Journal of Physics: Conference Series

Image Classification Algorithm Based on Improved AlexNet

Image Classification Algorithm Based on Improved AlexNet

Shaojuan Li, Lizhi Wang, Jia Li* and Yuan Yao

Department of Basic Sciences, Air Force Engineering University, Xi’an, China

*Corresponding author email: 34713439@qq.com

Abstract. Aiming at the problems that the traditional CNN has many parameters and a large proportion of fully connected parameters, a image classification method is proposed, which based on improved AlexNet. This method adds deconvolution layer to traditional AlexNet and classifies the images by full connection layer. Using Cifar-10 data set to test the classification algorithm. The results indicate that the method not only reduces the number of parameters and parameters proportion of the full connection layer, but also improves the classification accuracy compared with AlexNet.

Keywords: Image classification; Deep learning; Convolutional neural network; Feature extraction.

1. Introduction

Image classification is a popular study field in the computer vision field, and it could be applied to many fields such as video content recognition, face recognition [1], vehicle license plate recognition, etc. In the internet era, image data has increased dramatically, and image classification problems have become more and more important. Traditional image classification algorithms are mainly divided into two categories: spatial domain image classification algorithms [2,3] and transform domain image classification algorithms [4]. Images are mapped into the transform domain for classification by the transform domain algorithm, in the transform domain, but some images cannot be well extracted for classification, and the generality of this type of method is not good. Spatial domain-based image classification algorithms design various feature extraction methods in image space, which are used to classify the images, this type of method depends on the feature extraction of the image. Therefore, feature extraction is the difficulty of this method.

In recent years, image classification used machine learning algorithms widely and achieved good results. Deep learning is an important branch of machine learning, it is applied in computer vision [5], natural language processing, speech recognition [6]and so on, and significant progress has been made [7]. Deep learning is to construct neural network architecture with multiple hidden layers, through layer-by-layer feature extraction to combine to form more abstract features. Compared with traditional artificial feature extraction, features extracted by deep learning is more representative and more expressive [8]. Traditional neural networks perform complex non-linear transformations through multiple hidden layers to fit functions, but often because of too many parameters, the obtained model is prone to over fitting, at the same time, when there are too many hidden layers, the gradient will disappear or the gradient will explode.

In 2012, the AlexNet model first applied deep learning to image classification, which made historical breakthroughs in convolutional neural networks and greatly drove the development of convolutional neural networks [9]. Compared the convolutional neural network with traditional neural networks, the convolutional neural network greatly reduced the network complexity, reduces the number of parameters, and effectively solves the problem of overfitting.

However, in traditional convolutional neural network, the parameters proportion about the full connection layer is relatively large. For example, in AlexNet, the parameters proportion about the full connection layer is 96.2%. To reduce the parameter proportion about the full connection layer and reduce the amount of calculation, a deconvolution operation was introduced after the convolution operation, which greatly reduced the parameter proportion about the full connection layer, the results show that the parameters number is reduced, the overfitting degree is reduced, and good results are achieved by this method.

2. Convolutional Neural Network

Convolutional neural network is a feedforward neural network, a feature extractor was introduced, which is consist of a convolutional layer and a pooling layer, and can effectively recombine the extracted low-level features to extract higher-level features. Automatic feature extraction reduces a lot of manual investment, which is faster and more efficient, in the meantime, the number of parameters is reduced due to the sharing of weights [10].

A convolutional neural network usually consists of a convolutional layer, a pooling layer, a full connection layer, and an output layer [11]. The convolutional layer is the core part of a convolutional neural network and consists of multiple convolutional kernels, different features can be extracted with different convolutional kernels. In traditional neural networks, each neuron must be connected to all neurons in the previous layer, and the amount of calculation is too large; each neuron in the convolution neural network only extracts the features from the local perception of the previous layer, which effectively reduces the number of parameters. The number of convolution kernels corresponds to the number of output feature maps, various features can be learned by multi-core convolution. To a certain extent, increasing the number of convolution kernels can obtain more features and improve the expression ability of the model. The deconvolution layer is also called transposed convolution, which is usually used to restore the extracted feature image to the original image, which is widely used in image restoration and super-resolution reconstruction. This paper uses the deconvolution layer for image classification, restores the features extracted by the convolution layer, continuously reduces the number of feature maps, and finally outputs through the full connection layers. Because of deconvolution operation, the number of output characteristic graph is greatly reduced compared with the output of convolution layer, which makes the number of nodes in the full connection layer decrease, thus reducing the proportion of parameters in the full connection layer.

