International Journal of Swarm Intelligence Research
Volume 14 • Issue 2

A Multi-Feature Fusion Model Based on Denoising Convolutional Neural Network and Attention Mechanism for Image Classification

Jingsi Zhang, Faculty of Robot Science and Engineering, Northeastern University, Shenyang, Northeastern University, China*

Xiaosheng Yu, Faculty of Robot Science and Engineering, Northeastern University, Shenyang, Northeastern University, China

Xiaoliang Lei, Faculty of Robot Science and Engineering, Northeastern University, Shenyang, Northeastern University, China

Chengdong Wu, Faculty of Robot Science and Engineering, Northeastern University, Shenyang, Northeastern University, China

ABSTRACT

Spatial location features extracted by denoising convolutional neural network. At this time, an attention mechanism is introduced into denoising convolutional neural network. The dual attention model of local area is presented from two dimensions of channel and space—channel attention mechanism weights channel and spatial attention mechanism weights location. A variety of machine learning methods are used to classify and train different features. Multi-semantic features and heterogeneous features are fused by adaptive weighted fusion algorithm. Finally, the data sets Cifar-10, STL-10, Cifar-100 and GHIM-1OK are verified on the proposed method. Compared with a single semantic feature, the accuracy is improved by 10%-15%. Compared with several advanced algorithms, the performance has a significant advantage, which proves the complementarity of heterogeneous features and multi-network semantic features and the effectiveness of the adaptive weighted fusion algorithm.

KEYWORDS

color volume histogram feature, denoising convolutional neural network, heterogeneous feature, image classification, multi-feature fusion

1. INTRODUCTION

With the rapid development of computer science and Internet technology, a large amount of data is produced all over the world all the time, such as, text, audio, pictures, videos and so on. Big data contains a lot of information, but this information is difficult to identify and organize manually (Li et al. 2023; Huang et al. 2021). Therefore, how to classify and identify information has been

DOI: 10.4018/IJSIR.324074 *Corresponding Author

concerned and studied by many researchers. Among all kinds of information, image information is a very important information carrier. In the last century, researchers proposed many excellent image algorithms for texture information in images, which laid a solid foundation for computer vision and pattern recognition. In the 21st century, Convolutional Neural Network (CNN) (Guo et al. 2022) has made great breakthroughs in the field of image classification, so computer vision has entered a new era.

CNNs are a kind of Feed forward Neural Networks with deep structure and convolutional computation. It can better obtain the position space and shape information of image, which is beneficial to image classification. (Ghimire et al. 2022) discovered that Graphics Processing Unit (GPU) played an important role in machine learning (ML), and proposed an efficient CNN training method based on GPU, which greatly improved the computing ability of CNN. (Alex Krizhevsky et al. 2017) proposed AlexNetc network, which adopted ReLU activation function and GPU packet convolution for parallel training. (Teng et al. 2022) applied batch normalization to neural networks, which ensured that the output distribution of each layer in the network was basically stable. At the same time, the network greatly reduces the dependence on the initial parameters and improves the network performance. The following is the classification of feature fusion.

1) Heterogeneous Feature fusion. It is a fusion method based on traditional features and deep semantic features. (Kas et al. 2021) proposed a facial expression recognition method based on multi-direction gradient HOG features and deep learning features for generation selection fusion. (Bibin et al. 2017) extracted traditional features (LBP, SIFT and Color) and semantic features of deep belief networks of images, and used Boltzmann machines to mine the association between traditional features and deep features. The fusion of heterogeneous features proved that traditional features can effectively improved the diversity of linguistic and semantic feature spaces.
2) Multi-neural network feature fusion: A fusion method based on different convolutional network features. (Apicella et al. 2022) trained middle-level features of different neural networks by supervised learning method, and used late fusion strategy to process classifier prediction, which effectively improved image classification performance. (Chen et al. 2020) constructed the CFR-DenseNet and ILFRDenseNet dual networks and used end-to-end training to fuse the features of the two different networks. (Lopes et al. 2017) used three pre-trained convolutional networks as feature extractors, used a single hidden layer to transform high-level features into low-dimensional space, and fused the rich information of each feature. Compared with advanced methods, the multi-feature fusion method has sufficient competitiveness in the accuracy of image classification.
3) Multi-layer feature fusion of neural networks: A fusion method based on multiple different features of a single network. (Ferrara et al. 2021) introduced residual learning to establish a deep convolutional network, and used the deep feature fusion network to weighted the features of different layers of the convolutional network. (Feng et al. 2019) proposed MSLN-CNN, which enhanced images under local and non-local constraints and fused features at different levels, and performed extremely well in the case of limited training samples. (Guo et al. 2019) proposed a multi-classifier network (MCN) using semantic features of different layers of the same CNN to fuse through an adaptive method, and verified the effectiveness of the fusion method on multiple datasets. The multi-layer semantic features of neural network are extracted by multi-layer feature fusion. The less abstract features are regarded as the complementary features of deep features, and the mutual fusion can effectively improve the accuracy and reliability of classification results.

