Lightweight Deep Learning Model for Invoice Image Classification

Run Wang, Chonghu Cheng, Yang Meng, Jie Yang and Guan Gui
{1218012107@njupt.edu.cn, chengch@njupt.edu.cn, guiguan@njupt.edu.cn}

College of Telecommunications & Information Engineering
Nanjing University of Posts and Telecommunications
Nanjing 210003, China

Abstract. Deep learning based image classification usually needs a large amount of data and computing resources to achieve better classification results. Hence, image classification for small data sets has attracted more and more attention in recent years. we proposed a lightweight deep learning model in this paper, which can be used for image classification on small datasets. The experimental results show that this model can achieve high accuracy classification with small data sets.

Keywords: Image classification, deep learning, convolutional neural network, computer vision.

1 Introduction

In recent years, with the rapid development of deep learning [1], the application of deep learning has made breakthroughs in many aspects. For example, in the annual Image large-scale Visual Recognition Challenge(ILSVRC), the constantly updated deep convolution network model is used to classify 1000 categories of about 100,000 object images. When AlexNet won the 2012 ImageNet competition by an overwhelming margin [2], AlexNet only had 5 layers of convolutional layer. It developed to 19 layers of VGG-Net [3], 22 layers of GoogLeNet [4] [5], and ResNet [6], which created a new network structure. Its classification accuracy has been greatly improved every year.

The achievements of deep learning in recent years are mainly attributed to the significant improvement of data available for model training and the significant increase of computing resources (the continuous improvement of GPU performance) [7]. Along with the increasing depth of convolution model layer, that means need a large amount of data and computing resources, and deep learning in the practical application often need real-time and occupy less memory and computing resources of the lightweight model, such as hyperspectral images classification [8] [9], face recognition [10], water level observing [11]. Also a great deal of data for training model can be difficult to obtain, in order to solve this problem, how deep learning application on small data sets has attracted widespread attention.

As the depth of the model increases, the learning ability of the model does not increase. On the contrary, the deep model will produce higher training error rate than the shallow model. But it is not caused by overfitting. Because the neural network needs to propagate gradient continuously in the process of back propagation [12]，but when the network model becomes complex, the gradient will gradually disappear in the process of propagation, and optimization

EAI MOBIMEDIA 2020, August 27-28, Harbin, People's Republic of China

DOI 10.4108/eai.27-8-2020.2295049

will become more difficult, which leads to the so-called "degradation" problem. He proposed ResNet is to solve this problem.

This paper proposes a method of invoice image classification based on ResNet, which solves the problem of "degradation" caused by the increase of model depth. ResNet got the first place in the classification task of ILSVRC2015. However, when the data set is small, the model is still too large. Based on resnet-18, this paper modified the model and train in a data set with only 400 images for each type of invoice, and finally achieved a better classification effect in the test set.

2 DESCRIPTION OF THE MODEL STRUCTURE

In this section, the original structure of ResNet and the improved ResNet structure are described and analyzed in detail. The original model structure of ResNet is shown in Fig.1.

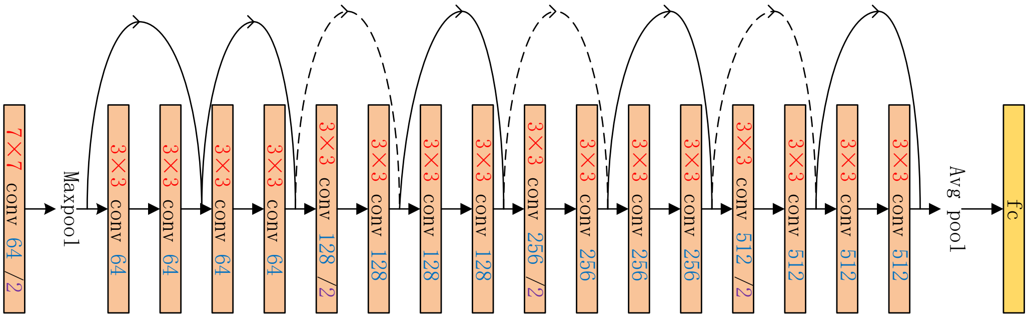


Fig. 1. Structure of ResNet.

As illustrated in Fig.1，ResNet-18 contains 18 layers, which including 17 convolution layers and a full connection layer. The first convolution layer is followed by the Max pooling layer, and there is an average layer before the final convolution layer. The red number in the rectangle represents the size of kernel, the blue number represents the number of kernel, and the purple number represents the convolution stride. ResNet is different from the general deep learning model in that it does not stack the convolution layer directly.