The pooling layer is also known as down sampling layer. Usually, after the convolution layer, the multidimensional feature map output from the convolution layer is sampled to reduce the size of the feature map, so as to maintain the image distortion invariance such as displacement, scaling and rotation [12], and reduce the complexity of the network and alleviate over fitting to a certain extent. It is generally divided into mean pooling, maximum pooling and random pooling. The maximum pooling can better preserve the texture information of the image, while the mean pooling can effectively preserve the background information of the image, while the random pooling is between the two, the corresponding probability selection is made according to the element value of the sampling area [13].

The output layer usually uses the Softmax classifier, which is suitable to solve multi-classification problems, the calculation amount is relatively small, and the training speed is fast [14]. Assume that the function is defined as in equation (1):

$$\mathrm{h}_{\theta}(x^{(i)}) = \begin{bmatrix} p(y^{(i)}=1 \mid x^{(i)};\theta) \\ p(y^{(i)}=2 \mid x^{(i)};\theta) \\ \vdots \\ p(y^{(i)}=\mathrm{K} \mid x^{(i)};\theta) \end{bmatrix} = \frac{1}{\sum_{\mathrm{j}=1}^{K} e^{\theta_j^T x^{(i)}}} \begin{bmatrix} e^{\theta_1^T x^{(i)}} \\ e^{\theta_2^T x^{(i)}} \\ \vdots \\ e^{\theta_K^T x^{(i)}} \end{bmatrix} \tag{1}$$

among them θ is the model parameter, K is the number of categories, and the total classification probability of each category is 1.

The loss function corresponding to the Softmax classifier is shown in equation (2):

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} l\{y^{(i)} = j\}\right]\log\frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_i^T x^{(i)}}} \tag{2}$$

Where m is the samples number,$l\{the\ expression\ is\ true\} = 1$, $l\{the\ expression\ is\ false\} = 0$, update the value of θ by minimizing the loss function to predict the category of new samples.

3. Improved AlexNet Image Classification Algorithm

3.1. AlexNet Model

AlexNet has about 650,000 neurons with a total of 60 million parameters, compared to Lenet-5, the network scale has been greatly improved. In the 2012 ImageNet image classification competition, it has achieved a huge advantage of 11% higher accuracy than the second place.

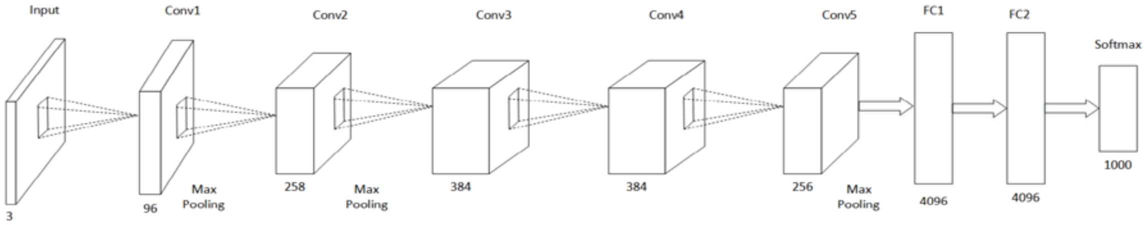


Figure 1. AlexNet model structure diagram.

