probability distribution between model outputs and observed results. The specific quantization method on the right side of the above equation depends on the problem and the model, but it is required to meet the general definition of loss function, namely the non-negative measurable function of the sample space. By this way, the network is able to correctly classify the handwritten digital into the right category.

3. Experiment

This section will mainly discuss the experiment of our module.

3.1 Dataset

This dataset contains 80,000 handwriting digital examples. Among those images, 60,000 examples are used for training the model and 20,000 examples exploited for testing. These Numbers have been dimensioned and are located in the center of the image, which is a fixed size (28x28 pixels) with values from 0 to 1. For simplicity, each image is flattened and transformed into a one-dimensional NUMPY array of 784(28 * 28) features in program.

Some examples of the dataset is shown in Fig. 4. It can be seen that the dataset suffers from the

large intra-class variation problem. For example, it's a little hard to recognize that the several handwritten digit seven belong to the same class. What's more, there are also severe inter-class variation problem. There are only two kinds of number and the is small, so it' much easier to distinguish.

3.2 Experiment Protocol

For experiment, we use the MATLAB for programming and modeling. The learning scheme is stochastic gradient descent (SGD). In machine learning / deep learning algorithms, the loss function of the objective function usually takes the average of the loss functions of each sample. If gradient descent method (batch gradient descent method) is used, the gradient of [formula] samples should be obtained during each iteration.

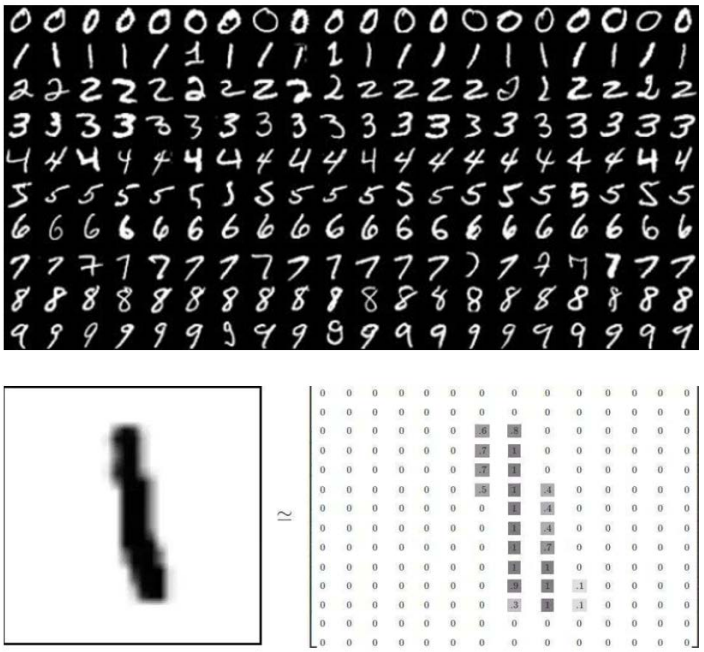

Fig.4 Example of the Mnist Handwritten Digital Dataset. Upside Picture: Examples of Each Class is Illustrated. Downside Picture: a Detail Example of Handwritten Digital Image.

In our experiment, the learning rate is set to 0.01, batch size is 96, and epoch is 100. The image is preprocessed by data augmentation. Specifically, horizontal flip, and image jittering are used in the preprocession of this experiment.

3.3 Experiment Results

Table 1 Table Captions Should Be Placed Above the Tables.

Algorithms	Accuracy	
LeNet	97.7%	
AlexNet	98.7%	
VGG 16	98.9%	
ResNet	86.5%	
Ours (LeNet+Attention)	99.3%	

Our experiment was carried out on the MNIST data set. The experimental result is reported in Table 1. If only the backbone network is used which is LeNet5, the performance will be 97.7 % accuracy. However, if our attention mechanism is added upon the LeNet backbone, the performance will be improved to 99.3% accuracy.

ResNet only achieves 86.5% accuracy in our experiment, which is a bad performance. That is because that ResNet is a large neural network. Therefore, it is easy to be overfitting in MNIST dataset.

3.4 Parameter Experiments

This subsection discusses the influence of different number of epochs, batch size and learning rate on the performance of our proposed algorithm.

