EURASIP Journal on Image and Video Processing

RESEARCH Open Access

Research on image classification model based on deep convolution neural network

Mingyuan Xin[1] and Yong Wang[2*]

Abstract

Based on the analysis of the error backpropagation algorithm, we propose an innovative training criterion of depth neural network for maximum interval minimum classification error. At the same time, the cross entropy and M^{3}CE are analyzed and combined to obtain better results. Finally, we tested our proposed M3 CE-CEc on two deep learning standard databases, MNIST and CIFAR-10. The experimental results show that M^3 CE can enhance the cross-entropy, and it is an effective supplement to the cross-entropy criterion. M3 CE-CEc has obtained good results in both databases.

Keywords: Convolution neural network, Image classification, M^{3}CE-CEc

1 Introduction

Traditional machine learning methods (such as multilayer perception machines, support vector machines, etc.) mostly use shallow structures to deal with a limited number of samples and computing units. When the target objects have rich meanings, the performance and generalization ability of complex classification problems are obviously insufficient. The convolution neural network (CNN) developed in recent years has been widely used in the field of image processing because it is good at dealing with image classification and recognition problems and has brought great improvement in the accuracy of many machine learning tasks. It has become a powerful and universal deep learning model.

Convolutional neural network (CNN) is a multilayer neural network, and it is also the most classical and common deep learning framework. A new reconstruction algorithm based on convolutional neural networks is proposed by Newman et al. [1] and its advantages in speed and performance are demonstrated. Wang et al. [2] discussed three methods, that is, the CNN model with pretraining or fine-tuning and the hybrid method. The first two executive images are passed to the network one time, while the last category uses a patch-based feature extraction scheme. The survey provides a milestone in modern case retrieval, reviews a wide selection of different categories of previous work, and provides insights into the link between SIFT and the CNN based approach. After analyzing and comparing the retrieval performance of different categories on several data sets, we discuss a new direction of general and special case retrieval. Convolution neural network (CNN) is very interested in machine learning and has excellent performance in hyperspectral image classification. Al-Saffar et al. [3] proposed a classification framework called region-based pluralistic CNN, which can encode semantic context-aware representations to obtain promising features. By combining a set of different discriminant appearance factors, the representation based on CNN presents the spatial spectral contextual sensitivity that is essential for accurate pixel classification. The proposed method for learning contextual interaction features using various region-based inputs is expected to have more discriminant power. Then, the combined representation containing rich spectrum and spatial information is fed to the fully connected network and the label of each pixel vector is predicted by the Softmax layer. The experimental results of the widely used hyperspectral image datasets show that the proposed method can outperform any other traditional deep-learning-based classifiers and other advanced classifiers. Context-based convolution neural network (CNN) with deep structure and pixel-based multilayer perceptron (MLP) with shallow structure are recognized neural network algorithms

* Correspondence: 2471739879@qq.com
[2]Heihe University, No. 1 Xueyuan Road education science and technology zone, Heihe, Heilongjiang, China
Full list of author information is available at the end of the article

Springer Open

© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

which represent the most advanced depth learning methods and classical non-neural network algorithms. The two algorithms with very different behaviors are integrated in a concise and efficient manner, and a rule-based decision fusion method is used to classify very fine spatial resolution (VFSR) remote sensing images. The decision fusion rules, which are mainly based on the CNN classification confidence design, reflect the usual complementary patterns of each classifier. Therefore, the ensemble classifier MLP-CNN proposed by Said et al. [4] acquires supplementary results obtained from CNN based on deep spatial feature representation and MLP based on spectral discrimination. At the same time, the CNN constraints resulting from the use of convolution filters, such as the uncertainty of object boundary segmentation and the loss of useful fine spatial resolution details, are compensated. The validity of the ensemble MLP-CNN classifier was tested in urban and rural areas using aerial photography and additional satellite sensor data sets. MLP-CNN classifier achieves promising performance and is always superior to pixel based MLP, spectral and texture based MLP, and context-based CNN in classification accuracy. The research paves the way for solving the complex problem of VFSR image classification effectively. Periodic inspection of nuclear power plant components is important to ensure safe operation. However, current practice is time-consuming, tedious, and subjective, involving human technicians examining videos and identifying reactor cracks. Some vision-based crack detection methods have been developed for metal surfaces, and they generally perform poorly when used to analyze nuclear inspection videos. Detecting these cracks is a challenging task because of their small size and the presence of noise patterns on the surface of the components. Huang et al. [5] proposed a depth learning framework based on convolutional neural network (CNN) and Naive Bayes data fusion scheme (called NB-CNN), which can be used to analyze a single video frame for crack detection. At the same time, a new data fusion scheme is proposed to aggregate the information extracted from each video frame to enhance the overall performance and robustness of the system. In this paper, a CNN is proposed to detect the fissures in each video frame, the proposed data fusion scheme maintains the temporal and spatial coherence of the cracks in the video, and the Naive Bayes decision effectively discards the false positives. The proposed framework achieves a hit rate of 98.3% 0.1 false positives per frame which is significantly higher than the most advanced method proposed in this paper. The prediction of visual attention data from any type of media is valuable to content creators and is used to drive coding algorithms effectively. With the current trend in the field of virtual reality (VR), the adaptation of known technologies to this new media is beginning to gain momentum R. Gupta and Bhavsar [6] proposed an extension to the architecture of any convolutional neural network (CNN) to fine-tune traditional 2D significant prediction to omnidirectional image (ODI). In an end-to-end manner, it is shown that each step in the pipeline presented by them is aimed at making the generated salient map more accurate than the ground live data. Convolutional neural network (Ann) is a kind of depth machine learning method derived from artificial neural network (Ann), which has achieved great success in the field of image recognition in recent years. The training algorithm of neural network is based on the error backpropagation algorithm (BP), which is based on the decrease of precision. However, with the increase of the number of neural network layers, the number of weight parameters will increase sharply, which leads to the slow convergence speed of the BP algorithm. The training time is too long. However, CNN training algorithm is a variant of BP algorithm. By means of local connection and weight sharing, the network structure is more similar to the biological neural network, which not only keeps the deep structure of the network, but also greatly reduces the network parameters, so that the model has good generalization energy and is easier to train. This advantage is more obvious when the network input is a multi-dimensional image, so that the image can be directly used as the network input, avoiding the complex feature extraction and data reconstruction process in traditional recognition algorithm. Therefore, convolutional neural networks can also be interpreted as a multilayer perceptron designed to recognize two-dimensional shapes, which are highly invariant to translation, scaling, tilting, or other forms of deformation [7–15].