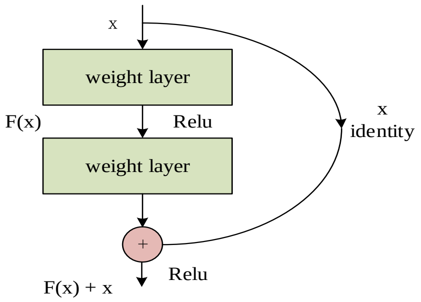


Fig. 2. A building block.

The curve above Figure 1 is the same as the curve in Figure 2 is called shortcut connection. It represents that the output H(x) is equal to the input x plus the F(x), When F(x) is inconsistent with the number of channels of x, it is represented by a dotted line, and a convolution operation needs to be added to make F(x) consistent with the number of channels of x. When F(x) is consistent with the number of channels of x, no operation is needed to connect directly with a solid line. Resnet-18 network performs well in many classification tasks, which is different from the simple stacking of convolution layer in traditional network. Shortcut connection is a revolutionary invention. Two layers of convolution make a building block. Such a structure can increase the depth of the network without causing degradation problems. With its unique model structure, the network ResNet has achieved convincing accuracy in many classification tasks. Nevertheless, the number of parameters contained in the 18-layer network still needs a lot of data for training. In the classified invoice task, there are only 400 invoice images of each category in the training data, which is far less than the number of images in ImageNet. It is difficult for the model to converge. If the model is not converge, the classification task cannot be well completed. Based on the resnnet-18 network, some improvements are made in this paper, so that this model has better classification effect than the original model and other models. Based on the existing advantages of ResNet, this paper modified the model parameters and structure to further improve the model performance. The structure is described in Fig. 3.

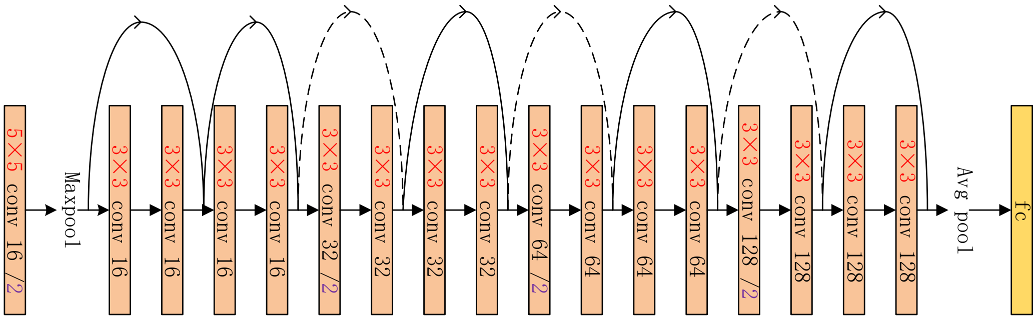


Fig. 3. Structure of proposed model.

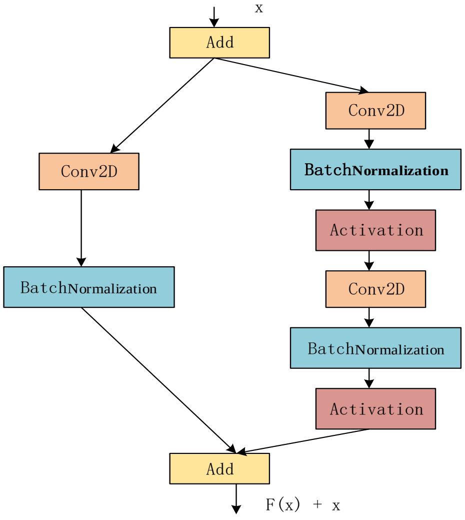


Fig. 4. Structure of proposed model.

