Neural network feature learning based on image self-encoding

Yangyang Liu¹, Minghua Tian², Chang Xu¹ and Lixiang Zhao¹

Abstract
With the rapid development of information technology and the arrival of the era of big data, people’s access to information is increasingly relying on information such as images. Today, image data are showing an increasing trend in the form of an index. How to use deep learning models to extract valuable information from massive data is very important. In the face of such a situation, people cannot accurately and timely find out the information they need. Therefore, the research on image retrieval technology is very important. Image retrieval is an important technology in the field of computer vision image processing. It realizes fast and accurate query of similar images in image database. The excellent feature representation not only can represent the category information of the image but also capture the relevant semantic information of the image. If the neural network feature learning expression is combined with the image retrieval field, it will definitely improve the application of image retrieval technology. To solve the above problems, this article studies the problems encountered in deep learning neural network feature learning based on image self-encoding and discusses its feature expression in the field of image retrieval. By adding the spatial relationship information obtained by image self-encoding in the neural network training process, the feature expression ability of the selected neural network is improved, and the neural network feature learning based on image coding is successfully applied to the popular field of image retrieval.

Keywords
Image self-encoding, deep learning, neural network feature learning, image retrieval

Date received: 7 January 2020; accepted: 18 March 2020

Introduction
The in-depth application of big data technology and mobile Internet has made the growth trend of image data become explosive. At the same time, humans generally acquire knowledge in different fields by perceiving external information. Therefore, natural images are a relatively intuitive way of information perception.¹ According to the relevant statistics, the amount of information that humans obtain through vision accounts for about 70%. Image retrieval is also a technique for studying how to accurately and quickly locate target image information. In the 1970s and 1980s, image retrieval was mainly based on text, that is, based on text information or related key words input by people and then matched related images. This relies on manual descriptions of the images in the retrieved database. Image retrieval based on text descriptions is therefore converted into retrieval of text content. Image retrieval based on image content has been widely used. Image retrieval is

¹ School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, Hubei, China
² College of Electronic Information Engineering, Inner Mongolia University, Hohhot, Neimenggu, China

Corresponding author:
Yangyang Liu, School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, Hubei 430070, China.
Email: liuyangyangcug@163.com
available in the fields of intellectual property, medical image processing, remote sensing image processing, and public safety. However, the traditional content-based image retrieval technology also has certain defects, such as low accuracy and insufficient speed on the large image. Excellent feature representation not only can represent the category information of the image but also can capture the relevant semantic information of the image. With the rapid development of computer technology, the performance of computers has made a qualitative leap. Deep learning technology and neural network feature learning have been paid more and more attention by researchers in the field of image processing, which shows the extremely high application value and broad prospects in the field of image retrieval. The research on machine recognition technology will definitely help to improve the performance of image retrieval and to promote the rapid improvement of image retrieval technology.

At present, in the field of image retrieval, there are still problems of slow retrieval speed, long time-consuming, and low retrieval accuracy, and users cannot find the images they need quickly and efficiently in a short time. The existence of these problems obviously does not meet the needs of users. Therefore, this article proposes the study of neural network feature learning based on image self-encoding and uses robot learning to perform image recognition to achieve satisfactory image retrieval results and to improve image retrieval performance and accuracy.

The first section of the article is an introduction to the theme of the article and an analysis of the form of the current field; the second section is an overview of the relevant literature; the third section is a detailed description of the methods used to achieve the innovation of the article; the fourth section is an experimental part, including experimental data and experimental procedure; the fifth section is experimental results and the experimental results are analyzed and discussed. The sixth section summarizes the whole article and points out the innovative points of the article.