With the rapid development of mobile Internet technology, more and more image information is stored on the Internet. Image has become another important network information carrier after text. Under this background, it is very important to make use of a computer to classify and recognize these images intelligently and make them serve human beings better. In the initial stage of image classification and recognition, people mainly use this technology to meet some auxiliary needs, such as Baidu's star face function can help users find the most similar star. Using OCR technology to extract text and information from images, it is very important for graph-based semi-supervised learning method to construct good graphics that can capture intrinsic data structures. This method is widely used in hyperspectral image classification with a small number of labeled samples. Among the existing graphic construction methods, sparse representation (based on SR) shows an impressive performance in semi-supervised HSI classification tasks. However, most algorithms based on SR fail to consider

the rich spatial information of HSI, which has been proved to be beneficial to classification tasks. Yan et al. [16] proposed a space and class structure regularized sparse representation (SCSSR) graph for semi-supervised HSI classification. Specifically, spatial information has been incorporated into the SR model through graph Laplace regularization, which assumes that spatial neighbors should have similar representation coefficients, so the obtained coefficient matrix can more accurately reflect the similarity between samples. In addition, they also combine the probabilistic class structure (which means the probabilistic relationship between each sample and each class) into the SR model to further improve the differentiability of graphs. Hyion and AVIRIS hyperspectral data show that our method is superior to the most advanced method. The invariance extracted by Zhang et al. [17], such as the specificity of uniform samples and the invariance of rotation invariance, is very important for object detection and classification applications. Current research focuses on the specific invariance of features, such as rotation invariance. In this paper, a new multichannel convolution neural network (mCNN) is proposed to extract the invariant features of object classification. Multi-channel convolution sharing the same weight is used to reduce the characteristic variance of sample pairs with different rotation in the same class. As a result, the invariance of the uniform object and the rotation invariance are encountered simultaneously to improve the invariance of the feature. More importantly, the proposed mCNN is particularly effective for small training samples. The experimental results of two datum datasets for handwriting recognition show that the proposed mCNN is very effective for extracting invariant features with a small number of training samples. With the development of big data era, convolutional neural network (CNN) with more hidden layers has more complex network structure and stronger feature learning and feature expression ability than traditional machine learning methods. Since the introduction of the convolutional neural network model trained by the deep learning algorithm, significant achievements have been made in many large-scale recognition tasks in the field of computer vision. Chaib et al. [18] first introduced the rise and development of deep learning and convolution neural network and summarized the basic model structure, convolution feature extraction, and pool operation of convolution neural network. Then, the research status and development trend of convolution neural network model based on deep learning in image classification are reviewed, and the typical network structure, training method, and performance are introduced. Finally, some problems in the current research are briefly summarized and discussed, and new directions of future development are predicted. Computer diagnostic technology has played an important role in medical diagnosis from the beginning to now. Especially, image classification technology, from the initial theoretical research to clinical diagnosis, has provided effective assistance for the diagnosis of various diseases. In addition, the image is the concrete image formed in the human brain by the objective things existing in the natural environment, and it is an important source of information for a human to obtain the knowledge of the external things. With the continuous development of computer technology, the general object image recognition technology in natural scene is applied more and more in daily life. From image processing technology in simple bar code recognition to text recognition (such as handwritten character recognition and optical character recognition OCR etc.) to biometric recognition (such as fingerprint, sound, iris, face, gestures, emotion recognition, etc.), there are many successful applications. Image recognition (Image Recognition), especially (Object Category Recognition) in natural scenes, is a unique skill of human beings. In a complex natural environment, people can identify concrete objects (such as teacups) at a glance (swallow, etc.) or a specific category of objects (household goods, birds, etc.). However, there are still many questions about how human beings do this and how to apply these related technologies to computers so that they have humanoid intelligence. Therefore, the research of image recognition algorithms is still in the fields of machine vision, machine learning, depth learning, and artificial intelligence [19–24].

Therefore, this paper applies the advantage of depth mining convolution neural network to image classification, tests the loss function constructed by M^3 CE on two depth learning standard databases MNIST and CIFAR-10, and pushes forward the new direction of image classification research.

2 Proposed method

Image classification is one of the hot research directions in computer vision field, and it is also the basic image classification system in other image application fields, which is usually divided into three important parts: image preprocessing, image feature extraction and classifier.

