From Artificial Neural Networks to Deep Learning: A Research Survey

Zhongwei Zhang
National Space Science Center, Chinese Academy of Sciences, NO.1 Nanertiao, Zhongguancun, Haidian District, Beijing, 100190, China
zhangzhongw@yeah.net

Abstract. Deep learning (DL) is one of the hot topics in the field of artificial intelligence. As an extension of artificial neural network (ANN), deep learning models have more powerful learning and adaptation capabilities and they develop more and more rapidly. At present, because of the great advantages compared with the traditional methods, the development of convolutional neural network (CNN) is the hottest and it is widely used. This article first introduces the development of DL in chronological order and then analyses and summarizes artificial neural networks, and then uses the example of the LeNet-5 to explain the basic network structure of CNN and describe its training methods comprehensively. Based on the CNN model, many new convolutional neural networks with improved structure are analysed, such as AlexNet, ZFNet, VGGNet, GoogleNet and so on. The principle of each method is introduced in detail and the accuracy is compared with each other to conclude the correlation between the depth and the accuracy of DL. Then, for each method, it discusses some current problems existing in the corresponding method of DL, and points out future research and application directions worth exploring of each method. Finally, it summarizes the history of DL and the application of specific methods and the existing problems and solutions.

1. Introduction

At present, China has approved the "New Generation Artificial Intelligence Development Plan", which puts artificial intelligence as a strategic goal and key task towards 2030 development. Deep learning is one of the hot topics in the field of artificial intelligence research, and its development has grown rapidly with the progress of the times.

In 1943, Warren McCulloch and Walter Pitts proposed the concept of MP neurons. This model constructed a simple mathematical model based on the structure and mechanism of biological neurons. In 1949, D. Hebb proposed an unsupervised Hebb learning rule. This rule trains neural networks by adjusting weights based on the activation level of the connections between neurons. In 1969, Minsky showed that even a simple XOR problem cannot be solved using a perceptron model, which is inconsistent with the definition of a true intelligent machine. In 1985, Rumelhart et al. re-presented the BP algorithm (Back Propagation Algorithm) for multi-layer perceptron weight training.

In 1989, Cybenco and Hornik et al. proved that both linear and non-linear functions can be approached infinitely by a three-layer neural network with arbitrary precision. LeCun et al. developed a convolutional neural network model, and the research of neural networks has entered a whole new stage. In 1991, it was pointed out that the gradient disappeared due to the saturation characteristic of the Sigmoid function during the back-propagation of the error gradient. In 2006, Hinton et al.
Proposed the use of a self encoder to reduce the dimension to suppress the disappearance of the gradient. In 2011, Glorot et al. replaced the Sigmoid function with the ReLU function, which alleviated the disappearance of the gradient to a certain extent. In 2015, Choromanska et al. proved that the local minima problem had little effect on the experimental results. In 2017, DCN, GNMT, JMT and other models were successively proposed, and the neural network glowed unprecedented vitality.

Deep learning is a kind of representation learning, which can automatically extract features from the data. It originated from the research on artificial neural networks. Multi-layer perceptron is a simple deep learning structure. The model discovers the distributed features of the data by extracting the features of the first few layers so that it can discriminate the types of attributes at a high level [1]. At present, deep learning excels in intelligent technology applications such as speech processing, computer vision and natural language processing, and robotics.

2. Artificial neural network

Artificial neural network (ANN) is the most basic network structure. ANN consists of an input layer, an output layer, and several hidden layers, and each layer contains several neurons. A neural network whose activation function uses a radial basis function (RBF) is called a radial basis function network. The forward propagation of ANN is shown in Figure 1.

\[ z_{i}^{l+1} = \sum W_{ji}^{l} y_{j}^{l} + b_{i}^{l} \]  \hspace{1cm} (1)

\[ y_{i}^{l+1} = f(z_{i}^{l+1}) \]  \hspace{1cm} (2)

\( y_{i}^{l} \) is the output of the \( j \)-th neuron in the \( l \)-th layer, \( z_{i}^{l+1} \) is the value before the \( i \)-th neuron in the \( l+1 \)th layer is activated by the activation function, and \( W_{ji}^{l} \) is the \( j \)-th neuron in the \( i \)-th layer and the \( l+1 \)th. The weight between the \( i \)-th neuron in the layer. \( b_{i}^{l} \) is bias, and \( f(\bullet) \) is the activation function. Commonly there are RBF, ReLU, Tanh, Sigmoid, etc.

The loss function of a neural network is usually measured by the mean square error, and its mathematical expression is:

\[ J = \frac{1}{2} \sum (y_{i}^{l} - y_{i})^{2} \]  \hspace{1cm} (3)
$y_i^k$ is the output of the i-th neuron in the network output layer, and $y_i$ is the true value of the i-th neuron. The ultimate goal of ANN is to minimize the loss function of the neural network, usually using gradient descent method or Newton method.

