Real time handwritten digital image recognition system based on edge computing equipment

Lei Zhang[1,3,*], *Lili* Duan[2], and *Jiasun* Suo[1,3]

[1]Department of Electrical and Electronic Information engineering, Hubei Polytechnic University, Huangshi, China
[2]Department of Computer and Information Engineering, Hubei Normal University, Huangshi, China
[3]Hubei Key Laboratory of intelligent transportation technology and device (Preparatory), Hubei Polytechnic University, Huangshi 435003, China

Abstract. With the development of edge computing technology, real-time handwritten numeral recognition system deployed in edge computing devices has a bright future. However, the edge computing equipment has weak computing power and limited storage space, so the mainstream image recognition neural network can not guarantee the real-time performance on low-performance edge devices. To solve this problem, this paper designs a handwritten digit recognition system which is suitable for low performance edge computing devices. The system extracts the effective information such as the position of the number in the input image through the image preprocessing module, and infers through a lightweight neural network specially designed for edge devices. In this paper, the proposed handwritten numeral image recognition system is deployed on the edge computing device Jetson Nano. The experimental data show that the inference speed of our model is 10 times faster than that of the original Tensorflow inference, and 60 times higher than the neural network Mobilenet specially designed for mobile devices. At the same time, with the increase of input video resolution, the FPS of the system does not decline significantly, which can meet the needs of most edge tasks. Finally, the system also provides design and deployment experience for other edge AI tasks.

Keywords: Edge computing, Edge intelligence, Digital image recognition, Neural network, Singular value decomposition.

1 Introduction

In recent years, handwritten numeral image recognition system is more and more widely used in real life, such as handwritten numeral statistics, test paper score scanning, automatic mail sorting, data entry and so on [1]. At present, handwritten numeral recognition is mostly implemented based on cloud computing, which will have problems such as network delay and data leakage. It is not suitable for application environments with high requirements for

* Corresponding author: zhanglei4180@gmail.com

real-time and security, such as real-time handwritten image recognition in the field of industrial Internet [1]. With the development of edge intelligence (EI) technology, the real-time handwritten numeral system deployed on edge devices is not limited by the transmission network, and the data could not be uploaded to the cloud. The system not only ensures the real-time performance, but also protects the data security. It can better meet the handwritten image recognition application environment with high requirements for real-time performance and security.

In [2], the development of EI and the importance of edge end in the field of artificial intelligence (AI) is described; In [3], an Edgent framework is proposed to optimize the reasoning of deep neural network on edge devices. At the same time, the experimental data show that deep neural network (DNN) is effective in image recognition; In [4], the authors proposed a VGg network model, which achieved high accuracy in image recognition; In [5], the authors proposed a RESNET network model, which uses the residual structure to solve the "degradation" problem that the error rate increases when the level of the model deepens; In [6], the authors proposed a wideresnet network model, which achieved better accuracy by widening RESNET and reducing its layers; In [7], the authors proposed proposed a mobilenet network model specially designed for mobile devices, which greatly reduces the amount of DNN calculation through separable convolution. However, whether the handwritten numeral recognition system can run effectively on low-performance edge computing devices is still a problem to be solved. This paper designs a lightweight handwritten numeral recognition neural network, and constructs a handwritten numeral recognition system based on it. The system runs on the low-cost edge computing device Jetson Nano. The experimental data show that the system designed in this paper can meet the needs of low-cost edge computing devices, and also has good accuracy and real-time on edge computing devices with low computing power.

2 System structure

As shown in Figure 1, the real-time handwritten numeral system based on edge meter device designed in this paper is composed of image preprocessing module, neural network inference model on edge device and image post-processing module. The image preprocessing module extracts the effective information such as the position of the number in the input image for neural network inference; In the process of image inference, because the edge computing equipment has low computing power and limited storage space, the current mainstream image recognition neural network can not operate. The system specially designs a lightweight neural network and selects an appropriate edge deployment framework to ensure the AI performance of the system; The picture post-processing module marks the prediction results and location information in the picture, and finally outputs the visual video stream.

In the practical application of the system, in order to solve the problem of inaccurate recognition caused by the difference between the actual handwritten digital picture style and the picture style in the MNIST dataset used in the training model, the system first uses a picture preprocessing method to process the captured picture into a style similar to the image in the MNIST dataset, and then sends it to the neural network for inference.