2.1 The ZCA process is shown as below

In this process, we first use PCA to zero the mean value. In this paper, we assume that X represents the image vector [25]: $\mu = \frac{1}{m}\sum_{j=1}^{m} x_j$

$$x_j = x_j - \mu, j = 1, 2, 3, \cdots, m, \tag{1}$$

Next, the covariance matrix for the entire data is calculated, with the following formulas:

$$\sum = \frac{1}{m}\sum_{j=1}^{m} x_j x_j{}^T \tag{2}$$

where I represents the covariance matrix, I is decomposed by SVD [26], and its eigenvalues and corresponding eigenvectors are obtained.

$$[U, S, V] = SVD\left(\sum\right) \tag{3}$$

Of which,U is the eigenvector matrix of ∑, and S is the eigenvalue matrix of ∑. Based on this, x can be whitened by PCA, and the formula is:

$$x_{\text{PCAwhiten}} = S^{-\frac{1}{2}}U^T x \tag{4}$$

So $X_{\text{ZCAwhiten}}$can be expressed as

$$x_{\text{ZCAwhiten}} = U x_{\text{PCAwhiten}} \tag{5}$$

For the data set in this paper, because the training sample and the test sample are not well distinguished [27], the random generation method is used to avoid the subjective color of the artificial classification.

2.2 Image feature extraction based on time-frequency composite weighting

Feature extraction is a concept in computer vision and image processing. It refers to the use of a computer to extract image information and determine whether the points of each image belong to an image feature extraction. The purpose of feature extraction is to divide the points on the image into different subsets, which are often isolated points, a continuous curve, or region. There are usually many kinds of features to describe the image. These features can be classified according to different criteria, such as point features, line features, and regional characteristics according to the representation of the features on the image data. According to the region size of feature extraction, it can be divided into two categories: global feature and local feature [24]. The image features used in some feature extraction methods in this paper include color feature and texture feature, analysis of the current situation of corner feature, and edge feature.

The time-frequency composite weighting algorithm for multi-frame blurred images is a frequency-domain and time-domain weighting simultaneous processing algorithm based on blurred image data. Based on the weighted characteristic of the algorithm and the feature extraction of target image in time domain and frequency domain, the depth extraction technique is based on the time-frequency composite weighting of night image to extract the target information from depth image. The main steps of the time-frequency composite weighted feature extraction method are as follows:

Step 1: Construct a time-frequency composite weighted signal model for multiple blurred images, as the following expression shows:

$$g(t) = \sqrt{s} f([t-\tau]) \tag{6}$$

Of which,$f(t)$ is original signal, $S = (c - v)/(c + v)$, called the image scale factor. Referred to as scale, it represents the signal scaling change of the original image time-frequency composite weighting algorithm.$\sqrt{S}$ is the normalized factor of image time-frequency composite weighting algorithm.

Step 2: Map the one-dimensional function to the two-dimensional function$y(t)$ of the time scale a and the time shift b, and perform a time-frequency composite weighted transform on the continuous nighttime image of the image time-frequency composite weighted 0 using the square integrable function as shown below:

$$W_\psi y(a,b) = \left\langle y, \psi_{a,b} \right\rangle = \int_{-\infty}^{+\infty} y(t) \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right) dt \tag{7}$$

Of which, divisor $1/\sqrt{|a|}$. The energy normalization of the unitary transformation is ensured. $\psi_{a,b}$ is $\psi(t)$ obtained by transforming $U(a,b)$ through the affine group, as shown by the following expression:

$$\psi_{a,b}(t) = [U(a,b)\psi(t)] = \frac{1}{\sqrt{|a|}} \psi\left(\frac{t-b}{a}\right) \tag{8}$$

Step 3: Substituting the variable of the original image $f(t)$by $a = 1/s$ and $b = \tau$ and rewriting the expression to obtain an expression:

$$fs, \tau(t) = [U(1/s, \tau) f(t)] = \sqrt{|s|} f(s(t-\tau)) \tag{9}$$

Step 4: Build a multi-frame fuzzy image time-frequency composite weighted signal form.

$$u(t) = \frac{1}{\sqrt{T}} \text{rect}\left(\frac{t}{T}\right) \exp\left\{-j\left[2\pi k \ln\left(1-\frac{t}{t_0}\right)\right]\right\} \tag{10}$$

Of which, rect(t) = 1 and $|t| \le 1/2$.

Step 5: The frequency modulation law of the time-frequency composite weighted signal of multi-thread fuzzy image is a hyperbolic function;

$$f_i(t) = \frac{K}{t_0 - t}, |t| \le \frac{T}{2} \tag{11}$$

among them, $K = Tf_{\max}f_{\min}/B, t_0 = f_0 T/B, f_0$ is arithmetic center frequency, and $f_{\max}$, $f_{\min}$ are the minimum and maximum frequencies, respectively.

Step 6: Use the image transformation formula of the multi-detector fuzzy image time-frequency composite weighted signal to carry on the time-frequency composite weighting to the image, the definition of the image transformation is like the formula.

$$w_u u(a,b) = e^{j2\pi k \ \ln a} \times \frac{K}{\sqrt{a}} \tag{12}$$

$$\left\{ \left[\frac{a e^{\frac{j2\pi f_{\min}}{a}(b-b_a)}}{f_{\min}} - \frac{e^{j2\pi f_{\max}(b-b_a)}}{f_{\max}} \right] + j2\pi(b-b_a)[Ei(j2\pi f_{\max}(b-b_a)) - Ei(\frac{j2\pi f_{\min}}{a}(b-b_a)] \right\} \tag{13}$$

Of which, $b_a = (1-a)(\frac{1}{afm_{ax}} - \frac{T}{2})$, and $Ei(\bullet)$ represents an exponential integral.

