Survey of image classification based on deep learning

Yiming Luo
Department of Computer Technology
Zhengzhou University
Henan Zhegnzhou,China
ieluoyiming@gs.zzu.edu.cn

***Abstract*—Image classification is an important basic research direction in computer vision. It has a wide range of applications, not only in image classification management and information extraction, but also in object detection, face recognition, image classification, and handwriting font recognition. Traditional machine learning methods, such as k nearest neighbors and Support vector machine, face the requirements of processing efficiency, performance, and intelligence in the case of increasingly complex real-world applications. This year, a series of studies on deep learning, especially based on convolutional neural networks, have demonstrated the outstanding performance of neural networks in this field. This article will discuss several typical deep learning network structures this year.**

Keywords—Image classification, convolutional neural network, deep learning (CNN), Deep Learning (DL)

I. INTRODUCTION

Hinton et al. Published a paper in "science" in 2016, the main points of which are: (1) the multi-hidden layer artificial neural network has excellent feature learning ability; (2) the "deep layer by layer" can effectively overcome the deep nerves The difficulty of the network in training has led to the study of deep learning, and has also set off another wave of artificial neural networks. In the layer-by-layer pre-training algorithm of deep learning, unsupervised learning is first applied to the pre-training of each layer of the network. For training, only one layer of unsupervised training is used at a time, and the training result of this layer is used as the input of its next layer. Finally, the supervised learning (BP) algorithm and chain rule are used to fine-tune the pre-trained network[1].

Due to the continuous development of computer computing performance and storage technology, convolutional neural networks have achieved great success in the field of image and video recognition, such as ImageNet datasets, large-scale distributed GPU clusters. Various competitions have also promoted the development of this field. In the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) challenge, many excellent deep visual recognition networks have appeared.

The rest of the paper is organised as follos. In Sect.2,we describe the basic structure of Convolutional neural network and components of a normal CNN.The detail of multiple model structures based on CNN are then presentd in Sect.3, and several commonly used data sets in Sect.4.Sect.5 will incroduce some tricks during the training of deep learning.

II. STRUCTURE OF CNN

Deep neural networks include a variety of network structures, such as Recurrent neural network, Graph neural network, and Convolutional neural network. CNNs can extract features in a certain area of the image because of its filter with a convolution window. It is widely used in computer vision tasks and has become a research focus in the field of image recognition.

A. Basic Network stucture

Most commonly used CNNs consist of an input layer, convolutional layers, pooling layers, a classifier such as Sigmod for logical classification or Softmax for multi-class, and an output layer.

CNNs use a matrix to represent the pixel values of a image，such as a input image with the resolution of 224×224，The pixel of each position can be expressed as x_ij，where i represent the pixel value of the i-th row and j represent the pixel value of the j-th columns. CNNs use weight matrix W and bias B to storage the features。

The feature map in the convolutional layer can be calculated by equations as in:

$$x_j^{l-1} = f(u_j^l)$$
$$u_j^i = \sum_{i \in M_j} x_i^{l-1} * W_{ij}^l + b_j^l$$

After the image has passed through the convolutional layer, in order to reduce the amount of data and increase the receptive field of the network, the maximum map or average pooling is generally used to down sample the feature map. The fully connected layer can be regarded as a linear network, which fuses the features extracted by each convolutional layer. Then get an output of one-dimensional matrix for classification assignment., or multidimensional matrix for other tasks such as image denoising or population density estimation etc.

B. Pre-processing for Images

Some image recognition task datasets, such as the minist, cifar10, and cifar100 datasets, have been processed by a large number of researchers to form a clean image dataset for easy use by engineers. For most real-world applications or competitions, the dataset contains a variety of interferences, such as the Imagenet Dataset, which contains 1,000 different

types of original images. Other datasets also have different scenes, different shooting equipment, and different resolutions Rate, noise image, and rotation, crop, scale and noisy images, will cause serious interference to the network, and even fail to converge. Commonly used image preprocessing includes normalization and changing the channel of the input image. In order to increase the generalization ability of the model, in a specific network structure, a large number of experiments have proven that after artificial rotation, resize, crop, scale, and image chroma, contrast, Saturation-transformed images can achieve better performance. For different tasks, different pre-processing schemes are generally used, which requires the rich experience of the experimenters, and recent research has published an automatic pre-processing scheme called "AutoAugment"[2].