3 Improved handwritten numeral recognition neural network

In order to adapt to the characteristics of weak computing power and small storage of edge computing devices, this paper improves the classical convolutional neural network (CNN) model Lenet-5 [8], and proposes an improved neural network model (Lenet-c) for

handwritten digit recognition. The improvement is mainly reflected in the optimization of activation function, model parameters, model generalization ability and model training. The improved neural network includes convolution layer 1 (C1), pooling layer 1 (P1), convolution layer 2 (C2), pooling layer 2 (P2), full connection layer 1 (F1), full connection layer 2 (F2) and output layer (out). Finally, singular value decomposition (SVD) is used to compress the model to reduce the amount of calculation and size of the model.

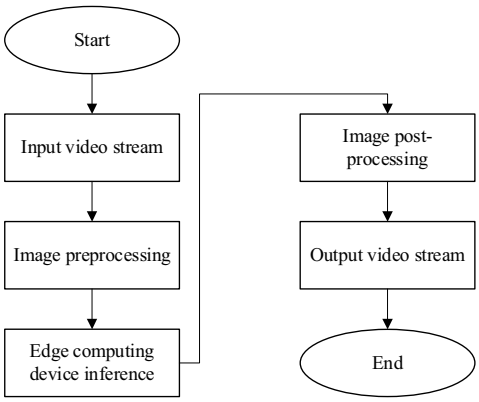

Fig. 1. System processing flow.

In the optimization of activation function, firstly, Relu activation function is used to replace the Sigmoid activation function in Lenet-5, so as to avoid the problem of exponential budget, improve the speed of calculation and avoid the disappearance of gradient. At the same time, the model parameters are optimized as follows: (1) reduce the size of convolution kernel to 3x3 and reduce one layer of convolution to accelerate the speed of convolution operation. (2) Increase the number of channels in the convolution layer (C1, C2) to improve the feature extraction ability of the network. (3) Add a full connection layer to improve the classification ability of the network. The specific parameters of the improved model are shown in Table 1. In order to increase the generalization ability of the model, L2 regularization and "Dropout" methods are added during training to improve the classification accuracy of the model when facing real data. In the training, the method of exponential attenuation learning rate is used to make the model easier to train to the optimal parameters.

Table 1. Parameters of the model.

Layer	Input	Output	Type
c1	28x28x1	26x26x32	-
p1	26x26x32	13x13x32	max
c2	13x13x32	11x11x64	-
p2	11x11x64	6x6x64	max
Reshape	6x6x64	1x2304	-
f1	1x2304	1x768	-
f2	1x768	1x10	-

Edge computing devices usually have limited storage space and are limited by the transmission network bandwidth. The size of the neural network model deployed on the edge platform should not be too large. In order to make the model more suitable for the edge platform, the model needs to be compressed within the allowable range of accuracy. The number of parameters in layer F1 accounts for 99.52% of the total number of Lenet-c model. As long as layer F1 is compressed, the size of the model can be effectively reduced. Based on the idea of singular value decomposition (SVD), this paper compresses the model,

which effectively reduces the size of the model while retaining most of the information of the model. Assuming any real matrix, it can be decomposed by SVD as:

$$W_{m\times n} = U_{m\times m} S_{m\times n} V^{\tau}_{n\times n} \tag{1}$$

$U_{m\times m} = [u_1, u_2, \cdots, u_m]$, $V_{n\times n} = [v_1, v_2, \cdots, v_n]$ are left singular value matrix and right singular value matrix, respectively. $S_{m\times n} = \begin{bmatrix} C_{q\times q} & O \\ O & O \end{bmatrix}, C_{q\times q} = diag(\sigma 1, \sigma 2, \cdots, \sigma q), q = min(m, n)$. $\sigma i (i = 1,2,\cdots, q)$ are the Singular value of W,and $\sigma 1 > \sigma 2 > \cdots > \sigma q > 0$. Take the largest first k singular values to approximately restore the matrix $W_{m\times n}$:

$$\begin{aligned} W_{m\times n} &\approx U'_{m\times k} S'_{k\times k} V'^{T}_{n\times k} \\ &= U'_{m\times k} (S'_{k\times k} V'^{T}_{n\times k}) \\ &= U'_{m\times k} T_{k\times n} \end{aligned} \tag{2}$$

$U'_{m\times k} = [u'_1, u'_2, \cdots, u'_k], V'_{n\times k} = [v'_1, v'_2, \cdots, v'_k]$ are left singular value matrix and right singular value matrix, respectively. $S'_{k\times k} = diag(\sigma 1, \sigma 2, \cdots, \sigma k)$.

