International Journal for Housing Science and Its Applications
Publish August 4, 2025. Volume 46, Issue 3 Pages 2756-2767

https://doi.org/10.70517/ijhsa463232

Analysis of optimization and efficient application of image recognition algorithm based on graph neural network

Ying Hu[1,*]
[1] Shanghai Institute of Commerce and Foreign Languages, Shanghai, 201399, China
Corresponding authors: (e-mail: huying8520@126.com).

Abstract In the context of the development of regional tourism, the planning pattern of accommodation facilities in tourist attractions plays a very important role in enhancing tourists' experience perception, which is conducive to the enhancement of tourists' goodwill towards tourist attractions. This paper is oriented to the theory of sustainable development, and selects GL city as the research case site to obtain the visitor evaluation data of accommodation facilities in tourist attractions. The evolution trend of tourist attraction accommodation facilities is analyzed by geographic concentration index, standard deviation ellipse, nearest neighbor index, combined with TF-IDF algorithm to extract high-frequency words of tourists' experience on the planning of tourist attraction accommodation facilities, and coarse-grained tourism review text sentiment classification model is introduced to analyze the tourists' sentiment evaluation of tourist attraction accommodation facilities. The geographic concentration index of accommodation facilities in tourist attractions increases from 20.19 to 53.72 from 1980 to 2023, and the relative frequency difference between "service" and "room" of hotel facilities in tourist attractions is 4.35 percentage points, while that of B&B is only 0.5 percentage points. The relative frequency of "service" and "room" of hotel facilities in tourist attractions differed by 4.35 percentage points, while that of B&Bs differed by only 0.77 percentage points, and more than 70% of the tourists had a positive attitude towards the experience of accommodation facilities in tourist attractions. Lodging facilities in tourist attractions need to enrich the sequence of lodging products, improve the layout of service and reception facilities, create relevant themed B&Bs, and enhance the service quality of lodging facilities by combining modern technology to further promote the sustainable development of lodging facilities planning.

Index Terms Multiscale Attention Mechanism, Twins Transformer, Graph Convolutional Network, Image Recognition

I. Introduction

With the continuous development of artificial intelligence technology, image recognition technology has been widely used in various fields such as industry, medicine, finance and so on [1], [2]. However, in practical applications, the traditional image recognition algorithms can no longer meet the actual needs due to the increasing data volume and model complexity [3], [4]. At this time, image recognition algorithms based on graph neural networks have become a hot topic [5].

Image recognition algorithm, refers to the algorithm that automatically analyzes and recognizes images by computer [6], [7]. It is a technique based on machine learning and artificial intelligence, which enables computers to automatically learn and understand the content of images and classify or describe them [8], [9]. And graph neural network (GNN) is a new method that has emerged in the field of deep learning in recent years [10]. By combining graph structures with neural networks, GNN can show its unique advantages in processing complex data, especially in the fields of social networks, bioinformation, and recommender systems [11]-[13]. In the optimization of image recognition algorithms, graph neural network algorithms are usually divided into three steps, i.e., preprocessing, feature extraction and decision classification. Preprocessing is to transform the raw data into a format that can be processed by the neural network, including image normalization, graying, noise reduction, filtering and other processing [14], [15]. Feature extraction is the extraction of features from the data by means of convolution and pooling. Decision classification is to classify the features and determine the classification or recognition result of the image [16]-[18]. In practical applications, image recognition algorithms of graph neural networks can be applied in the fields of face recognition, license plate recognition, object recognition, etc. to achieve more accurate, stable and reliable image recognition results [19]-[21].

The image recognition algorithm designed in this paper consists of four modules: multi-scale image feature extraction, attention mechanism, graph convolutional network, and feature fusion. Firstly, the multi step mechanism of Twins Transformer is utilized to capture different parts of image features, and the multi-scale attention mechanism framework is designed for deep learning of image semantic features. The three-layer standard Transformer decoder is used as a component for further fusion of image semantic features and image features. The cross attention mechanism of decoder is then utilized to complete the interaction between features. After comparative experiments on PASCAL VOC2007 dataset and NUS-WIDE dataset, the image recognition method of this paper is applied to YogaPose Recognition Dataset yoga action dataset, which proves the effectiveness and efficiency of the algorithm proposed in this chapter.

II. Optimization of image recognition algorithm based on graph neural network

II. A. Image recognition algorithm optimization techniques

II. A. 1) Deep Learning Models for Multi-Labeled Images

Convolutional Neural Network (CNN) [22] is currently the most efficient image processing technique, the network model consists of three layers: convolution, activation function and aggregation, and can represent the feature space of each image. In image recognition, the output of the CNN is utilized as the input and the classification of the image is achieved using the fully connected layer.

In practical applications, deep learning networks have been shown to be effective in capturing more complex information and features. However, when the network hierarchy is deepened, its optimization efficiency instead weakens and the accuracy on the validation and training sets slips in tandem. This is mainly due to the challenges such as gradient explosion and gradient vanishing that are encountered when the network structure is deepened. To cope with these challenges, an innovative network architecture, ResNet, has been proposed [23].