This paper combines the advantages of heterogeneous feature fusion and multi-convolutional network feature fusion, and proposes an adaptive weighted fusion algorithm, which can effectively and concisely fuse multiple semantic features and traditional features, and greatly improve the accuracy of classification.

The contributions of this paper are as follows:

1) A manual feature is designed according to the color and edge information of the image, which can strengthen the diversity of the feature space and effectively complement the semantic features;
2) A variety of semantic features and learning methods of different convolutional neural networks are studied and compared, and it is found that the complementarity between semantic features of different networks is better than the features of different layers of the same network. Heterogeneous features and multi-neural network features are fused to further strengthen the diversity of feature space.
3) A new fusion algorithm is proposed to integrate various semantic features of different convolutional neural networks and traditional manual features. Compared with the end-to-end training and fusion algorithm, it has better classification performance and greatly reduces the time cost required by the algorithm.

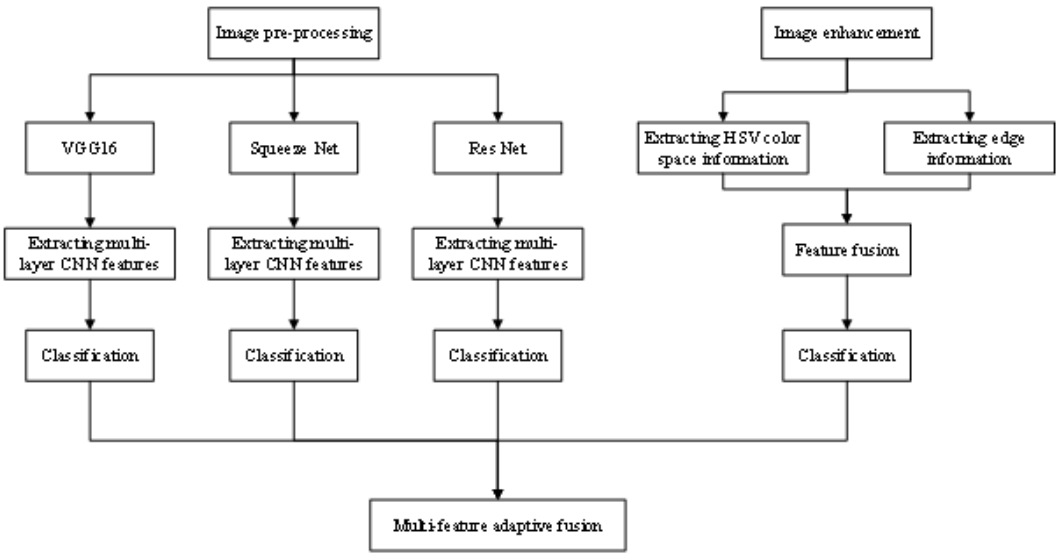
This paper is organized as follows. In section 2, we introduce the direction gradient and color volume histogram and adaptive weighted fusion. Experimental results and analysis are conducted in section 3. There is a conclusion in section 4.

2. DIRECTION GRADIENT AND COLOR VOLUME HISTOGRAM AND ADAPTIVE WEIGHTED FUSION ALGORITHM

The image classification method consists of three steps:

1) image preprocessing (including image denoising, illumination normalization, etc.);
2) Image feature extraction (including training images and test images);
3) Use classification to learn extracted features (Yin et al. 2021). In this paper, according to the differences and complementarity between different network features, weights are used to dynamically fuse multiple features (Figure 1).

Figure 1. Flow chart of multi-feature fusion image classification algorithm



2.1 HOG and Color Volume Histogram

HSV color space is close to the visual perception of human eyes and is very sensitive to the difference between colors. The parameters of HSV color space to measure color information are: hue (H), saturation (S), brightness (V). The hue, saturation and brightness are quantified into 6, 4 and 3 respectively, and the dimension of color volume feature is 6×4×3=72. In this paper, the color volume feature is utilized, and the color space feature marking matrix is added to improve the distinction between colors. Moreover, the cumulative calculation of color space volume features is optimized and improved, which improves the classification performance and makes it more complementary to semantic features.

The RGB image is converted into HSV color space, and the volume feature matrix VFM of the image color space is calculated.

$$VFM = \left[\frac{1}{3}\bullet\pi\bullet S^2\bullet V\bullet\frac{H}{360}\right] \tag{1}$$

According to the component size of image H, S and V, the color space feature label matrix K of pixels in different spatial positions of the image is calculated.

$$K = F\bullet\frac{6H}{360} + 6F\bullet\frac{4S}{1.001} + 24F\bullet\frac{3V}{1.001} \tag{2}$$

In Equation (2), F is rounded downward. Suppose there are multiple 3×3 pixel blocks in an image, and the coordinate of the central pixel of each pixel block is $\left(x_0, y_0\right)$, the other eight pixel coordinates are named $\left(x_i, y_i\right)$. Where, $i \in \left[1, 8\right]$, it calculates the average value of the volume of the color space of 9 pixels, and takes it as the color space feature $h\left(x_0, y_0\right)$ in the center. Finally, the volume features of color space marked with equal k in the color space feature labeling matrix are accumulated and counted, which is the color volume histogram feature of the current image.

