



Research article

Exploration of machine algorithms based on deep learning model and feature extraction

Yufeng Qian*

School of Science, Hubei University of Technology, Wuhan 430068, China.

* **Correspondence:** Email: yfqian@aliyun.com.

Abstract: The study expects to solve the problems of insufficient labeling, high input dimension, and inconsistent task input distribution in traditional lifelong machine learning. A new deep learning model is proposed by combining feature representation with a deep learning algorithm. First, based on the theoretical basis of the deep learning model and feature extraction. The study analyzes several representative machine learning algorithms, and compares the performance of the optimized deep learning model with other algorithms in a practical application. By explaining the machine learning system, the study introduces two typical algorithms in machine learning, namely ELLA (Efficient lifelong learning algorithm) and HLLA (Hierarchical lifelong learning algorithm). Second, the flow of the genetic algorithm is described, and combined with mutual information feature extraction in a machine algorithm, to form a composite algorithm HLLA (Hierarchical lifelong learning algorithm). Finally, the deep learning model is optimized and a deep learning model based on the HLLA algorithm is constructed. When $K = 1200$, the classification error rate reaches 0.63%, which reflects the excellent performance of the unsupervised database algorithm based on this model. Adding the feature model to the updating iteration process of lifelong learning deepens the knowledge base ability of lifelong machine learning, which is of great value to reduce the number of labels required for subsequent model learning and improve the efficiency of lifelong learning.

Keywords: deep learning model; feature extraction; machine learning; genetic algorithms; mutual information

1. Introduction

From the birth of the world's first general-purpose computer, "ENIAC" in 1946, people began to

explore the possibility of autonomous learning by computers without the limitation of preset modes [1]. Machine learning is a small branch of artificial intelligence, which can predict and infer unknown sample categories or information about new emerging things by learning some regular characteristics of the known data or event forms. When the tasks encountered are very complex or need computer adaptive adjustment, this needs to be done with the help of the powerful data analysis ability of machine learning algorithms [2]. Therefore, it is proposed to investigate whether self-learning can be continued through computers or related algorithms to achieve the purpose of summing up experience, for the system to constantly improve itself. With continuous exploration and the development of technology, the application of computers in various fields is more and more extensive, and artificial intelligence has been rapidly developed to a high level [3]. In daily life, commonly used smart phone applications can provide personalized services according to the user's historical search data; in the medical field, smart phone applications can accumulate experience from medical big data to assist doctors in updating and learning disease treatment methods [4–6].

Machine learning is the core concept in the field of artificial intelligence. There is no clear conclusion about its specific mechanism up to now [7], but it is clear that the main process in machine learning is the selection of a group of representative features to build a model for a learning task, and to achieve self-improvement through continuous learning. Deep learning is a type of machine learning algorithm, which is essentially a multi-layer neural network structure. Compared with the traditional machine learning algorithm, its principle is to increase the network layer to better select data characteristics and improve the ability to handle variability [8]. The calculation process needs to carry out a back propagation (BP) calculation for each layer of the network to achieve the optimal calculation of the objective function, and finally obtain the network model parameters. With the accumulation of experience, how to automatically improve the performance of computer algorithms is the focus of current research. It has been found that the nature of data itself plays an important role in a learning task. If the data is adapted to the learning process, the efficiency of machine learning will be improved and the learning task will be simpler. However, if the data itself has little important information related to the learning task, machine learning may often fail [9–11]. From the perspective of biological neuroscience theory, the deep learning model can better simulate the principle of the human brain signal transmission mechanism. Each output neuron is a more deep-seated neural node through weighted calculation and nonlinear transmission, so as to realize the effective flow of original data in the network. Successful feature extraction plays an important role in the field of machine learning, such as in data mining, information classification and retrieval.

With the explosive growth of multi-source heterogeneous data, big data has become the focus of attention. It is the focus of current research into how to learn efficiently and continuously from massive and complex information. Based on the deep learning model and feature extraction technology, the study analyzes several representative machine learning algorithms, and compares the results of the optimized deep learning model with other algorithms. The research is of great value for extracting representative features from high-dimensional and unlabeled data, thus greatly reducing the amount of labeled data required for learning new tasks.

2. Literature review

Almabdy and Elrefaei (2019) constructed an eight-layer deep neural network model, AlexNet, which reduced the error rate of image recognition to 16.4% and made further breakthroughs in deep

learning in the field of image recognition [12]. Basha et al. (2020) invented a deeper VGG deep neural network model, including 16 convolution layers and 3 full connection layers [13]. Wang and Zhang (2020) further analyzed the deep convolution network and constructed DenseNet's deep neural network model, based on the idea of reusing the characteristics of different layers [14]. Du et al. (2018) found that the application of optimization algorithms, such as gradient descent, in the deep learning model can implicitly balance the unbalanced parameters between positive homogeneous layers [15].

In research into feature extraction, text classification is a challenging field with high-dimensional feature space. Onan et al. (2016) found it very useful to extract keywords of document content and use them as features [16]. On this basis, Onan et al. (2016), based on the prediction performance of the classification algorithm, combined with Bayesian logical regression, and with appropriate weight values assigned to the classifier and each output category, enhanced the prediction performance of emotion classification [17]. Kang (2019) analyzed the basic concepts of machine learning algorithms, obtained the important means to deal with big data at present, and put forward theoretical guidance for the development of the big data era [18]. Fuentes (2018) constructed a deep neural network and completed the verification phase accordingly. Compared with the previous technology, it has better performance and improves the accuracy significantly [19].

The above research reveals that the algorithm in deep learning has been optimized in continuous improvement, but the deep learning model needs to be further explored in terms of feature extraction from images, speech, and other information. In view of this deficiency, this study discusses several classical machine learning algorithms based on deep learning models and feature extraction. A genetic algorithm is combined with a mutual information feature extraction algorithm in machine learning to form a combined algorithm which improves and optimizes the deep learning model, and realizes an efficient and long-term task learning process.