3. Convolutional neural network
Convolutional neural network (CNN)[2] is widely used in the field of artificial intelligence, especially computer vision. Its model architecture is derived from the study of visual nerves. CNN is mainly composed of a convolutional layer and a pooling layer. The convolutional layer is mainly responsible for extracting local features of the image. The pooling layer can use max pooling or mean pooling. The main function of the pooling layer is to reduce the dimension and reduce the amount of operations.

The schematic of convolution and pooling operations is shown in Figure 2:

![Figure 2. The illustration for convolution and pooling](image)

The convolutional neural network model is derived from LeNet-5 proposed by LeCun et al. in 1998. The input image size is $32 \times 32$. After a series of convolutions, a $5 \times 5$ image is obtained. It then passes through fully connected layers (FCs) with 120,84,10 neurons, and finally passes through the Softmax function to obtain the probability of each class and takes the maximum probability as the CNN prediction result. With the increase of convolutional layers and pooling layers, the larger the number of network layers, the smaller the image. Use a visual way to understand CNN [3]: the second layer of CNN can recognize the boundary and color; the third layer can recognize more complex feature information such as texture and text; the fourth layer can identify the specific details of living things Parts; the fifth layer can identify specific objects such as cardboard boxes, glass. CNN's local connection, weight sharing, and pooling operations reduce model parameters, reduce network complexity, and also provide translation, distortion, rotation, and scaling invariance.

4. Convolutional neural network structure improvements
ImageNet [4] (ImageNet large scale visual recognition competition, ILSVRC) is one of the most authoritative computer vision competitions. The competition's image classification and target location (CLS-LOC), target detection (DET), video target detection (VID), and scene classification (Scene) have greatly stimulated the development of convolutional neural networks. The ImageNet championships results and related CNN models from 2012 to 2017 have been illustrated in Table1.
Table 1. ImageNet championships and related CNN models

| Year | CNN models  | Test top-5 |
|------|-------------|------------|
| 2012 | AlexNet     | 15.32%     |
| 2013 | ZFNet       | 13.51%     |
| 2014 | GoogleNet   | 6.67%      |
| 2014 | VGG         | 6.8%       |
| 2015 | ResNet      | 3.57%      |
| 2016 | Trimps-Soushen | 2.99%  |
| 2017 | SENet       | 2.25%      |

AlexNet for the first time shows the application of deep learning in computer vision. ZFNet is the result of visual understanding of convolutional neural networks. VGGNet shows that the deeper the neural network layer, the higher the accuracy of the classification results. GoogleNet[5] broke the convolutional layer pooling layer stacking model for the first time, ResNet innovatively introduced the residual term and successfully brought the network layer to 152 layers. One type of CNN successfully applied to object detection is R-CNN [6] and later evolved Fast R-CNN, Faster R-CNN. The other is the YOLO series of algorithms[7]. In addition, NIN [8] proposed the idea of nesting networks in the network; the spatial transformation network [9] shows that the model effect can also be improved by transforming the input data.

1) AlexNet. The AlexNet network connects 5 convolutional layers, max pooling layers, and dropout layers after the input layer, and then connects 3 fully connected layers. The final output layer corresponds to 1,000 classifications, and the probability of each class is obtained after the Softmax function. AlexNet first uses the image processing method to translate, flip or intercept a part to increase the training samples, and at the same time uses dropout to randomly remove neurons to prevent overfitting, and uses ReLU as the activation function. This series of techniques is still widely used now.

2) ZFNet. ZFNet is the 2013 ILSVRC champion, with top-5 being only 13.51%. Zeiler and Fergus visualized CNN to understand the role of each layer. ZFNet requires much less training samples than AlexNet. At the same time, the size of the convolution kernel is changed. The smaller convolution kernel makes ZFNet retain more information in the first layer.

3) VGGNet. Simonyan and his team compared 6 different depth networks and found that the deeper the neural network, the better the effect. When it is increased to 16, 19 layers, the effect is significantly improved. VGGNet strictly uses a 3 × 3 convolution kernel with stride and padding of 1; 2 × 2 max pooling with a step size of 2 makes the model parameters less and increases training with AlexNet. The methods of samples are different. VGG uses image dithering to increase the number of samples, which has a good effect on image classification and object positioning tasks.

4) GoogleNet. GoogleNet only has a 6.67% error rate in top5 with 22 layers. GoogleNet uses the Inception module, and the convolution layer and pooling layer in the module are parallel, so there is no need to choose whether this layer uses a convolution layer or a pooling layer. This module reduces the parameter amount to 1/49. GoogleNet creatively outputs the Softmax probability of multiple deformed pictures of the same picture as the probability of this picture.

5) ResNet. ResNet won three championships in image classification, object positioning, and object detection in 2015, and the image classification task error rate was 3.57%. The number of ResNet network layers reaches 152 layers. For the disappearance of the gradient, ResNet directly adds a linear communication path between two or more layers, which constitutes a residual
module, so that the information at the input layer can be directly retained to the subsequent network layer.

