Research and Development of Deep Learning

Ruihua Zhao

Henan Shangqiu Minquan County Vocational and Technical Education Center, Henan Shangqiu, 476800, China

Keywords: Shallow Learning; Deep Learning; Hierarchical Structure; Artificial Intelligence; Machine Learning

Abstract. In view of the shortcomings of shallow learning in the past, such as lack of feature expression and excessive dimensionality, the depth learning solves these problems well by the unique hierarchical structure and the ability to extract high-level features from low-level features, and brought new hope for artificial intelligence. The development of deep learning during different periods was introduced. The basic models of restricted boltzmann machines (RBM),auto encoder (AE) and convolutional neural networks (CNN) were analyzed to present the deep hierarchical structures of deep belief networks (DBN), deep boltzmann machine(DBM) and stacked auto encoders (SAE).The applications of deep learning in the fields of speech recognition, computer vision,natural language processing and information retrieval in recent years were introduced to illustrate the superiority and flexibility of deep learning compared with other shallow learning algorithms. Some future research directions were predicted based on the analysis, and some conclusions were made according to the improvement of deep learning on algorithm generalization, adaptation of big data and modifying on deep structure.

Introduction

Since 2006, deep learning has become a new field of machine learning, as a method of modeling (audio, image, text, etc.) in the field of machine learning. Deep learning aims at making machine learning closer to its original goal - artificial intelligence [1].

In recent years, with the emergence of deep learning, many researchers have devoted themselves to the study of the principle and application of deep learning, which is mainly reflected in the major conferences. The meeting include: 2013 acoustics speech and signal processing International Conference (International Conference on acoustics speech and signal processing, ICASSP) to discuss a new type of learning about the depth of the neural network and related applications of speech recognition;2010、2011and 2012 neural information processing systems (neural information processing systems, NIPS) discussed in depth learning and unsupervised feature learning; 2011、2013 years of International Conference on machine learning (International Conference on machine, ICML) to discuss the audio, speech and visual information processing learning structure, representation and optimization[2].

The Development of Deep Learning

The development of machine learning can be roughly divided into 2 stages: shallow learning and deep learning. In recent years, most of the machine learning methods have been used to deal with the data of the shallow structure, which has only 1 or 2 layers of nonlinear features. The typical shallow structures are: Gauss mixture model (GMMs) [5], support vector machine (SVM) [6], logistic regression and so on. In the shallow layer model, SVM is the most successful model, a linear model of shallow separation model using SVM, when the data vector of different categories can't be divided in the low dimensional space, SVM will put them through a kernel function mapping into a high dimensional space and find the optimal classification hyper plane.

In 2006, Hinton proposed the deep confidence network (deep belief network, DBN) [8], the

network can be seen as a result of the (restricted Boltzmann machines) (RBM) [9] superposition. From the perspective of structure, the depth of confidence little multi-layer perceptron network and the difference between the traditional supervised learning, but the training needs to unsupervised learning and training, and then learned parameters as the initial value of supervised learning. It is this kind of learning method of change makes the depth structure now can solve the problem of BP can not solve the past the problem.

The other algorithm models of the depth structure are then presented, and some of the best records are refreshed on many data sets. For example, the 2013 Wan Li [10] proposed drop connect network standard, the model in the data set on the CIFAR-10 error rate was 9.32%, 9.55% lower than the previous best results, and the error rate of 1.94% in SVHN, 2.8% lower than the previous best results and so on.

The Basic Model and Improvement of Deep Learning

The time of deep learning is not long, so most of the models are based on several core models. For example, RBM, AE (atuo encoders)[11], convolutional neural network (convolutional neural networks, CNN), such as improved[12]. In this paper, we firstly introduce some basic models, and then introduce the depth structure model or the improved model.

Restricted Boltzmann machine

RBM has a rich theoretical framework, which is a statistical neural network [13] (Boltzmann machines, BM) developed by D.H.Ackley and other statistical mechanics proposed in 1985. BM has strong ability of unsupervised learning, and can learn the complex rules of data. However, it is impossible to calculate exactly the distribution of BM. In order to solve this problem, Smolensky introduced a restricted Boltzmann machine. He made the BM with the original interlayer connection limit, which making the same layer in different nodes are independent of each other, only nodes between the layers are connected.