ResNet is a residual network, which is a sub-network that can be stacked in multiple layers to form a depth-adjustable network structure; therefore, a new method based on residual network is proposed in this paper. This residual network is composed of a set of residual data blocks, and its residual data block generalization can be expressed as:

$$x_{l+1} = x_l + F\left(x_l, W_l\right) \tag{1}$$

where the residual block consists of two parts: the direct mapping part and the residual part. $H\left(x_l\right)$ represents the direct mapping and $F\left(x_l, W_l\right)$ denotes the residual term, and the residual part generally includes two to three convolution operations.

Here, the weight layer represents the convolution operation and $\oplus$ is the unit addition operation. In convolutional neural networks, the number of feature maps represented by x_l and x_{l+1}, respectively, is usually different, and thus 1×1 convolution operations are needed to increase or decrease the dimensionality. In this case, the residual block can be expressed as:

$$x_{l+1} = W_l' x_l + F\left(x_l, W_l\right) \tag{2}$$

where W_l' is the 1×1 convolution operation. For a specific block of data, the operation can be performed by first studying the residuals $H\left(x\right) - x$, rather than letting the network $F\left(X\right)$ learn the underlying image directly.

The ResNet network structure has the property of being easy to adapt. It can be adjusted and various functions can be realized by simple correction and expansion of the number of block channels and the number of stacks. Without paying too much attention to "degradation", the performance of the network gets better and better as the amount of training grows and time passes.

In multi-label classification problems, each data often contains several target labels. In many cases, the missing labels are filled with zeros, so one-hot coding cannot be used. The classification of labeled nodes in the network is viewed as a binomial distribution, which converts the multi-label classification problem into a binary classification problem of label-by-label recognition in images. For this purpose, a binary classification cross entropy loss function is used as a loss function:

$$L = \sum_{c=1}^{C} y^c \log\left(\sigma\left(\hat{y}^c\right)\right) + \left(1 - y^c\right)\log\left(1 - \sigma\left(\hat{y}^c\right)\right) \tag{3}$$

where $\sigma\left(\cdot\right)$ represents the sigmoid activation function, y^c is the true label of the sample (0 or 1), and $\hat{y}^c$ is the predicted probability.

II. A. 2) Label Relevance

In real life, in many cases, images often contain multiple labels, e.g., an image can contain both humans and animals. However, manually classifying the samples one by one would consume a lot of human and material resources. The goal of multi-label learning is to train from a given library of samples and labels to obtain the set of labels applicable to the samples. The application of label correlation is usually effective in improving the performance of multi-label classification.

II. A. 3) Modal attention drive

Modal attention-driven is the use of an attention mechanism that enables a model to selectively focus on the most informative and relevant parts of different modalities, thereby improving the efficiency and accuracy of task execution.

The mathematical representation of modal attention-driven is usually in the form of an attention mechanism, which contains the steps of weight allocation and information integration. Looking at the attention mechanism in deep learning models:

Its modal attention weights are calculated as:

$$Attention\left(Q,K,V\right)= soft\max\left(\frac{QK^{T}}{\sqrt{d_{k}}}\right)V \tag{4}$$

In the attention mechanism, the attention weights are obtained by measuring the similarity between keys and keys and standardizing them. Then according to the weights of each indicator, the weights of each indicator are summed up to obtain the precise expression effect of each indicator under the attention mechanism.

II. A. 4) Graph Convolutional Neural Networks

The key to convolutional neural networks is the convolutional kernel, which is utilized to perform convolutional operations on them using sliding windows to achieve feature extraction from images. Graph Convolutional Neural Networks (GCN) is similar to Convolutional Neural Networks (CNN) in that both have feature extraction capabilities, with the difference being that it takes graph data as its object.

The algorithm takes a series of features as input nodes, and realizes the modification of the eigennode vector A Î ¡ $^{C' C}$ and state update weight matrix W Î ¡ $^{D' D_u}$ to the eigennode vector V update. The updated nodes V_u Î ¡ $^{C' D_u}$ can be represented by a single-layer graph convolutional neural network:

$$V_u = d\left(AVW\right) \tag{5}$$

In the above equation, the matrix A is usually pre-determined and the matrix W is obtained by continuous training of the model. The $d(\times)$ denotes the activation function, which uses either $ReLU(\times)$ or $Sigmoid(\times)$ for the model training, and makes the system exhibit non-linear characteristics by introducing the activation function. The correlation matrix A embodies the interconnection between the characteristics of the nodes in the graph network. The method uses the correlation matrix A to propagate the correlation information to each node, which obtains the required information and then updates its state with the linear transformation W .

The graph convolutional network model has the following three deep learning properties:

1. Hierarchical architecture: by extracting features layer by layer, each layer is more abstract and advanced than the previous one.

2. Nonlinear transformations: enhancement of model representation.

3. end-to-end training: there is no need to define any rules, just label the nodes of the graph, allowing the model to learn on its own, fusing features and structural information.