C. Convolutional layer

The core of the convolutional layer mainly consists of a conv filter with kernel size of c. The convolution window uses the step size s to slide on the input image, extract information at different positions to form a feature map, and use it as the input of the next layer. The convolution process is shown in the following figure. In order for a convolution kernel to perform a complete convolution of an image, it is usually necessary to fill in one or more pixel values around the image. A convolution kernel can perform convolution operations on various parts of an image, and each part shares a convolution kernel. Compared to a fully connected network, weight sharing can greatly reduce the amount of network parameters.

D. Activation function

For the linear feature map output by each convolution kernel, in order to enhance the representation ability of the model, it is generally necessary to go through a non-linear mapping, which can enable the model to form a better fitting effect in complex realistic image recognition tasks. The choice of different activation functions has a great impact on network performance and convergence speed. Improper activation functions can even lead to the phenomenon of gradient dispersion and gradient explosion, making the network unable to converge to the local best advantage through back propagation. The commonly used activation functions are Sigmod, tanh, ReLU, Leaky-ReLU. Different activation function formulas and function images are shown in the figure below. In the field of image recognition, the most popular activation function is the ReLU activation function or Leaky-ReLU activation function which solved the ReLU activation function. The gradient dispersion phenomenon may occur. In the NLP field, tanh and Sigmod activation functions are commonly used for long-term and short-term memory information. Filter.

E. Pooling layer

For the output of the convolution layer, you need to reduce the amount of data and increase the receptive field of the network through the pooling layer. Common pooling layers include maximum pooling and average pooling. The pooling layer uses a pooling similar to the convolution layer. The window slides on the feature image, and the maximum or average value of multiple feature values in the pooled window is extracted to represent the feature information in the area. By alternately using convolutional layers and pooling layers, highly abstract feature images can be obtained. However, recent research shows that the use of pooling layers will inevitably cause the loss of context information. Therefore, some tasks use dilated convolution kernels instead of convolution layers and pooling layers, which increases the network experience. It reduces the amount of data and cannot lose the context information of the image.

F. Fully connected layer

In a convolutional neural network, after multiple convolutional layers and pooling layers perform feature extraction on the input image, usually one or more fully connected layers are connected, and each neuron in the fully connected layer integrates the input features to form a whole. Feature information. However, there is a huge flaw in the fully connected layer. When there are too many neurons, a large number of parameters will be generated, and overfitting may be caused. Therefore, a dropout method is adopted for the fully connected layer, and a part of the neurons are randomly disconnected each time, suppressing the output of some neurons, and improving the generalization ability of the model. Recent research shows that using global maximum pooling or global average pooling can achieve a similar effect to the fully connected layer, but it will greatly reduce the amount of parameters, and there is a small trick that uses GAP and GMP at the same time, and two global pooling with Concat operation of the output can achieve better results.

III. CLASSIC NETWORK STRUCTURE

A. LeNet

LeNet's[3] excellent performance in handwritten fonts and character recognition has made people pay attention to the powerful performance of convolutional neural networks. LeNet consists of a 7-layer network, including 4 convolutional layers, 2 fully connected layers, and 1 output layer. The resolution of the input image is 32 * 32pixel. The size of the convolution kernel used in the convolutional layer is 5 * 5. , The step size is 1, and the Sigmod activation function is used for non-linear mapping; the pooling layer uses a 2 * 2 pooling window with a step size of 2; the output layer uses the RBF network connection method, assuming x is the previous layer Input, y is the output of RBF, then the output of RBF is calculated as:

$$y_i = \sum_j \left(x_j - w_{ij}\right)^2.$$

The value of I is from 0 to 9, the closer the value of the RBF output is to 0, the closer it is to i, that is, the closer to the ASCII encoding figure of i, indicating that the current network input recognition result is the character i.

Although LeNet has achieved good results in handwriting, in the application process of other image recognition fields, in addition to the difficult hardware to support, overfitting problems have also limited the application of convolutional neural networks.

B. VGGNet

The VGG[4] network achieved the first place in image classification in the ImageNet Challenge 2014. The core idea

of the VGG network is to achieve better performance by increasing the depth of the model. At the same time, a 3 * 3 convolution kernel is used instead of a 5 * 5 or 7 * 7 convolution kernel. In some structures, a 1 * 1 convolution kernel is also used to linearly transform the number of input channels. Experiments show that two 3 * 3 convolution kernels are similar in performance to 5 * 5 convolution kernels, but will reduce the number of parameters that need to be saved in the network. Using the ReLU activation function instead of the Sigmod activation function can avoid partial gradient dispersion or gradient explosion problems; in addition, an LRN layer is used to mitigate the occurrence of overfitting.