$$CV\left(K\left(m,n\right)\right) = \sum_{x=2}^{M-1}\sum_{y=2}^{N-1} h\left(x_0, y_0\right) = \sum_{x=2}^{M-1}\sum_{y=2}^{N-1} log\left(\sum_{i=0}^{8}\frac{VFM_i}{9\left(M-1\right)\left(N-1\right)}\right) \tag{3}$$

Currently, the Histogram of Oriented Gradient (HOG) is a commonly used image edge feature in the field of computer vision and pattern recognition (Wang et al. 2022). HOG constructs features by calculating and counting the local gradient direction histogram of image. Compared with other feature description methods, HOG feature is formed on the local grid element, so it has good invariance to the geometric and optical deformation of the image, and the deformation of the two can only appear in a larger spatial neighborhood.

The extraction process of HOG feature is as follows: the color image is transformed into gray image, and the Gamma correction method is used to reduce the influence of illumination on image information.

$$I\left(x,y\right) = \left[I\left(x,y\right)\right]^{Gamma} \tag{4}$$

It calculates the horizontal gradient G_x and vertical gradient G_y of image A. According to G_x and G_y size, the gradient amplitude value G and gradient direction ， of each pixel in the image are determined.

$$G_x\left(x,y\right)=\left[-1\ \ 0\ \ 1\right]*A \tag{5}$$

$$G_y\left(x,y\right)=\begin{bmatrix}-1\\0\\1\end{bmatrix}*A \tag{6}$$

$$G\left(x,y\right)=\sqrt{G_x\left(x,y\right)^2+G_y\left(x,y\right)^2} \tag{7}$$

$$\theta=arctan\frac{G_y\left(x,y\right)}{G_x\left(x,y\right)} \tag{8}$$

The image is divided into multiple 4×4 units. $\left[0^\circ,360^\circ\right]$ is divided into 9 bins, and the gradient amplitude values of the gradient direction in the unit are added together to form the HOG feature vector of 9×1. The 4 units are combined into a larger square interval. In order to further reduce the influence of light and shadow, the feature vectors of different units are normalized. The HOG features of all intervals in the image are counted and synthesized into the final feature vector.

$$HOGCV=\left[\lambda*HOG,\left(1-\lambda\right)*VFM\right] \tag{9}$$

Through Principal Component Analysis (PCA), the HOG feature dimension is reduced to 128, and the color volume histogram feature is normalized respectively. According to Formula (9), a new feature descriptor is obtained by fusion, named as Histogram of Oriented Gradient and Color Volume. The parameter in Equation (9) is recommended to use 0.95. HOGCV features have geometric invariance and rotational invariance, which are effectively complementary to the semantic features extracted by convolutional neural networks.

2.2 Dual Attention Mechanism

The convolution features directly obtained from VGG-16 do not have strong discriminative power, so this paper introduces the attention mechanism to improve the representation ability of the network. Before discussing the attention mechanism, we first introduce the different dimensions of the feature graph. The feature graph after convolution and pooling has two dimensions, one is channel dimension and the other is space dimension. The relationship between channels has been widely studied in fine-grained image classification. For example, MA-CNN groups feature maps to obtain key features of different parts. The spatial dimension represents the relationship between different pixels. The attention mechanism in this paper mainly applies to the two dimensions of channel and space, which can be divided into the following two modules.

2.2.1 Channel attention module

The contribution value of the feature maps of different channels in image classification is different, but many studies believe that the feature maps of different channels have the same influence on the classification results. For example, (Lee et al. 2020) directly added the feature graphs of different channels, and the result was seriously disturbed by the chaotic image background.

We take any feature graph of 4 channels in Conv5_3 for visualization. Different feature channels have different concerns. The feature maps of the last three channels all focus on the target body part, while the feature maps of the 108th layer focus on the background noise.

Therefore, in order to highlight the target subject area and suppress the interference of background noise, we use the channel attention mechanism to learn the weight according to the importance of different channels. By increasing the weight of the feature graph containing the target subject area and reducing the weight of the noise response graph, we can enhance the effective channel information and suppress the invalid channel information. For deep convolutional neural networks, after multiple convolution and pooling, the last layer of the convolutional layer contains the most sufficient spatial and semantic information. Therefore, we only use the attention mechanism after the feature diagram output by the last convolutional layer, where the structure of the channel attention mechanism.

Global average-pooling can make full use of the spatial information of each channel, and it has strong robustness and is not easy to over-fit. Global max pooling can reflect the global maximum response and represent the key information in the channel to a certain extent. In addition, (Li et al. 2019) demonstrated that the combined results of global average pooling and global maximum pooling are more efficient than that of using only one global pooling method. Therefore, we use the fusion of GAP and GMP information to learn channel weights, and the pooled feature vectors are represented as FGAP and FGMP respectively. Then, in order to realize the information interaction between channels and reduce the amount of data, two layers of 1×1 convolution were joined and applied to FGAP and FGMP. The number of 1×1 convolution kernel in the previous layer is set to c / r, where c is the channel dimension of FGAP and FGMP. r is a variable parameter, representing the reduction multiple of feature dimension c. The number of 1×1 convolution kernels in the last layer is adjusted to c to ensure that the channel dimension of the output feature graph is the same as that of the input.