**Related work**

In the image retrieval, the earliest image retrieval method is attribute-based retrieval. In the literature, the image is characterized by the attribute set and the image query function is completed. Then the text retrieval method is used. For example, in the literature, Pramod Sankar uses an interactive text retrieval idea to establish a retrieval system using image annotation. In this article, the image annotation is improved, and the idea of using the probability theory for reverse labeling is proposed, which effectively solves the problems of long labeling time and low labeling performance for massive data in the prior art. Since 2004, content-based image retrieval has gradually become a popular research discipline. It covers a wide range of research areas, such as computer vision, image processing, and pattern recognition. By analyzing image visual information and related content, and using new data models, it has designed an effective, reliable, and user-friendly system. In the current commercial system, the requirements for database retrieval are different from those of the traditional database, which means that the traditional retrieval method is not suitable for the current system. Therefore, many meaningful systems for mass data services use content-based image retrieval. Next, we will introduce some of the more stable typical image retrieval systems. Take the classic content-based image retrieval (CBIR) system from abroad as an example. The US “Digital Library” project began in 1994, and the second joint autonomy program was implemented in 1998, which led to a global wave of research on digital libraries, which also promoted CBIR as a hot research discipline. Virage is the CBIR engine developed by the US company VIRAGE. The system can query visual information through texture, features, color, and color layout and any combination of these four methods. Query by image content (QBIC) is the first commercial image retrieval system developed by IBM. The system can be queried by example graphs, outlines, sketches, specific colors, and other information. Photo-book is an interactive tool developed by a multimedia lab at Massachusetts Institute of Technology (MIT) in 1994. The tool is mainly for browsing and querying images. In addition, there are many image retrieval systems such as MARS, Netra, Web Seek, and Retrieval Ware. In large-scale image retrieval, it is usually necessary to construct a low-dimensional indexing method for feature vectors after extracting the global features of images. Most indexing methods based on descriptions of global features use nearest neighbors or nearest neighbors search method. One is the more classical k-dimensional (KD) tree method, which is a classical nearest neighbor retrieval method. This method has better retrieval efficiency when the dimension is lower, but the data dimension is higher. The retrieval complexity is relatively high. The other is to perform clustering and quantitative analysis on feature vectors. For example, Fukunaga and others use k-means to cluster the established tree structure to obtain a lower dimension. On this basis, Hou proposed a tree-like index structure of hierarchical k-means and compared the similarity between leaf nodes, which is suitable for large-scale image retrieval. Another method is the locality-sensitive hashing (LSH)-based method that has been used in image retrieval in recent years, which is an approximate nearest neighbor search method. The main method is to map the original space to the Hamming space and to construct multiple hash functions under the corresponding conditions, that is, after the two adjacent or similar vectors in the original space are hashed into the Hamming space. However, the similar probability is large enough. The high-dimensional vector can be encoded into a binary form by the hash function. Compared with the KD tree method, the large-scale image retrieval is more obvious. The LSH method occupies less storage space and has a larger data
dimension. Obvious advantage. In addition, the LSH method is constantly being improved. The image retrieval technology used by these systems cannot meet all the needs of users, but they have made meaningful exploration in some aspects, laying the foundation for future research. Since the 1990s, the field of image retrieval has experienced more than 20 years of development and achieved fruitful research results. However, there are still many problems to be solved in this field, which still attracts the continuous attention of researchers in different fields.

The neural network learning method has been used in various neighborhoods and has achieved many achievements. Over the years, along with the vigorous development of research directions such as machine learning, artificial intelligence, and data mining, scholars have been paying close attention to the development of neural network algorithms. Professor Hiton proposed the concept of “deep learning” in 2006. In fact, the deep learning method is a deep neural network framework, which has achieved good results in terms of image, target detection, and speech. Autoencoder, as a basic module often used in deep learning, can effectively extract features from big data. Autoencoder is an unsupervised learning algorithm. To improve the feature extraction performance, some scholars have continuously improved and applied this. For example, Wanli et al. proposed a noise reduction autoencoder (AE) that randomly destroys some data in the input features, the output data still use the original data, and the self-encoding network estimates the corrupted values based on the original data, so that the network is generalized more capable. Haibo combined the multilevel noise reduction self-encoding neural network with the undersampling local update meta-cost algorithm to further apply to clinical classification diagnosis and to improve the auxiliary diagnostic performance of the classification model. Vincent et al. combined sparse autoencoder and deep belief network (DBN) to establish SD algorithm for text classification. A DBN is combined to establish an SD algorithm for text classification. Moreover, since deep learning self-encoding and neural network have good target information representation ability, robustness, and generalization, complex function approximation can be realized. Therefore, domestic and foreign scholars have successfully used it for image processing, natural language processing, and so on. The field has made breakthrough progress. Shuai are divided into several data sets according to the feature A, an information entropy is calculated for each data set and then they are weighted and summed according to the data ratio, so that the entropy of the divided total data set is obtained, and the entropy is recorded. Take the entropy as $H(X|A)$. Then the information gain of feature A can be expressed as $H(X) - H(X|A)$. The greater the information gain, the more useful the feature is. The feature selection of the filtering method is carried out in a greedy manner, and the relationship between the features cannot be considered, so the selected feature subset is not optimal.

