Multimedia Tools and Applications (2022) 81:25363–25381
https://doi.org/10.1007/s11042-022-12397-y

# Label graph learning for multi-label image recognition with cross-modal fusion

**Yanzhao Xie**[1] · **Yangtao Wang**[2] · **Yu Liu**[1] · **Ke Zhou**[1]

Received: 6 April 2021 / Revised: 4 January 2022 / Accepted: 25 January 2022 /
Published online: 23 March 2022


## Abstract

It has become popular to learn the correlation between labels in most existing multi-label image recognition tasks. Existing approaches begin to construct a label graph to learn the label dependencies but they suffer from a low convergence efficiency when fusing image features and label embeddings, and also limit the performance improvement on multi-label images. To overcome this challenge, we propose a label graph learning model (termed as LGLM) for multi-label image recognition, which integrates a multi-modal fusion component to efficiently fuse cross-modal embeddings. First, LGLM uses convolution neural network to learn the feature for each image. Second, LGLM first constructs a label graph according to the word vector of each object and then adopts graph convolution network to learn the label correlations to generate label co-occurrence embeddings. Finally, the multi-modal fusion component efficiently fuses image features and label co-occurrence embeddings to generate an end-to-end image recognition model. We conduct extensive experiments on MS-COCO and FLICKR25K and the experimental results demonstrate the superiority of LGLM compared with the state-of-the-art image recognition methods. The code of LGLM has been released on GitHub: https://github.com/lzHZWZ/LGLM.

**Keywords** Multi-label image recognition · Label graph · Graph convolution network · Multi-modal fusion

✉ Yangtao Wang
ytaowang@gzhu.edu.cn

Yanzhao Xie
yzxie@hust.edu.cn

Yu Liu
liu_yu@hust.edu.cn

Ke Zhou
zhke@hust.edu.cn

[1] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, 1037 Luoyu Road, Wuhan, China

[2] School of Computer Science and Cyber Engineering, Guangzhou University, 230 Wai Huan Xi Road, Guangzhou, China

## 1 Introduction

Recently, multi-label image recognition has been widely studied in multiple fields such as human attribute recognition [21], scene understanding [30] and fashion attribute recognition [16], which becomes a popular task in computer vision. Different from single-label image classification that naively allocates one label for each image, multi-label image recognition tasks aim to learn all objects that co-appear in the same image and predicts a set of labels for this image, thus it brings greater challenges to capture rich semantic information in multi-label images. Early approaches divide this multi-label image recognition task into multiple individual single-label image classification problems, which train a classifier for each label. With the rapid development of deep convolution neural network (CNN) [13], the image classification performance of these binary approaches has been greatly promoted. Nevertheless, above methods have not taken the correlations between these objects into consideration at all, resulting in low classification performance on multi-label images. Note that we list the commonly-used abbreviations and their explanations in Table 1 for convenience.

To capture the co-occurrence correlation between objects, various works have been explored in the past few years. Some researchers model the label dependencies via attention mechanisms. For instance, Zhu et al. [35] propose SRN that uses image-level supervisions to learn an attention map for each given label, which focuses on the related image region corresponding to each label. Combined with a long short-term memory (LSTM) unit, Wang et al. [33] efficiently perceive attentional regions via a spatial transformer to capture the label dependencies. However, they only associate each attentional region with its corresponding label within an image, but fail to take the global label distribution over the label set into consideration. For example, "book" and "desk" will co-occur in an image with a high possibility, while "book" and "ocean" will hardly co-appear in the same image. To capture the global label dependencies on the whole dataset, Chen et al. [3] design a multi-label recognition model ML-GCN, which utilizes graph convolution network (GCN) [19] to generate label co-occurrence embeddings based on the label statistical information. Although this approach has achieved good performance on multi-label images, it needs to manually collect the mutual conditional probability between different labels and thus becomes not flexible enough in the design of its label graph. Furthermore, to get rid of this inflexible correlation graph problem, Li et al. [22] propose A-GCN to model the label correlations by designing an adaptive label graph. Nevertheless, these two methods adopt dot product (DP) to complete image features and label co-occurrence embeddings fusion process, which severely limits the model convergence and prevents the multi-label image classification performance improvement.

**Table 1** Commonly-used abbreviations and their explanations used in this paper

| Abbreviation | Explanation |
|---|---|
| CNN | convolution neural network |
| LSTM | long short-term memory |
| GCN | graph convolution network |
| DP | dot product |
| MFC | multi-modal fusion component |
| RNN | recurrent neural network |
| VQA | visual question answering |
| SGD | stochastic gradient descent |

To overcome this challenge, this paper integrates a multi-modal fusion component (MFC) [37] and proposes a label graph learning model (termed as LGLM) for multi-label image recognition. LGLM mainly contains three key modules: an image feature extraction module, a label graph learning module and a cross-modal fusion module. In the feature extraction module, LGLM employs CNN (ResNet-101 [13]) to learn each image feature. In the label graph learning module, following A-GCN [22], LGLM first obtains each label vector via the word embedding technique [27], then constructs a label graph to learn the correlation between these labels, and finally adopts GCN to learn the label correlations to obtain the label co-occurrence embeddings. At last, different from previous works, a multi-modal fusion component (MFC) has been integrated into LGLM in the cross-modal fusion module to efficiently complete cross-modal vectors fusion in an end-to-end manner. Extensive experiments on MS-COCO [23] and FLICKR25K [15] demonstrate not only the convergence speed of LGLM has been greatly promoted but also the multi-label image recognition performance has been further boosted than the state-of-the-art methods.