3. Overview of machine learning system

3.1. Theoretical basis and present situation of machine learning

Continuous machine learning, also known as Learning to Learn, can not only update the deep learning model online, but also update learning skills according to changing learning progress. Based on this, the related algorithms of machine learning must satisfy the migration of knowledge or structure. It has been concluded that machine learning should possess the following elements: discovering high-level abstract representation of knowledge from raw data; transferring knowledge from completed learning tasks to improve learning ability and complete new learning tasks; possessing the ability to quickly learn new tasks and accumulate and renew knowledge to enhance the level of knowledge; ability to develop knowledge under human guidance and feedback; and to share learning with other systems [20,21]. The structure of a machine learning system is shown in Figure 1.

Knowledge migration: Select and migrate useful knowledge in the knowledge warehouse, which is helpful in promoting the effective learning of new models.

Knowledge storage: This refers to storing the knowledge that the system learns and which will be used later, playing the function of knowledge warehouse. In terms of knowledge storage, the first method is to store directly the data samples that have been learned, which can preserve the complete data information. However, the high spatial complexity will affect the reading efficiency of the stored information when acquiring knowledge, so the second method is to extract and compress the original

information, only preserving the feature representation of the knowledge. The second method is faster than the first method when extracting information, but some detailed data may be omitted.

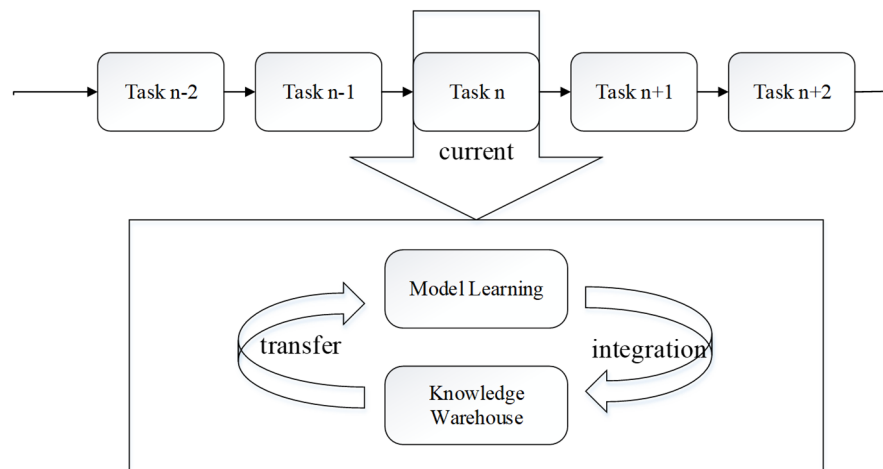


Figure 1. The structure of a machine learning system.

Model learning: This mainly involves rapid learning of new tasks and interacting with the knowledge warehouse. It can improve the learning speed of new tasks, while integrating the new learning content with the original knowledge, which makes the knowledge base more abundant. The process of model learning is independent of that of the machine learning system. Knowledge in the knowledge warehouse is not always reflected in the process of learning new models.

Knowledge integration: The integration of knowledge is a key link in a machine learning system. Its purpose is to ensure that the knowledge warehouse is continuously updated, and to realize the effective transfer of knowledge after learning new tasks. In this process, a machine learning system should screen knowledge, eliminate invalid knowledge and integrate effective knowledge without losing the original key information.

Guided learning: This process focuses on analyzing the impact of the sequence of learning tasks on the learning performance and efficiency of the system, and emphasizes the importance of learning efficiency. In today's big data environment, because of the huge amount of complex data involved, it will not be able to complete all knowledge learning. By guiding learning, it is possible to arrange learning tasks reasonably and improve the learning efficiency of the system.

Realization of human-computer interaction and machine-computer interaction: This process is the complement of the whole system, and it can obtain additional relevant information, improve knowledge storage, as well as find errors in the system itself and correct them in a timely manner.

3.2. Machine-learning-related algorithms

The ELLA efficient machine learning algorithm was proposed by Paul Ruvolo and Eric Eaton in 2013 on the basis of the GO-MTL (overlapping grouping multi-task) algorithm. The algorithm establishes a prediction model, $f^{(n)}(x) = f^{(n)}(x; \theta^{(n)})$, for each task n by taking $\theta^{(n)} \in R^d$ as the parameter, which can represent the proposed parameter hypothesis through a set of linear combinations of shared bases. When learning a set of tasks simultaneously, a model based on the ELLA algorithm can selectively share information among tasks, such that irrelevant tasks will not affect each other [22].

The structure of the ELLA algorithm is shown in Figure 2. The ELLA algorithm stores a group of shared knowledge bases of potential tasks. When new tasks appear, it helps to learn knowledge from the new tasks by transferring the knowledge learned, and refines the knowledge bases of potential tasks on this basis. Based on this process of maintaining and sharing a knowledge base, the performance of the original learning task model can be effectively improved to meet various requirements of the machine learning algorithm.

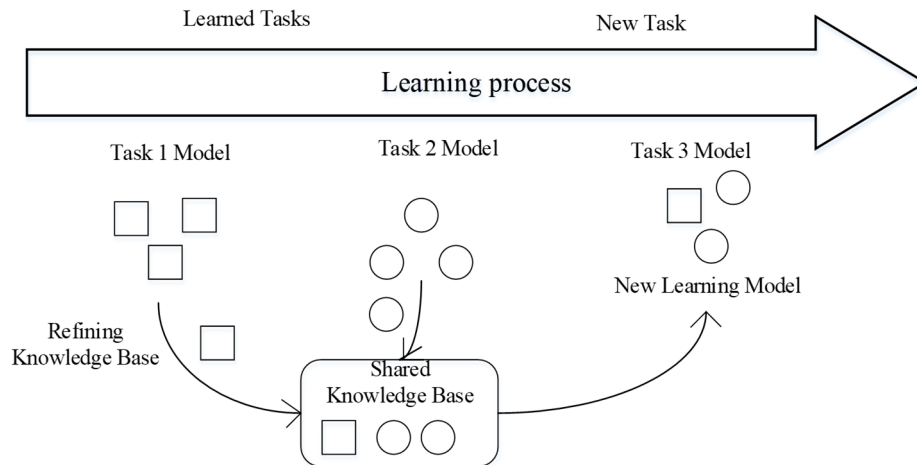


Figure 2. The structure of the ELLA algorithm.