(2) Wrapper

The selection of the feature of the packaging method refers to comparing the selection of feature subsets to a search optimization problem, generating different feature subsets, evaluating the subsets, comparing them with other subsets, and finally selecting the optimal features subset. Because the time complexity of the violent search feature subset is extremely high, many heuristic optimization algorithms can be used here. For example, common ant colony algorithms, genetic algorithms, and other optimization algorithms have been used for feature selection. Obviously, the packaging method can theoretically find the optimal

Method

Feature selection method

In most cases, the number of features in the data may be much larger than the amount of training data required, in which case this will inevitably make training of the model extremely difficult. Therefore, measures must be taken to screen features. There are three types of commonly used feature selection algorithms, which can be divided into filter, wrapper, and embedded.

(1) Filter

Filter feature selection refers to scoring each feature separately and then selecting the top scores based on the score. Common scoring algorithms have correlation coefficients, information gains, and $\chi^2$ tests. Here we introduce the information gain methods that are often used. Information gain refers to the degree of change of data entropy after dividing data according to a certain feature. It is assumed that the target variable $y$ may have $m$ categories, and the data belonging to each category in the data set $X$ are $p_1, p_2, \cdots, p_m$, respectively, and the information entropy defining the data is as follows

$$H(X) = -\sum_{i=0}^{m} p_i \log(p_i)$$

After the data $X$ Shuai are divided into several data sets according to the feature $A$, an information entropy is calculated for each data set and then they are weighted and summed according to the data ratio, so that the entropy of the divided total data set is obtained, and the entropy is recorded. Take the entropy as $H(X|A)$. Then the information gain of feature $A$ can be expressed as $H(X) - H(X|A)$. The greater the information gain, the more useful the feature is. The feature selection of the filtering method is carried out in a greedy manner, and the relationship between the features cannot be considered, so the selected feature subset is not optimal.

(2) Wrapper

The selection of the feature of the packaging method refers to comparing the selection of feature subsets to a search optimization problem, generating different feature subsets, evaluating the subsets, comparing them with other subsets, and finally selecting the optimal features subset. Because the time complexity of the violent search feature subset is extremely high, many heuristic optimization algorithms can be used here. For example, common ant colony algorithms, genetic algorithms, and other optimization algorithms have been used for feature selection. Obviously, the packaging method can theoretically find the optimal
feature subset, but in reality, due to the running time problem, it is usually necessary to select a good subset and only use the heuristic algorithm.

(3) Embedded

The feature selection of embedding method is to embed the process of feature selection into the learning of the model. The most commonly used embedding feature selection method is the L1 regularization term, which is also called ridge regression. Here, logarithmic regression is used as an example to illustrate the L1 regularization term. The loss function of logistic regression is expressed as follows

$$L(w,x,y) = y \times \log \left( \frac{1}{1 + e^{-w^T x}} \right) + (1 - y) \log \left( \frac{e^{-w^T x}}{1 + e^{-w^T x}} \right)$$

After optimizing this loss function, what is obtained will be dense. For the feature selection, we hope that the learned $w$ is sparse. The L1 regularization term (ridge regression) adds a penalty for $w$ sparsity to the model loss, resulting in a new loss function, as shown below

$$L(w,x,y) = y \times \log \left( \frac{1}{1 + e^{-w^T x}} \right) + (1 - y) \log \left( \frac{e^{-w^T x}}{1 + e^{-w^T x}} \right) + \alpha \sum_{i=0}^{n} |w_i|$$

In the formula, $w_i$ is the value of the $i$th position of the $w$ vector, corresponding to the weight of the $i$th feature. After adding the L1 regularization term (ridge regression), the learned $w$ will be sparse, and the feature corresponding to the item with 0 in $w$ will be deleted, so that the features selected by the remaining model can be retained. The advantages of the embedding method are (1) no additional work is required and (2) incorporating feature selection into the model learning process, so that the selected feature subset is suitable for the model itself.

