A Survey of Image Classification Algorithms based on Convolution Neural Network

Ruofan Mo
Shanghai New York University, China
rm3265@stern.nyu.edu

Abstract. With the deep learning (DL) sweeping the world. Traditional image classification methods are difficult to process huge image data and cannot meet people's requirements for image classification accuracy and speed. The image classification method based on convolutional neural network (CNN) breaks through the bottle neck of traditional image classification methods and becomes the mainstream algorithm of image classification at present, how to effectively use convolutional neural network to classify images has become a hot research topic in the field of computer vision at home and abroad. Convolutional neural network (CNN) has performed well in image classification and segmentation, target detection and other applications, and its powerful feature learning and feature expression capabilities are increasingly respected by researchers. However, CNN still has a few problems, such as incomplete feature extraction and overfitting of sample training. In view of these problems, after in-depth research on the application of convolutional neural network in image processing, this paper gives the mainstream structure model, advantages and disadvantages, time/space complexity, problems that may be encountered in the model training process and corresponding solutions used in image classification based on convolutional neural network. Through the overview of the research status of CNN model in image classification, it provides suggestions for the further development and research direction of CNN.

Keywords: Deep Learning; Convolutional Neural Network; Image Classification; Feature Extraction.

1. Introduction

Image classification is a key research field in computer vision, in which target classification and target location are important criteria for image classification. At present, convolutional neural network (CNN) is developing rapidly and has become the mainstream research method of image classification. CNN extracts a large amount of feature information from images through multi-layer neural network and classifies objects. It has strong robustness and generalization ability and has good application prospects in various fields of Internet applications, such as aerial remote sensing and marine remote sensing image analysis and face recognition. The early image recognition and classification technology mostly took people as the object of design features. For different recognition scenarios, the features of most applications need to be manually recognized by corresponding experts. Its principle mainly depends on the prior knowledge of the designer, and then manually codes according to specific data types and domain characteristics. In this way, it will be difficult to process massive data, and there are bottlenecks such as extremely low efficiency. In addition, the design of artificial features only supports a limited number of parameters, and the extracted features will directly affect the performance of the system, which may lead to unsatisfactory experimental results. In the age of big data, it is not appropriate to extract features manually. The new method based on deep learning and CNN can effectively reduce the task of developing and optimizing new feature extractors by automatically extracting and self-learning features.

2. Development of CNN

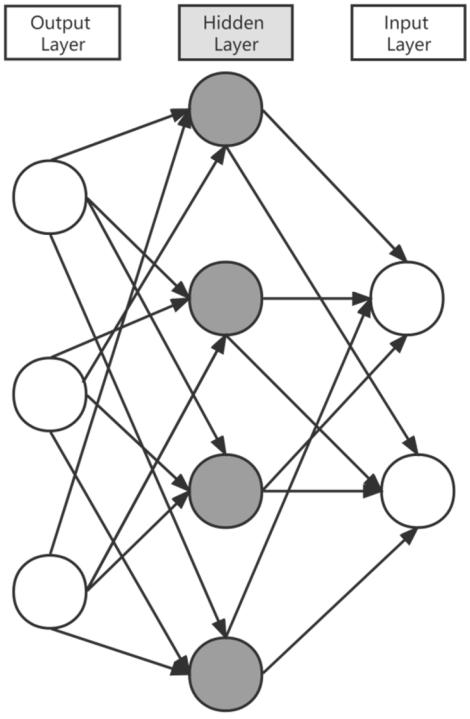
In the 1980s, because the multi-layer perceptron model was proposed, the computer showed excellent processing ability in digital recognition. However, due to the limitation of the computer ability, especially the processing ability of CPU, storage and other resources at that time, the data that can be processed is small in scale and the model expression ability is average, and usually cannot

handle complex picture problems. In 2006, Hinton proposed a layer-by-layer pre-training algorithm for the network model. By increasing the number of layers of the artificial neural network, the artificial neural network with multiple hidden layers has a strong feature learning ability. They reconstruct high-dimensional input vectors by training a multi-layer neural network with a small center layer, and effectively reduce the difficulty of deep level training of the neural network by coding dimensionality reduction. In addition, other researchers used support vector machines to overcome some of the difficulties encountered in training deep CNN. After that, the concept of deep learning and the rapid development of CNN received extensive attention from researchers.