The SUPART algorithm was proposed by M. Mishra on the basis of the DG-MTL (non-intersecting multi-task learning) algorithm in 2015. It is also called the Learning to Learn algorithm for supervising task space partitioning [23]. A supervised learning algorithm is used to obtain the segmentation function of task grouping and clustering. When new learning tasks occur, both the segmentation function and the inference function will undergo online updating of model parameters. The objective function of the model based on SUPART algorithm is expressed as:

$$R = \min \frac{1}{N} \sum_n l(\omega_n, D) + \lambda F(W) \quad (1)$$

In the above equation, ω_n represents the parameters of task n ; λ denotes the normalized parameters of the SUPART model, and F refers to the normal constraints on the whole-task parameter matrix, W . In addition, a segmentation function is introduced to divide the task into A and B regions in space. The corresponding task parameters ω_n can be expressed as $u_A + v_n$ and $u_B + v_n$, u represents the common model in the corresponding region, and v_n is the private model of the task. Therefore, the objective function of the SUPART model can be improved as follows:

$$R = \min \left(\frac{1}{N} \sum_n \Pi(g=0) l_n^A + \frac{1}{N} \sum_n \Pi(g=1) l_n^B + \lambda F(W) \right) \quad (2)$$

In Eq (2), $l_n^A = \frac{1}{N} \sum_i l(u_A, v_n, x_{i,n}, y_{i,n})$, $l_n^B = \frac{1}{N} \sum_i l(u_B, v_n, x_{i,n}, y_{i,n})$. The SUPART model is optimized by a single variable method. After determining the segmentation function, all task parameters are optimized until the model converges iteratively.

4. Feature extraction algorithms in machine learning

4.1. Wrapper method based on genetic algorithms

The genetic algorithm was first proposed by Professor Holland of the University of Michigan in the United States, and then summarized to form a simulated evolutionary algorithm [24,25]. The genetic algorithm is an adaptive, probabilistic search algorithm based on Darwin's theory of biological natural selection and evolution. It generates new populations through genetic operations such as selection, crossover and mutation, and then realizes the gradual evolution of populations.

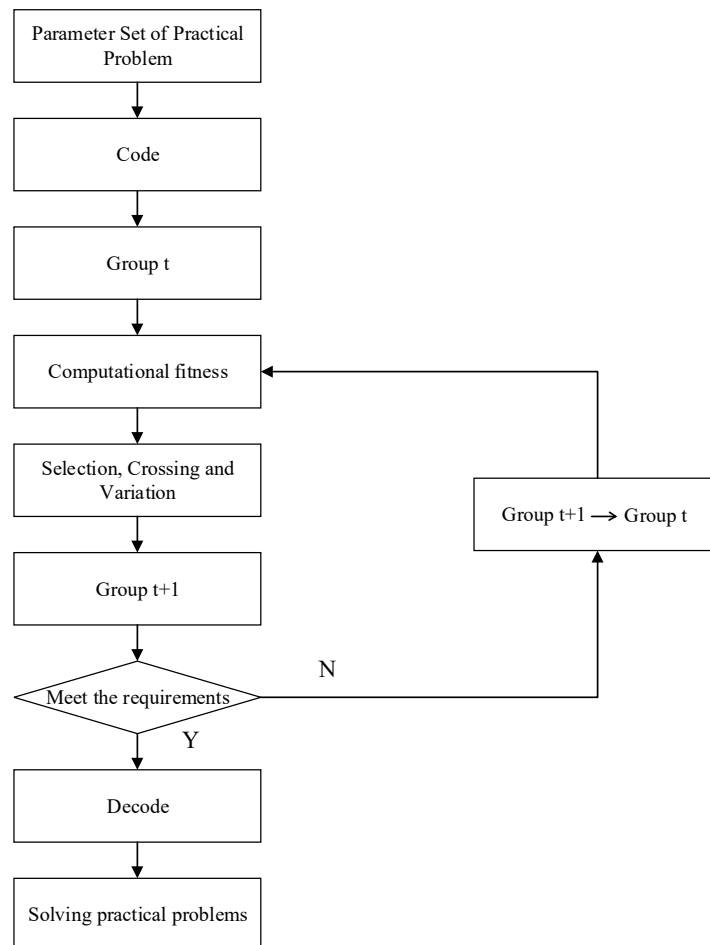


Figure 3. Steps of a genetic algorithm.

The genetic algorithm takes the coding of decision variables as the object of operation. In the process of optimization and calculation, it can imitate the genetic evolution mechanism of organisms in nature, search for groups composed of multiple groups, and then carry out selection, crossover and mutation operations to generate new groups containing distinct sets of group information. With the continuous evolution of the population, the new population will produce more high-quality individuals. The main components of a genetic algorithm are:

- 1) coding: the common methods are binary coding and floating-point coding, which refer to the process of describing the feasible solution of the problem;
- 2) population initialization: as a random search method, the common method of population

initialization is: random generation of a population containing multiple individuals, with reasonable individuals generated by using algorithms or a priori knowledge to avoid too close similarity between individuals in initialization, so as to converge the evolutionary process;

3) fitness function: mainly converting the objective function of the problem into a fitness function, and then determining the selection probability of the individual, which is also the core of the genetic algorithm;

4) selection, mutation and intersection. A crossover operator has strong global search ability, so, as the main operation, the mutation operator is an auxiliary operation. The basic step diagram of a genetic algorithm is shown in Figure 3.

In Wrapper feature selection, based on a genetic algorithm, each individual of the genetic algorithm represents a subset of features. When the population is initialized, a group of different individuals will be generated randomly. The fitness function includes two parts: the classification accuracy of the classifier and the size of the feature subset. The higher the classification accuracy of the classifier, the smaller the dimensions of the feature subset can be; the larger the fitness value of an individual is, the more chance of its features being inherited by the next generation. The genetic algorithm has low dependence on the problem of evaluating individuals by the fitness function. Moreover, the genetic algorithm has strong global search ability, and is widely used in various fields where feature selection is needed. However, when feature analysis is needed for large-scale data, the Wrapper structure based on a genetic algorithm is inefficient and will not generally be applied.