After two layers of 1×1 convolution, $F'GAP$ and $F'GMP$ are obtained, and their combined results are passed to the sigmoid function. This function maps each element of the merged result between 0 and 1 to get the weight M_C, which represents the importance of each feature channel. Finally, M_C is multiplied by the original feature graph F and weighted to get the attention feature graph F'_C, which means that weight distribution is performed on different feature channels to suppress useless information and increase the proportion of useful information. The formula can be expressed as:

$$F'_C = M_C \otimes F = \sigma\left(F'GAP + F'GMP\right) \otimes F \tag{10}$$

As part of the channel information may be lost during GAP and GMP transmission, inspired by ResNet (Li et al. 2021) residual learning, we superimpose the convolution feature of the attention module with the original output to obtain the best representation. The channel attention mechanism is added to the original feature channel as a side branch so that the network only needs to learn the attention module and not the entire output. The final output feature diagram F_C is expressed as:

$$F_C = F \oplus F'_C = \left(1 + \sigma\left(F'GAP + F'GMP\right)\right) \otimes F \tag{11}$$

Where σ represents the sigmoid function. $\oplus$ means add pixel by pixel and $\otimes$ means multiply element by element.

2.2.2 Spatial attention module

Different pixels in the feature map have different contributions to the classification results. The function of spatial attention module is to assign weight value to each pixel in the feature map. By increasing

the weight of discriminant region and reducing the weight of noise region and background region, the key region is enhanced and useless region is suppressed.

Same as the channel attention mechanism, GAP and GMP are first used to obtain FSGAP and FSGMP along the channel dimension, the size is $h \times w \times 1$. Then, after adding FSGAP and FSMP element by element, 3×3 convolution is carried out, and the convolution result is passed to sigmoid function to obtain the weight graph M_S. Finally, we multiply the weight graph element by element with the original feature graph F to get the spatial attention feature graph F'_S. The formula can be expressed as:

$$F'_S = M_S \otimes F = \sigma\left(f\left(FSGAP + FSGMP\right)\right) \otimes F \tag{12}$$

Similarly, we add the spatial attention mechanism as a side branch to the original feature channel, and the final output feature graph F_S is expressed as:

$$F_S = F \oplus F'_S = \left(1 + \sigma\left(f\left(FSGAP + FSGMP\right)\right)\right) \otimes F \tag{13}$$

Where σ denotes sigmoid function. f denotes 3×3 convolution operation. $\oplus$ means add pixel by pixel and $\otimes$ means multiply element by element.

Channel and spatial attention analyze image features from two dimensions to facilitate the network to learn more feature information. The dual attention mechanism combines two different dimensions of attention feature graphs. The dual attention mechanism is output to FDA, and the formula is expressed as:

$$FDA = F_C + F_S \tag{14}$$

2.3 Kernel Bilinear Aggregation Module

In the previous module, the dual attention mechanism fuses the attention feature graphs of the two mechanisms in the way of matrix addition to get FDA, with dimension $w \times h \times d$. (Wang et al. 2021) proposed a kernel bilinear convolutional network to solve the problem that B-CNN could only model nonlinear relations among feature channels. Therefore, in order to more fully excavate the rich information contained between channels, we also use sigmoid kernel function to nucleate the cross product matrix of FPA to model the nonlinear relationship between channels and enhance the characterization ability of the network.

The kernel bilinear aggregation module first normalizes the output feature map FDA of the dual attention mechanism by channel binormal, and expands the result by channel into the feature matrix $X \in R^{d \times N}$, where $N = h \times w$. d is the number of feature channels. Then, the matrix X is accumulated $XX^T \in R^{d \times d}$, and the sigmoid kernel function is used to nucleate the outer product matrix to model the nonlinear relationship between channels, and the image expression P is obtained. Finally, exponential power operation is carried out on the matrix P, and the power exponent is set as $\pm$, and the upper triangle part of the result is taken as the final expression of the image for image classification.

In this paper, sigmoid kernel function is used to nucleate the outer product matrix $A = XX^T$. The sigmoid kernel function formula K can be expressed as:

$$K\left(x_i, y_i\right) = tanh\left(\theta \bullet \left\langle x_i, y_i \right\rangle + \gamma\right) \tag{15}$$

Where θ is the amplitude adjustment parameter, $\theta > 0$. γ is the displacement parameter, $\gamma < 0$. The matrix P is denoted by:

$$P = K\left(XX^T\right) = tanh\left(\theta \bullet XX^T + \gamma \bullet 1_{d\times d}\right) = tanh\left(\theta \bullet A + \gamma \bullet 1_{d\times d}\right) = \left[tanh\left(\theta \bullet \left\langle x_i, y_i \right\rangle + \gamma\right]_{d\times d}\right. \tag{16}$$

Where $1_{d\times d}$ represents the D-dimensional square matrix with all the elements being 1. The back propagation formula of cross entropy loss function l for matrix A is shown below, where " ∘ " is the Hadamard product.

$$\frac{\partial l}{\partial A} = \theta \bullet \left(1 - P^2\right) \circ \frac{\partial l}{\partial P} \tag{17}$$

2.4 Denoising Convolutional Neural Network Architecture

In order to better read image content information, the overall framework of CNNs based on deep multi-feature fusion is shown in Figure 2.