The above three feature selection algorithms, such as filtering, packaging, and embedding, can be used to eliminate redundant features.

**Neural network feature learning**

As a deep learning model, convolutional neural networks have the ability to layer learning features. The features obtained by convolutional neural network learning have stronger discriminative ability and generalization ability than the artificial design features. As the research foundation of computer vision, feature representation can learn, extract, and analyze the feature expression of information through convolutional neural network, so as to obtain the general features of stronger discriminant performance and better generalization performance.

Because the deep neural network contains much more network layers than the previous shallow network, the network model is quite complex. A large number of parameters must be adjusted during the training process. At the same time, this model structure is also designed to handle large amounts of input. Deep learning uses the input data to perform unsupervised training methods for bottom-up layer-by-layer training to generate network parameters and supervised training methods for top-down layer-by-layer adjustment of parameters to avoid the phenomenon of falling into local optimum and overfitting when training the network.

(1) Unsupervised learning from the bottom up

This part of the training process is different from the traditional multi-layer perceptron (MLP), which is different from the training method using tagged data. Deep neural network uses the unlabeled data to train the network from the bottom layer, and the output of the current layer is used as the input of the next layer. The whole network is trained layer-by-layer according to this rhythm, and the main parameters such as offset and weight of each layer are obtained. After the training, the layers of the neural network have forward and backward weights, wherein the weight from the lower layer to the upper layer is the wake weight and the weight from the upper layer to the lower layer is the sleep weight. You can then use the wake–sleep algorithm to adjust the weight of each layer: in the wake phase, set the state of each layer of neurons by combining input data and weights and then use the gradient descent algorithm to update the weights between layers; in the sleep phase, based on the wake phase, the uppermost state and the wake weight are determined, and the state of each neuron in the remaining layer is obtained, and the sleep weight is updated again based on these state values.

(2) Top-down supervised learning

This part of the training process uses the already labeled data to fine-tune the parameters of each layer obtained in the previous part of the training and once again update the two-way weight and neuron status of each layer.

In the learning process of the input data, initial parameters of the deep neural network are generated. The initial parameters obtained in this way can find the optimal solution better than the random initial parameters of the traditional MLP and then improve the effect. Therefore, the main reason for the excellent performance of deep neural networks in various data processing is due to the unsupervised training from the bottom up.

**Self-encoding algorithm**

The AE was proposed by Rumelhart in 1986. It is a neural network consisting of three layers: input layer, hidden layer, and output layer. The self-encoder is an unsupervised
neural network model. Its goal is to obtain the original input data itself by learning a mapping, so that the original input data can be output after multilayer transformation. Therefore, the neurons in the middle hidden layer contain the intrinsic information of the data, so that it can be used as a feature of the data. The encoding and decoding process of the AE is shown in Figure 1. One of the outstanding advantages of the self-encoding algorithm is that its model has a strong expressive ability and can automatically learn image features. Data with high-dimensional sparse features can be handled well here, and unsupervised learning can reduce extraneous and redundant data, thereby reducing the number of dimensions.

Suppose the input layer of the autoencoder and the dimension of the output layer are \( n \), the dimension of the hidden layer is \( m \), and the set of samples is \( S = \{ \mathbf{x}^{(i)} \}_{i=1}^{N} \).

Let the encoding function be \( f \), the encoding process refers to the input layer to the hidden layer, that is, the input expression \( X \) code to get a new expression \( Y \); let the decoding function be \( g \), the decoding process refers to the hidden layer to the output layer, that is, the encoded \( Y \) is decoded back to \( X \).