As shown in figure 1, AlexNet is an 8-layer convolutional neural network, it is consisted of 5 convolutional layers and 3 full connection layers, of which the maximum pooling operation is performed after three convolutional layers. Different from previous neural networks, AlexNet uses ReLU as the activation function to instead of the traditional sigmoid and tanh functions. ReLU is a non-saturated activation function, which not only effectively improves the training speed of the model, but also better controls the problem of gradient disappearance and gradient explosion, it is easy to train a deeper network [15]. The form of ReLU function is shown in equation (3):

$$\mathrm{ReLU}(x) = \max(0, x) \tag{3}$$

In AlexNet, dropout is used to reduce the degree of over fitting, that is, training process of the model , neurons are stopped with a certain probability, thus the dependence on local nodes is reduced, and the generalization ability of the model is improved[16].But the introduction of large convolution kernel, on the one hand, increases the number of parameters, on the other hand, it makes it easy to lose local features in the feature extraction process; at the same time, the proportion of full connection layer parameters are relatively large, and the features extracted from convolution part have great influence on the results, so it is particularly important to enhance the proportion of convolution parameters.

3.2. Algorithm

Based on an in-depth study of the AlexNet model, a deconvolution layer is added to AlexNet. As shown in figure 2 is the designed network structure.

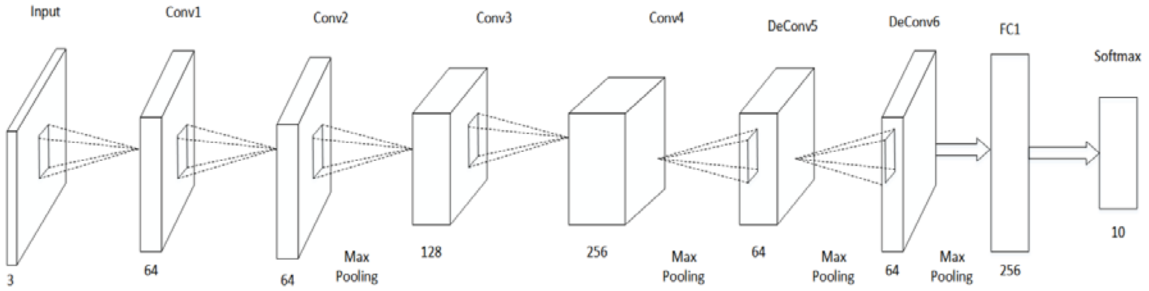


Figure 2. Schematic diagram of the algorithm structure in this paper.

The network is made of 4 convolutional layers, 2 deconvolutional layers, and 2 full connection layers. A 3×3 convolution kernel is used for feature extraction to avoid missing some detailed features. At the same time, the overlapped maximum pooling is used for down sampling to decrease the feature map size. The network structure detailed parameters are given in table 1.

Table 1. The network structure detailed parameters

Network layer	Enter	Convolution kernel size / moving step size	Number of feature maps	Activation function	Dropout	Pooling layer / moving step	Output
Conv1	28×28×3	3×3/1	64	ReLU	No	No	28×28×64
Conv2	28×28×64	3×3/1	64	ReLU	0.8	3×3/2	14×14×64
Conv3	14×14×64	3×3/1	128	ReLU	No	No	14×14×128
Conv4	14×14×128	3×3/1	256	ReLU	0.8	3×3/2	7×7×256
DeConv1	7×7×256	3×3/1	128	ReLU	0.5	3×3/2	4×4×128
DeConv2	4×4×128	3×3/1	64	ReLU	0.5	3×3/2	2×2×64
Fc1	256	-	-	ReLU	0.5	-	256
Output	256	-	-	Softmax	-	-	10

The biggest difference between this model and AlexNet is that the deconvolution layer is introduced between the convolution layer and the full connection layer. The feature maps number is simplified by deconvolution operation, and then the output is carried out through the full connection layer. Due to the simplification of the feature maps number, the input of the full connection layer is greatly reduced, and the node number in the full connection layer is greatly reduced, the parameter proportion of full connection layer in the whole network is reduced.