(1) Restricted Boltzmann machine principle

RBM is a special case of a Markov random field with 2 layers of structure [17] (see Figure 1). It consists of M visual units $V = (V_1, V_2,... V_M)$, a visual layer that is generally subject to Bernoulli or Gauss distributions; N hidden units $H = (H_1, H_2,... H_N)$ the hidden layer, which is generally subject to Bernoulli distribution. Figure 1 shows the hidden (output) layer of the N hidden units, and the lower layer represents the visible (input) layer of the M visual elements.

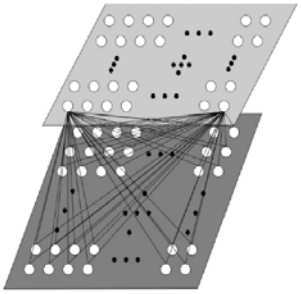


Fig.1. Restricted Boltzmann machine

As shown in Figure 1, there is a weight connection between the visual element layer and the hidden layer of the RBM, but there is no connection between them.

In statistical mechanics, energy function [8-9,11] can estimate the energy of a system. When the system evolves according to its internal dynamic rules, its energy function always changes in the direction of decreasing, or stays in a fixed value, and finally tends to be stable.

(2) Depth structure based on restricted Boltzmann machine

Figure 1 is a RBM structure. The lower layer is the input layer and the upper layer is the output layer. When the same RBM structure is added again, a part of the DBN structure is formed, that is, the output of an RBM is used as the input of another RBM during the pre training phase. Then use the BP fine-tuning to better weight training. This section will introduce DBN and DBM based on RBM, and a brief analysis of the differences between the DBN and DBM.

Automatic encoder

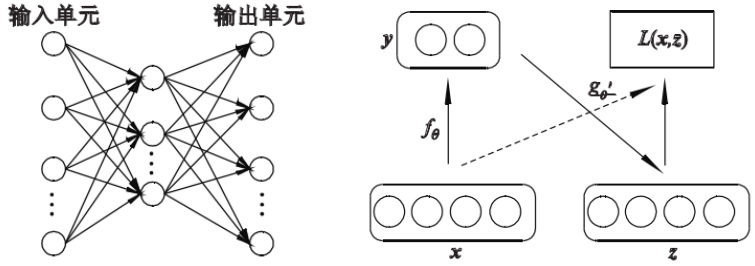
Y.Bengio[11]in 2007 by understanding the success of the DBN training strategy, that is, through the unsupervised pre training to better initialize the weights of all layers to slow down the depth of network optimization problems. And this idea is verified by replacing the RBM building blocks in the DBN structure with AE. This section first introduces the basic principles of AE, and then describes the AE stack based on the automatic encoder (stacked auto encoders, SAE) [23].

(1) Principle of automatic encoder

AE transforms the input of the visual layer to the hidden output layer, and then reconstructs the hidden layer so that the target output of the auto encoder is almost equal to that of the original input, as shown in Figure 2a.The objective function of AE is

$$J(\theta,\theta') = \frac{1}{n}\sum_{i=1}^{n} L(x^{(i)}, z^{(i)}) + \frac{\lambda}{2}(\| \theta \| + \| \theta' \|), \quad (1)$$

The first is to minimize the reconstruction error of the model; the second is the weight decay term. Firstly, suppose an auto encoder input for d' dimensional vector $x \in [0,1]^d$, Through a function mapping, mapping to the output layer for d' dimensional characterization of vector $y \in [0,1]^{d'}$, the mapping function is $y = f_\theta(x) = s(Wx+b)$,the structural parameters of the model are $\theta = \{W, b\}$,W is a d'*d weight matrix, b is the offset vector, s is a logical sigmoid function calculated by element, $s(t) = \frac{1}{1+\exp(-t)}$,$t \in \{1, m\}$,m is the number of units needed to propagate the next layer. Output characterization of y obtained was subsequently mapped to the "Reconstruction" of the vector $z \in [0, 1]^d$, $z = g_{\theta'}(y) = s(W'x+b')$, Model reconstruction parameter $\theta' = \{W', b'\}$,W' is a d x d' weighting matrix. Figure 3b is a simple representation of an automatic encoding process.



(a) basic structure of automatic encoder (b) basic principle of automatic encoder

Fig.2. the basic structure and basic principle of automatic encoder

The optimization of the parameters of the model $\{\theta,\theta'\}$ is the average reconstruction error minimization model:

$$\frac{1}{n}\sum_{i=1}^{n} L(x^{(i)}, z^{(i)}) = \frac{1}{n}\sum_{i=1}^{n} L(x^{(i)}, g_{q'}(f_q(x^{(i)}))), \quad (2)$$

In the type: n sample size; x is the original input vector; z is the reconstruction vector. According to the difference of input and output, the loss function L can be a continuous value of the traditional variance loss function L (x, z) =12||x - z||2 or two value of the cross entropy loss function $L(x,z) = -\sum_{j=1}^{d}[x_j \log(z_j) + (1-x_j)\log(1-z_j)]$[24].