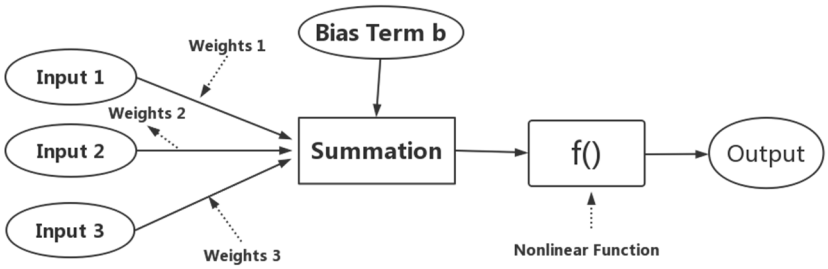
2.1 Convolutional Neural Network (CNN)

The most prominent feature of convolutional neural network (CNN) compared with general neural network is the addition of cony layer, normalized layer, activation layer and pooling layer. Other hierarchical structures are still consistent with general neural network. The flow process of data in the convoluted layer can be illustrated in Figure 1: suppose an RGB color image (5 × 5) Enter the convolution layer, and the numerical value in brackets represents the resolution. Then the corresponding input is no longer three values, but three-pixel matrices corresponding to the three-color channels of the color image. Therefore, let the three-color pixel matrices be x_1, x_2 and x_3 respectively, and their sizes are 5 × 5. The weight value of the convolutional neural network is no longer a value. It is generally a matrix smaller than the size of the input pixel matrix. Its action process on the input image is consistent with the filter convolution process in image processing, so it is called the convolution kernel. Let the convolution kernel size be w1, w2 and w3 respectively, and w represents (2 × 2) Weight matrix. The non-linear function is represented by G (·), the Bias-matrix is b, and its output is set as the pixel matrix y, which is generally equal to the size of the input image matrix. Therefore, the data flow process in the convoluted layer can be expressed as follows:

$$y=g(w_1 \cdot x_1+w_2 \cdot x_2+w_3 \cdot x_3 \cdot w_3+b)$$



(a) Classical neural network structure



(b) Data flow in neurons

Fig 1. Neural network structure

The biggest feature of the convolution layer is that it uses the parameter sharing mechanism. The weight of the convolution kernel is obtained through training, and the weight of the convolution kernel will not change in the process of convolution. This shows that we can extract the same features at different positions of the original image through the operation of a convolution kernel. To put it simply, the characteristics of the same target at different positions in an image are basically the same. The parameter sharing mechanism greatly reduces the number of training parameters, reduces the risk of overfitting, and improves the generalization ability of the model.

2.2 Active Layer

In 1943, based on the physiological characteristics of neurons, psychologist McClloch and others established a mathematical model of a single neuron, in which the concept of activation function was mentioned. The application of activation function increases the non-linearity of neural network model. The commonly used activation functions are linear rectification unit (ReLU), random linear rectification unit (RReLU), exponential linear unit (ELU), etc. ReLU is one of the most significant unsaturated activation functions, as shown in Figure 2. Its mathematical expression is as follows:

$f(x) = \max(0, x)$

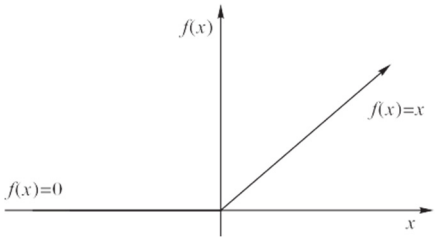


Fig 2. ReLU function

Although discontinuity at ReLU 0 may impair reverse propagation performance, studies have shown that ReLU is more effective than the Sigmoid and tanh activation functions.

2.3 Normalized Layer

Gradient descent is a simple method for training neural networks, but parameter selection needs to be done artificially, which results in a lot of time wasted by researchers in uncertain parameter adjustment. In 2015, the Google team proposed the idea of a Batch Normalization (BN). By using this method, researchers can choose a higher learning rate, so that the training speed of the model increases rapidly, and at the same time, the model has fast convergence.

