The Development of Image Classification Algorithms Based on CNNs

Qichang Liu *

Department of Advanced Computing Major Data Science Software Development, University of Sydney, Sydney, Australia

* Corresponding Author Email: qliu7188@uni.sydney.edu.au

Abstract. Image classification is a method to process the given images by picking up the distinct features of different kinds of images to distinguish different targets in the image. The classification of images analyzes the given images quantitatively by using computers and divides the image or areas contained in the image into several categories instead of explaining them by human's visualization. Image classification is an important research direction in computer vision technology and the basis for other applications such as image detection, behavior analysis, and object tracking. With the arrival of big data and the improvement of computer power, Deep Learning (DL) has swept the world, and the Convolutional Neural Network (CNN) image classification method has broken down the limitations of traditional image classification methods and has become a currently mainstream image classification algorithm. Thereinto, some typical architectures e.g., MobileNet, ResNet and VGG are attracted a lot of attention. This paper will review the development of the image classification and introduce some typical CNNs.

Keywords: Image classification, CNN, Deep Learning

1. Introduction

Image Classification Process means that when machine is given an input image, it uses some classify algorithm to determine which kind of image it belongs to. The Image Classification Process mainly contains the following steps which are the preprocess to the image, extract and describe the features of the given image and finally design the classifier of the algorithm [1].

The traditional Image Classification Process is handling separately following the above steps, and the difference between the classification methods is dependent on extracting image features and the design of the classifier. Traditional Image Classification Methods are all selected the features of the image manually. The commonly used features not only contain some low-level visual features, but also include some local invariant features. However, even though these features are universal for a given image, they cannot deal with some complex image as they cannot find the target features which human need accurately.

In the 1940s, the rise of artificial neural networks promoted the birth of image classification technology. The earliest applied method for image classification is the nearest neighbor method (KNN), which satisfies the datasets with small number of samples and is a typical and concise non-parametric classification method. Later, the backpropagation algorithm appeared. It uses a larger amount of training samples and the artificial neural network model to learn statistical laws of the samples, and can predict unknown events, but there are still problems of local minimization and slow convergence [2]. It is difficult for traditional image classification algorithms to deal with a dataset with a large amount. Besides, the traditional methods cannot satisfy the requirements of the accuracy and speed when classifying the images.

Therefore, in order to solve the problems existing in traditional image classification algorithms and make them better meet people's needs, classification algorithms based on convolutional neural networks emerge as the times require [1]. Instead of selecting and describe the features of the target images manually, the algorithm learns the features from the training samples itself by using neural network so that it can solve the problems related to the human-selected features and classifier.

Section 2 of this paper introduces the basic structure of convolutional neural networks; Section 3 describes the classic model based on CNN; Section 4 introduces the still-existing problems of CNN; Finally, it introduces the prospect and future development of CNN.

2. The Basic Architecture of CNNs

Convolutional Neural Networks are a subset of Feedforward Networks that have deep structures and convolutional computation. They are one of the deep learning algorithms that best exemplify the field. Convolutional Neural Networks are a unique type of Neural Networks because they replace general matrix multiplication with convolution in at least one layer. Convolutional Neural Network (Figure 1) is main made up of convolutional layer, linear rectification layer, pooling layer and fully connected layer.

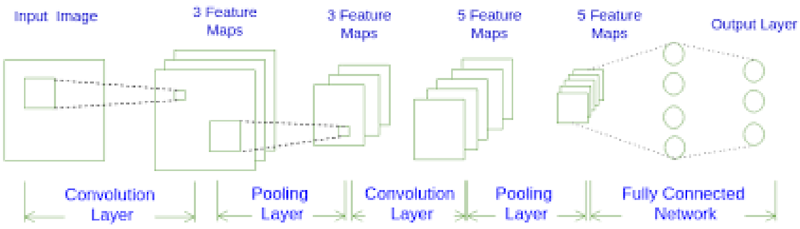

Figure 1. Convolutional Neural Network Structure Model [3].

2.1. Convolutional Layer

Convolutional layer: A convolutional layer is made up of various convolutional units, and the back-propagation algorithm optimizes the parameters of each unit. Some low-level features, such as lines, edges, and so forth, can be extracted by the first convolutional layer. The convolutional operation's goal, which is to extract various types of features from the provided image, can be satisfied by the deeper network by extracting more complicated information through integrative.