Final output image time-frequency composite weighted image signal $W_u u(a,b)$. Therefore, compared with the traditional time-domain, c extraction technique of image features can be better realized by the time-frequency composite weighting algorithm.

2.3 Application of deep convolution neural network in image classification

After obtaining the feature vectors from the image, the image can be described as a vector of fixed length, and then a classifier is needed to classify the feature vectors.

In general, a common convolution neural network consists of input layer, convolution layer, activation layer, pool layer, full connection layer, and final output layer from input to output. The convolutional neural network layer establishes the relationship between different computational neural nodes and transfers input information layer by layer, and the continuous convolution-pool structure decodes, deduces, converges, and maps the feature signals of the original data to the hidden layer feature space [28]. The next full connection layer classifies and outputs according to the extracted features.

2.3.1 Convolution neural network

Convolution is an important analytical operation in mathematics. It is a mathematical operator that generates a third function from two functions f and g, representing the area of overlap between function f and function g that has been flipped or translated. Its calculation is usually defined by a following formula:

$$z(t)^{\text{def}} = f(t) * g(t) = \sum_{\tau=-\infty}^{+\infty} f(\tau) g(t-\tau) \tag{14}$$

Its integral form is the following:

$$z(t) = f(t) * g(t) = \int_{-\infty}^{+\infty} f(\tau) g(t-\tau) d\tau = \int_{-\infty}^{+\infty} f(t-\tau) g(\tau) d\tau \tag{15}$$

In image processing, a digital image can be regarded as a discrete function of a two-dimensional space, denoted as f (x, y). Assuming the existence of a two-dimensional convolution function g (x, y), the output image z (x, y) can be represented by the following formula:

$$z(x,y) = f(x,y) * g(x,y) \tag{16}$$

In this way, the convolution operation can be used to extract the image features. Similarly, in depth learning applications, when the input is a color image containing RGB three channels, and the image is composed of each pixel, the input is a high-dimensional array of 3 × image width × image length; accordingly, the kernel (called "convolution kernel" in the convolution neural network) is defined in the learning algorithm as the accounting. Computational parameter is also a high-dimensional array. Then, when two-dimensional images are input, the corresponding convolution operation can be expressed by the following formula:

$$z(x,y) = f(x,y) * g(x,y) = \sum_t \sum_h f(t,h) g(x-t, y-h) \tag{17}$$

The integral form is the following:

$$z(x,y) = f(x,y) * g(x,y) = \iint f(t,h) g(x-t, y-h) dt dh \tag{18}$$

If a convolution kernel of $m \times n$ is given, there is

$$z(x,y) = f(x,y) * g(x,y) = \sum_{t=0}^{t=m} \sum_{h=0}^{h=n} f(t,h) g(x-t, y-h) \tag{19}$$

where f represents the input image G to denote the size of the convolution kernel m and n. In a computer, the realization of convolution is usually represented by the product of a matrix. Suppose the size of an image is M × M and the size of the convolution kernel is $n \times n$. In computation, the convolution kernel multiplies with each image region of $n \times n$ size of the image, which is equivalent to extracting the image region of $n \times n$ and expressing it as a column vector of $n \times n$ length. In a zero-zero padding operation with a step of 1, a total of $(M - n + 1) * (M - n + 1)$ calculation results can be obtained; when these small image regions are each represented as a column vector of $n \times n$, the original image can be represented by the matrix $[n^* n^* (M - n + 1)]$.

Assuming that the number of convolution kernels is K, the output of the original image obtained by the above convolution operation is$k^*(M-n+1)*(M-n+1)$. The output is the number of convolution kernels × the image width after convolution × the image length after convolution.

2.3.2 M^3CE constructed loss function

In the process of neural network training, the loss function is the evaluation standard of the whole network model. It not only represents the current state of the network parameters, but also provides the gradient of the parameters in the gradient descent method, so the loss function is an important part of the deep learning training. In this paper, we introduce the loss function proposed by M^3 CE. Finally, the loss function of M^3 CE and cross-entropy is obtained by gradient analysis.

According to the definition of MCE, we use the output of Softmax function as the discriminant function. Then, the error classification metric formula is redefined as.

$$d_k(Z) = -p_k + p_q = -\frac{\exp Z_k}{\sum_{j=1}^{K} \exp Z_j} + \frac{\exp Zq}{\sum_{j=1}^{K} \exp Z_j} \tag{20}$$

Where k is the label of the sample, $q = \arg\max_{l \neq k}$, P_l represents the most confusing class of output of the Softmax function. If we use the logistic loss function, we can find the gradient of the loss function to Z.

$$\frac{\partial \ell_k(d_k)}{\partial_Z} = \frac{\partial \ell_k(d_k)}{\partial d_k(z)} \cdot \frac{\partial d_k(z)}{\partial z} = a\ell_k(1-\ell_k) \cdot \frac{\partial d_k(z)}{\partial_Z} \tag{21}$$

This gradient is used in the backpropagation algorithm to get the gradient of the entire network, and it is worth noting that if z is misdivided,$\ell_k$will be infinitely close to 1, and $a\ell_k(1-\ell_k)$ will be close to 0. Then, the gradient will be close to 0, which will cause almost no gradient to be reversed to the previous layer, which will not be good for the completion of the training process [29].