2.4 Pooling Layer

The pooled layer is generally connected to the continuous convolution layer and the input is downsampled. There are many ways to downsample, such as maximum pooling, average pooling, and so on. Maximum pooling is the area corresponding to the filter size on the image, in which the

maximum value of pixel points is taken to obtain the feature data. Generally speaking, the feature data obtained by this method better preserves the texture characteristics of the image. Average pooling is to get the feature data by averaging all non-zero pixels in the above area. This method better preserves the extraction of image background information. It is important to note that the selected pixel points in average pooling do not contain 0 pixels. If 0-pixel points are added, the denominator will be increased and the overall data will be lower (large area of O pixels is not conducive to feature extraction). An example of average pooling is given below to illustrate the downsampling process, as shown in Figure 3, by entering a 4 × 4 Image Matrix, a 2x2 sliding filter is constructed to slide on the image matrix with a step of 2. The purpose of this sliding filter is to calculate the average of the pixels in its filter range, and ultimately to downsample the original pixel matrix to 2 × 2-pixel matrix. The role of this layer is also obvious. On the one hand, it reduces the dimensionality of the image matrix (similar to the principal component analysis pull) and reduces the computational load required by the model. On the other hand, invariance is achieved, including scale invariance, shift invariance, and rotation invariance.

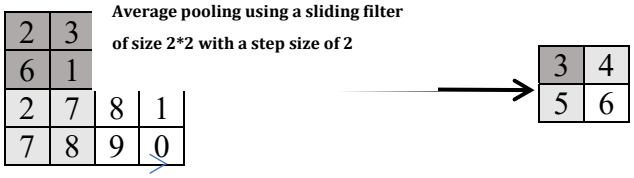


Fig 3. Downsampling Process

3. Image Classification based on Convolution Neural Network

CNNs based on deep learning can be used for image recognition and classification. This method automatically learns features from a large amount of data to improve the performance of the pattern recognition system. Most of the current methods of conventional image classification network directly use the common deep convolution network to directly classify images, such as AlexNet, VGG, GoogleNet, ResNet and MobileNet. This section first gives the common models for image classification based on convolution neural network, then gives an analysis of the advantages and disadvantages of the convolution neural network models commonly used in image classification, time/space complexity, problems that may be encountered in model training and corresponding solutions, as well as the future development direction and trend of the models in image classification tasks.

3.1 LeNet-5

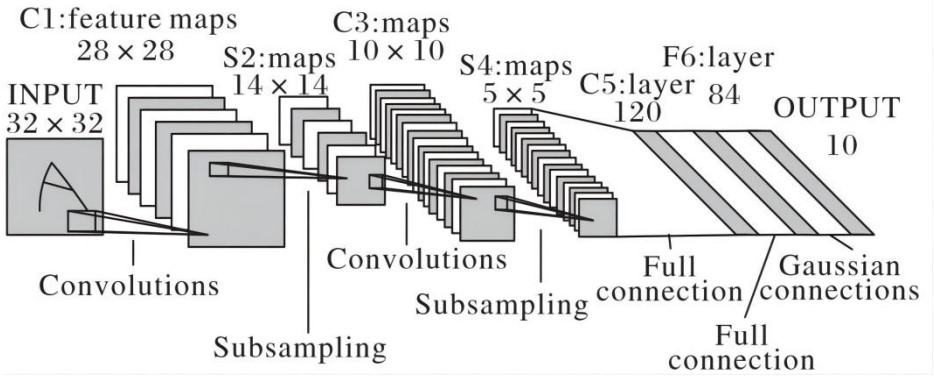


Fig 4. Structure of LeNet-5 model

LeCun was the first CNN method in the history, LeNet-5, proposed in 1998 and improved to recognize handwritten numbers. LeNet-5 defines the basic structure of a CNN and can be regarded

as the ancestor of a convolution network model. LeNet-5 uses fewer parameters to extract similar features in multiple locations, which not only reduces the number of learning parameters, but also automatically learns features from the original pixels. It was able to classify handwritten numbers without minor distortion, which many banks at the time used to identify handwritten numbers on checks. The LeNet-5 network model has seven layers, which are C1 convolution layer, S2 pooling layer, C3 convolution layer, S4 pooling layer, C5 convolution layer, F6 full connection layer and output layer, as shown in Figure 4.