VGG networks increase the accuracy of image classification by increasing the depth of the network, but this increase is not unlimited. Too many layers will lead to network degradation. For example, when using the chain rule for back propagation, an excessively deep network The path will cause the gradient information of the later layers to not be transmitted to the previous layers. The final number of VGG layers is determined in two versions of 16 layers and 19 layers.

C. *InceptionNet*

In the ILSVRC 2014 competition, InceptionNet[5] achieved the first place with a top5 error rate of 6.67%, while VGGNet had an error rate of 7.3%. Deeper networks tend to get better accuracy, but they also produce a large number of parameters. The more complex the network needs to use more data, but currently it is very expensive to obtain a large amount of high-quality data for model training, and it requires more computing resources. Inception Net network uses global average pooling to replace the full connection of the last layer, which reduces the amount of parameters and avoids overfitting of the fully connected layer. Inception Net uses a structure that increases the width of the network, uses multiple 4 branches to extract features from the output of the previous layer, combines 1 * 1, 3 * 3, and 5 * 5 convolution windows to increase the network's adaptability to different size features.

The Inception network also uses Batch Normalization, which solves the problem of Internal Covariate Shift. The data in each batch is normalized using Equation 1, so that the output of each layer is normalized to the N (0,1) normal distribution. At the same time, to a certain extent, the BN layer also plays a regular role, which can prevent overfitting and accelerate the convergence of the network.

$$\widehat{x}^{(k)} = \frac{x^{(k)} - \mathrm{E}[x^{(k)}]}{\sqrt{\mathrm{Var}[x^{(k)}]}}$$

D. *ResNet*

In order to solve the problem of gradient dispersion when the network layer is too deep, He Kaiming et al. Proposed a ResNet[6]. The residual neural network won the championship in ILSVRC 2015 with a top-5 error rate of 3.57%. The residual neural network uses shortcut to directly map the low-level feature map to the high-level, and the output of the high-level is the low-level mapping and the superposition of the original image. The structure of the residual unit is shown in figure 1.

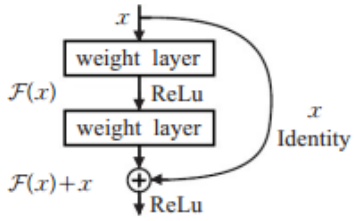

Fig.1 Structure of residual unit

After the residual structure is introduced, the existence of the output term x can make the overall gradient not vanishing.

E. *SENet*

Squeeze and Excitation Network (SENet[7]) won the championship with a top-error rate of 2.251% in ILSVRC 2017 and the last quarter of ILSVRC. The core idea of SENet is to learn the feature weights according to the loss through the network, so that effective feature maps get larger weights, and methods with smaller or ineffective feature maps have reduced weights to achieve better results. Embedding SE blocks in some original networks, although inevitably increasing some parameters and calculations, the trade-off effect is also acceptable. For the input x with the number of channels c1, after a series of transformations such as convolution, a feature map with the number of channels 2 is obtained, and the previous features are recalibrated through three operations. First, after the Squeeze operation, the input of H * W * C is compressed to the output of 1 * 1 * C, which is generally implemented using GAP, which is equivalent to this one-dimensional parameter to obtain the global field of view of H * W before, and the field of perception is larger. Second, use the Excitation operation to pass the 1 * 1 * C feature map through a fully connected layer, predict the importance of each channel, and obtain the importance of different channels before activating it on the corresponding channel of the previous feature map. Before proceeding. Figure 2 shows the process of applying SE block to ResNet.