The sigmoid function is used in the traditional neural network activation function. But this is also the case during training. The observation formula shows that when the activation value is high the backpropagation gradient is very small which is called saturation. In the past, the influence of shallow neural networks was not very large, but with the increase of the number of network layers, this situation would affect the learning of the whole network. In particular, if the saturated sigmoid function is at a higher level, it will affect all the previous low-level gradients. Therefore, in the present depth neural networks, an unsaturated activation function linear rectifier unit (Rectified Linear Unit, Re LU) is used to replace the sigmoid function. It can be seen from the formula that when the input value is positive, the gradient of the linear rectifying unit is 1, so the gradient of the upper layer can be reversely transmitted to the lower layer without attenuation. The literature shows that linear rectification units can accelerate the training process and prevent gradient dispersion.

According to the fact that the saturation activation function in the middle of the network is not conducive to the training of the depth network, but the saturation function in the top loss function, has a great influence on the depth neural network.

$$\ell_k(d_k) = \max(0, 1 + d_k) \tag{22}$$

We call it the max-margin loss, where the interval is defined as $\in_k = -d_k(z) = P_k - P_q$.

Since P_k is a probability, that is, $P_k \in [0, 1]$, then $d_k \in [-1, 1]$, when a sample gradually becomes misclassified from the correct classification, d_k increases from − 1 to 1, compared to the original logistic loss function, and even if the sample is seriously misclassified, the loss function still get the biggest loss value. Because of$1 + d_k \geq 0$, it can be simplified.

$$\ell_k(d_k) = 1 + d_k \tag{23}$$

When we need to give a larger loss value to the wrong classification sample, the upper formula can be extended to

$$\ell_k(d_k) = (1 + d_k)^{\gamma} \tag{24}$$

where γ is a positive integer. If $\gamma = 2$ is set, we get the squared maximum interval loss function. If the function is to be applied to training deep neural networks, the gradient needs to be calculated and obtained according to the chain rule.

$$\frac{\partial \ell_k(d_k)}{\partial_Z} = \gamma(1 + d_k)^{\gamma-1} \cdot \frac{\partial d_k(z)}{\partial z} \tag{25}$$

Here, we need to discuss (1) when the dimension is the dimension corresponding to the sample label, (2) when the dimension is the dimension corresponding to the confused category label, and (3) when the dimension is neither the sample label nor the dimension corresponding to the confused category label. The following conclusions have been drawn:

$$\frac{\partial d_k(z)}{\partial_{Z_j}} = \begin{cases} p_j \in_k, j \neq k, q \\ p_j(\in_k - 1), j = k \\ p_j(\in_k + 1), j = q \end{cases} \tag{26}$$

3 Experimental results

3.1 Experimental platform and data preprocessing

MNIST (Mixed National Institute of Standards and Technology) database is a standard database in machine learning. It consists of ten types of handwritten digital grayscale images, of which 60,000 training pictures are tested with a resolution of 28×28.

In this paper, we mainly use ZCA whitening to process the image data, such as reading the data into the array and reforming the size we need (Figs. 1, 2, 3, 4, and 5). The image of the data set is normalized and whitened respectively. It makes all pixels have the same mean value and variance, eliminates the white noise problem in the image, and eliminates the correlation between pixels and pixels.

At the same time, a common way to change the results of image training is a random form of distortion, cropping, or sharpening the training input, which has the advantage of extending the effective size of the training data, thanks to all possible changes in the same image. And it tends to help network learning to deal with all distortion problems that will occur in the real use of classifiers. Therefore, when the training results are abnormal, the images will be deformed randomly to avoid the large interference caused by individual abnormal images to the whole model.

3.2 Build a training network

Classification algorithm is a relatively large class of algorithms, and image classification algorithm is one of them. Common classification algorithms are support vector machine, *k*-nearest algorithm, random forest, and so on. In image classification, support vector machine (SVM) based on the maximum boundary is the most widely used classification algorithm, especially the support vector machine (SVM) which uses kernel techniques. Support vector machine (SVM) is based on VC dimension theory and structural risk minimization theory. Its main purpose is to find the optimal classification hyperplane in high dimensional space so that the classification spacing is in maximum and the classification error rate is minimized. But it is more suitable for the case where the feature dimension of the image is small and the amount of data is large after extracting the special vector.

Another commonly used target recognition method is the depth learning model, which describes the image by hierarchical feature representation. The mainstream depth learning networks include constrained Boltzmann machine, depth belief network foot, automatic encoder, convolution neural network, biological model, and so on. We tested the proposed M3 CE-CEc. We design different convolution neural networks for different datasets. The experimental settings are as follows: the weight parameters are initialized randomly, the bias parameters are set as constants, the basic learning rate is set to 0.01, and the impulse term is set to 0.9. In the course of training, when the error rate is no longer decreasing, the learning rate is multiplied by 0.1.

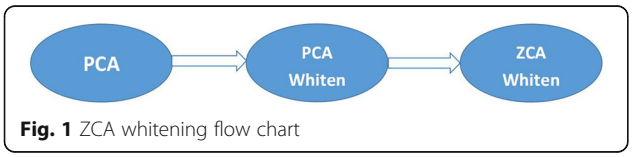

Fig. 1 ZCA whitening flow chart

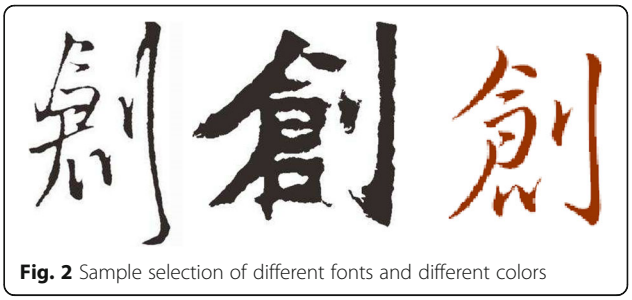

Fig. 2 Sample selection of different fonts and different colors

3.3 Image classification and modeling based on deep convolution neural network

The following is a model for image classification based on deep convolution neural networks.