4 Experimental results and analysis

On the edge computing device Jetson Nano, the Tensorflow1 13.1, Tensorrt5 and Opencv4 1.1are employed.The experimental results of three framework deployment models are shown in Table 2. Opencv and Tensorrt have faster inference speed and smaller memory footprint than Tensorflow. At the same time, on this kind of lightweight neural network, Opencv is nearly twice as fast as Tensorrt's inference speed, and the memory occupation is only 8.08%. The experimental results show that Opencv is more suitable for deploying the improved Lenet-c network model proposed in this paper, and it has better performance on edge computing devices.

Table 2. Comparison of deployment models of various frameworks.

Framework	Time of 1000 times (s)	Memory occupation (MiB)
Tensorflow	6.007	1244.9
TensorRT	1.174	996.1
OpenCV	0.604	80.5

In order to test the system performance, input video streams with different resolutions are used to test the real-time performance of the system. As shown in Table 2, the system has good adaptability to videos input with different resolutions on edge computing devices with low performance, and its FPS can better meet the requirements of the system.

At the same time, we analyzed the time consumption on the Jetson nano through the inferred time in Table 2 and the FPS during system operation, and found that its computing bottleneck mainly focused on image processing. Because of its weak CPU capacity, the time of image processing is long; In the aspect of neural network inference, because the system proposed in this paper uses GPU inference, its time consumption is far lower than that of image processing. It is proved that the GPU performance of some simple neural

networks running on edge computing devices such as Jetson Nano can meet the requirements.

Table 3. FPS of input video with different resolutions.

Resolution	FPS
2400x2400	7
2000x2000	8
1600x1600	9
1200x1200	9
1000x1000	10
800x800	10
600x600	10
400x400	10

The practical application results of the system are shown in Figure 2. Under the input resolution of 1000X1000, the system can accurately locate and recognize the handwritten digits in the input video, and the FPS reaches 10.

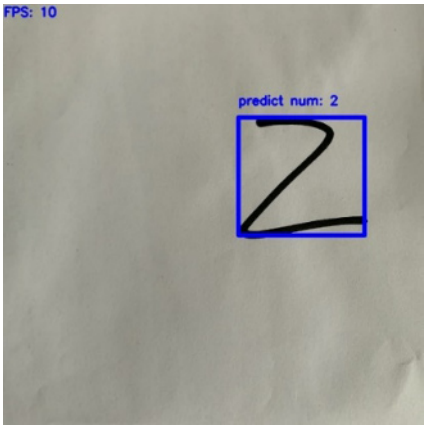

Fig. 2. Experimental results of practical application of the system.

5 Conclusion

A handwritten numeral recognition system which is suitable for the edge computing devices is proposed in this paper. In order to ensure its performance in edge computing equipment, some improvements suitable for edge computing equipment are made in image preprocessing, neural network design, model inference framework selection and so on, so as to ensure the practicability of the system. The design of this paper also provides some experience for AI applications on other edge computing devices: including accelerating the full connection layer of neural network in a way similar to singular value decomposition to make it perform better on edge devices, and using open CV as the network inference framework on edge devices may have better results When executing lightweight AI tasks in low-cost edge computing devices such as Jetson Nano, the CPU limit is greater than GPU.

This research was funded by the National Science Foundation of China with Grant 61871178 and the Open project of Hubei Key Laboratory of intelligent transportation technology and equipment (Preparatory) (2020xz110).

References

1. Fujisawa H. Forty years of research in character and document recognitionan industrial perspective [J]. Pattern Recognition, 2008, 41(8): 2435-2446.
2. Zhou Z, Chen X, Li E, et al. Edge intelligence: Paving the last mile of artificial intelligence with edge computing [J]. Proceedings of the IEEE, 2019, 107(8): 1738-1762.
3. Li E, Zhou Z, Chen X. Edge intelligence: On-demand deep learning model co-inference with device-edge synergy[C]//Proceedings of the 2018 Workshop on Mobile Edge Communications. 2018: 31-36.
4. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Sep. 4, 2014[J]. arXiv preprint arXiv:1409.1556, 2019.
5. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
6. Zagoruyko S, Komodakis N. Wide residual networks [J]. arXiv preprint arXiv:1605.07146, 2016.
7. Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications [J]. arXiv preprint arXiv:1704.04861, 2017.
8. Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324