2.2. Rectified Linear Units Layer

Rectified Linear Units Layer (ReLU layer) use rectified linear units (ReLU) f(x)=max (0, x) as the activate function of the neural layer. The activate function can enhance the non-linearity of the whole neural network but doesn't change the condition of the convolutional layer. Besides the ReLU function, some other functions are also used to achieve the same purpose such as hyperbolic tangent function f(x)=|tanh(x)|, and the sigmoid function. However, because of its faster training speed and affecting the model slightly, ReLU function is more likely for people to use as an activate function.

2.3. Pooling Layer

Pooling is another important point in the Convolutional Neural Network. The essence of pooling is actually sampling, the pooling reduces the dimension and compress to make the calculation faster in a certain way. The most popular method is called "max pooling," which divides the provided image into many rectangular sections and outputs the highest value from each area. The mostly used pooling layer is shown in Figure 2 which stride is 2 and the window is a 2*2 two-dimension layer. The layer can divide the image into 2*2 pieces and extract maximum of four numbers in every pieces.

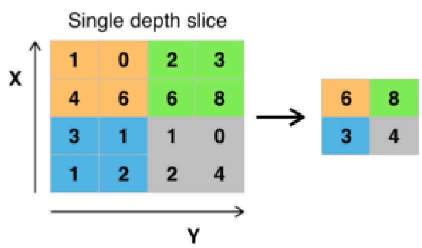

Figure 2. A max-pooling layer with stride 2 and pooling window of 2x2

The pooling layer can also use additional functions, such as "average pooling" and "L-2 norm pooling," among others, in addition to the max pooling approach. Average pooling was utilized extensively in the past, but as a result of the max pooling method's superior performance, its use has decreased.

2.4. Fully Connected Layer

Finally, after the convolutional and pooling layers, then it comes to fully connected layer where the algorithm performs the high-level inference, The feature expression vector acquired from numerous feature extraction layers before to this layer is mapped to the next layer or the final layer using the fully connected layer.

3. Classical CNN models

The convolutional neural network model automatically learns features from a large amount of data to improve the accuracy of the pattern recognition system. At present, most of the conventional image classification methods directly use the common deep convolutional network to directly perform image classification, such as AlexNet, VGG, GooLeNet, ResNet and MobileNet etc. This section will introduce these classic convolutional neural network models.

3.1. LeNet-5

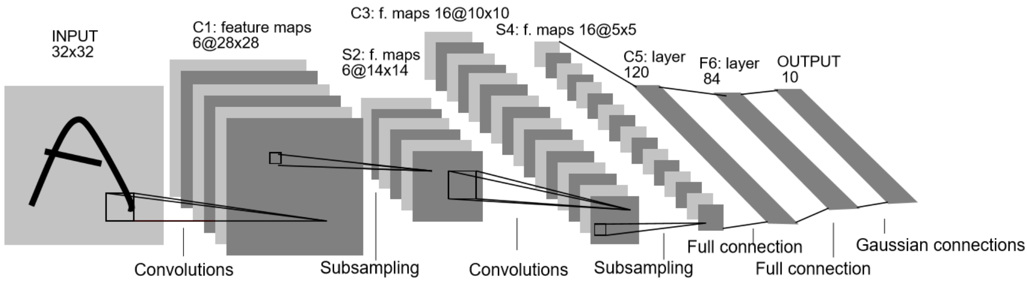

Figure 3. LeNet-5 Network Model [4].

In 1998, LeCun et al. proposed the first CNN method LeNet-5 in history and use it to do the recognition of handwritten digits. Even though seeing from now LeNet-5 is a small network, it doesn’t have any specific characteristics or obvious advantages, but it is the earliest ancestor of CNN and defines the Convolutional Neural Network structure for the first time. The LeNet-5 network model has a total of 7 layers, as shown in Figure 3 [4]. However, the model has a shallow depth, few parameters, and a single structure due to the limited computational efficiency at the time, making it unsuitable for difficult picture classification tasks. [1]. However, it also has some problems due to the old period such as the low computational efficiency, shallow depth and so on. These disadvantages determine that the model cannot be used for doing complex tasks [1].