1. Input: Input is a collection of N images; each image label is one of the K classification tags. This set is called the training set.
2. Learning: The task of this step is to use the training set to learn exactly what each class looks like. This step is generally called a training classifier or learning a model.
3. Evaluation: The classifier is used to predict the classification labels of images it has not seen and to evaluate the quality of the classifiers. We compare the labels predicted by the classifier with the real labels of the image. There is no doubt that the classification labels predicted by the classifier are consistent with the true classification labels of the image, which is a good thing, and the more such cases, the better.

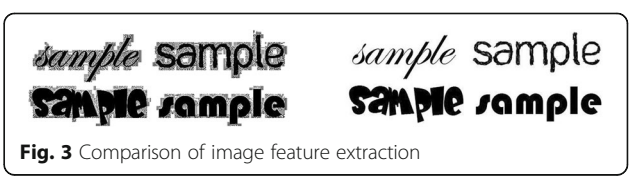

Fig. 3 Comparison of image feature extraction

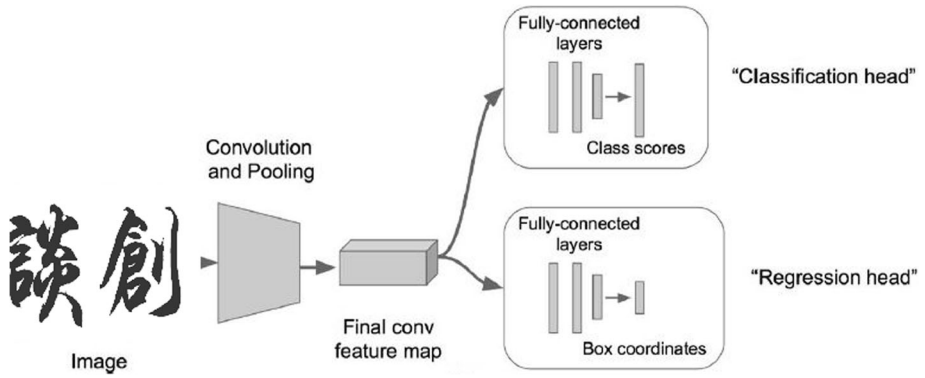

Fig. 4 Image classification and modeling based on deep convolution neural network

3.4 Evaluation index

In this paper, the image recognition effect is mainly divided into three parts: the overall classification accuracy, classification accuracy of different categories, and classification time consumption. The classification accuracy of an image includes the accuracy of the overall image classification and the accuracy of each classification. Assuming that *nij* represents the number of images in category *I* divided into category *j*, the accuracy of the overall classification is as follows:

$$accu_{all} = \sum_i n_{ii} / \sum_{i,j} n_{ij} \tag{27}$$

The accuracy of each classification is as follows:

$$accu_i = n_{ii} / \sum_j n_{ij} \tag{28}$$

Run time is the average time to read a picture to get a classification result.

4 Discussion

4.1 Comparison of classification effects of different loss functions

We compare the images of the traditional logistic loss function with our proposed maximum interval loss function. It can be clearly seen that the value of the loss function increases with the increase of the severity of the misclassification, which indicates that the loss function can effectively express the error degree of the classification.

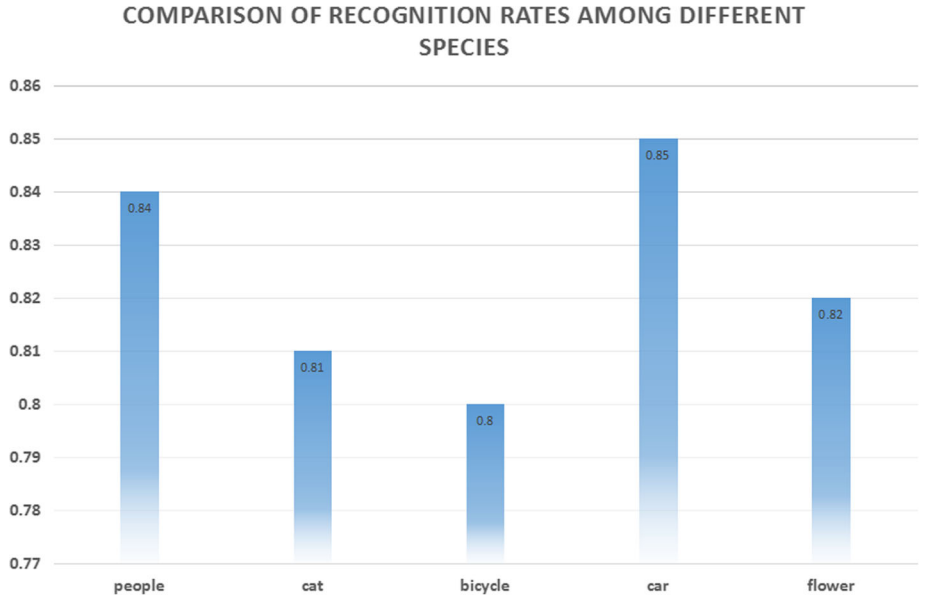

Fig. 5 Comparison of recognition rates among different species

4.2 Comparison of recognition rates between the same species

Classification	Bicycle	Car	Bus	Motor	Flower
Definition	0.82	0.84	0.81	0.80	0.85

4.3 Comparison of recognition rates among different species

As can be seen from the following table, the recognition rate of this method is generally the same among different species, reaching more than 80% level, among which the accuracy of this method is relatively high in classifying clearly defined images such as cars. This may be due to the fact that clearly defined images have greater advantages in feature extraction.