The main idea of the graph convolution operation is to use the data in the nodes to pass through the network and achieve an update of the representation of the nodes in the network. In graph data, assuming that there are C nodes, each node has a feature of dimension d , so that a feature matrix H^l Î ¡ $^{C' d}$ can be formed, where $\left(l = 0,L\ P\right)$, and an adjacency matrix that describes the relationships between nodes. Also known as the adjacency matrix $\overset{+}{M}$ Î ¡ $^{C' C}$. This set of data represented by H^l and $\overset{+}{M}$ is the input to the graph convolution network. Each graph convolutional network layer can be represented as a nonlinear functional form:

$$H^{l+1} = f_{gcn}\left(H^{l}, \overset{+}{M}\right) \tag{6}$$

After using the convolution operation, $f_{gcn}(\times)$ can be expressed as:

$$H^{l+1} = o_{gcn}\left(\overset{+}{M}_{nor} H^{l} W^{l}\right) \tag{7}$$

In the above algorithm, the transformation matrix to be trained, $W^{l}\ \text{Î}\ \text{¡}^{d'd}$, is normalized to obtain the normalized matrix $\overset{+}{M}_{nor}\ \text{Î}\ \text{¡}^{C'C}$. The implementation of $o_{gcn}(\times)$ is a nonlinear operation that introduces a known graphic convolutional neural network into a multilayered convolutional neural network, obtaining a change in the properties of each node from H^{0} to H^{P}. However, the association between each node (adjacency matrix correlation) remains regardless of the number of layers.

II. B.Image recognition based on multi-scale attention mechanism

II. B. 1) Design of the overall model structure

In this paper, a multi-scale attention mechanism framework based on Transformer [24] is proposed, which mainly consists of a multi-scale feature extraction module, an attention module, a graph convolutional network module for constructing label correlation, and a feature fusion module, and the structure of the model is shown in Fig. 1. The cross-entropy loss function is used as shown in the following equation:

$$L_{loss} = \overset{n}{\underset{i=1}{\text{å}}} y^{i}\log\left(s\left(\hat{y}^{i}\right)\right) + \left(1 - y^{i}\right)\log\left(1 - s\left(\hat{y}^{i}\right)\right) \tag{8}$$

where $y^{i} = \{0,1\}$ indicates whether label i appears in the image, $\hat{y}^{i}$ is the output of the fully connected layer, and $s(\times)$ is the sigmoid function.

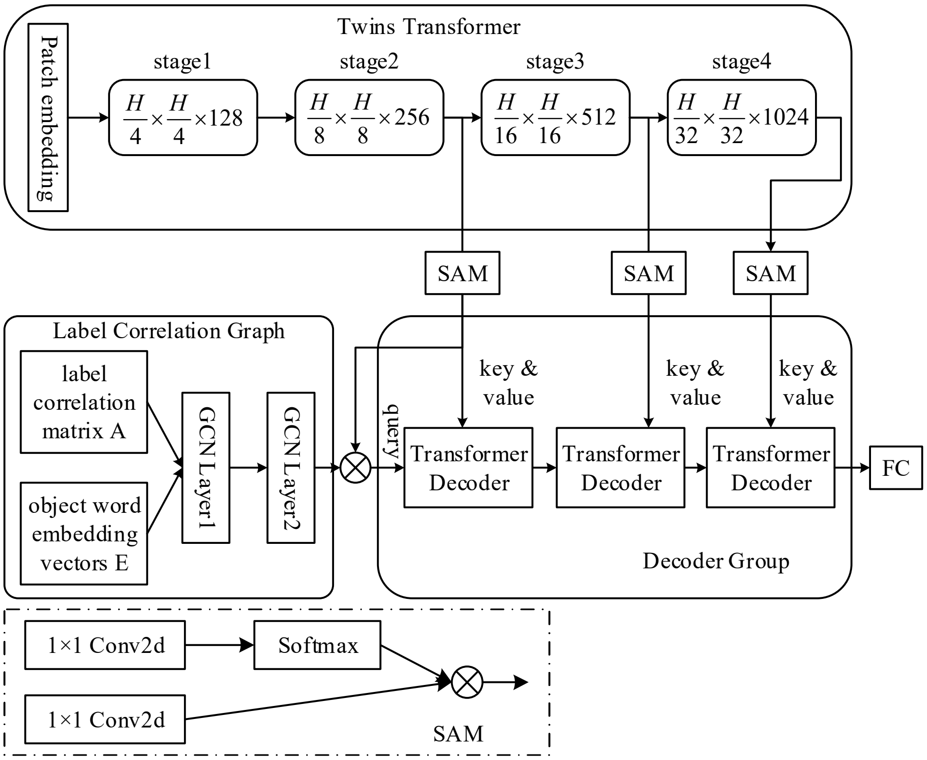

Figure 1: Model architecture

II. B. 2) Multi-scale image feature extraction

In this paper, image feature extraction is performed by Twins Transformer and the architecture achieves excellent performance on a wide range of visual tasks.

Similar to CNN, the model has four stages to generate feature maps at different scales, each using a similar architecture that consists of a Patch embedding layer, a Transformer encoder and a position encoder.