3.2. AlexNet

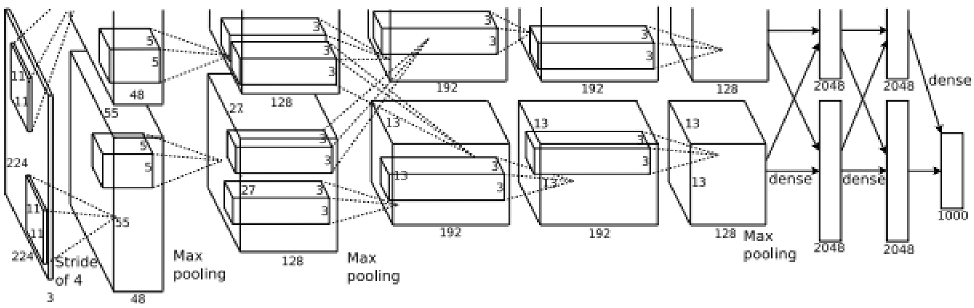

Figure 4. AlexNet Model Structure [5]

In 2012, AlexNet as shown in Figure 4 was designed by Alex Krizhevsky and Ilya Sutsker and Hinton. In the same year, it joined the ImageNet Large Scale Visual Recoginiton Challenge and won the first place of the competition without any doubt that its top-5 error rate was lower than the second one by 10.8 percent [5].

The Features of AlexNet: 1) AlexNet uses a dual-GPU network structure, so that a larger and deeper network can be designed; 2) ReLU is used instead of Tanh, which slightly solves the problem of gradient disappearance and speeds up network convergence. The author of AlexNet also proposed the concept of local corresponding normalization; 3) Make the pooling operation's stride lower than the pooling kernel's size to create overlap between the adjacent pooling sections. This operation is called Overlapping Pooling. 4) Random Crop the training data. And during the test, crop from the four corners and the center of the image, and perform mirror flipping, so that 10 Patches can be obtained, and the results of these Patches are averaged to obtain the final prediction result. 5) Perform PCA (Principal Component Analysis) on the training images and perturb the principal components with random variables that follow a Gaussian distribution of (0,0.1). The author points out that this operation can reduce the indicator Top-1 Error Rate by 1% [5].

3.3. VGG

Actually, VGG has some similarities with AlexNet. However, the improvement of VGG which is better than AlexNet is that VGG uses several 3*3 convolutional kernels continuously instead of larger convolutional kernel used by AlexNet. Besides, when it comes to the given receptive fields, VGG is better than AlexNet because it can learn more complex pattern by increasing the depth of the network with less parameters [6].

3.4. ResNet

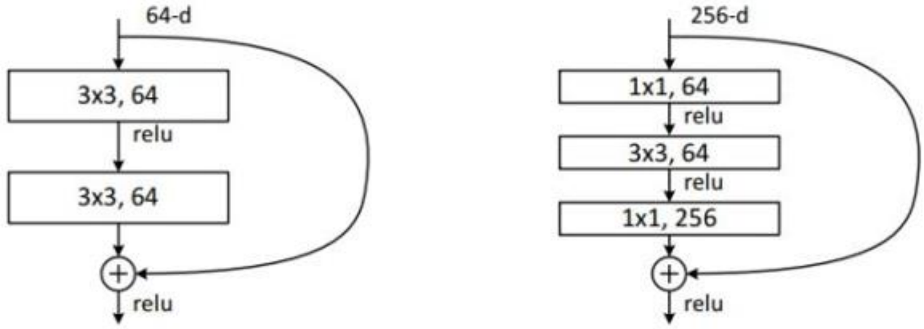

Figure 5. Residual Block in ResNet Model

In 2015, Kaiming He designed a type of CNN called ResNet which solve the degradation problem of the deep network. The degradation problem means that when the depth of the network increasing, the accuracy of the network is full even decreasing. The problem shows the fact that deep network is not that easy to train. However, Kaiming proposed the residual structure to solve the problem in his paper, which shown on Figure 5. In his assumption, he changed the function H(x) into F(x) where the original function H(x) is equal to F(x)+x. The changes made it easier to optimize and converge. As there is no additional parameters and calculations, Kaiming did some experiments in his paper to prove the effectiveness. In addition, in order to increasing the depth of the model, he changed the original Residual Block into Bottleneck Block so that there will be less parameters and calculations [7].