3.2 AlexNet

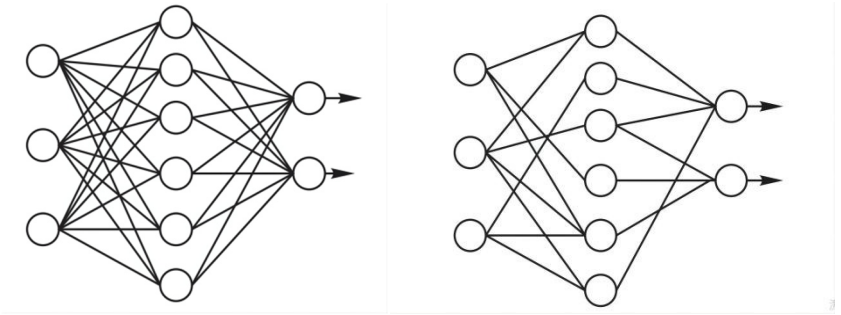
AlexNet, known as the first modern deep convolution network model, first used many modern deep convolution network technologies. It is a model proposed by Hinton Research Group in Large Scale Visual Recognition Challenge in 2012. They achieved their best results of the year, ranking first with a 15.3% error rate, which is 10.8 percentage points lower than the second algorithm (26.1%). The main contributions are as follows:

1) The first parallel training using multiple GPUs overcomes the hardware limitations of deep CNN architecture learning ability;

2) ReLU is used as the non-linear activation function to alleviate the problem of gradient disappearance and improve the convergence speed of the network model.

3) Use the Dropout layer to prevent overfitting. Figure 4 shows the difference between the full connection layer and the Dropout layer.

4) Enhance the learning ability of CNN through parameter optimization strategies.



Fully connected layer Dropout layer

Fig 5. Principle model of fully connected layer and Dropout layer

3.3 GoogLeNet Model

The network involves a total of 5M parameters, ILSVRC2014 champion network. The main feature of this model is the introduction of the Inception module, which has four branches as shown in Figure 6. The first branch convolutes the input by 1x1, which can organize information across channels and improve the expressive ability of the network. The second branch uses 1 first × 1 convolute, then connect 3 × 3 Convolution, equivalent to two characteristic transformations; The third branch is similar, starting with 1 × 1 convolution, then join 5 × 5 Convolution; The last branch is 3 × 3 Use directly after maximum pooling 1 × 1 Convolution. This Inception module dramatically improves the efficiency of parameters because, in general, the convolution layer improves expression by increasing the number of output channels, but the side effects are increased computational effort and overfitting. Each output channel corresponds to a filter, and the same filter shares parameters. Only one type of feature can be extracted, so only one type of feature processing can be done for an output channel. This model allows information to be combined between output channels, so the effect is obvious. The module also uses 1 × 1 convolution kernel reduces the dimension of the input and greatly reduces the number of parameters. GoogLeNet is much deeper than the previous network model, reaching an unprecedented 22 layers. Because its parameters are only 1/12 of Alexnet's, the computational complexity of the model is greatly reduced, but the accuracy of image classification

has risen to a new level. Although the GoogLeNet model hierarchy reaches 22 levels, it is extremely difficult to go deeper because as the model hierarchy gets deeper, the problem of gradient diffusion becomes more serious, making the network difficult to train.