4.4 Time-consuming comparison of SVM, KNN, BP, and CNN methods

On the premise of feature extraction using the same loss function method constructed by M^3 CE, the selection of classifier is the key factor to affect the automatic detection accuracy of human physiological function. Therefore, this paper discusses the influence of different classifiers on classification accuracy in this part (Table 1). The following table summarizes the influence of some common classifiers on classification accuracy. These classifiers include linear kernel support vector machine (SVM-Linear), Gao Si kernel support vector machine (SVM-RBF), and Naive Bayes (NB) (NB) *k*-nearest neighbor (KNN), random forest (RF), and decision. Strategy tree (DT) and gradient elevation decision tree (GBDT).

The experimental results show that the accuracy of CNN classifier is higher than that of other classifiers in training set and test set. Although the speed of DT is the fastest when it is used for automatic detection of human physiological function in the classifier contrast experiment, its accuracy on the test set is only 69.47% unacceptable.由In this paper, the following conclusions can be drawn in the comparison experiment of classifier: compared with other six common classifiers, CNN has the highest accuracy, and the spending of 6 s is acceptable in the seven classifiers of comparison.

First, because each test image needs to be compared with all the stored training images, it takes up a lot of storage space, consumes a lot of computing resources, and takes a lot of time to calculate. Because in practice, we focus on testing efficiency far higher than training efficiency. In fact, the convolution neural network that we want to learn later reaches the other extreme in this trade-off: although the training takes a lot of time, once the training is completed, the classification of new test data is very fast. Such a model is in line with the actual use of the requirements.

5 Conclusions

Deep convolution neural networks are used to identify scaling, translation, and other forms of distortion-invariant images. In order to avoid explicit feature extraction, the convolutional network uses feature detection layer to learn from training data implicitly, and because of the weight sharing mechanism, neurons on the same feature mapping surface have the same weight. The *ya* training network can extract features by *W* parallel computation, and its parameters and computational complexity are obviously smaller than those of the traditional neural network. Its layout is closer to the actual biological neural network. Weight sharing can greatly reduce the complexity of the network structure. Especially, the multi-dimensional input vector image WDIN can effectively avoid the complexity of data reconstruction in the process of feature extraction and image classification. Deep convolution neural network has incomparable advantages in image feature representation and classification. However, many researchers still regard the deep convolutional neural network as a black box feature extraction model. To explore the connection between each layer of the deep convolutional neural network and the visual nervous system of the human brain, and how to make the deep neural network incremental, as human beings do, to compensate for learning, and to increase understanding of the details of the target object, further research is needed.

Table 1 Comparison before different classifiers

Categorizer	Accuracy on training set	Accuracy on the test set	Time-consuming(s)
CNN	99.68%	83.67%	6.003
SVM-RBF	89.41%	87.63%	9.334
NB	90.78%	89.36%	7.392
KNN	81.25%	72.59%	7.348
RF	85.26%	79.31%	1.203
DT	100%	69.47%	3.137
GBDT	87.79%	76.23%	6.947

Abbreviations
Ann: Artificial neural network; BP: Backpropagation; called NB-CNN: Convolutional neural network and Naive Bayes; CNN: Convolutional neural network; MLP: Multilayer perceptron; ODI: Omnidirectional image; VFSR: Very fine spatial resolution; VR: Virtual reality

Acknowledgements
The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

About the author
Xin Mingyuan was born in Heihe, Heilongjiang, P.R. China, in 1983. She received the Master degree from harbin university of science and technology, P.R. China. Now, she works in School of computer and information engineering, Heihe University, His research interests include Artificial intelligence, data mining and information security.
Wang yong was born in Suihua, Heilongjiang, P.R. China, in 1979. She received the Master degree from qiqihaer university, P.R. China. Now, she works in School of Heihe University, His research interests include Artificial intelligence, Education information management.

Funding
This work was supported by University Nursing Program for Young Scholars with Creative Talents in Heilongjiang Province (No.UNPYSCT-2017104). Scientific research items of basic research business of provincial higher education institutions of Heilongjiang Provincial Department of Education (No.2017-KYYWF-0353).

Availability of data and materials
Please contact author for data requests.

Authors' contributions
All authors take part in the discussion of the work described in this paper. XM wrote the first version of the paper. XM and WY did part experiments of the paper. XM revised the paper in a different version of the paper, respectively. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]School of Computer and Information Engineering, Heihe University, No. 1 Xueyuan Road education science and technology zone, Heihe, Heilongjiang, China. [2]Heihe University, No. 1 Xueyuan Road education science and technology zone, Heihe, Heilongjiang, China.