The denoising deep convolutional feature in figure 2 (abbreviated as f_c) is denoising spatial position and shape relationship information after three-layer convolutional neural network, as shown in figure 3. CNN's convolution layer contains multiple convolution kernels, each of which can extract shape-related features. Each neuron in the convolutional layer is connected with multiple neurons in the region adjacent to the previous layer, also known as "receptive field" (A et al. 2021), and thus relies on the network to learn context-invariant features: shape and spatial position feature information, which is particularly useful for image classification. In addition,CNN also has the function of denoising the input original image features. Considering the above features of CNN, the structure of de-noising convolutional neural network in Figure 3 outputs the denoising deep convolutional features.

The CNN network model framework with deep multi-feature fusion in Figure 2 has four convolutional layers. The first convolution layer is called the data feature fusion layer, which fully integrates the denoising depth convolution and the main color feature. This is followed by a pooling layer, which serves to reduce network parameters and speed up fusion. The second convolution layer is called the deep feature fusion layer for further feature fusion, followed by a pooling layer. The third convolutional layer is called the feature abstraction representation layer, and the size of the convolutional kernel in this layer is changed from 5×5 in the first two convolutional layers to 3×3, which is helpful to eliminate the noise in the feature and improve the abstract representation of the feature, and then it is also a pooling layer. The fourth convolutional layer, called the feature high-level representation layer, helps to eliminate redundant features and improve representativeness, followed by a pooling layer. The three-layer full-connection layer is used for feature classification and parameter optimization in the process of back propagation. At the end of the network, the Softmax layer is used for classification. Softmax is a supervised learning approach for multi-classification problems, providing important confidence levels for classification, where 0 means the lowest confidence level and 1 means the highest.

Convolution layer, nonlinear activation transformation and pooling layer are the three basic components of CNN. By stacking multiple convolutional layers with nonlinear operations and multiple pooling layers, a deep CNN can be formed, and input features can be extracted in layers with invariance and robustness. Using specific architectures such as local joins and shared weights, CNN tends to have good generalization capabilities. The convolution layer 7 with nonlinear operations is as follows.

Figure 2. Overall framework of deep multi-feature fusion CNNs network model

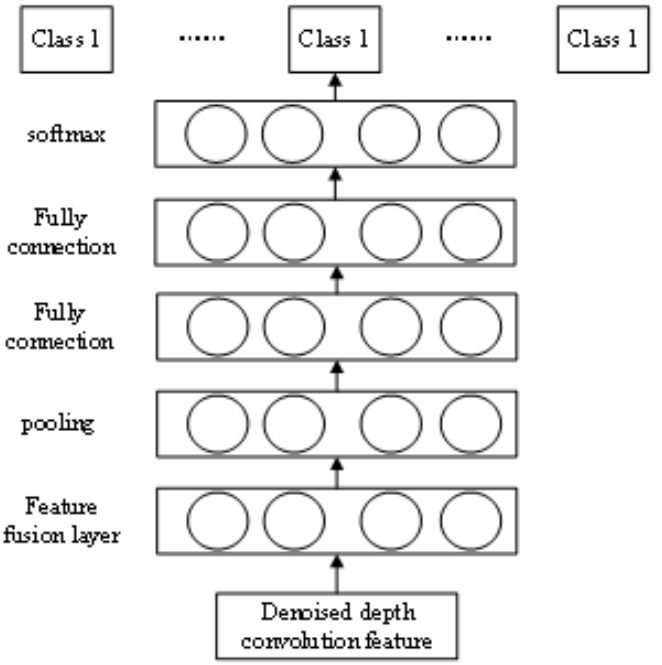
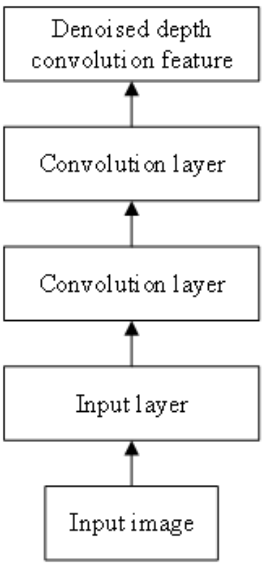


Figure 3. Denoising convolutional neural network architecture



$$x_j^l = f\left(\sum_{i=1}^{M} x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{18}$$

The matrix x_i^{l-1} is the $i-th$ feature graph of layer $l-1$. x_j^l is the $j-th$ feature graph of the current l layer. M is the number of input feature graphs. k_{ij}^l and b_j^l are randomly initialized and set to zero, respectively. Then it is fine-tuned by back-propagation. $f(\cdot)$ is a nonlinear activation function, and $*$ is a convolution operation.

3. EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we conduct experiments on public data sets Cifar-10, STL-10, and GHIM-10K. Features extracted from the full connection layer of VGG and ResNet are better than those of other layers, and the features of conv10 layer are better in SqueezeNet. VGG-16 network has better classification effect than the other two deep convolutional networks. It achieves 85% accuracy without training stage. Among the four learning methods, SVM and LDA generally outperform the other two methods in the three data sets, as shown in Figure 4, Figure 5 and Figure 6.

Based on the experimental results of the single layer features of convolutional neural networks, the adaptive weighted fusion algorithm is verified by convolutional neural network features. The fusion results are shown in Tables 1~3. Here, DCNN is denoising convolutional neural network and AM is attention mechanism.

The experimental results show that the performance improvement of different network features is limited after more than three kinds of features or multiple features of the same network. On the basis of multi-CNN feature fusion, further use HOG feature to strengthen feature space diversity. HOG features can improve the result of multi-feature fusion by about 25.5% (table 1), which indicates that HOG features can effectively strengthen the diversity of high-level semantic features and have good complementarity with high-level semantic features. After the fusion of three different CNN features and HOG features on Cifar-10, STL10 and GHIM-10K data sets, the final classification accuracy reaches 96.7%, 97.2%, 96.6%, respectively. Compared with the optimal result of single feature, the comparison accuracy is improved by 10% to 30%, and the comparison result of double feature fusion is improved by about 15.6%, which proves the effectiveness of the adaptive weighted fusion algorithm.

In this paper, the pre-training model is used to extract semantic features, the fusion weight is adjusted through performance adaptation, and the attention mechanism features are used to enhance