To reduce the model over fitting degree, increasing a penalty factor to the loss function to avoid that parameters are too large or too small, to keep the model relatively simple. If the loss function after regularization is $\tilde{J}$, then

$$\tilde{J}(\theta) = J(\theta) + \alpha\Omega(\theta) \tag{4}$$

as in equation (4), $\alpha \in [0, +\infty)$， if α is 0, there is no regularization, the larger α is, the greater the penalty is, and the larger the proportion of regularization is. In this paper, L2 regularization is used to enhance the model generalization ability, and the model weight is attenuated to approach 0, as in equation (5):

$$\tilde{J}(\omega) = J(\omega) + \frac{1}{2}\alpha\|\omega\|_2^2 \tag{5}$$

The batch normalizing (BN) layer is increased after the convolution layer to normalize the data, which not only speeds up the network convergence, but also helps to resolve the problems of gradient disappearance and gradient explosion [17]. BN layer normalizes the data by calculating the mean and variance of input samples, and introduces two parameters, γ and β, to recover the learned feature distribution by training these two parameters continuously, the formula is shown in equation (6) and (7): among them μ is the sample mean, σ is the sample variance, ε is a constant close to 0.

$$\hat{x}_i = \frac{x_i - \mu}{\sqrt{\sigma^2 + \varepsilon}} \tag{6}$$

$$y_i = \gamma \hat{x}_i + \beta = \mathrm{BN}_{\gamma,\beta}\left(x_i\right) \tag{7}$$

4. Experiment and Analysis

4.1. Selection of Data Set

As shown in figure 3, the data set Cifar-10 is used, the size of the image is 32×32, and it is divided into 50,000 training samples and 10,000 test samples, including 10 kinds of objects, such as aircraft, car, boat, truck, cat, dog, bird, deer, horse and frog. To improve the model generalization ability, the training samples were randomly cut into 28×28 images and expanded 6 times to 300,000 training samples, and then the data is standardized.

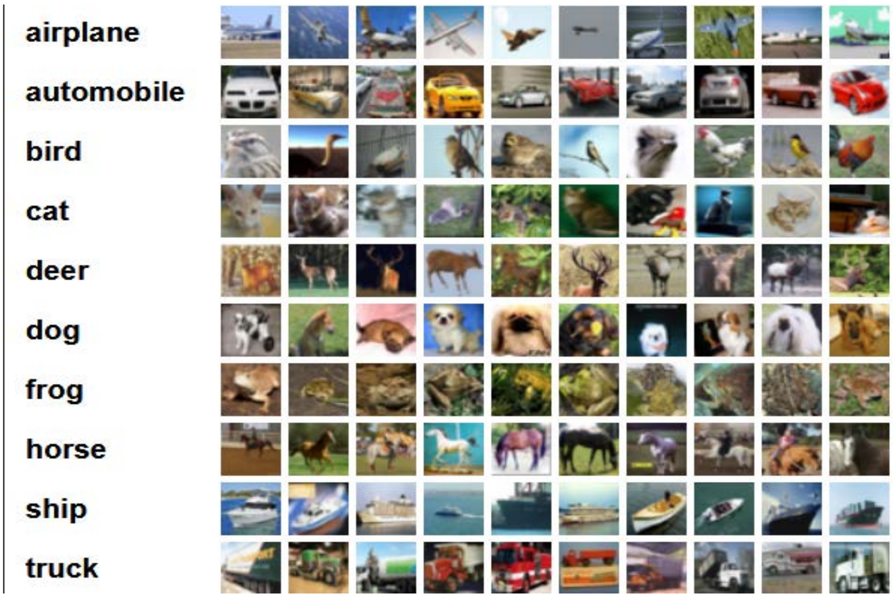


Figure 3. Cifar-10 partial data.