Received: 17 October 2018 Accepted: 7 January 2019
Published online: 11 February 2019

References
1. E. Newman, M. Kilmer, L. Horesh, *Image classification using local tensor singular value decompositions* (IEEE, international workshop on computational advances in multi-sensor adaptive processing. IEEE, Willemstad, 2018), pp. 1–5.
2. X. Wang, C. Chen, Y. Cheng, et al, Zero-shot image classification based on deep feature extraction. United Kingdom: IEEE Transactions on Cognitive & Developmental Systems, **10**(2), 1–1 (2018).
3. A.A.M. Al-Saffar, H. Tao, M.A. Talab, *Review of deep convolution neural network in image classification* (International conference on radar, antenna, microwave, electronics, and telecommunications. IEEE, Jakarta, 2018), pp. 26–31.
4. A.B. Said, I. Jemel, R. Ejbali, et al., *A hybrid approach for image classification based on sparse coding and wavelet decomposition* (Ieee/acs, international conference on computer systems and applications. IEEE, Hammamet, 2018), pp. 63–68.
5. Huang G, Chen D, Li T, et al. Multi-Scale Dense Networks for Resource Efficient Image Classification. 2018.
6. V. Gupta, A. Bhavsar, Feature importance for human epithelial (HEp-2) cell image classification. J Imaging. **4**(3), 46 (2018).
7. L. Yang, A.M. Maceachren, P. Mitra, et al., Visually-enabled active deep learning for (geo) text and image classification: a review. ISPRS Int. J. Geo-Inf. **7**(2), 65 (2018).
8. Chanti D A, Caplier A. Improving bag-of-visual-words towards effective facial expressive image classification Visigrapp, the, International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. 2018.
9. X. Long, H. Lu, Y. Peng, X. Wang, S. Feng, Image classification based on improved VLAD. Multimedia Tools Appl. **75**(10), 5533–5555 (2016).
10. B. Kieffer, M. Babaie, S. Kalra, et al., *Convolutional neural networks for histopathology image classification: training vs. using pre-trained networks* (International conference on image processing theory. IEEE, Montreal, 2018), pp. 1–6.
11. J. Zhao, T. Fan, L. Lü, H. Sun, J. Wang, Adaptive intelligent single particle optimizer based image de-noising in shearlet domain. Intelligent Automation & Soft Computing **23**(4), 661–666 (2017).
12. Mou L, Ghamisi P, Zhu X X. Unsupervised spectral-spatial feature learning via deep residual conv-Deconv network for hyperspectral image classification IEEE transactions on geoscience & Remote Sensing. 2018,(99):1–16.
13. Newman E, Kilmer M, Horesh L. Image classification using local tensor singular value decompositions IEEE, international workshop on computational advances in multi-sensor adaptive processing. IEEE, 2018:1–5.
14. S.A. Quadri, O. Sidek, Quantification of biofilm on flooring surface using image classification technique. Neural Comput. & Applic. **24**(7–8), 1815–1821 (2014).
15. X.-C. Yin, Q. Liu, H.-W. Hao, Z.-B. Wang, K. Huang, FMI image based rock structure classification using classifier combination. Neural Comput. & Applic. **20**(7), 955–963 (2011).
16. Z. Yan, V. Jagadeesh, D. Decoste, et al., HD-CNN: hierarchical deep convolutional neural network for image classification. Eprint Arxiv 4321-4329 (2014).
17. C. Zhang, X. Pan, H. Li, et al., A hybrid MLP-CNN classifier for very fine resolution remotely sensed image classification. Isprs Journal of Photogrammetry & Remote Sensing **140**, 133–144 (2018).
18. Chaib S, Yao H, Gu Y, et al. Deep feature extraction and combination for remote sensing image classification based on pre-trained CNN models. International Conference on Digital Image Processing. 2017: 104203D.
19. S. Roychowdhury, J. Ren, *Non-deep CNN for multi-modal image classification and feature learning: an azure-based model* (IEEE international conference on big data. IEEE, Washington, D.C., 2017), pp. 2893–2812.
20. M.Z. Afzal, A. Kölsch, S. Ahmed, et al., *Cutting the error by half: investigation of very deep CNN and advanced training strategies for document image classification* (Iapr international conference on document analysis and recognition. IEEE computer Society, Kyoto, 2017), pp. 883–888.
21. X. Fu, L. Li, K. Mao, et al., in *Chinese High Technology Letters*. Remote sensing image classification based on CNN model (2017).
22. Sachin R, Sowmya V, Govind D, et al. Dependency of various color and intensity planes on CNN based image classification. International Symposium on Signal Processing and Intelligent Recognition Systems. Springer, Cham, Manipal, 2017:167–177.
23. Shima Y. Image augmentation for object image classification based on combination of pre-trained CNN and SVM. International Conference on Informatics, Electronics and Vision & 2017, International sSymposium in Computational Medical and Health Technology. 2018:1–6.
24. J.Y. Lee, J.W. Lim, E.J. Koh, A study of image classification using HMC method applying CNN ensemble in the infrared image. Journal of Electrical Engineering & Technology **13**(3), 1377–1382 (2018).
25. Zhang C, Pan X, Zhang S Q, et al. A rough set decision tree based Mlp-Cnn for very high resolution remotely sensed image classification. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2017:1451–1454.
26. M. Kumar, Y.H. Mao, Y.H. Wang, T.R. Qiu, C. Yang, W.P. Zhang, Fuzzy theoretic approach to signals and systems: Static systems. Inf. Sci. **418**, 668–702 (2017).
27. W. Zhang, J. Yang, Y. Fang, H. Chen, Y. Mao, M. Kumar, Analytical fuzzy approach to biological data analysis. Saudi J Biol Sci. **24**(3), 563, 2017–573.

28. Z. Sun, F. Li, H. Huang, *Large scale image classification based on CNN and parallel SVM. International conference on neural information processing* (Springer, Cham, Manipal, 2017), pp. 545–555.
29. Sachin R, Sowmya V, Govind D, et al. Dependency of various color and intensity planes on CNN based image classification. 2017.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com