3.4.1 Learning Rate

During training procedure, learning rate of the model has a huge effect on the training result and the inference performance of the model. Therefore, we conducted experiments to choose the best learning rate. The result is shown in Table 3. It can be seen that when learning rate is set to 0.01, the test result is 99.3%, which is the best accuracy. Therefore, we choose learning rate as 0.01.

Table 2 Performance of The Model under Different Learning Rates

Learning rate	Training accuracy	Test accuracy
0.1	88.3%	80.9%
0.01	99.7%	99.3%
0.001	81.1%	79.8%

3.4.2 Batch Size

Batch size is also very important to the performance of our model. Therefore, we conducted experiments to choose the best batch size. The result is shown in Table 3. It can be seen that when batch size is set to 96, the result is already very good. To save the memory, we choose batch size as 96.

Table 3 Performance of the Model under Different Batch Sizes

Batch size	Training Accuracy
48	95.9%
64	98.2%
96	99.7%
128	99.7%
256	99.8%

3.4.3 Number of Epoch

In addition to learning rate and batch size, the number of epoch also is very important to the accuracy of the model. Therefore, we conducted experiments to choose the best number of epoch. The result is shown in Table 4. It can be seen that when batch size is set to 100, the result is already very good. To save the training time, we choose epoch number as 100.

Table 4 Performance of the Model under Different Epoch

Number of Epoch	Training accuracy	Test accuracy
50	94.2%	93.4%
100	99.7%	99.3%
150	99.9%	99.3%

4. Conclusion

In this paper, an attention mechanism based deep neural network is proposed, which is able to enhance the information of regions of interest, and filter out those noises. The proposed model is evaluated in the predominate handwriting digital dataset, and achieves state of the art performance.

References

[1] Shady S. Al-Atrash, Ibrahim Abuhaiba. Robust Face Recognition. (2012)Shady S. Al-Atrash, & Ibrahim Abuhaiba.. Robust face recognition.

[2] Biglari M, Mirzaei F, Neycharan J G (2014). Persian/Arabic handwritten digit recognition using

local binary pattern International Journal of Digital Information and Wireless Communications (IJDIWC), vol.4, no.4, pp.486-492.

[3] Bannigidad P, Gudada C (2018). Identification and classification of historical Kannada handwritten document images using LBP features. International Journal of Intelligent Systems Design and Computing, vol.2, no.2, pp.176-188.

[4] Kamble P M, Hegadi R S (2015). Handwritten Marathi character recognition using R-HOG Feature. Procedia Computer Science, vol.45, no.1, pp.266-274.

[5] Jebril N A, Al-Zoubi H R, Al-Haija Q A (2018). Recognition of handwritten arabic characters using histograms of oriented gradient (HOG). Pattern Recognition and Image Analysis, vol.28, no.2, pp.321-345.

[6] Qu X, Wang W, Lu K, et al (2018). In-air handwritten Chinese character recognition with locality-sensitive sparse representation toward optimized prototype classifier. Pattern Recognition, vol.2018, no.78, pp.267-276.

[7] Tsourounis D, Theodorakopoulos I, Zois E N, et al (2018). Handwritten signature verification via deep sparse coding architecture, 2018 IEEE 13th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). IEEE, pp.1-5.

[8] Wright J, Ganesh A, Zhou Z, et al (2009). Demo: Robust face recognition via sparse representation, 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition. IEEE, pp.198-201

[9] Wen Y, Shao Y, Zheng D (2019). A Novel Deep Convolutional Neural Network Structure for Off-line Handwritten Digit Recognition,Proceedings of the 2nd International Conference on Big Data Technologies. pp. 216-220.

[10] Wright, John, Ganesh, Arvind, Zhou, Zihan,. Demo (2008): Robust face recognition via sparse representation [M]. IEEE, Wright, John, Ganesh, Arvind, Zhou, Zihan, Wagner, Andrew, & Ma, Yi. (2008). Demo: Robust face recognition via sparse representation. IEEE.

[11] Chen Y, Lu X., Wang S (2020). Deep Cross-Modal Image-Voice Retrieval in Remote Sensing . IEEE Transactions on Geoscience and Remote Sensing, vol.1, no.1, pp.132.

[12] Yandong Wen, Kaipeng Zhang, Zhifeng Li, (2016). A Discriminative Feature Learning Approach for Deep Face Recognition. Computer Vision – ECCV 2016. Springer International Publishing,

[13] Yubero F, Tougaard S, Elizalde E, et al (1993). Dielectric loss function of Si and SiO2 from quantitative analysis of REELS spectra. Surface & Interface Analysis, vol.20, no.8, pp.719-726.