In addition, in order to prevent over fitting, the weight attenuation term is added to the objective function as the regularization term, which is the second of the formula (1).The attenuation parameter λ indicates the importance of the reconstruction error and the weight decay term.

Convolutional neural network

In 1989, Yan Lecun, they proposed a structure that can be used to successfully train the depth of BP network based on previous work: CNN, which combines local receptive fields, weight sharing, and spatial or temporal subsampling of these 3 structures to ensure translation and deformation invariance.

re are 6 feature maps in the hidden layer, each of which corresponds to the small box in the input layer, which is a local receptive field, or a sliding window, which is called the "sliding window".

Convolution: The activation value of the first j feature map in L is a_j^l

$$a_j^l = f(b_j^l + \sum_{i \in M_j^l} a_i^{i-1} * k_{ij}^j). \quad (3)$$

Weight sharing: the f is a nonlinear function, which usually is tanh function and sigmoid function, b_j^l is the L layer which is the offset J of the unit value index. M_j^l is vector mapping of I L-1 layer, l layer and features in the map J is accumulated, a_i^{i-1} is a 2 dimensional is the core and k_{ij}^l convolution operation feature mapping i function in the L-1 layer in the input part accumulation can generate feature mapping J in the L layer. A volume layer usually consists of several characteristic map, and k_{ij}^l is the weight, in the same feature map is the same, so it reduces the number of free parameters.

Sub sampling: if the input will shift the output volume of this translation layer, but it will not change, and once a feature is detected, the exact location will be less important, as long as compared to the approximate location of other features are preserved. Therefore, each roll laminated behind there will be an additional the mean of the local layer to execute, namely sub sampling 30-31 to reduce the output of translation and deformation sensitivity. For a sub sampling feature mapping J layer in L, there are

$$a_j^l = down(a_j^{i-1}, N^l), \quad (4)$$

In the type: down is a function of factor Nl down sampling based on Nl; for the L sub sampling window size boundary layer is needed, then the size of Nl * Nl window non overlapping area is calculated. The mean output layer hypothesis of neurons in the C dimension, you can identify the C class, the output layer is the output characterization of feature mapping before layer connection:

$$output = f(b^o + W^o f_v), \quad (5)$$

In the type: b^o is the bias vector; W^o is weight matrix; f_v is feature vector, the model parameters are $\{k_{ij}^l, b_j^l, b^o, w^o\}$, convolution layer and subsampling layer are usually layer alternately, and the number of feature maps is increased with decreasing spatial resolution.

Application of Deep Learning

Deep learning from the beginning of 2006 in speech recognition, computer vision, information retrieval and Natural Language Processing above all has achieved good results in different data sets and industrial applications have shown far more than the previous shallow learning can achieve the best results.

Speech recognition

In the past few decades, researchers in the field of speech recognition have focused their efforts on HMM-GMM based systems, while ignoring the original structural features of the original speech data. The deep neural network (DNN) was introduced to deal with the problem of speech recognition in 2010, because the correlation between the DNN data has a greater tolerance, so that when the GMM is replaced by DNN, the effect has been a leap.

2012, Microsoft Corp, a deep learning based voice and video retrieval system (Microsoft audio video service, MAVIS) successfully came out, the word error rate was reduced by 30% (from 27.4% to 18.5%) [36]. In 2014, IBM Watson Research Center's T.N.Sainath[37] results show that DNN than in the past, the GMM-HMM model has improved 8%~15%.Compared with the general DNN CNN can have stronger adaptability to the strong correlation between the data, and the network has the characteristics of translation invariance.

Computer vision

The successful application of deep learning in computer vision is mainly in the field of object recognition [38]and face recognition[39]. For a long time, object recognition in machine vision has

always depended on the characteristics of artificial design. For the more complex past small samples which cannot in the real environment of the information, 2010 people into a larger data set, such as the Image Net data set with 15 million markers of high resolution images and more than 22 thousand categories. A.Krizhevsky et al.[33] In 2012 trained a large depth neural network to classify the 1 million 200 thousand high resolution images of the Image Net LSVRC-2010, which contained over 1000 different categories. In the test data, their error rates on top-1 and top-5 are 37.5% and 17%, which refreshes the best record for this dataset.