4.2. Combinatorial feature selection algorithm based on mutual information and a genetic algorithm

With the deepening of information theory and the rapid development of information engineering, the field that it can cover goes far beyond the narrow sense of communication engineering. It has been combined with artificial intelligence, automatic control, system engineering, and other disciplines, and developed into a comprehensive discipline-Information Science.

In information theory, the general source is represented by X , and information entropy is represented by $H(X)$, which is a number representing the overall information measure of the source in an average sense. The information entropy can represent the impurity of X . From the point of view of information theory, if the source is represented by X and the destination is represented by Y , then $H(X)$ is a measure of the prior uncertainty of input variable X before it receives Y , so it is also called a priori entropy; $H(X|Y)$ represents the average uncertainty of input variable X after receiving output Y , which is mainly caused by noise. $H(X|Y)$ is called channel ambiguity. It can be seen that, after receiving Y , some information is obtained and some uncertainties are eliminated. In information theory, the mutual information between X and Y is expressed by $I(X; Y)$, which can be expressed as:

$$I(X; Y) = H(X) - H(X|Y) \quad (3)$$

$I(X; Y)$ can reflect the interdependence between input and output random variables and the degree of statistical constraints between them. Generally, conditional entropy $H(X|Y) \leq$ unconditional entropy $H(X)$, so mutual information $I(X; Y)$ will take a non-negative value. It should be noted that mutual information $I(X; Y)$ and $I(Y; X)$ are equal because of the symmetry of mutual information, which is very valuable for measuring the interdependence between the two attributes.

Mutual information-based feature prescreening is actually a category of feature selection methods based on correlation. The main route of feature selection based on correlation is to evaluate the

correlation between features and category labels on the basis of a specific definition of correlation. After sorting the features in descending order, a feature subset which correlates highly with category labels is selected to achieve the goal of dimensionality reduction. This selection method is very important for feature selection. Besides mutual information, there are linear correlation coefficients to measure the correlation of two attributes. The linear correlation coefficients of X and Y attributes are expressed as follows:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (4)$$

In the equation, \bar{x} and \bar{y} indicate the mean value of X and Y, respectively, and the range of r is $-1 \sim 1$. If X and Y are completely linear, then r is -1 or 1; if X and Y are not linear, then r is 0. The linear correlation coefficient is limited to the calculation of the correlation between numerical attributes, and it cannot capture non-linear correlation between attributes.

The second stage of the combinatorial feature selection algorithm (MI-GA) is to search the final feature subset through a genetic algorithm in the feature set after mutual information screening. Using binary coding, the length of coding is the number of features in the subset after the output of the algorithm. The result of feature ranking based on mutual information is used to guide the population initialization, and the result of feature ranking based on mutual information provides the prior knowledge of the problem for the genetic algorithm.

5. Deep learning model

5.1. Overview of deep neural network

A deep learning model belongs to a kind of non-linear machine learning model. Because its concept is inspired by neural networks in the biological brain, it is also called a deep neural network model. Neuron cells are the basic units of the brain's biological mechanism. Each neuron cell is differentiated into dendrites and axons upstream and downstream, respectively, as shown in Figure 4. When the nerve signal starts to transmit under stimulation, the upstream neurons release chemicals, and the dendrites receive the signals and process them, before the axons release them to the downstream neurons. Based on this structure of the brain's neural network, researchers constructed an artificial neuron structure, as shown in Figure 5. The artificial neuron can receive more than one signal input, and then a nonlinear activation function can be driven by weighted summation results.

A convolutional neural network (CNN) in deep learning is a kind of neural network that can deal with time series data or image data and other data with a grid structure. In other words, at least one layer of the network uses a convolution operation to replace the general matrix multiplication operation. Time series data can be regarded as one-dimensional data on the time axis; image data is two-dimensional data with two dimensions (height and width); and video data can be regarded as three-dimensional data. Compared with the traditional full-connection layer, a CNN is characterized by small parameters and more efficient calculation.

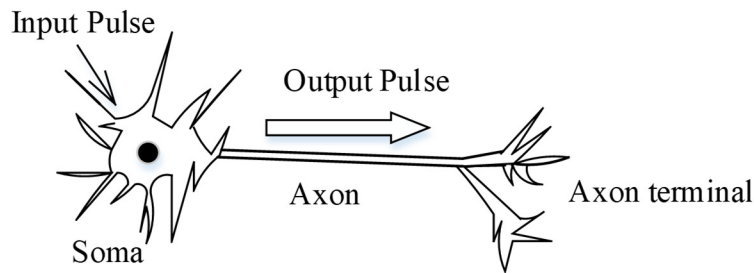


Figure 4. Neuronal cell structure.

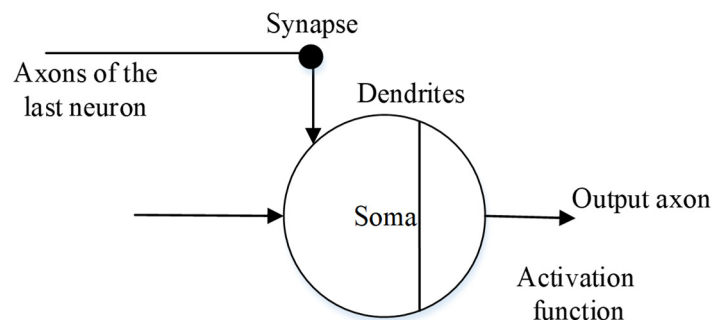


Figure 5. Cell structure of artificial neurons.

A recurrent neural network (RNN) is a kind of neural network algorithm. It can be seen as a process of generating a directed graph by connecting the edges of different nodes. This structure enables it to deal dynamically with different behaviors in a time series. Compared with the traditional feedforward neural network, this network can use the internal state to process the input of the sequence. In addition, the structure also enables it to be applied to tasks such as handwriting recognition and speech recognition.

Reinforcement learning is a widely used machine learning algorithm in complex scenes. This algorithm has been widely used in machine control tasks. In recent years, with the rapid development of deep learning, the reinforcement learning method has been widely used in computer vision, automatic driving, and other new fields. Recently, there are many reinforcement learning algorithms based on deep learning. In 2015, a deep Q network (DQN) was born, in which the Q learning method was used to solve the action value function problem in reinforcement learning. Unlike the traditional manual feature extraction method, the model can carry out end-to-end learning.