Figure 4. Single network optimal layer accuracy on the Cifar-10 data set

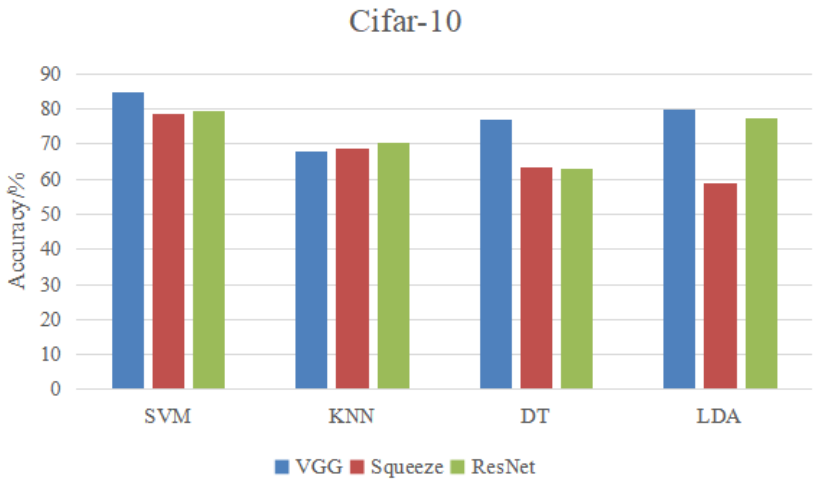


Figure 5. Single network optimal layer accuracy on the STL-10 data set

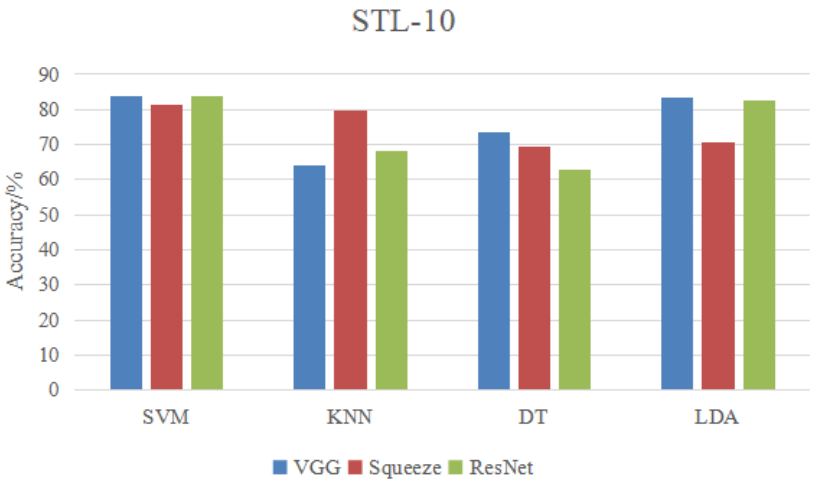


Figure 6. Single network optimal layer accuracy on the GHIM-10K data set

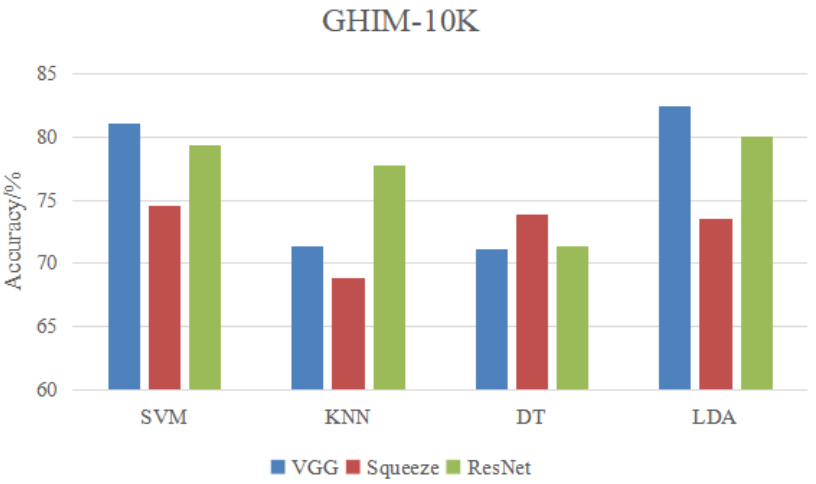


Table 1. Ablation experiment on Cifar-10

Method	Accuracy/%
HOG	59.6
DCNN	76.4
HOG+DCNN	82.1
DCNN+AM	85.6
HOG+DCNN+AM	96.7

Table 2. Ablation experiment on STL-10

Method	Accuracy/%
HOG	58.1
DCNN	77.9
HOG+DCNN	83.6
DCNN+AM	86.1
HOG+DCNN+AM	97.2

Table 3. Ablation experiment on GHIM-10K

Method	Accuracy/%
HOG	60.2
DCNN	75.3
HOG+DCNN	82.7
DCNN+AM	88.2
HOG+DCNN+AM	96.6

the diversity of fusion features, which further enhances the classification performance. Experimental results on various data sets show that the proposed algorithm is universal. Compared with the research results of many advanced methods (including Crossvit (Chen et al. 2021), Resmlp (Touvron et al. 2022), TARDB-Net (Cai et al. 2021), ATT (Sitaula et al. 2021)) as shown in table 4, the proposed method has more significant performance, and because the models used are lightweight and do not require further training, the efficiency is high, which has great significance for the practical application of image classification field.

4. CONCLUSION

The rapid development of convolutional neural networks has greatly promoted the progress in the field of computer vision, bringing new opportunities and challenges to the traditional manual features. In this paper, a manual feature based on image color and edge combined with attention mechanism is proposed to improve the diversity of semantic feature space. An adaptive weighted fusion algorithm based on accuracy is proposed to dynamically control various semantic features and traditional features, and a pre-trained network model is used to extract semantic features, which not only guarantees classification results, but also greatly speeds up the computation speed of the algorithm. The experimental results prove that the semantic features of different networks are more complementary,

Table 4. Comparison with different methods/%

Crossvit	Cifar-10	STL-10	GHIM-10K
Resmlp	79.8	80.4	81.4
TARDB-Net	88.7	86.4	85.8
ATT	91.4	93.2	90.7
Proposed	95.7	96.1	96.5

and verify the robustness of SVM to various convolutional neural networks, the strengthening effect of HOG features on semantic features and the effectiveness of the adaptive weighted fusion algorithm. In this paper, the complex fusion algorithm has obtained better performance, but it is not excellent in fine-grained image classification. The future research goal is to optimize the fusion algorithm and structure, and strengthen its performance in fine-grained image classification.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant nos. U20A20197, 61973063, Liaoning Key Research and Development Project 2020JH2/10100040, Natural Science Foundation of Liaoning Province 2021-KF-12-01 and the Foundation of National Key Laboratory OEIP-O-202005.

REFERENCES

Apicella, A., Giugliano, S., Isgrò, F., & Prevete, R. (2022). Exploiting auto-encoders and segmentation methods for middle-level explanations of image classification systems. *Knowledge-Based Systems*, *255*, 109725. doi:10.1016/j.knosys.2022.109725

Bibin, D., Nair, M. S., & Punitha, P. (2017). Malaria parasite detection from peripheral blood smear images using deep belief networks. *IEEE Access : Practical Innovations, Open Solutions*, *5*, 9099–9108. doi:10.1109/ACCESS.2017.2705642

Cai, W., Liu, B., Wei, Z., Li, M., & Kan, J. (2021). TARDB-Net: Triple-attention guided residual dense and BiLSTM networks for hyperspectral image classification. *Multimedia Tools and Applications*, *80*(7), 11291–11312. doi:10.1007/s11042-020-10188-x

Chen, C., & Tong, Y. (2020). Research on System Architecture Based on Deep Learning Convolutional Neural Network[M]//Artificial Intelligence in China. Springer. Singapore.

Chen, C F R., Fan, Q., & Panda, R. (2021). Crossvit: Cross-attention multi-scale vision transformer for image classification[C]. *Proceedings of the IEEE/CVF international conference on computer vision,* (pp. 357-366). IEEE.