The coding part of the AE is to map the input data to the hidden layer unit through the nonlinear mapping function, and let \( h \) be the hidden layer of the neural unit activation and then have the following expression

\[
h = f(x) = S_f(wx + p)
\]

where \( w \) represents a weight matrix consisting of the weights between the input layer and the hidden layer. \( S_f \) is the encoder activation function, usually replaced by the Sigmoid function, that is

\[
f(x) = \frac{1}{1 + e^{-x}}
\]

The principle of the decoding part of the AE is similar to the coding principle described above, and the hidden layer obtained after encoding is reconstructed from the previous original input data. Then there is an expression of the following decoding function

\[
y = g(h) = S_g(\tilde{w}h + q)
\]

where \( y \) represents the decoder reconstructing the original input data, \( S_g \) is the decoder activation function, usually replaced by a Sigmoid function or an identity function, \( w \) represents a weight matrix, and the matrix element is between the hidden layer and the output layer. The value elements are composed. And the parameter of the AE is \( \theta = \{ w, p, q \} \).

The input data \( x \) are predicted to obtain the output data \( y \). If the proximity of the output data \( y \) to the input data \( x \) is acceptable, the AE retains most of the original data, and therefore, it can be considered that the automatic compiler has been trained.

The reconstruction error function \( L(x, y) \) is used to describe the closeness of \( y \) and \( x \).

In the case where the decoder activation function \( S_g \) takes an identity function

\[
L(x, y) = ||x - y||^2
\]

In the case where the decoder activation function \( S_g \) is taken as a Sigmoid function

\[
L(x, y) = -\sum_{i=1}^{n} [x_i \log y_i + (1 - x_i) \log (1 - y_i)]
\]

When the sample set used for training is \( S \), the representation of the overall loss function of the self-encoder is as follows

\[
J_{AE}(\theta) = \sum_{x \in S} L(x, g(f(x)))
\]

Finally, iteratively calculates the minimum value of \( J_{AE}(\theta) \). The calculation method is to repeat the gradient descent method, so that the parameter \( \theta \) of the self-encoding network can be solved. Therefore, the training of the self-encoder is completed.

The overall structure of the training model based on self-encoding neural network is shown in Figure 2. In this figure, the autoencoder (AE) consists of an encoder and a decoder. The input data are encoded into a lower-dimensional representation, and then the encoded data are decoded back to the original input data. The reconstruction error is calculated and minimized through an iterative process until the AE is trained.
article, the segmentation image corresponding to the original training image in the training sample database is used for self-encoding, and the intermediate hidden layer corresponding to self-encoding is used as the implicit variable of the segmentation image. At the same time, the implicit variable is used as another training target of the training convolutional neural network image for real-value regression, so that the training target of the trained convolutional neural network not only contains classification information but also contains spatial position information. The purpose of autoencoder in this article is to map the input information to a manifold area. It is expected that the feature expression can be improved by this simple strategy.

**Experiment**

**Data sources**

To evaluate the performance of the proposed algorithm in image retrieval, we conducted experiments on two general image databases. The databases are ImageNet and COCO. Among them, there are a total of 80 categories of COCO data set categories, a total of more than 80,000 images for training and about 40,000 images for test. These images can be well applied to the research and application of image self-encoding neural network feature learning and image retrieval.

**Experimental evaluation criteria**

In this article, neural network feature learning based on image self-encoding is applied to large-scale image retrieval. Similarly, two indicators are used to measure the accuracy of query results: accuracy and recall.

(1) Accuracy is the ratio of the correct query results to the total query results in all query results.

(2) The recall rate is the ratio of the total number of similar samples in the image library in all the query results in the sample image similar to the query image in the image library.

When the total number of similar images in the image library and the query image is fixed, the precision and recall rate of the query result will be related to how many query results are selected as the benchmark. When selecting more query results, the precision rate tends to be smaller, and the recall rate tends to increase.