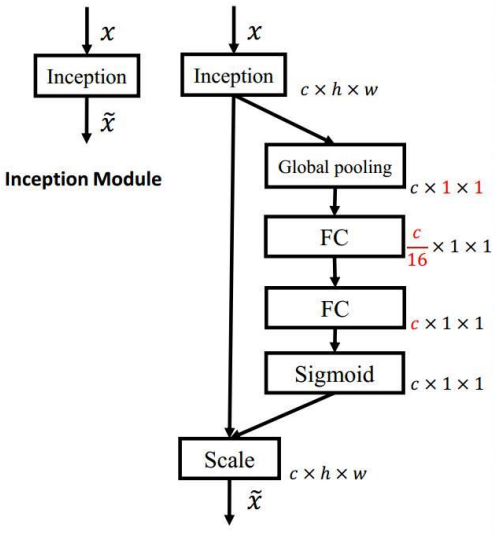

Fig.2 process of applying SE block to InceptionNet

F. *EfficientNet*

The above network and a large number of experiments have proven that scaling the depth, width, and number of channels of the network can improve the recognition accuracy of the network. But how to balance the three dimensions to achieve the best results, the compound model scaling algorithm proposed by the EfficientNet[8] solves the problem of determining the scaling of each dimension. Use AutoML to search for the correlation coefficients α, β, and γ using the grid search form, which represent the depth, width, and resolution of the zoom network as in:

$$\text{depth: } d = \alpha^{\phi}$$
$$\text{width: } w = \beta^{\phi}$$
$$\text{resolution: } r = \gamma^{\phi}$$
$$\text{s.t. } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$$
$$\alpha \geq 1, \beta \geq 1, \gamma \geq 1$$

Where Φ is a user-defined correlation coefficient

The Fig.3 shows the compartion of performance comparison of the efficientNet family with other networks. It is obvious that efficient obtains better accuracy and uses fewer parameters.

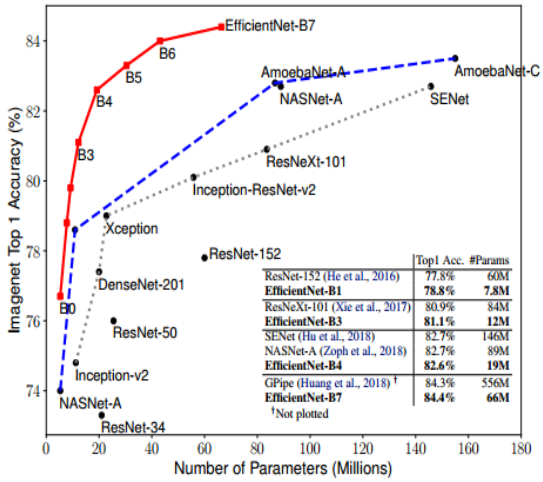

Fig.3 Deffrent Networks' preformance

IV. CONCLUSION

Deep learning has achieved great success in the field of image recognition, but the research on convolutional neural networks has just begun. There are four problems that need to be solved urgently: (1) The design of convolutional neural networks and the optimization of parameters are more dependent on There is no complete set of mathematical theories to improve by artificial experience, which has caused certain restrictions on the development of CNN. (2) In order to obtain better results, it is necessary to use a large amount of sample data for training during the network training process, which results in three problems. One is that it is very expensive to collect a large number of suitable samples, and the other is that a large number of samples will occupy more Large memory. Third, it requires huge hardware resources for training. (3) The current network continues to try to increase the width and depth of the network to achieve better results, and even some networks reach a depth of thousands of layers. Huge models are practically used, especially on ordinary devices. Impossible, so we need to optimize the design structure of the network and seek more efficient neurons and structural units. (4) The gradient descent method is commonly used in deep learning to make the network converge, but some practical problems cannot use this method. Therefore, it is also very important to find a better method for training network parameters.

REFERENCES

[1] Zhou Feiyan, Jin Linpeng and Dong Jun, "Review of Convolutional Neural Networks,卷积神经网络研究综述",CHINESE JOURNAL OF COMPUTERS, pp. 1230–1234, June 2017.

[2] Ekin D.Cubuk,Barret Zoph,Dandelion Mane,Vijay Vasudevan,V.Le, "AutoAugment:Learning Augmentation Strategies from data,",cs.CV, Apr 2019

[3] Yann Lecun,Y.Bengio,Patrick Haffner, "Gradient-Based learning Applied to Document Recognition,"Proceedings of the IEEE, December 1998.

[4] Karen Simonyan,Andrew Zisserman, "Very Deep Convolutional Networks for Large-scale image recognition,"ICLR, 2015.

[5] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, "Going Deeper with Convolutions," Computer Vision and Pattern Recognition, Sep 2014.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Squeeze-and-Excitation Networks," Computer Vision and Pattern Recognition, Dec 2015.

[7] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, Enhua Wu, "Deep Residual Learning for Image Recognition," Computer Vision and Pattern Recognition, May 2019.

[8] Mingxing Tan,Quoc V.Le, "EfficientNet:Rethinking Model Scaling for Convolutional Neural Networks,"cs.LG, Nov 2019.