5.2. Deep learning model for feature information sharing

At present, most machine learning models are unable to cope with inconsistent input distribution and high dimensionality. Therefore, an algorithm is proposed that combines a deep learning model with lifelong machine learning, based on the ELLA algorithm; on this basis, a new deep learning model (HLLA, Hierarchical lifelong learning algorithm) is ultimately formed. By adding the feature extraction model of unsupervised deep learning, the information data after feature extraction can better fit the same distribution, which is helpful for knowledge transfer of different distributions in a lifelong machine learning system, while input feature representation can improve the performance of traditional

supervised machine learning models when the number of tags is small and the dimensionality of data is high.

The process of machine learning based on deep learning model optimization is divided into four main steps: unsupervised pretraining; maintenance and updating of the feature model; feature extraction; and supervised machine learning. This deep learning model has the following advantages:

1) the deep learning model has been proved to be effective in preventing the subsequent classification model falling into a local optimal solution under the condition of insufficient labels, which can greatly improve the efficiency of machine learning.

2) the deep learning model generates neural network weights in the unsupervised pre-training process, which will be a significant leap forward in machine learning in the context of big data.

3) a hierarchical feature extraction network can increase the complexity of the model and greatly improve the performance of the learning model in dealing with non-linear separable data and high-dimensional data.

4) data will contain the features trained and learned by many unlabeled data before entering the deep learning model, which is similar to the process of knowledge transfer and has a significant effect on subsequent integration and refinement of knowledge.

The HLLA algorithm based on DBN and ELLA assumes that there are a total of T_{max} supervised learning tasks $Z^{(1)}, Z^{(2)}, Z^{(T_{max})}$, and many unlabeled data x_u , in accordance with the distribution of learned tasks. The label of each learning task $Z^{(t)} = (X^{(t)}, y^{(t)})$ is determined by a real implicit function $f^{(t)}$. At any time point, the learning system receives a batch of labeled samples from task t , which may be from a new task or a task that has been learned. T represents the total number of tasks learned before. After the training data are received, the learning system aims to learn the learning model $\hat{f}^{(t)}$ of each task through the given training samples to approach the real objective function $f^{(t)}$. Supposing that the DBN model has n_L layers, the weight connection network of the l -th layer is W_l ; the bias vector is b_l ; the number of hidden neurons in each layer is d_l ; the response function of neurons is the sigmoid function; and the objective function of the HLLA model can be expressed as follows:

$$e_T(L, W_l) = \frac{1}{T} \sum_t \min_{s^{(t)}} \left\{ \frac{1}{n_t} \sum_i^{n_t} \text{loss} \left(\hat{f} \left(h_{i, n_L}^{(t)}; Ls^{(t)} \right), y_i^{(t)} \right) + \mu \|s^{(t)}\|_1 \right\} + \lambda \|L\|_F^2 \quad (5)$$

$h_{i, n_L}^{(t)}$ shows the feature representation of the i -th input of the task t in the l -th layer, and n_t represents the number of training samples of task t . The parameters $\theta^{(t)}$ of each task model are composed of column vectors in matrix L by linear combination. The HLLA model has two main optimization objectives: knowledge base matrix L and network matrix W_l for layer connection of the depth model. If L is optimized first and then W_l , the characteristics of the last layer will be changed due to the change of network weights. For each new task, the deep learning feature network W_l needs to be optimized first, and then the knowledge base matrix L is optimized.

Before the learning process starts, it effectively uses information from several unlabeled data X_u , and constructs a DBN feature extraction model through unsupervised training, so that the labeled training data $((X^{(t)}, y^{(t)}))$ is more consistent with a distribution after feature extraction, making the initial value of the ELLA knowledge base better represent and fit the distribution of task parameters, as well as enabling the subsequent training process to achieve convergence more easily.

The HLLA model consists of a fully shared feature extraction model based on deep learning and supervised lifelong machine learning. The learning process of the new model includes four main steps:

1) before accepting the continuous learning task, sample extensive task-related unlabeled data from the spatial distribution of the input data, and then train the DBN network model by using the unsupervised reconstruction pretraining method in the DBN model training step described in the previous section;

2) fine-tune and update the generated DBN neural network to adjust the feature network to avoid affecting the representation of features by distribution drift;

3) extract the layer by layer features of the current DBN network pair by a task that is about to enter the lifelong learning system;

4) supervised lifelong machine learning (i.e., the process of parameter sharing and hypothesis integration between tasks).

5.3. Feature extraction of CNN based on transfer learning

In order to bridge the gap between low-level visual features and high-level semantics, a CNN is applied to image feature extraction. In addition, by introducing the migration learning idea and using the pre-training model obtained from the ImageNet data set, the problem of model over-fitting caused by the initial training can be solved. A sample image is used to adjust the model parameters through BP, so as to generate a new model suitable for the image classification task.

The AlexNet network has the characteristics of local connection, weight sharing, pooling operation and so on. As an example of a CNN, AlexNet is composed of five convolution layers, three pooling layers, and three full-connection layers, including 60 million parameters and 650,000 neurons, which can gradually transform low-level features into high-level features. Among them, the first several layers mainly obtain low-level features, such as color, edge, texture, and so on, while the second several layers can produce high-level features for direct recognition and classification.

Transfer learning can transfer knowledge from the original domain to the target task. In order to improve the generalization ability of AlexNet to the sample image, the activation function of the training CNN is selected as the rectification linear unit to avoid gradient disappearance during the training process of the model; local response normalization is used to smooth the feature image, and multi-layer output feature combination can improve the image classification accuracy.

5.4. Experimental methods and data analysis

Application in an image classification and recognition task: in order to evaluate the proposed model, this experiment uses a large-scale image description data set, MSCOCO, including 81,903 training images, 39,893 verification images, and 40,015 test images. Finally, 10,000 samples are extracted from the training set as the development set, and the performance of unsupervised feature learning is evaluated.

First, many small patches are randomly extracted from the original training image on MSCOCO. Each patch is a 6×6 -pixel image, which is represented by 36-dimensional vector RD. A total of 40,000 patches are extracted and used for unsupervised feature learning. The classification error rate of the linear support vector machine (SVM) classifier, based on unsupervised learning and feature training in the MSCOCO database, is shown in Table 1.