[14] Basu S, Markov S (2004) . Loss function assumptions in rational expectations tests on financial analysts' earnings forecasts . Journal of Accounting & Economics, vol.38, no. 1/3, pp.171-203.

[15] Chu W, Keerthi S S, Ong C J (2004). Bayesian support vector regression using a unified loss function. IEEE Transactions on Neural Networks,

[16] Xiao H, Rasul K, Vollgraf R (2017). Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms .

[17] Вадим Васильович Романюк (2016). Training Data Expansion and Boosting of Convolutional Neural Networks for Reducing the MNIST Dataset Error Rate .vol.22, no.12, pp.12-23.

[18] Zhang Y , Farrell S , Crowley M , et al (2020). A Molecular-MNIST Dataset for Machine Learning Study on Diffraction Imaging and Microscopy, Clinical and Translational Biophotonics.

[19] Hailong, Xi, Haiyan, et al (2018). Recognition and Optimization Algorithm of MNIST Dataset Based on LeNet5 Network Structure.

[20] Wang Z , Wu S , Liu C , et al (2019). The Regression of MNIST Dataset Based on Convolutional Neural Network, International Conference on Advanced Machine Learning Technologies and Applications. Springer, Cham.

[21] Lejeune E (2020). Mechanical MNIST: A benchmark dataset for mechanical metamodels. Extreme Mechanics Letters, pp.100659.

[22] Pereira De Freitas, Daniel, Shkunov, Alexander (2014). Error rates reduction in handwritten digits classification using the MNIST data with Artificial Neural Networks. Bragantia, vol.69, sup. L, pp.121-129.

[23] Shi F. Learn About Convolutional Neural Networks in Python With Data From the MNIST Dataset [M]. 2019.

[24] Shisheie R, Galun B A , Kim J (2018). Implementation and Analysis of Different Digit Recognition Methods on Reduced MNIST Dataset,

[25] Paras (2014). Stochastic Gradient Descent. Optimization.

[26] Theodoridis S (2017). Stochastic Gradient Descent. Deep Learning with Python. Apress,

[27] Niu F, Recht B, Re C, et al (2011). HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent. Advances in Neural Information Processing Systems, vol.24, pp.693-701.

[28] Zhang, Tong (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. pp.116.

[29] Gardner W A (1984). Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique. Signal Processing, vol.6, no.2, pp.113-133.

[30] Loshchilov I , Hutter F (2016). SGDR: Stochastic Gradient Descent with Restarts.

[31] Xu W (2011). Towards Optimal One Pass Large Scale Learning with Averaged Stochastic Gradient Descent. Computer Science,

[32] Meuleau N , Dorigo M (2014) . Ant colony optimization and stochastic gradient descent. Artificial Life, vol.8, no.2, pp.103-121.

[33] Heye A (2019). Scaling deep learning without increasing batchsize. Concurrency and Computation: Practice and Experience, vol.31. no.16, pp.e5147.1-e5147.8.

[34] Schmeiser B (1982). Batch Size Effects in the Analysis of Simulation Output. Operations Research, vol.30, no.3, pp.556-568.

[35] Lamb, H. H (1965). The early medieval warm epoch and its sequel. Palaeogeography Palaeoclimatology Palaeoecology, vol. 1965, no.1, pp.13-37.

[36] Batur A U , Iii M H H (2004) . Segmented Linear Subspaces for Illumination-Robust Face Recognition. International Journal of Computer Vision, vol.57, no.1, pp.49-66.

[37] Ou W , You X , Tao D , et al (2014). Robust face recognition via occlusion dictionary learning. Pattern Recognition, vol.47, no.4, pp.1559-1572.

[38] Wang J , Lu C , Wang M , et al. Robust Face Recognition via Adaptive Sparse Representation. IEEE Trans Cybern, 2014, vol.44, no12, pp.2368-2378.

[39] Tao, Dacheng, Ding, et al. Robust Face Recognition via Multimodal Deep Face Representation. IEEE Transactions on Multimedia, 2015.

[40] Stephen Balaban. Deep learning and face recognition: the state of the art, Spie Defense + Security. International Society for Optics and Photonics, 2015.