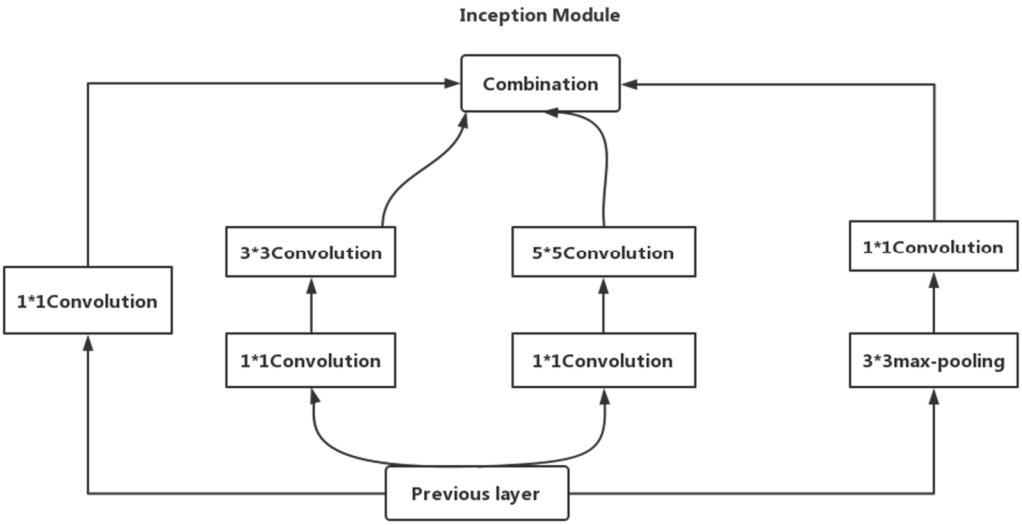


Fig 6. Inception module

3.4 VGGNet Model

This model is developed from the AlexNet model and mainly modifies the following two aspects: (1) Using several convolution layers with small filters instead of one with large filters, that is, the convolution layer uses a smaller convolution core but increases the depth of the model; (2) multi-Scale training strategy is used. Specifically, first, the original image is scaled equally to ensure that the short edge is larger than 224. Then, 224x224 windows are randomly selected on the processed image, because the scale of the object varies, this training strategy can better identify objects. These two improvements are very helpful to the learning ability of the model. However, the network uses too many parameters and the training speed is slow. Further research can continue to optimize on this issue.

3.5 ResNet

From the experience of VGG-19 and GoogLeNet models, the more network layers there are, the richer the features that can be extracted to different levels, and the deeper the features extracted, the more abstract the features are, the more semantic information they have. However, it has been proved that the training effect of deep network training model is worse with the increase of network model layers. Suppose that in a deep network, a non-linear unit (which can be one or more layers of convolution) is expected $f(x,m)$ To approximate an objective function $H(x)$ so as to divided the objective function is into two parts: the identity function and the residual function.

Where:

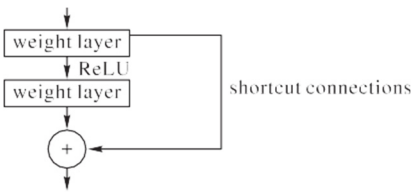


Fig 7. Residual module structure

$H(x)$ is the residual module; x is Identity Function, $f(x,m)$ is Residual Function. The principle is shown in Figure 7.

3.6 MobileNet

Embedded devices cannot use complex and bulky models. How to study small and efficient convolution network models is very important. The Google team launched MobileNet in 2017, a lightweight CNN focused on mobile or embedded devices. Then, in 2018, Google introduced MobileNet V2, and in 2019, Google introduced MobileNet V3. Compared with traditional CNN, MobileNet V3 significantly reduces the number of model parameters at the expense of small accuracy. Test results on ImageNet show that the accuracy is reduced by 0.9% compared to the VGG-16 model, but its model parameters are only 3.1% of the VGG-16 model. MobileNet uses a deep detachable convolution layer, as shown in Figure 8, where Depthwise Convolution is applied to each channel of the signature graph, followed by point-by-point 1 × 1 Convolution to reduce the amount of calculation and model parameters.

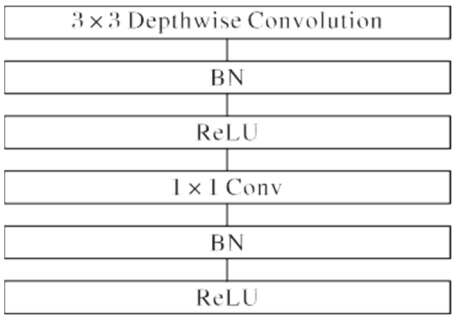


Fig 8. Deep separable convolution layer structure

4. Summary

Using CNN to process images is currently one of the hot research directions. This paper discusses some models of CNN. However, CNN still faces the following problems: 1) Theoretically, a key issue in applying CNN to new fields is that it requires a lot of prior experience to select appropriate hyperparameters, which are internally dependent, making their adjustment costs especially high. How to better handle these parameters remains a problem. 2) From the application point of view, the network model's ability to compute hardware, the dependence of large-scale data, and the dynamic adaptability of training model are all the problems that need to be solved urgently. In addition, new research methods and ideas have emerged in the field of image recognition.