Table 1. Error rate of linear SVM classifier based on unsupervised learning and feature training in MSCOCO database.

Model/K	400	800	1200	1600
Kmeans	1.40	1.30	1.15	1.12
Spkmeans	1.08	0.90	0.87	0.83
movMF	0.90	0.81	0.82	0.83
PCA-movMF	0.87	0.76	0.74	0.75
Optimization model	0.75	0.70	0.63	0.67

Each feature extractor performs a convolution operation on the image to get the feature expression of each local patch of the image. Then, the image is divided into four quadrants, and each quadrant's feature gets a feature vector by accumulating. In order to verify the quality of the feature expression extracted by different feature extractors, a simple linear SVM is trained in the back-end to classify the image by taking the extracted feature vector as the input, and the quality of the extracted feature expression is judged by the accuracy of the image classification. In the case of the same back-end classifier, the lower the error rate of image classification, the better the feature expression extracted by the corresponding front-end feature extractor. The experimental results show that the deep learning model based on feature extraction can further improve the performance of image classification. When $K = 1200$, it can achieve a classification error rate of 0.63%, which reflects an excellent performance by the unsupervised database algorithm based on the model.

Application in speech recognition task: a speech recognition database is selected to verify the impact of various algorithms on the performance evaluation of data sets. A total of 90 subjects' pronunciation of 26 English letters is collected. All subjects are divided into three groups according to the similarity of pronunciation: A, B and C, 30 people in each group. Sample feature extraction includes information such as the spectrum coefficient feature and the acoustics feature, with a total of 367 dimensions. An overview of the experimental voice database is shown in Table 2.

Table 2. Experimental speech database information.

Name	Number of tasks	Dimensions	Total sample size
Speech recognition	3	367	4019

Table 3. Classification accuracy (%) of four algorithms in speech recognition database.

Algorithm	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
STL	65.5	71.2	74.7	76.8	78.3	80.3	82.0	82.4	82.8	83.6
DG-MTL	67.8	73.6	80.3	83.3	84.8	84.5	84.9	85.7	86.2	85.9
ELLA	73.1	77.3	82.2	86.0	87.6	89.4	91.3	92.8	93.3	94.0
HLLA	79.1	80.5	84.6	87.9	89.8	90.9	93.0	93.7	94.8	95.5

By comparing the STL algorithm, DG-MTL algorithm, ELLA algorithm, and HLLA algorithm, the reliability of the deep model based on the HLLA algorithm was analyzed. The experimental results show that the HLLA algorithm optimized in this paper performed best in all cases on the speech recognition database. Especially at the key points of 50 and 100%, its classification accuracy is higher than that of other classification algorithms. The classification accuracy of the four algorithms in the

speech recognition database is shown in Table 3, and the comparison of the classification accuracy of each algorithm in the speech database is shown in Figure 6.

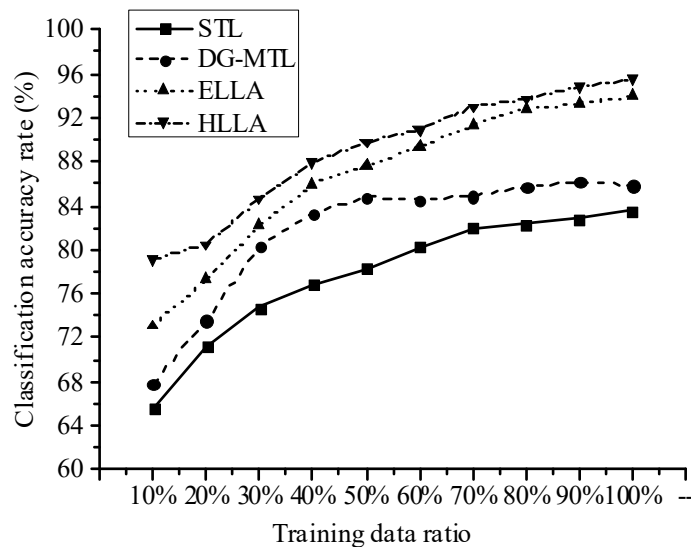


Figure 6. Comparison of classification accuracy of speech database based on different algorithms.

According to the experimental results of four different training set sizes in the speech recognition multi-task database, the abscissa in Figure 6 is the proportion of the training set used (100% indicates using all training data), and the ordinate is the classification accuracy (ACC) index of various algorithms in various situations; the higher the value, the better the performance. The HLLA algorithm proposed performs the best in all cases on this database. Even when the number of tags in each task is reduced, the new HLLA algorithm still maintains good performance because the features extracted from feature extraction carry more structured information than the original data and are more consistent with one distribution. On the other hand, the performance of the ELLA algorithm alone is seriously affected because the amount of information is not enough to fit the overall distribution.

Figure 7 describes the task correlation matrix learned by the HLLA algorithm on two databases. The abscissa and ordinate of the matrix represent the task number, and the depth of the grid colour is proportional to the correlation of the corresponding coordinate task. The Animals dataset contains 50 categories and 30,475 images, each of which has only one category label. Two kinds of dogs and four kinds of birds are selected as the normal group, and the remaining animals are randomly selected as the parent group, comprising the two-class multi-task database. According to the learning of task correlation based on reconstruction error of the HLLA algorithm, in the Landmine database, the first 14 tasks are highly correlated, belonging to the first group, while tasks 15–30 are highly-correlated, belonging to the second group, and the correlation between the two groups is low. This learning result is also consistent with the actual situation that the database itself comes from two different surface areas. On the Animals database, the HLLA algorithm also successfully divided bird and dog into three independent groups, dog 1, bird 1 ~ bird 4, and dog 2. The statistical results of animal image classification accuracy in Figure 8 reveal that, the HLLA algorithm model is the optimal algorithm in almost all cases, especially in the case of fewer tags, there is a considerable gap between HLLA and other algorithms.