The attention used in the Transformer encoder layer is an optimized and improved Spatially Separable Self-Attention (SSSA) mechanism, which consists of two types of attention operations, the local grouped self-attention mechanism and the global subsampling attention mechanism. The SSSA can efficiently extract the features while reducing the computational cost, which is computed as shown in the following equation.

$$\begin{aligned} \hat{z}_{ij}^{l} &= LSA\left(LayerNorm\left(z_{ij}^{l-1}\right)\right)+z_{ij}^{l-1} \\ z_{ij}^{l} &= FFN\left(LayerNorm\left(\hat{z}_{ij}^{l}\right)\right)+\hat{z}_{ij}^{l} \\ \hat{z}^{l+1} &= GSA\left(LayerNorm\left(z^{l}\right)\right)+z^{l} \\ z^{l+1} &= FFN\left(LayerNorm\left(\hat{z}^{l+1}\right)\right)+\hat{z}^{l+1} \end{aligned} \tag{9}$$

where i Î $\{1,2,\mathrm{L}\ ,m\}, j$ Î $\{1,2,\mathrm{L}\ ,n\}$, j Î $\left\{1,2,\mathrm{L}\ ,n\right\}$, LSA is the local grouped self-attention within the sub-window, and GSA is the global sub-sampled attention obtained by interacting with each sub-window, which have a multi-head structure in the standard self-attention mechanisms all have multi-head structures.

In the first stage, given an input image of size $H\ ' \ W\ ' \ 3$, it is first divided into $\frac{HW}{4^2}$ patches image blocks. Then, the image blocks are linearly projected, followed by the embedding of the image blocks into the Transformer encoder, and after the first encoder block, the conditional position encoding is performed in the position encoder, and the final output of the feature map.

II. B. 3) Attention module

The weights of image pixels are calculated to obtain the relationship between image pixels and pixels, and the module consists of two $1\ ' \ 1$ convolutional layers and a dot product operation. The image features obtained in the latter three stages are convolved twice respectively, and the output of one of the convolutions is passed through the softmax function to obtain the corresponding attention feature map, and then a dot product is done with the output of the other convolution before finally obtaining the output of each stage through the attention module $f_s¢$, as shown in the following equation:

$$f_s¢= s\left(\left(W_j * f_s\right)^T\left(W_g * f_s\right)\right) \tag{10}$$

where $s\left(\times\right)$ denotes the softmax function. W_j and W_g denote convolution kernels. * is the convolution operation. f_s is the image feature of the s th stage.

II. B. 4) Graph Convolutional Networks for Multi-Label Recognition

Since graphs have a unique topological structure, this structure can model the correlation between any two labels, in order to explore the dependency between labels, this paper uses an ML-GCN-like approach to construct a graph convolutional network GCN network model consisting of two stacked layers.

In order to construct the label co-occurrence relation matrix A , it is necessary to obtain the label co-occurrence relation matrix from the multi-label image dataset, which records the co-occurrence of all labels in the image. In this paper, the dependency relationship between labels is defined as conditional probability, and the i th label is defined as l_i , then the conditional probability formula shown in the following equation can be obtained:

$$P_{ij} = P(l_j \mid l_i) = \frac{M_{ij}}{M_i} \tag{11}$$

where M_{ij} denotes the number of times label i and label A j co-occur in the dataset, M_i denotes the total number of times label i occurs in the training set, and P_{ij} denotes the probability that label j occurs when label

i occurs. Generally speaking, P_{ij} and P_{ji} are not equal, for example, when "car" appears, "person" is likely to appear, and conversely, when "person" appears, "car" may not appear, so the constructed label relationship matrix is not a symmetry matrix. This gives the label co-occurrence relation matrix for each element.

$$A_{ij} = P_{ij} \tag{12}$$

where A_{ij} denotes the element in the i th row and j th column of the label co-occurrence relation matrix A. However, there are two problems with this simple correlation relation matrix; one, the co-occurrence patterns among labels may exhibit long-tailed distributions, so that some rare co-occurrences may be incorrectly regarded as noise. Second, the number of labels that co-occur in the training and test sets may not be exactly the same, which can lead to overfitting of the correlation matrix on the training set and thus affect its generalization performance. Therefore, here the matrix A needs to be binarized, and for a set threshold t, the elements in the label co-occurrence relation matrix A become:

$$A_{ij} = \begin{cases} 0,(P_{ij} < t) \\ 1,(P_{ij} \geq t) \end{cases} \tag{13}$$

where t takes the value range of $[0,1]$, in this paper, the value of t is consistent with the ML-GCN and takes the value of 0.4. At this point, matrix A is a binary-valued matrix. However, an immediate problem with binary correlation matrices is that they can lead to excessive smoothing when performing update calculations in GCN networks. In other words, node features may be over-smoothed so that nodes from different clusters may become indistinguishable from each other. To mitigate this problem, the same reweighting strategy as ML-GCN is adopted, which is constructed as follows:

$$A'_{ij} = \begin{cases} \left(\lambda / \sum_{j=1, i \neq j}^{c} A_{ij} \right) A_{ij},(i \neq j) \\ 1-\lambda,(i = j) \end{cases} \tag{14}$$

where A'_{ij} is the reweighted label co-occurrence matrix and λ takes values in the range $[0,1]$.

The GCN is utilized to propagate information between individual nodes as a way to update the node representation. According to the definition of a single-layer GCN in a graph convolutional neural network, the forward propagation process in a one-layer graph convolutional network can be expressed as the following equation.

$$H^{l+1} = \delta\left(H^{l},A\right) \tag{15}$$

where H^{l} denotes the features of all nodes in the previous layer, which is the input to the next layer H^{l+1}.