4.2. Analysis of Results

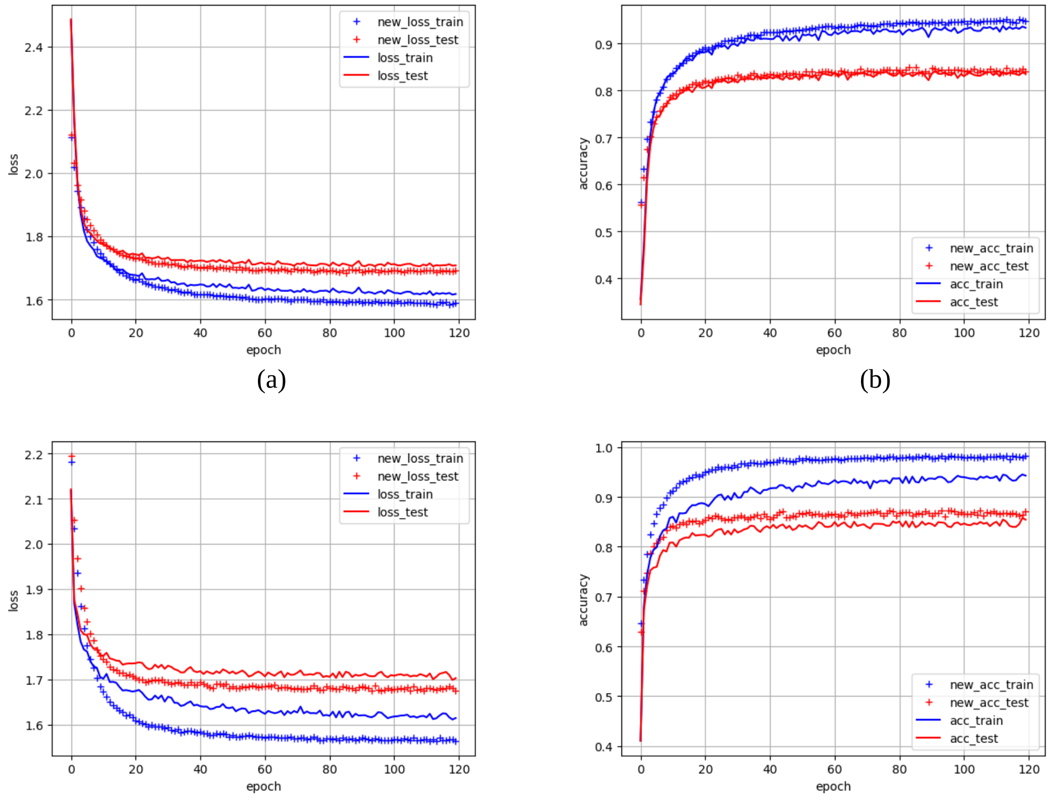
In the process of training, each batch size is 128, a total of 120 cycles of training, and each training cycle, the order of training samples will be disordered, to avoid in each training cycle, data loss function, accuracy distribution is roughly the same.

The comparison results are shown in table 2 between the used model and the AlexNet parameters. The model used is new_AlexNet_2, and AlexNet_1 and AlexNet_2 are obtained by reasonable adjustment of parameters for small-scale image classification according to the AlexNet model, the difference is that the feature maps number of the AlexNet_2 model is twice as much as AlexNet_1; based on their respective models, new_AlexNet_1 and new_AlexNet_2 are compared with AlexNet_1 and AlexNet_2, and deconvolution layers are introduced between the convolution layers and the full connection layers. The table 2 shows that when the feature maps number is constant, the deconvolution operation introduced into the convolution layer and the full connection layer, it will reduce the total parameters number greatly and reduce the parameter proportion of full connection layer in the model.

Table 2. Comparison of model parameters.

Network name	Is there a deconvolution layer	Number of feature maps	Total number of parameters	full connection layer parameter proportion
AlexNet_1	no	Halved	2211k	95.4%
new_AlexNet_1	Have	Halved	148k	12%
AlexNet_2	no	constant	8819k	95.4%
new_AlexNet_2	Have	constant	586k	11.6%

The experimental simulation results are shown in figure 4. The abscissa is the training cycle, and the ordinate is the loss function or the model accuracy, the loss function is smaller, the higher the model accuracy, and the more reliable the prediction result of the model. Figure (a) and figure (b) show that the experimental results of AlexNet_1 and new_AlexNet_1, in the case of halving the feature maps number. It can be seen that after the introduction of the deconvolution layer, the accuracy rate has been improved to a certain degree, and the test set accuracy rate is about 1% higher. Figure (c) and figure (d) show the experimental data of AlexNet_2 and new_AlexNet_2 under the condition that the feature map is unchanged, through comparison, it can be seen that the feature maps number has doubled, and the accuracy rate has been significantly improved, the test set accuracy is about 3% higher.