In clustering experiments, features are clustered first, followed by image retrieval. And the retrieval value $F$ after clustering is calculated at the same time. The search worthy formula is expressed as follows

$$F = \frac{2PR}{P + R}$$

where $P$ represents the accuracy and $R$ represents the recall rate. The appearance of the $F$ value allows us to consider the two at the same time considering the accuracy and recall rate and can reflect the whole more comprehensively and truly. $P$ and $R$ are calculated as follows

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

In the formula:

(1) $TP$ indicates that the prediction is a positive sample and is actually the number of features of the positive sample, that is, the number of two images that are clustered together are correctly classified. If at least one of the two images has the same category and the two images are divided into the same category, the $TP$ value is incremented by 1.

(2) $FP$ indicates that the prediction is a positive sample and the actual number of features of the negative sample, that is, the number of images that should not be grouped together into one category. If two images do not have the same category and the two images are assigned to the same category, the value of $FP$ is incremented by 1.

(3) $FN$ indicates that the prediction is a negative sample, which is actually a feature of the positive sample, that is, the number of images that should not be separated are erroneously separated. If two images

![Figure 2. Overall structure of training model based on self-encoding neural network.](image-url)
have at least one of the same categories and the two images are not assigned to the same category, the value of $FN$ is incremented by 1.

**Experiment platform**

The server used is a Dell workstation with GPU Tesla K20c. The server trains with 20,000 training materials, which takes about 4 h. The steps to build the system platform are as follows:

1. install and deploy Caffe on Windows 64 and test it easily,
2. install and deploy Django and the database,
3. develop the corresponding program.

**Results and discussion**

**Result 1: Feature selection**

In this feature selection experiment, we recorded that a variable subset is selected sequentially or randomly from the original pixel representation as a way to reduce dimensionality. When the selected subset is used as a new expression, the performance of the softmax classifier changes as the new expression dimension changes. When the subset is randomly selected, for each dimension, the code is run 200 times through the experiment, and the following average result is obtained. The experimental results obtained are shown in Figure 3. Blue lines indicate sequential selection subsets and red lines indicate random selection subsets. The abscissa indicates the change in the dimension, and the ordinate indicates the level of feature selection performance.

As shown in Figure 3, as the dimension increases, the performance of the sequential selection subsets increases rapidly and eventually remains at a high level close to 1, while the randomly selected subsets show a monotonously slow increasing trend throughout the process. Its performance eventually tends to stabilize at around 0.75. The specific changes of each stage are described as follows:

1. When the dimension is lower than 100, the performance corresponding to the sequential selection subset is almost unchanged, while the randomly selected subset has a slight increase;
2. When the dimension varies from 100 to 250, as the dimension increases, the performance of the sequential selection subsets increases monotonously, and the growth rate increases sharply, while the randomly selected subset still maintains the previous growth trend and a little slower;
3. When the dimension is in the range of 260–400, as the dimension increases, the performance of the sequential selection subsets increases slowly and tends to be stable, while the randomly selected subset only grows a little, basically not monotonously slow increase;
4. When the dimension is higher than 400, the performance corresponding to the sequential selection of subsets and the randomly selected subsets of dimension increasing no longer grows with their respective growth trends but tend to be stable, and their respective corresponding performances are maintained at levels close to 1 and 0.75, respectively.

**Result 2: Cluster analysis**

The process of clustering is as follows: firstly, the image needs to be retrieved, the convolutional neural network (CNN) feature is extracted, then the cluster center label is obtained according to $k$-means, and then the topN is found in the corresponding sample of the cluster center. If the cluster center topN does not reach N, then the search sample is selected from the sample corresponding to the cluster center closest to the sample, and so on, until the topN is obtained. By referring to the experiments on the COCO database of the model mentioned in the reference, the $F$ values obtained by the methods proposed in this article and Alex, VGGF, VGGS, MIN, Zeiler, and so on, are carried out. The results obtained are summarized in Table 1. The difference in visual results is shown in Figure 4.

**Result 3: Returns the recall and precision in the case where the number of search results is different**

After 30 simulation experiments, different precision and recall ratios were calculated according to the difference in the number of image retrieval results. The number of images returned to the image search result is 2, 5, 8, and 10. The recall and precision obtained are listed in Table 2.
It can be seen from Table 2 that with the gradual increase in the number of returned search results, the precision of image retrieval shows an overall downward trend, while the recall rate is constantly increasing.

| The number of returned search results | Precision rate | Recovery rate |
|--------------------------------------|----------------|---------------|
| 2                                    | 0.20           | 0.20          |
| 5                                    | 0.40           | 0.40          |
| 8                                    | 0.375          | 0.60          |
| 10                                   | 0.40           | 0.80          |

Table 2. Returns the precision and recall rate in the case where the number of search results is different.