Although the convolution-based neural network method achieves good results for some simple image classification tasks, its performance for some complex image classification needs to be improved. Secondly, image classification is not only an independent task, but also the basis of many images processing tasks. The research on semi-supervised or even unsupervised image classification and overlapping image classification tasks has just started. How to make better use of convolution neural network in this field (such as combining capsule network, generation antagonistic network, etc.) is a hot spot in the future. Therefore, more in-depth research on image classification based on convolution neural network is needed, such as the structure of the Transformer model. Many researchers attempt to introduce the Transformer model into the image. For example, try to introduce Transformer's self-attention mechanism into the CNN, or replace the convolution block directly with the Transformer model structure. These attempts have also achieved good results, such as DeiT, Pyramid Vision Transformer, Swin Transformer networks, and so on. In addition to the above research ideas, the combination of in-depth learning and enhanced learning in the field of image classification is also one of the important research directions in the future.

References

[1] HUANG B, HE B Y, WU L N, et al. A deep learning approach to detecting ships from high-resolution aerial remote sensing images[J]. Journal of Coastal Research,2020,111(SI):16-20.

[2] FU K S, ROSENFELD. Pattern recognition and image processing[J]. IEEE Transactions on Computers, 1976, C-25(12):1336-1346.

[3] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science,2006,313(5786):504-507.

[4] HE K M, ZHANG X Y, REN S Q, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2015,37

[5] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE,2016:770-778.

[6] HOWARD A G, ZHU M L, CHEN B, et al. Mobile Nets: efficient convolutional neural networks for mobile vision applications [EB/OL]. (2017-04-17) [2021-06-20]. https://arxiv. org/pdf/1704. 04861. pdf.

[7] ZHANG L, WANG X S, YANG D, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation[J]. IEEE Transactions on Medical Imaging, 2020,39(7):2531-2540.

[8] Kong Lingjun, Wang Qiwen, Bao Yunchao, etc. A review of medical image segmentation based on in-depth learning [J]. Radio Communications Technology,2021,47(2):121-130. (KONG L J, WANG Q W, BAO Y C, et al. A survey on medical image segmentation based on deep learning[J]. Radio Communications Technology,2021,47(2):121-130.)

[9] Tian Jin, Yuan Jiazhen, Liu Hongzhe. Lane line detection and adaptive fitting algorithm based on instance segmentation [J]. Computer application, 2020,40(7):1932-1937(TIAN J, YUAN J Z, LIU H Z. Instance segmentation-based lane line detection and adaptive fitting algorithm[J]. Journal of Computer Applications, 2020,40(7):1932-1937)

[10] Fan Wei, Liu Ting, Huang Rui, etc. Image Instance Segmentation Method Assisted by Low Level Features of Convolution Neural Network [J]. Computer Science,2020,47(11):186-191(FAN W, LIU T, HUANG R, et al. Low-level CNN feature aided image instance segmentation[J]. Computer Science, 2020,47 (11): 186-191.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY: Curran Associates Inc. 2017:6000-6010.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale [EB/OL]. (2021-06-03) [2021-06-20]. https://arxiv. org/pdf/ 2010. 11929. pdf.

[13] TOUVRON H, CORD M, DOUZE M, et al. Training dataefficient image transformers & distillation through attention[C]// Proceedings of the 38th International Conference on Machine Learning. New York: JMLR. org,2021:10347-10357.

[14] XU B, WANG N Y, CHEN T Q, et al. Empirical evaluation of rectified activations in convolutional network [EB/OL]. (2015-11- 27) [2021-06-20]. https://arxiv. org/pdf/1505. 00853. pdf.

[15] CLEVERT D A, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by Exponential Linear Units (ELUs)[EB/OL]. (2016-02-22) [2021-06-20]. https://arxiv. org/ pdf/1511. 07289. pdf.

[16] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models [C/OL]// Proceedings of the 30th International Conference on Machine Learning. [2021-06- 20. https://ai. stanford. edu/~amaas/papers/relu_hybrid_icml201 3_final. pdf.