II. B. 5) Feature Fusion Module

Previous work mainly focuses on capturing the correlation between labels, while ignoring the effective fusion of image features and label embeddings, which seriously affects the convergence efficiency of the model. To overcome this drawback, this paper introduces Transformer decoder, whose built-in cross-attention mechanism can effectively fuse image features and tag semantic information, thus it serves as an effective component for fusing image features and tag semantic features.

The cross-attention mechanism in Transformer decoder can generate specific features for each tag, after multiplying query and key, there is an attention mechanism between each tag feature and image feature, and the result of the multiplication of the two is then applied to the image feature representation itself, so that the image feature vector fuses the tag semantic information. The computation process in decoder is as follows.

$$O_{i}^{(1)} = MultiHead\left(Q_{i},F,F\right) \tag{16}$$

$$O_{i} = FFN\left(O_{i}^{(1)}\right) \tag{17}$$

where $O_i^{(1)}$ denotes the output after multi-head attention, where both the $MultiHead$ (query, key, value) and $FFN(x)$ functions are the same as those defined in the standard Transformer decoder, and then the obtained result O_i is fed into the feedforward network.

$$Atten(Q,K,V) = soft\ \max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{18}$$

where Q, K, and V are matrices of query, key, and value, respectively. d_k denotes the dimension of k. The dot product operation returns the similarity of each query and key value.

In order to fuse with the multi-scale features extracted by Twins Transformer, a three-layer decoder decoder is used in this paper to obtain the final fused features of the MSAT model.

III. Image recognition algorithm optimization results analysis

III. A. Data sets and assessment criteria

The PASCAL Visual Object Classes Challenge (VOC 2007) dataset is a widely used dataset in multi-label image recognition tasks, and it is also one of the commonly used datasets in tasks such as target detection and semantic segmentation. The dataset contains the $trainval$ set (5011 images) and the $test$ set (4952 images), totaling 9963 images and covering 20 object categories.

The Microsoft COCO 2014 dataset is another multi-label image recognition dataset and it is also one of the commonly used datasets in tasks such as target detection and semantic segmentation, where each image contains more than three object labels on average. This dataset contains a total of 80 object categories, and the publicly available part includes the $train$ set containing a total of 82,081 images and the val set containing a total of 40,504 images.

The main evaluation criteria used during the experiment were the average precision (AP) of each category and the mean of the average precision of all categories (mAP), and the overall precision (OP), the overall recall (OR), Overall F1-Score (OF1), Precision per Class (CP), Recall per Class (CR), and F1-Score per Class (CF1) are added to the experiments on the COCO data set, and are calculated as follows:

$$\begin{aligned} OP &= \frac{\sum_i N_i^c}{\sum_i N_i^p}, CP = \frac{1}{C}\sum_i \frac{N_i^c}{N_i^p} \\ OR &= \frac{\sum_i N_i^c}{\sum_i N_i^g}, CR = \frac{1}{C}\sum_i \frac{N_i^c}{N_i^g} \\ OF1 &= \frac{2\times OP\times OR}{OP+OR}, CF1 = \frac{2\times CP\times CR}{CP+CR} \end{aligned} \tag{19}$$

where N_i^c is the number of images correctly predicted for the i th label, N_i^p is the number of images predicted for the i th label, and N_i^g is the number of true images for the i th label.

III. B. Experimental setup

The hardware environment for all experimental procedures in this paper used an Intel Core i9-9900K CPU processor, 3.60GHz 16GB of RAM, and a single NVIDIA GeForce RTX 2080Ti graphics processing unit, and was based on the Ubuntu 16.04 software system platform in combination with PyTorch 1.40 framework for network training and testing. During the training process, the data enhancement method was used to scale the image size to 512×512, and the image size was 448×448 at the time of testing. The initial learning rate is 0.001 and the learning rate is updated with a learning decay rate of 0.1 every 30 epochs until the training is terminated after completing 80 epochs.

III. C. Experimental results

Because the PASCAL VOC 2007 dataset contains a small number of images and in order to facilitate the comparison with other methods, the experimental process uses the trainval data for network training, and the test data for the final effect of the test, and its loss convergence process is shown in Figure 2. The experimental results of the comparison between this paper's model and several other multi-label image recognition algorithms on the PASCAL VOC 2007 dataset are shown in Table 1, which mainly compares the average precision (AP) of each category and the mean average precision (mAP) of the overall category between the algorithms. The overall performance demonstrated by this paper's model on the PASCAL VOC 2007 dataset is improved by 0.7% over ML-GCN.