The above experiments show that the application of neural network feature learning based on image self-encoding has good stability in image retrieval technology and can improve the retrieval effect. This method is feasible.

Conclusions

In the rapidly evolving field of computer vision, characterization has always played a significant role. But how to get image features that are more in line with human visual mechanisms has always been an important and difficult problem to describe. In addition, in the face of large-scale images with complex visual information on the network, how to help users find the required pictures accurately and efficiently has become a hot research topic in the field of multimedia information retrieval. The development of image retrieval is in the ascendant, and there is still a lot of room for the future. Therefore, consider combining neural network feature learning with image retrieval techniques.

In this thesis, the feature learning based on neural network is studied, and the neural network feature learning model based on image self-encoding is proposed and applied to the field of image retrieval. Through the deep analysis of existing work and the related research on the field of image retrieval, this article believes that the key to image retrieval problem lies in feature learning and expression. The traditional neural network model is a common feature expression model. Because of the lack of supervised information, its feature learning ability and expressive ability are limited, and the high performance and high-quality characteristics brought by people’s desired retrieval technology cannot be realized. In the experiment, the
segmentation training image set corresponding to the training image set is constructed by the segmentation label of the multilabel image data set. Then, the self-encoding neural network is trained, and the implicit variables of the segmentation images corresponding to each training sample are extracted and normalized. The convolutional neural network is then trained by using the implicit variable as the training target corresponding to the original image of the training set. Subsequently, the performance of feature selection was compared and analyzed by clustering experiments. The results show that the research and application of neural network feature learning based on image self-encoding has great feasibility and rationality.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs
Yangyang Liu https://orcid.org/0000-0002-8148-2528
Minghua Tian https://orcid.org/0000-0003-4441-8065
Lixiang Zhao https://orcid.org/0000-0003-2096-8345