(a)

(b)

(c)

(d)

Figure 4. Performance comparison between models.

The results show that the used method effectively reduces the model parameters number, simplifies the model scale, not only increases the network training speed, but also makes the accuracy rate improved significantly. When the convolution layer features number is large, the effect is obvious, and the accuracy rate reaches 87.2% finally, which is improved by about 3%.Therefore, it has great significance to improving the model accuracy by reducing the parameters number and reducing the parameters proportion of the full connection layer in the model.

5. Conclusion

Based on AlexNet, the deconvolution layer is introduced between the convolution layer and the full connection layer, which reduces the parameters proportion of the full connection layer, and simplifies the parameters number of the model, it not only increases the network training speed, but also ensures the model classification accuracy. During the subsequent work, we can improve the convolution layer by dividing a 3×3 convolution operation into 3×1 and 1×3 convolution operations. At the same time, we use 1×1 convolution kernel to realize the operation of increasing and decreasing dimensions, so as to further reduce the network parameters number.

Acknowledgments

The work was supported by Research Fund Department of Basic Sciences at Air Force Engineering University (JK2019210).

References

[1] Prithaj B, Kumar B A, Avirup B, et al. Local Neighborhood Intensity Pattern – A new t exture feature descriptor for image retrieval[J]. Expert Systems with Applications, 2018:S0957417418303919-
[2] Xie Dong, Zhang Xing, Cao Renxian. Islanding Detection Technology Based on Wavelet Transform and Neural Network [J]. Proceedings of the CSEE, 2014 (4)
[3] Cheung B. Convolutional Neural Networks Applied to Human Face Classification[C]// International Conference on Machine Learning & Applications. IEEE Computer Society, 2012
[4] Yang D W, Wu H. Three-dimensional temperature uniformity assessment based on gray level co-occurrence matrix[J]. Applied Thermal Engineering, 2016, 108:689-696
[5] Lu Hongtao, Zhang Qinchuan. A Review on the Application of Deep Convolutional Neural Networks in Computer Vision [J]. Data Acquisition and Processing, 2016, 31 (1): 1-17
[6] Hinton G , Deng L , Yu D , et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[J]. IEEE Signal Processing Magazine, 2012, 29(6):82-97
[7] Yin Baocai, Wang Wentong, Wang Lichun. A Review of Deep Learning Research [J]. Journal of Beijing University of Technology, 2015 (1): 48-59
[8] Guo Y, Liu Y, Oerlemans A, et al. Deep learning for visual understanding: A review[J]. Neurocomputing, 2016, 187(C):27-48
[9] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2014, 115(3):211-252
[10] Li Yandong, Hao Zongbo, Lei Hang. Review of Convolutional Neural Network Research [J]. Journal of Computer Applications, 2016, 36 (9): 2508-2515
[11] Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks[J]. 2013
[12] Scherer D, Muller A, Behnke S. Evaluation of Pooling Operation in Convolutional Architecture for Object Recognition[C]//20 International Conference on Artificial Neural Networks (ICANN) .Spring, 2010:92-101
[13] Liu Wanjun, Liang Xuejian, Qu Haicheng. Study on learning performance of convolutional neural networks with different pooling models [J]. Journal of Image and Graphics, 2016, 21 (9): 1178-1190
[14] Krizhevsky A, Sutskever I, Hinton G E . ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. South Lake Tahoe , NV:MIT,2012: 1097-1105

[15] Hinton G E. Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair[C]// International Conference on International Conference on Machine Learning. Omnipress, 2010

[16] Hinton G E , Srivastava N , Krizhevsky A , et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2012

[17] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]// International Conference on International Conference on Machine Learning. JMLR.org, 2015