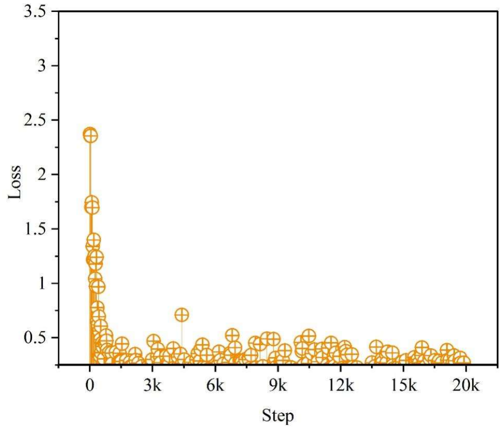

Figure 2: Loss convergence process on the PASCAL VOC 2007 dataset

Table 1: Performance comparisons in PASCAL voc 2007 data set(%)

Model	mAP	aero	bike	bird	boat	bottle	bus	car	cat
CNN-RNN	86.6	96.7	83	93.8	92.6	61.4	81.9	89.5	94.2
ResNet-101	92.8	99.9	97.5	97.3	96.5	65.7	91.9	96.6	97.3
HCP	93.8	98.4	96.9	98.0	95.5	74.7	94.1	95.8	96.9
RNN-Atten	93.7	98.5	97.2	96.2	96.0	75.4	91.8	96.2	97.9
VGG	93.2	98.6	97.0	96.3	96.5	74.8	91.5	95.8	94.8
ML-GCN	95.1	98.6	97.2	97.4	98.2	79.2	95.0	96.8	97.8
A-GCN	91.8	97.6	95.6	95.5	93.0	72.2	89.2	95.3	95.8
This model	95.8	99.8	97.5	98.7	98.9	79.9	95.5	97.2	98.5

Both in terms of the number of object categories and the number of images, the MS-COCO 2014 dataset is much larger than the PASCAL VOC 2007 dataset. Therefore, the experiments on the MS-COCO 2014 dataset are more in line with the actual scenario setting and also more challenging. The same TRAIN dataset as other methods is used for network training during the experiment, and the TEST dataset is utilized for the final verification of the recognition effect after completing the training, and the loss convergence process is shown in Fig. 3. The experiments include the results of this paper's model running on the MS-COCO 2014 dataset, as well as the comparison results with ResNet-101, ML-GCN, A-GCN and other methods, and the comparison results are shown in Table 2, the experimental results of this paper's model on the MS-COCO 2014 dataset also have a significant advantage, and its overall performance improves with respect to ML-GCN by 0.6%, which is due to the fact that this paper's method more fully utilizes the multi-label information to learn a classifier with stronger performance, which in turn proves the effectiveness of the proposed method.

The final multi-scale graph convolutional network used in this paper is a two-layer network, and the dimensions of its output features are 1024 and 2048, respectively, and the features can be observed using the t-SNE downscaling method. t-SNE is a downscaling method to downscale high-dimensional data to two-dimensional or three-dimensional data, which can be used to downscaling the high-dimensional data and then observe its distribution. When high-dimensional data are downscaled to 2D or 3D and visualized, the more categories there are, the more distinguishable they are. Since the PASCAL VOC 2007 dataset has only 20 categories, while the MS-COCO 2014 dataset has 80 categories, in this paper, we will carry out t-SNE dimensionality reduction and visualization of the label relationship graph of the MS-COCO 2014 dataset, and the results of the initial label visualization and the final label visualization results are shown in Fig. 4 and Fig. 5, respectively. Each point in the graph indicates each category, and the same color indicates the categories with relevant connections, and it can be seen that the multiscale graph convolutional network can make the categories with correlations obviously close to each other.

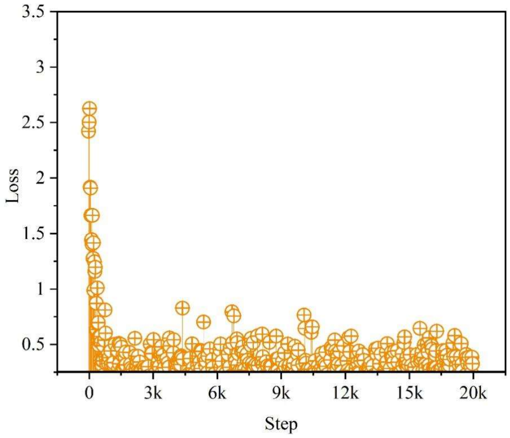

Figure 3: Loss convergence process on the MS-COCO 2014 dataset

Table 2: Performance comparison of MS-COCO 2014 data set

Model	mAP	CP	CR	CF1	OP	OR	OF1
CNN-RNN	61.0	66.0	55.7	60.8	69.1	66.4	67.8
SRN	77.2	81.6	65.2	71.3	82.7	69.7	75.8
ResNet-101	77.1	80.1	66.5	72.7	84	70.6	77.0
ML-GCN	82.5	83.3	71.4	76.3	84.5	74.3	78.9
A-GCN	82.0	81.2	71.3	75.8	81.7	73.9	77.6
This model	83.1	83.9	71.5	77.0	82.7	74.6	78.0