References
1. Zhao Q and Cao JF. Joint optimization of feature selection and parameters for multi-class SVM in skin symptomatic recognition. In: The 2009 international conference on artificial intelligence and computational intelligence (AICI ’09), Shanghai, China, 7–8 November 2009, 1, pp. 407–411. IEEE.
2. Deng J, Dong W, Socher R, et al. ImageNet: A large-scale hierarchical image database. In: IEEE conference on computer vision and pattern recognition (CVPR 2009), Miami, FL, USA, 20–25 June 2009, pp. 248–255. IEEE.
3. Swain MF and Ballard DH. Color indexing. Int J Comput Vis 1991; 7(1): 11–32.
4. Sankar PK and Jawahar CV. Probabilistic reverse annotation for large scale image retrieval. In: 2009 IEEE conference on computer vision and pattern recognition, Minneapolis, MN, USA, 20–25 June 2009, pp. 1–6. IEEE.
5. Sun Y, Wang XG, and Tang XO. Deep learning face representation by joint identification-verification. In: 2014 IEEE conference on computer vision and pattern recognition (CVPR), Columbus, OH, 23–28 June 2014, pp. 1891–1898. IEEE.
6. Li WJ, Wang S, and Kang WC. Feature learning based deep supervised hashing with pairwise labels. In: IJCAI’16 Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, New York, US, 9–15 July 2016, pp. 1711–1717. Gerhard Brewka: AAAI Press.
7. Drouin S and Collins DL. Prism: an open source framework for the interactive design of GPU volume rendering shaders. PLoS ONE 2018; 13(3): e0193636.
8. Pentland AP, Picard RW, and Sclaroff S. Photobook: tools for content-based manipulation of image databases. In: Proceedings of SPIE the International Society for Optical Engineering 1994; 2185: 34–47.
9. Niblack W, Ron B, Will E, et al. QBIC project: querying images by content using color, texture and shape. In: Proceeding of SPIE storage and retrieval for image and video databases, Vol. 1908 (ed Niblack W). 1993, pp. 173–187. San Jose, CA: SPIE Press.
10. Pentland AP, Picard R, and Sclaroff S. Photobook: content-based manipulation of image databases. Int J Comupt Vis 1996; 18(3): 233–254.
11. Rao A, Srihari RK, and Zhang Z. Spatial color histogram for content-based image retrieval. In: 11th IEEE international conference on tools with AI, Chicago, IL, 9–11 November 1999, pp. 183–186. IEEE.
12. Lin G, Shen C, Shi Q, et al. Fast supervised hashing with decision trees for high-dimensional data. In: 2014 IEEE conference on computer vision and pattern recognition (CVPR), Columbus, OH, 23–28 June 2014, pp. 1971–1978. IEEE.
13. Ma WY and Manjunath BS. NeTra: a toolbox for navigating large image databases. In: Proceeding of international conference on image processing, Santa Barbara, CA, 26–29 October, 1997, 1, pp. 568–571.
14. Allem JP, Escobedo P, Chu KH, et al. Images of little cigars and cigarillos on instagram identified by the hashtag #swisher: thematic analysis. J Med Int Res 2017; 19(7): e255.
15. Douze M, Jegou H, Sandhawalia H, et al. Evaluation of gist descriptors for web-scale image search. In: International conference on image and video retrieval, Fira, Greece, July 2009. ACM.
16. Friedman JH, Bently JL, and Finkel RA. An algorithm for finding best matches in logarithmic expected time. ACM T Math Software 1977; 3(3): 209–226.
17. Fukunaga K and Narendra PM. A branch and bound algorithm for computing k-nearest neighbors. Comput IEEE T 1975; C-24(7): 750–753.
18. Hou Y, Zhang H, and Zhou S. Bofcnf: efficient image matching with bag of convnet features for scalable and robust visual place recognition. Auton Robots 2017; 42(9): 1–17.
19. Lee J, Sun J, Wang F, et al. Privacy-preserving patient similarity learning in a federated environment: development and analysis. JMIR Med Inf 2018; 6(2): e20.
20. Andoni A and Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Commun ACM 2008; 51(1): 117.
21. Har-Peled S and Kumar N. Robust proximity search for balls using sublinear space. Algorithmica 2018; 80(1): 279–299.
22. Huang Y and Lin Z. Binary multidimensional scaling for hashing. *IEEE Trans Image Process* 2017; 27(1): 406–418.
23. Datta R and Joshi D. Image retrieval: ideas, influences, and trends of the new age. *ACM Comput Surv* 2008; 40(2): 1–60.
24. Feiyun Y. Codebook generation based on self-organizing incremental neural network for image classification. *J Comput Appl* 2013; 33(7): 1976–1979.
25. Meng C, Hongyan L, and Rongrong Z. A pitch detection method based on deep neural network. *Microelectronics & Computer* 2016; 33(6): 143–146.
26. Wanli O, Xingyu Z, and Xiaogang W. Joint deep learning for pedestrian detection. In: *Proceeding of the 2013 IEEE international conference on computer vision*, Sydney, 1–8 December, 2013, pp. 2056–2063. Washington: IEEE Computer Society.
27. Haibo W. *Research of face recognition based on deep learning*. Hefei: Hefei University of Technology, 2014.
28. Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders. In: *Proceedings of the 25th international conference on machine learning*, Helsinki, 5–9 June, 2008, pp. 1096–1103. New York: ACM.
29. Shuai H, Zhiyong Y, Ling X, et al. Stacked denoising autoencoders applied to clinical diagnose and classification. *Appl Res Comput* 2015; 32(5): 1417–1420.
30. Shengjun Q and Zhiping L. Research of text categorization based on sparse autoencoder algorithm. *Sci Technol Eng* 2013; 13(31): 9422–9426.
31. Ding J, Chen B, Liu H, et al. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geosci Remote S* 2016; 13(3): 1–5.
32. Guo D, Sun F, Kong T, et al. Deep vision networks for real-time robotic grasp detection. *Int J Adv Robot Syst* 2016; 14(1): 1729881416682706.