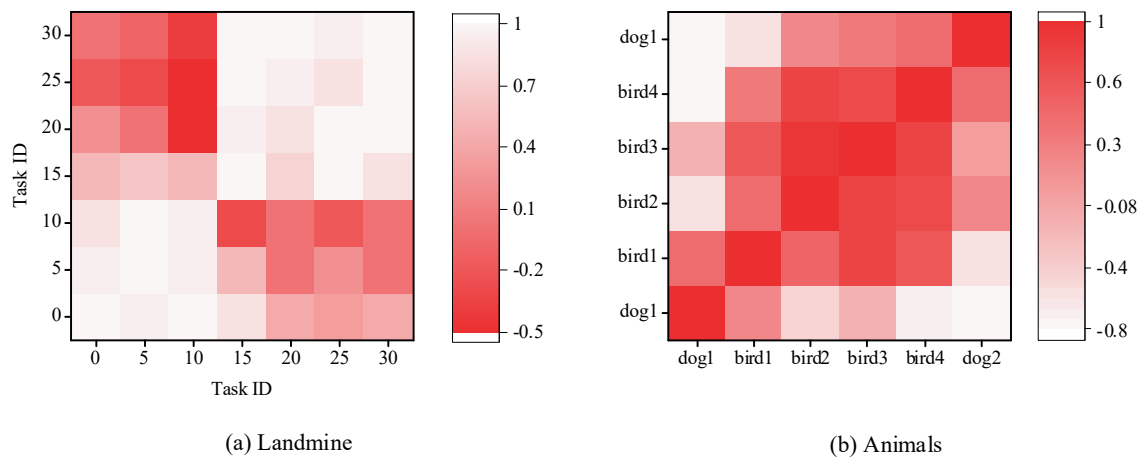


Figure 7. Task correlation matrix learned by HLLA algorithm.

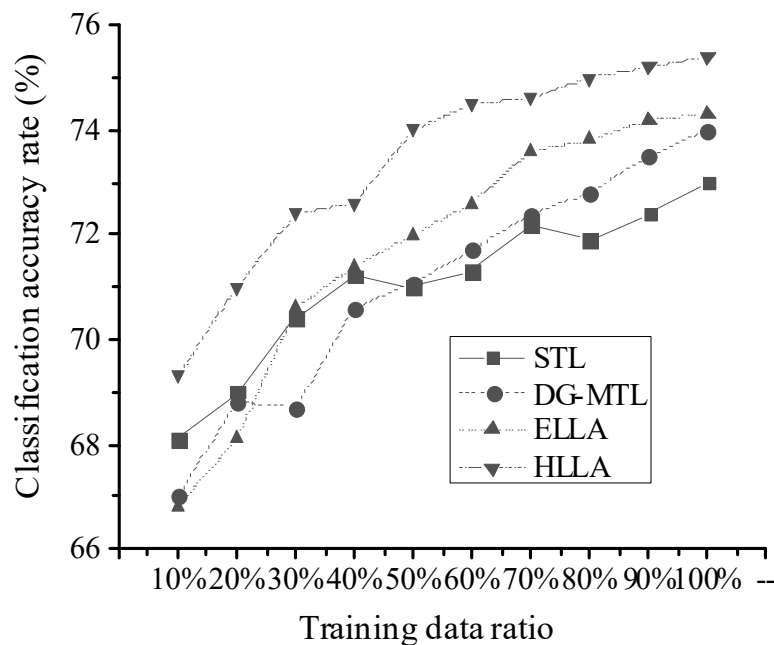


Figure 8. Accuracy of animal image classification.

6. Conclusions

After the ELLA and HLLA algorithms are learned, a deep learning model based on ELLA is constructed and optimized, so that the learning equipment can carry out more lasting learning, and can learn from the huge and complex data centralized databases and groups in the big data era. The relevant information and research results can improve the learning equipment for learning. It can provide the design concept for the new learning equipment combined with learning technology in the future, which is a major achievement to open up practice and connect the future. The great success of deep learning

depends on its powerful ability to fit data, the explosive growth of data available for training, and the development of distributed optimization algorithms. Deep learning has brought unexpected success in many complex tasks. The main concept of machine learning is to select a group of representative features to build a model for a learning task, and to achieve self-improvement through continuous learning. In recent years, many machine learning algorithms have been shown to be negatively affected by irrelevant and redundant features. Therefore, selecting a representative feature subset from the original feature set is of positive significance to many problems in the field of machine learning. With the explosive growth of multi-source heterogeneous data, big data has become the focus of attention. How to learn efficiently and continuously in huge and complex information is the focus of current research. The existing trend in lifelong machine learning algorithm for high-dimensional data is to use a traditional dimensionality reduction method (such as principal component analysis) to reduce the data dimensionality first and then input it into the lifelong machine learning system. This method, however, does not fully utilize the data information and knowledge of previous tasks, so the effect of some high-dimensional multi-task databases is not ideal. On the other hand, if the dimensionality is not reduced, the multi-task high-dimensional database can be directly input into the existing lifelong machine learning system, which will lead to excessively long system learning time, so that the system does not have practicability.

Based on the deep learning model and feature extraction, several representative machine learning algorithms are analyzed, and the results of the optimized deep learning model in practical applications are compared with those of other algorithms. First, the basic theory of machine learning and deep learning is introduced, and the structure of a machine learning system for dealing with multiple tasks in different distributions is simply analyzed. Based on the model structure of a deep neural network, a more effective acoustic model and language model of speech recognition is studied. How to further optimize the structure of the existing deep neural network to improve the performance of the model is discussed, aiming to explore the working mechanism of the deep neural network. In addition, a deep learning model for feature information sharing is proposed, based on the combined feature selection algorithm of mutual information and a genetic algorithm. A feature learning network can be a kind of deep learning feature extraction model which can update network parameters by using a back propagation algorithm. A deep sharing network can extract highly representative features from high-dimensional and unlabeled data, thus greatly reducing the amount of tag data needed for learning new tasks.

The experiments show that the optimized HLLA algorithm performs best in all cases of the speech recognition database, and the classification accuracy is significantly higher than that of other algorithm models. The deep learning model based on feature extraction can further improve the performance of image classification. When $K = 1200$, it can achieve a classification error rate of 0.63%, which reflects excellent performance of the unsupervised database algorithm based on the model. In the model experiment stage, the extraction of subjects' voices is carried out in a relatively quiet environment, without considering the interference of noise. In a practical application, it is easy to be disturbed by the external environment, which leads to inaccurate information extraction, which will be the part to be optimized in the follow-up study.