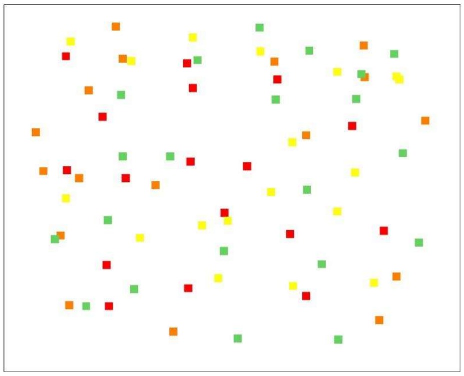

Figure 4: Initial label vision

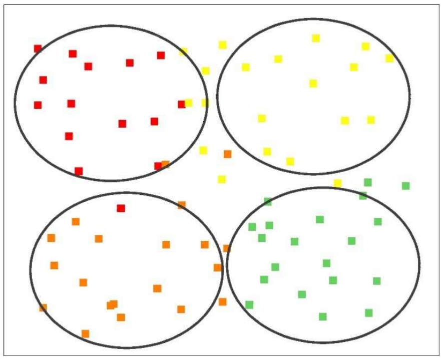

Figure 5: Final label vision

IV. Image recognition optimization algorithm application and analysis

IV. A. Experimental data and pre-processing

YogaPose Recognition Dataset is a publicly available yoga pose recognition dataset containing videos of multiple yoga poses and annotation information. Each video records a complete yoga pose process with accompanying pose category labels. The content of the dataset includes video data of yoga poses and corresponding annotation information. The video data are stored in a standard video format and contain sequences of consecutive frame images to show the complete process of a yoga pose. The annotation information provides action category labels for each video to indicate the category of the displayed yoga pose. The labels can be predefined categories of yoga poses such as Tree Pose, Downward Facing Dog, Cat Pose, etc. or customized categories.

IV. B. Experimental results and analysis

In order to delve into the performance of image recognition algorithms based on multi-scale attention mechanisms in practical applications, the aforementioned datasets are used. These datasets cover a wide range of yoga poses, which are recognized in detail according to the different categories of yoga poses in order to ensure accuracy. The algorithm was compared with other popular models in order to evaluate its performance more comprehensively. Multiple models were trained in the experiments in order to evaluate it from multiple perspectives. First, a traditional machine learning model, the Support Vector Machine (SVM), was chosen. Next, two base models were chosen: the Resnet50 and the LSTM.Then, to further improve the performance, the RNN-CNN was used, which is a network that captures both temporal and spatial information. While LSTM-CNN combines the temporal processing capability of LSTM and the image processing capability of CNN, the training results of all these models are shown in Table 3.

In the experiments, the performance of the machine learning algorithm SVM shows a moderate trend. Specifically, its performance metrics are located between the deep learning model and the deep learning combined model. This means that although SVM may be a powerful tool in some tasks, it did not perform as well as it should have in this particular yoga pose recognition task. The deep learning model Resnet50 also did not perform well in this classification task, despite its in-depth feature extraction of the action images in an attempt to extract more information from them, Resnet50 still appeared to be overpowered in processing these features. This may be due to the fact that Resnet50 mainly targets conventional image recognition tasks, while yoga action recognition involves features that may be different from conventional tasks. Comparatively, the model in this paper, along with LSTM-CNN, shows significant advantages in all performance metrics. This finding suggests that graph neural networks seem to be more advantageous when dealing with data that have complex feature dimensions and strong inter-correlations between different dimensions. This inter-correlation may be due to the fact that some parts of the yoga pose are closely related to other parts, and the graph neural network is able to capture these relationships effectively. It is worth mentioning that the F1 value of this paper's model is the highest among all seven sets of comparison models, which is 91.76%. This result further proves the excellent performance and efficient application of this paper's image recognition algorithm based on graph neural network in the task of yoga pose image recognition, which not only performs well in a single metric, but also shows the best results overall.

Table 3: Comparison of experimental results with different models adopted

Model	Accuracy/%	Recall rate/%	F1/%
SVM	78.20	72.76	79.31
Resnet50	70.94	73.02	73.62
LSTM	85.34	86.38	87.51
RNN-CNN	86.94	86.52	86.14
LSTM-CNN	86.75	89.75	88.68
This model	90.05	92.97	91.76

Table 4 shows the accuracy results of different algorithms for recognizing different yoga pose categories. The accuracy of this paper's method for image recognition of multiple yoga poses in meditation and tree pose reaches 90% and above, especially in the task of downward dog yoga pose recognition, the recognition accuracy reaches 98%, which obtains the highest performance compared to all the comparative algorithms, and further proves the robustness and feasibility of this paper's algorithms for the task of image recognition of yoga poses with multi-variable nature.

Table 4: Different actions are compared in different models

Model	Meditation accuracy/%	Tree posture accuracy/%	Dog accuracy/%	Cat accuracy/%
SVM	83	74	95	83
Resnet50	91	81	95	84
LSTM	88	84	87	82
RNN-CNN	90	87	90	90
LSTM-CNN	96	84	91	91
This model	97	93	98	94

V. Conclusion

In this paper, we propose a multi-label image recognition method based on Transformer's multi-scale attention mechanism.

The method makes full use of the multi-label information to learn a classifier with stronger performance, and obtains excellent performance on both datasets PASCAL VOC 2007 and MS-COCO. Comparing the recognition effect of this method with the sub-optimal ML-GCN method, it is found that the overall performance of this paper's method is improved by 0.7% and 0.6% on the two datasets compared to the ML-GCN method, which fully proves the effectiveness of the proposed method in this paper, respectively.