6) R-CNN. R-CNN uses selective search to generate about 2,000 boxes, and uses the trained CNN model to extract features from the pictures in each box, and then puts the features into the SVM for classification. Into the return device to further adjust the position.

7) Fast R-CNN. Fast R-CNN first inputs the image and processes the entire picture with CNN to obtain a feature map; and uses a region of interest pooling layer to process each box to obtain a fixed-size feature map. Then connect several fully connected layers, and finally output the probability of a certain class at the same time.

8) Faster R-CNN. Faster R-CNN also replaced the method of generating boxes with a deep learning model, and changed from generating on the entire map to generating on smaller feature maps, which further accelerated the model training speed.

9) YOLO series. YOLO is another type of algorithm proposed for target detection. The core idea is to solve object detection as a regression problem. YOLO divides the full image into $S \times S$ grids, and each grid is responsible for object detection centered on the grid. It uses direct prediction of the bounding box, confidence, and all class probability vectors of objects contained in the grid.

10) Network in network (NIN). The network-in-network structure uses a micro-neural network to replace the convolution kernel in the CNN, forming a structure in which the micro-neural network is nested in the neural network. This greatly reduces model parameters, prevents overfitting, and increases interpretability. NIN has 29 million parameters.

11) Spatial transformer networks (STNs). The spatial transformation network improves the accuracy by transforming the input pictures, rather than changing the network structure. STNs are composed of a localization network, a grid generator, and a sampler. STNs are very robust and have good spatial invariance.

12) Other improved convolutional neural networks. Including MR-CNN, FV-CNN, Deep Edge, Box Sup, T-CNN, cascade CNN, deep parsing network, 3D CNN and so on.

5. Problems and trends of deep learning

Although deep learning has made breakthroughs in many fields, there are still some problems that need to be resolved.

1) The data training time is too long, and the proportion of required computing resources is too large. At the same time, a large number of parameter adjustment tasks are also very heavy, so technical support is needed to accelerate the training of the model.

2) Rely on large-scale labeled training data to train the model. Manual data labeling is time-consuming, labor-intensive, expensive, and poorly accurate; it is almost impossible to collect enough labeled data in specific cases, such as special cases in the medical field.

3) Distributed training problem. When there are too many training samples, the training time of the model on a single machine is very long, and the model that is too large cannot be put into one machine. Therefore, large-scale distributed training of deep learning models is needed. Distributed parallel training can be divided into data parallel, model parallel and hybrid parallel.

4) Poor robustness [10]. Even though the average accuracy of deep learning is high, the prediction effect on some test cases may be poor. The technical fields of remote surgery, satellite launch, etc. do not allow some obvious wrong results, so the robustness of deep learning needs to be improved to make the model's application area wider.

5) Poor interpretability. The deep interpretability of deep learning models is a typical black box algorithm. The model is complex and usually contains hundreds of millions of parameters.

6) Deep learning requires large-scale training data, and sample utilization is not high, but human learning requires only a few samples.

7) The model is too large. Deep learning models are very large and are not convenient to use on mobile. Especially for language models, the vocabulary is very large and there are many output neurons, resulting in a very large model. Therefore, under the premise of ensuring accuracy, the model
is compressed to make the model smaller. At present, there are four types of model compression methods: parameter pruning and sharing, low rank decomposition, compressed convolution filter, and knowledge refining.

6. Summary
This article first briefly introduces the development history of neural networks, and then analyzes and summarizes the topology of artificial neural networks, the basic network structure of convolutional neural networks (CNN), training methods, and their characteristics. Based on the CNN model, this paper analyzes many new types of convolutional neural networks such as AlexNet, ZFNet, VGGNet, GoogleNe, and ResNet with improved structure. Finally, it expounds some problems in current deep learning methods, and points out future research and application directions worth exploring.

References
[1] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. Cambridge, USA: MIT Press, 2016.
[2] Zhou Feiyan, Jin Linpeng, Dong Jun. Review of Convolutional Neural Network Research [J]. Chinese Journal of Computers, 2017, 40 (6): 1229-1251.
[3] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural network[C] Annual Conference on Neural Information Processing Systems. Cambridge, USA: MIT Press, 2012; 1097-1105.
[4] LeCun Y, Bottou L, Bengio y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.
[5] Sutskever I. Training recurrent neural networks[D]. Toronto, Canada: University of Toronto, 2013.
[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C] IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2016; 770-778.
[7] He K, Gkioxari G, Dollár P, et al. Mask R-CNN[C] International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2017:2980-2988.
[8] Choromanska A, Henaff M, Mathieu M, et al. The loss surfaces of multilayer networks[C] International Conference on Artificial Intelligence and Statistics., Piscataway, NJ, USA; IEE, 2015: 192-204.
[9] Liu L, Chuang Ding Y J, Zhou M, et al. The IFLYTEK system for Blizzard challenge[R]. 2017.
[10] CAILKIMT. Context-driven hybrid image inpainting[J]. IET Image Processing, 2015, 9(10): 866-873.