In 2014 Sun Yi et al.[42] Proposed a method of deep hiding identity (deep hidden identity feature, ID Deep) to learn high-level feature representation for face recognition. The face part of the region as each convolution network input, the local low level feature extraction in the bottom, and the last layer of hidden layer neurons in the depth of the convolutional network activation value form Deep ID features, test results show that Yi obtained 97.45% accuracy in LFW.

Information retrieval

Information retrieval (IR) is that a user enters a query into a computer system that contains many documents and obtains the closest document required by the user, [2]. Deep learning in IR application is mainly through semantic feature extraction useful to the sub sequence document ranking, proposed by R.Salakhutdinov [25] in 2009, they at that time, the most widely used in the analysis of the document retrieval system TFIDF[25], that TFIDF system has the following defects: the similarity of the documents directly calculated in the word count space, which makes the vocabulary will be very slow; do not use semantic similarity between words. In 2014, Shen Yelong [45] proposed the deep structure semantic convolution version of the model (convolutional deep-structured semantic modeling, C-DSSM, C-DSSM) can be in the context of the semantic similarity of words through a convolution structure projection to the context feature vector space, improve the accuracy rate of from top 43.1% to 44.7%.

Different from the previous shallow structure, it can only solve many simple or many constraints, the depth structure can handle many complex real-world problems, such as the human voice, sounds of nature and natural language, image, visual scene and other issues, they can extract features contained in the data directly from the data without the specific model constraints, thus has the generalization ability.

Prospect of Deep Learning

With the deepening of the research, deep learning has become an indispensable field in machine learning. However, the study on deep learning is still in its infancy, and many problems still haven't found a satisfactory answer [46]. Such as the ability to improve online learning, as well as the ability to adapt to big data and improve the depth of the hierarchy.

Online learning: The depth of structure training current deep learning almost all by the application of the algorithm are first carried out on the basis of the structure layer training, and fine tune the parameters fitted the data better set with a global layer after training. his training algorithm in a purely online environment is not very suitable, because the online data set is constantly expanding, once in the online environment introduces global fine-tuning, it is likely to fall into local minimum.

On the improvement of deep structure: Although the hierarchical model of deep structure has a breakthrough in structure than the shallow model, the hierarchical structure of the biological visual system is simulated, but it can not completely match the information processing structure of the cortex. For example, the researchers found that the depth of the existing structure of the mainstream does not take into account the impact of time series on learning, but as the real biological cortex in processing the information, the information data of the independent learning is not static, but as time has a context.

The human information processing mechanism indicates that the depth structure can extract the complex structure from the rich perceptual information and establish the internal representation of the data. Because the depth of learning is still in the initial stage, many problems have not been solved, so it can not really achieve the standard of artificial intelligence. But the success and

development of deep learning shows that deep learning is a big step towards artificial intelligence.

Conclusion

1) In this paper, firstly, we classify the depth structure used in the existing depth learning. This paper introduces the principles and characteristics of several basic models used in deep learning, such as RBM, AE, CNN and so on. And then we analyze how to get the true depth hierarchy model of DBN, DBM, SAE and so on.

2) Through the introduction of deep learning application in speech recognition computer vision, Natural Language Processing and several areas of information retrieval, illustrates the deep learning in the field of machine learning are compared to other shallow structure learning superiority and less has better error rate compared.

3) This paper makes a summary and Reflection on the problems faced by the current deep learning through analyzing it`s adaptability of online learning and big data; improving its depth structure. The deep learning is not yet mature, there is still a lot of work needs to be studied. However, its powerful learning ability and generalization ability show that it will be the focus of research in the field of machine learning in the future.

References

[1] Sun Z J, Xue L, Xu Y M, et al. Overview of deep learning[J]. Jisuanji Yingyong Yanjiu, 2012, 29(8): 2806-2810.

[2] Yu D, Deng L. Deep learning and its applications to signal and information processing [exploratory dsp][J]. IEEE Signal Processing Magazine, 2011, 28(1): 145-154.

[3] Hu Xiaolin, Zhu Jun. Deep Learning-new hot spot in the field of maching earning [J]. Communications of the CCF,2013,9(7);64-69.(in Chinese)

[4] Bengio Y. Learning deep architectures for AI[J]. Foundations and trends in Machine Learning, 2009, 2(1): 1-127.

[5] Duarte-Carvajalino J M, Yu G, Carin L, et al. Task-driven adaptive statistical compressive sensing of Gaussian mixture models[J]. IEEE Transactions on Signal Processing, 2013, 61(3): 585-600.