Feng, J., Chen, J., Liu, L., Cao, X., Zhang, X., Jiao, L., & Yu, T. (2019). CNN-based multilayer spatial–spectral feature fusion and sample augmentation with local and nonlocal constraints for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *12*(4), 1299–1313. doi:10.1109/JSTARS.2019.2900705

Ferrara, M., Della Santa, F., Bilardo, M., De Gregorio, A., Mastropietro, A., Fugacci, U., Vaccarino, F., & Fabrizio, E. (2021). Design optimization of renewable energy systems for NZEBs based on deep residual learning. *Renewable Energy*, *176*, 590–605. doi:10.1016/j.renene.2021.05.044

Ghimire, D., Kil, D., & Kim, S. (2022). A Survey on Efficient Convolutional Neural Networks and Hardware Acceleration. *Electronics (Basel)*, *11*(6), 945. doi:10.3390/electronics11060945

Guo, J., Han, K., & Wu, H. (2022) Cmt: Convolutional neural networks meet vision transformers. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12175-12185. doi:10.1109/CVPR52688.2022.01186

Guo, Z., Ma, X., Li, H. (2019). Self-Adaption Multi-classifier Fusion Networks for Image Recognition[C]//2019 IEEE International Conference on Multimedia and Expo (ICME), (pp. 399-405). IEEE.

Huang, W., & Looi, C. K. (2021). A critical review of literature on "unplugged" pedagogies in K-12 computer science and computational thinking education. *Computer Science Education*, *31*(1), 83–111. doi:10.1080/08993408.2020.1789411

Jisi, A., & Yin, S. (2021). A New Feature Fusion Network for Student Behavior Recognition in Education. *Journal of Applied Science and Engineering*, *24*(2), 133–140.

Kas, M., Ruichek, Y., & Messoussi, R. (2021). New framework for person-independent facial expression recognition combining textural and shape analysis through new feature extraction approach. *Information Sciences*, *549*, 200–220. doi:10.1016/j.ins.2020.10.065

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. doi:10.1145/3065386

Lee, H., Park, J., & Hwang, J. Y. (2020). Channel attention module with multiscale grid average pooling for breast cancer segmentation in an ultrasound image. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, *67*(7), 1344–1353. PMID:32054578

Li, B., & Lima, D. (2021). Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering*, *2*, 57–64. doi:10.1016/j.ijcce.2021.02.002

Li, P., Laghari, A. A., Rashid, M., Gao, J., Gadekallu, T. R., Javed, A. R., & Yin, S. (2023). A Deep Multimodal Adversarial Cycle-Consistent Network for Smart Enterprise System. *IEEE Transactions on Industrial Informatics*, *19*(1), 693–702. doi:10.1109/TII.2022.3197201

Li, Z., Wang, S. H., Fan, R. R., Cao, G., Zhang, Y.-D., & Guo, T. (2019). Teeth category classification via seven-layer deep convolutional neural network with max pooling and global average pooling. *International Journal of Imaging Systems and Technology*, *29*(4), 577–583. doi:10.1002/ima.22337

Lopes, U. K., & Valiati, J. F. (2017). Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. *Computers in Biology and Medicine*, *89*, 135–143. doi:10.1016/j.compbiomed.2017.08.001 PMID:28800442

Sitaula, C., & Hossain, M. B. (2021). Attention-based VGG-16 model for COVID-19 chest X-ray image classification. *Applied Intelligence*, *51*(5), 2850–2863. doi:10.1007/s10489-020-02055-x PMID:34764568

Teng, L., & Qiao, Y. (2022). BiSeNet-oriented context attention model for image semantic segmentation. *Computer Science and Information Systems*, *19*(3), 1409–1426. doi:10.2298/CSIS220321040T

Touvron, H., Bojanowski, P., Caron, M., Cord, M., El-Nouby, A., Grave, E., Izacard, G., Joulin, A., Synnaeve, G., Verbeek, J., & Jegou, H. (2022). Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–9. doi:10.1109/TPAMI.2022.3206148 PMID:36094972

Wang, K., Chen, G., & Chu, H. (2021). Finger vein recognition based on multi-receptive field bilinear convolutional neural network. *IEEE Signal Processing Letters*, *28*, 1590–1594. doi:10.1109/LSP.2021.3094998

Wang, X. (2022). Crowd Density Estimation Based On Multi-scale Information Fusion And Matching Network In Scenic Spots. *Journal of Applied Science and Engineering*, *26*(6), 865–875.

Yin, S., Li, H., Laghari, A., Karim, S., & Jumani, A. (2021). A Bagging Strategy-Based Kernel Extreme Learning Machine for Complex Network Intrusion Detection. *EAI Endorsed Transactions on Scalable Information Systems*, *21*(33), e8. doi:10.4108/eai.6-10-2021.171247

Xiaosheng Yu received Ph.D degree from Northeastern University in 2014. He is currently an associate professor with the Faculty of Robot Science and Engineering, Northeastern University. His interest is image processing and computer vision. Jingsi Zhang is with Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110819, China. Research interests are image processing and AI.

Xiaosheng Yu is with Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110819, China. Research interests are image processing and Robot control.

Xiaoliang Lei is with Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110819, China. Research interests are image processing and AI.

Chengdong Wu is with Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110819, China. Research interests are image processing and Robot control.