When dealing with more complex feature dimensions, the graph neural network in this paper's method is able to capture the correlations that exist between different objects. The strong capture ability of the graph neural network allows the F1 value of this paper's model to reach 91.76% among the compared models, which is the best performance. It further proves the excellent ability of this paper's model in the task of yoga pose image recognition, which can quickly categorize the characteristics of different yoga poses and ensure the accurate recognition of diverse yoga poses.

Although the multi-label image recognition method proposed in this paper has better results in improving image recognition accuracy, its more complex structure increases the number of parameters in the model itself, which in turn affects the training speed of the model. Therefore, future research can focus on the research areas of lightweight models and achieving real-time image recognition.

References

[1] Song, J., Lee, S. B., & Park, A. (2020). A study on the industrial application of image recognition technology. The Journal of the Korea Contents Association, 20(7), 86-96.

[2] Zhang, X. (2022). Application of artificial intelligence recognition technology in digital image processing. Wireless Communications and Mobile Computing, 2022(1), 7442639.

[3] Zuo, K. J., Saun, T. J., & Forrest, C. R. (2019). Facial recognition technology: a primer for plastic surgeons. Plastic and reconstructive surgery, 143(6), 1298e-1306e.

[4] Chen, H., Geng, L., Zhao, H., Zhao, C., & Liu, A. (2022). Image recognition algorithm based on artificial intelligence. Neural Computing and Applications, 1-12.

[5] Hijazi, S., Kumar, R., & Rowen, C. (2015). Using convolutional neural networks for image recognition. Cadence Design Systems Inc.: San Jose, CA, USA, 9(1).

[6] Zhang, S., Wu, Y., & Chang, J. (2020, June). Survey of image recognition algorithms. In 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (Vol. 1, pp. 542-548). IEEE.

[7] Traore, B. B., Kamsu-Foguem, B., & Tangara, F. (2018). Deep convolution neural network for image recognition. Ecological informatics, 48, 257-268.

[8] Liu, Y. H. (2018, September). Feature extraction and image recognition with convolutional neural networks. In Journal of Physics: Conference Series (Vol. 1087, p. 062032). IOP Publishing.

[9] Zhang, Z., Xu, Y., Shao, L., & Yang, J. (2017). Discriminative block-diagonal representation learning for image recognition. IEEE transactions on neural networks and learning systems, 29(7), 3111-3125.

[10] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1), 4-24.

[11] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., ... & Sun, M. (2020). Graph neural networks: A review of methods and applications. AI open, 1, 57-81.

[12] Corso, G., Stark, H., Jegelka, S., Jaakkola, T., & Barzilay, R. (2024). Graph neural networks. Nature Reviews Methods Primers, 4(1), 17.

[13] Veličković, P. (2023). Everything is connected: Graph neural networks. Current Opinion in Structural Biology, 79, 102538.

[14] Cappart, Q., Chételat, D., Khalil, E. B., Lodi, A., Morris, C., & Veličković, P. (2023). Combinatorial optimization and reasoning with graph neural networks. Journal of Machine Learning Research, 24(130), 1-61.

[15] Zhao, S., & Gu, S. (2024). A neural network algorithm framework based on graph structure for general combinatorial optimization. Neurocomputing, 587, 127670.

[16] Goel, L., Gupta, S., Gupta, A., Rajan, S. N., Gupta, V. K., Singh, A., & Gupta, P. (2024). Advancing ASD detection: novel approach integrating attention graph neural networks and crossover boosted meerkat optimization. International Journal of Machine Learning and Cybernetics, 15(8), 3279-3297.

[17] Shi, M., Tang, Y., Zhu, X., Huang, Y., Wilson, D., Zhuang, Y., & Liu, J. (2022). Genetic-gnn: Evolutionary architecture search for graph neural networks. Knowledge-based systems, 247, 108752.

[18] Guo, K., Hu, Y., Qian, Z., Liu, H., Zhang, K., Sun, Y., ... & Yin, B. (2020). Optimized graph convolution recurrent neural network for traffic prediction. IEEE Transactions on Intelligent Transportation Systems, 22(2), 1138-1149.

[19] Zhou, K., Huang, X., Song, Q., Chen, R., & Hu, X. (2022). Auto-gnn: Neural architecture search of graph neural networks. Frontiers in big Data, 5, 1029307.

[20] Kasar, M. M., Bhattacharyya, D., & Kim, T. H. (2016). Face recognition using neural network: a review. International Journal of Security and Its Applications, 10(3), 81-100.

[21] Weihong, W., & Jiaoyang, T. (2020). Research on license plate recognition algorithms based on deep learning in complex environment. IEEE Access, 8, 91661-91675.

[22] Zengxi Feng,Xiuming Ji,Hua Zou,Jincong Lu,Junhao Yan & Xiuying Yan. (2025). Exploring crowd counting methodology by integrating CNN and transformer: performance optimization under weak supervision. Signal, Image and Video Processing,19(6),483-483.

[23] Jari Isohanni. (2025). Customised ResNet architecture for subtle color classification. International Journal of Computers and Applications,47(4),341-355.

[24] Abhinav Suresh,Henning Schlömer,Baran Hashemi & Annabelle Bohrdt. (2025). Interpretable correlator Transformer for image-like quantum matter data. Machine Learning: Science and Technology,6(2),025006-025006.