ResNet solves the degradation problem caused by the deepening of neural network, but it is undeniable that it also has the problem of overfitting. As illustrated in Fig.4, each convolutional layer is followed by a Batch Normalization layer, this is the improvement of the original network. Batch Normalization(BN) can solve gradient vanishing with gradient explosion. The operator of BN can be described as follows [13].

$$\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \tag{1}$$

$$\sigma_B^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2 \tag{2}$$

$$\widehat{x_\iota} \leftarrow \frac{x_i-\mu_B}{\sqrt{\sigma_B^2+\epsilon}} \tag{3}$$

$$y_i \leftarrow \gamma\widehat{x_\iota} + \beta \equiv BN_{\gamma,\beta(x_i)} \tag{4}$$

Among these four equations, Eq. (1) calculate the mean of the samples in a mini-batch. Eq. (2) calculate the variance of the samples in a mini-batch. Eq.(3) calculates the normalized results, Eq.(4) trains the parameter γ and β ,and throw a linear transformation get a new y_i. During forward propagation, new distribution values are obtained from learnable γ and β parameters. In the case of back propagation, by taking a chain derivative can find γ and β and their associated weights.

The BN layer is followed by the activation layer. The activation function used here is Rectified Linear Unit (ReLU). The vanishing gradient is particularly obvious when the number of network layers is large, and it is one of the main obstacles to deepen the network structure. The gradient of the ReLU function is constant in most cases, which is helpful to solve the convergence problem of the deep network. Moreover, it has the biological principle, which is usually better than other activation functions in practice. The operator of ReLU function can be described as follows.

$$\mathrm{f}(x) = \max(0,x)$$

where x is the input vector and $\mathrm{f}(x)$is the output of ReLU as the activation function of neurons.With the introduction of activation function, the output of deep neural network is no longer a linear combination of input. The activation function can make the output of some neurons equal to 0, resulting in the sparsity of the network, reducing the interdependence of parameters and alleviating the overfitting problem.

3 EXPERIMENTAL RESULTS

3.1 Data set

In this invoice classification task, both the train set and the test set are photos taken by the mobile phone, with a total of 1200 images to training model. Instead of directly putting the data into the model for training, we first carry out the grayscale processing, then convert it into an array, and finally carry out the normalization processing. This is a supervised learning task, so each kind of picture needs to be labeled as "one hot". Within each epoch, 1,200 images are divided into two parts, 840 of which are used for training and 360 for verification. To test the fitting degree of the model, 42 images different from the training set and the verification set are used for test model. If the various invoices can be accurately classified then the fitting is better.

The data set includes three types of invoices, namely value-added tax , train , and taxi invoices. Data imbalance will lead to deviation in the process of training model. Therefore, it is necessary to ensure that each type of images in the train set accounts for 1/3 of the total number when conditions permit, this means that there are 400 invoices images for each type of invoice. Similarly, the 42 images in the test set also account for 1/3 of each type. Considering that the model is used to classify the images taken by the mobile phone, the image will have various angles, so the train set need to perform data augmentation. Data augmentation is achieved by rotating, flipping, scaling, color jittering and adding noise to the train image. The significance of data augmentation is to increase the robustness of the model, so that the model has better performance for image classification tasks. Three of the raw Image after grayscale processing are provided in Fig.5.

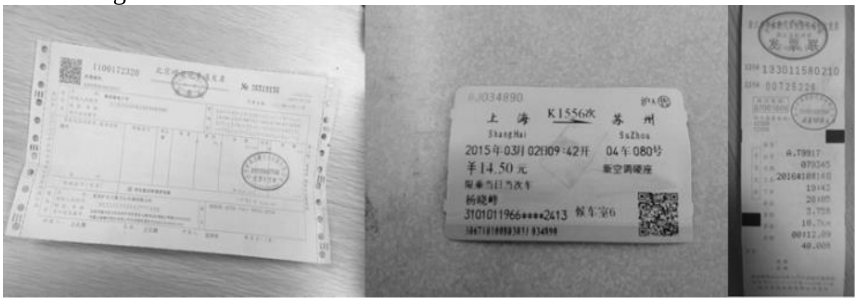


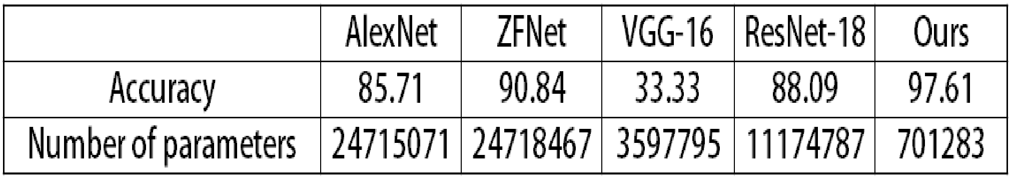
Fig. 5. Samples of the training images after grayscale processing

3.2 COMPARISION OF MODELS IN CLASSIFICATION ACCURACY

In order to fairly compare the classification accuracy of the proposed model and other models, all experiments were performed on the same equipment and in the same data set, and the same test set is used for model testing.