3.5. MobileNet

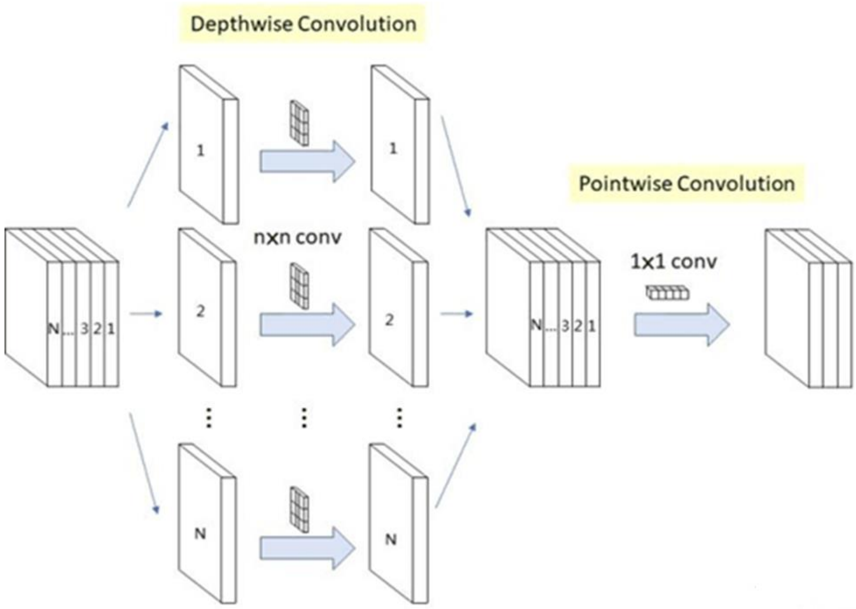

Figure 6. Schematic diagram of the detailed structure of DSC

The Google team proposed MobileNet in 2017, which is a lightweight CNN focused on mobile or embedded devices. The Google team subsequently proposed MobileNetV2 in 2018 and MobileNetV3 in 2019. MobileNet uses a depthwise separable convolutional layer, as shown in Figure 6, that is, firstly perform depthwise convolution (Depthwise Convolution) on each channel of the feature map, and then perform pointwise 1×1 convolution (Pointwise Convolution), In order to achieve the purpose of reducing the amount of calculation and model parameters [8].

The methods described above are mainly based on CNN. In addition, some new architectures are also attracted much attention in recent years. For example, in the field of image recognition, there are many workers trying to introduce the Transformer model into the image [9]. In addition, workers have also tried to introduce Transformer's self-attention mechanism into CNN, or directly replace the convolution block with the Transformer model structure [10]. These attempts have also achieved certain results. Not only that, the combination of deep learning and reinforcement learning for image classification is also one of the main research directions in the future.

4. Conclusion

This paper summarizes the overview of CNN and introduces some classic models based on CNN, but CNN still has some problems that need to be solved. 1) In theory, a key issue in applying CNNs to new domains is that much prior practical experience is required to select appropriate hyperparameters, and these hyperparameters have internal dependencies that make them particularly expensive to tune, so how to better handle these parameters is still a question. 2) Application-related issues such as the network model's reliance on powerful hardware, massive amounts of data, and the training models' dynamic adaptability all require immediate attention. CNN and deep learning are still active research areas today, and there are many development potentials waiting to be explored and tapped in the future.

References

[1] Yang Z et al. Review of Image Classification Algorithms Based on Convolutional Neural Networks [J]. JOURNAL OF SIGNAL PROCESSING, 2018,34(12):1475-1489.

[2] Hu M et al. Image Classification Method based on Modified Convolutional Neural Network [J]. Communications Technology,2018,51(11):2594-2600.

[3] Han B Zero-based entry deep learning (4) - convolutional neural network [Z]. (2014). https://www.zybuluo.com/hanbingtao/note/485480

[4] Ji C et al. Review of image classification algorithms based on convolutional neural network [J]. Journal of Computer Applications, 2022, 42(4): 1044-1049.

[5] Alex K et al. ImageNet Classification with Deep Convolutional Neural Networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[6] Szegedy C et al. Going Deeper with Convolutions [R]. arXiv, (2014).

[7] He K et al. Deep Residual Learning for Image Recognition [R]. arXiv. 2015.

[8] Howard A G et al. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications [R]. 2017

[9] Vaswani A 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.

[10] Dosovitskiy, A et al. An image is worth 16x16 words: Transformers for image recognition at scale [R]. arXiv preprint arXiv:2010.11929, 2020.