The study focuses on a feature selection algorithm for the data in large-scale integrated deep learning according to features, but it is limited to the joint update sequence of vector combination in feature framework and knowledge database. When the vector groups in the feature framework and knowledge base are updated separately and sequentially, the change of the feature network may no

longer be suitable for distributed learning tasks. In the future research, the focus will be how to carry out effective feature selection and feature network division in the unsupervised learning field, to avoid the negative transfer problem in the grouping of related learning tasks. Moreover, in the integrated deep learning based on feature selection, the proposed algorithm does not deliberately increase the difference degree of a classifier. In the feature selection of integrated deep learning, the focus of future research will be how to improve the accuracy of a single classifier, thus further increasing the differences between them.

Acknowledgments

We would like to thank you for following the instructions above very closely in advance.

Conflict of interest

The authors declare no conflict of interest.

References

1. A. Esteva, K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, et al., Deep learning-enabled medical computer vision, *NPJ Digital Med.*, **4** (2021), 1–9.
2. D.T. Nguyen, M. B. Lee, T. D. Pham, G. Batchuluun, M. Arsalan, K. R. Park, Enhanced image-based endoscopic pathological site classification using an ensemble of deep learning models, *Sensors*, **20** (2020), 5982.
3. T. Higaki, Y. Nakamura, J. Zhou, Z. Yu, T. Nemoto, F. Tatsugami, et al., Deep learning reconstruction at CT: phantom study of the image characteristics, *Acad. Radiol.*, **27** (2020), 82–87.
4. A. Hakim, Y. Mor, I. A. Toker, A. Levine, M. Neuhof, Y. Markovitz, et al., WorMachine: machine learning-based phenotypic analysis tool for worms, *BMC Biol.*, **16**(2018), 1–11.
5. C. Wang, Z. Xiao, B. Wang, J. Wu, Identification of autism based on SVM-RFE and stacked sparse auto-encoder, *IEEE Access*, **7** (2019), 118030–118036.
6. A. N. Aicha, G. Englebienne, K. S. Schooten, M. Pijnappels, B. Kröse, Deep learning to predict falls in older adults based on daily-life trunk accelerometry, *Sensors*, **18** (2018), 1654.
7. A. Fc, B. Ky, B. JI, Deconvolutional neural network for image super-resolution, *Neural Networks*, **132** (2020), 394–404.
8. J. Sun, D. I. Liping, Z. Sun, et al. Estimation of GDP using deep learning with NPP-VIIRS imagery and land cover data at the county-level in CONUS, *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, **13** (2020), 1400–1415.
9. M. A. Khan, I. Ashraf, M. Alhaisoni, R. Damaševičius, R. Scherer, A. Rehman, et al., Multimodal brain tumor classification using deep learning and robust feature selection: A machine learning application for radiologists, *Diagnostics*, **10** (2020), 565.
10. F. P. An, Human action recognition algorithm based on adaptive initialization of deep learning model parameters and support vector machine, *IEEE Access*, **6** (2018), 59405–59421.
11. M. Heidarysafa, K. Kowsari, D.E. Brown, K. J. Meimandi, L. E. Barnes, An improvement of data classification using random multimodel deep learning (rmdl), *Int. J. Mach. Learn. Cybern.*, **8** (2018), 298–310.

12. S. Almabdy, L. Elrefaei, Deep convolutional neural network-based approaches for face recognition, *Appl. Sci.*, **9** (2019), 4397.
13. S. H. S. Basha, S. R. Dubey, V. Pulabaigari, S. Mukherjee, Impact of fully connected layers on performance of convolutional neural networks for image classification, *Neurocomputing*, **378** (2020), 112–119.
14. S. H. Wang, Y. D. Zhang, DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification, *ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, **60** (2020), 1–19.
15. S. S. Du, W. Hu, J. D. Lee, Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced, preprint, arXiv: 1806.00900.
16. S. Duari, V. Bhatnagar, Complex network based supervised keyword extractor, *Expert Syst. Appl.*, **140** (2020), 112876.
17. Y. Hua, X. Sui, S. Zhou, Q. Chen, G. Gu, H. Bai, et al., A novel method of global optimisation for wavefront shaping based on the differential evolution algorithm, *Optics Commun.*, **481** (2021), 126541.
18. L. Kang, C. Wu, B. Wang, Principles, approaches and challenges of applying big data in safety psychology research, *Front. Psychol.*, **10** (2019), 1596.
19. A. F. Fuentes, S. Yoon, J. Lee, D. S. Park, High-performance deep neural network-based tomato plant diseases and pests diagnosis system with refinement filter bank, *Front. Plant Sci.*, **9** (2018), 1162.
20. S. B. Dias, S. J. Hadjileontiadou, J. Diniz, L. J. Hadjileontiadi, DeepLMS: a deep learning predictive model for supporting online learning in the Covid-19 era, *Sci. Rep.*, **10** (2020).
21. S. Hizlisoy, S. Yildirim, Z. Tufekci, Music emotion recognition using convolutional long short term memory deep neural networks, *Eng. Sci. Technol., Int. J.*, **24** (2021), 760–767.
22. V. G. V. Vydiswaran, Y. Y. Zhang, Y. S. Wang, H. Xu, Special issue of BMC medical informatics and decision making on health natural language processing, *BMC Med. Inf. Decis. Making*, **19** (2019), 76.
23. K. Stuburi, M. Gaiduk, R. Seepold, A deep learning approach to detect sleep stages, *Procedia Comput. Sci.*, **176** (2020), 2764–2772.
24. G. Yang, S. Yu, Synthesized fault diagnosis method reasoned from rough set-neural network and evidence theory, *Concurrency Comput.: Pract. Exper.*, **31** (2019), e4944.
25. S. Hizlisoy, S. Yildirim, Z. Tufekci, Music emotion recognition using convolutional long short term memory deep neural networks, *Eng. Sci. Technol., Int. J.*, **24** (2020), 760–767.



AIMS Press

©2021 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)