Compared with the original model, the improved model is already lightweight, so the experiment only compares the classification accuracy of the model. In this experiment, in addition to comparing the model proposed in this paper with the original model, it will also be compared with other models that have won the ImageNET competition, they are Alex Net, ZFNet and VGG-16. Table 1 shows the classification performance of each model in the test set and the amount of parameters for each model. The source of these data is obtained in the same experimental environment through the same steps.

Table 1. COMPARISON OF MODELS



	AlexNet	ZFNet	VGG-16	ResNet-18	Ours
Accuracy	85.71	90.84	33.33	88.09	97.61
Number of parameters	24715071	24718467	3597795	11174787	701283

Results illustrate that the ResNet has natural advantages, compared with Alex Net, ZFNet, VGG-16, ResNet has more layers, but fewer parameters. The model proposed in this paper is based on ResNet. This model has the least parameters, but its classification accuracy is the highest. Only one of the forty-two test images was misclassified. In summary, in the invoice image classification task with high classification accuracy requirement, the accuracy of classification is the first consideration. The model proposed in this paper can accurately classify the images in the test set. And an image that is not accurately classified in the test set, even a person can't determine which type of invoice belongs to. Therefore, this model is far superior to other models.

4 CONCLUSION

This paper proposes an invoice image classification model based on deep learning model ResNet, which can accurately classify value-add tax invoice images, train invoice images and taxi invoice images. In this paper, The kernel size and number of hidden neurons are modified based on the original model, and BN is added to 'shortcut'. The experimental results show that these modifications make the model perform better in the invoice classification task and perform better than other models, so the model is the best choice for invoice classification tasks. There are many types of invoices in the market. To better meet the actual needs, future research will increase the category of invoices and further optimize the model structure.

References

[1] D. Zhang, X. Han, and C. Deng, “Review on the research and practice of deep learning and reinforcement learning in smart grids,” CSEE J. Power Energy Syst., vol. 4, no. 3, pp. 362–370, Sep. 2018.

[2] Alex krizhevsky, Ilya Sutskever, Geoffrey. "ImageNet classification with deep convolutional neural networks. International Conference on Neural Information Processing Systems (NIPS), vol.1, pp. 1097-1105, 2012.

[3] X. Liu, M. Chi, Y. Zhang, and Y. Qin, “Classifying High Resolution Remote Sensing Images by Fine-Tuned VGG Deep Networks,” IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Valencia, Spain, 2018, pp. 7137-7140.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 2016, pp. 2818-2826.

[5] C. Szegedy et al., "Going deeper with convolutions," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1-9.

[6] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), LasVegas, NV, 2016, pp. 770-778.

[7] J. Wu, B. B. Gao, X. S. Wei, J. H. Luo, “Resource-Constrained Deep Learning: Challenges and Rractices,” SCIENTIA SINICA Informationis, vol.48, no.5, pp. 501–510, 2018.

[8] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, “Deep Feature Extraction and Classification of Hyperspectral Images Based on Convolutional Neural Networks,” IEEE Trans. Geosci. Remote Sens., vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[9] Z. Ma, Z. H. Tan, and J. Guo, “Feature selection for neutral vector in EEG signal classification,” Neurocomputing, vol. 174, no. PB, pp. 937–945, Jan. 2016.

[10] T. Zhou, S. Yang, L. Wang, J. Yao, and G. Gui, “Improved Cross-Label Suppression Dictionary Learning for Face Recognition,” IEEE Access, vol. 6, pp. 48716–48725, Aug. 2018.

[11] J. Pan, Y. Yin, J. Xiong, W. Luo, G. Gui, and H. Sari, “Deep Learning-Based Unmanned Surveillance Systems for Observing Water Levels,” IEEE Access, vol. 6, pp. 73561–73571, Nov. 2018.

[12] G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai, and Z. He, “Spatially supervised recurrent convolutional neural networks for visual object tracking,” in 2017 IEEE International Symposium on Circuits and Systems (ISCAS), Baltimore, MD, USA, 2017, pp. 1–4.

[13] Ioffe, Sergey, and C. Szegedy . "Batch normalization: accelerating deep network training by reducing internal covariate shift," International Conference on International Conference on Machine Learning, vol. 37, pages 448-456, july.2015.