The main contributions of this paper can be summarized as follows.

- We propose LGLM that utilizes GCN to learn the label dependencies for multi-label image recognition.
- We design a cross-model component MFC to effectively fuse image features and label co-occurrence embeddings to greatly speed up the model convergence as well as further promote the classification performance.
- Extensive experiments on MS-COCO and FLICKR25K demonstrate LGLM can achieve better performance compared with the state-of-the-art multi-label image recognition methods.

## 2 Related works

In this section, we mainly review existing studies related to our work including multi-label image recognition, GCN and cross-modal fusion.

### 2.1 Multi-label image recognition

Single-label image recognition has achieved great success with the rapid development of CNN [13]. Based on the powerful feature extraction ability of deep CNN, many researchers transform multi-label recognition task into multiple individual single-label classification problems for each label. Gong et al. [8] annotate multi-label images by integrating a top-k ranking objective function into its convolution structures. Razavian et al. [28] apply off-the-shelf features extracted from deep network pre-trained on ImageNet [29] to recognize multi-label images. These binary approaches have not taken the correlations between these objects into consideration at all. To better capture the label dependencies, several conventional graph models like dependency network [9], conditional random field [7] and recurrent neural network (RNN) [31] are widely incorporated. Recently, some works begin to associate the image region with its corresponding label via attention mechanisms. Zhu et al. [35] propose SRN that uses image-level supervisions to learn an attention map for each given label. Combined with an LSTM unit, Wang et al. [33] efficiently perceive attentional regions via a spatial transformer to capture the label dependencies. Guo et al. propose VACIT [10] to learn the attention map distances between the original image and its transformed one. However, these methods only focus on the local label correlations within each image but ignore

the global label dependencies between different objects over the data distribution. To this end, Chen et al. [3] present ML-GCN to learn the label co-occurrence embeddings based on the label statistical information, but this method requires to collect the label statistical information by hand. To overcome this, Li et al. [22] propose A-GCN that constructs an adaptive label graph rather than the hand-crafted correlation graph to capture label dependencies. Both of these schemes use DP to complete image features and label co-occurrence embeddings fusion process, so they severely limit the model convergence and prevents further performance improvement on multi-label image recognition.

### 2.2 Graph convolution network

GCN [19] takes the node features and the correlation matrix between nodes as input, and then outputs the updated node features after multiple graph convolution operations. Formally, this process can be described as follows:

$$H^{l+1} = a^l(\hat{A}H^l W^l), \tag{1}$$

where $H^l$, $W^l$, $a^l$ respectively denote the input features, weights, non-linear activation function of the $l$-th graph convolution layer and $\hat{A}$ denotes the normalized version of correlation matrix $A$. Monti et al. [26] utilize GCN to classify research papers for their citation network. In addition, Defferard et al. [4] apply GCN to N-grams for text categorization. More recently, GCN begins to play an important role in image generation, image classification, emotion learning, *etc*. Johnson et al. [17] apply GCN to the image generation tasks from scene graphs. With the development of deep GCN, ML-GCN [3] has been proposed to learn the label co-occurrence embeddings based on the label statistical information, but this method requires to collect the label statistical information by hand. To overcome this, Li et al. [22] propose A-GCN that constructs an adaptive label graph rather than the hand-crafted correlation graph to capture label dependencies. Both of these schemes use DP to complete image features and label co-occurrence embeddings fusion process, so they severely limit the model convergence and prevents further performance improvement on multi-label image recognition. EmotionGCN [14] is a similar work that adopt CNN for image feature and GCN for emotion distribution learning. Furthermore, SS-GRL [2] generates class-specific image representations via semantic graph networks learning to pay more attention to the semantic regions of images. These methods aim to capture the co-occurrence probability of training samples, which may reduce the model generalization ability, especially for those rare co-occurrence objects. To overcome this, ADD-GCN [34] designs an attention driven dynamic GCN, which can dynamically generate a specific graph for each image to model the relationships between content perception categories generated by semantic attention module. The above works inspired us to effectively explore the correlations between different objects.

### 2.3 Cross-modal fusion

In recent years, modeling textual or visual information with vector representations trained from large language or visual datasets has been successfully explored in the visual question answering (VQA) problem [24], which aims to effectively fuse vectors from different modalities. Most existing schemes simply use linear models [36] like concatenation or element-wise addition to integrate cross-modal embeddings, but these methods fail to obtain expressive image-text features and cannot fully capture the complex correlations between these features. To effectively address these problems, a deep multimodal attentive fusion

(DMAF) [11] approach has been proposed to automatically draws focuses on affectional regions and words, which can capture the complementary and non-redundant information for more effective sentiment classification. Besides, Multi-modal Compact Bilinear (MCB) [5] pooling and Multi-modal Low-rank Bilinear (MLB) [18] pooling have been developed to reduce the computational complexity of the original bilinear pooling model and make it practicable for VQA. However, MCB needs very high-dimensional feature to guarantee good performance and MLB needs a great many training iterations to converge to a satisfactory solution. Furthermore, Zhou et al. propose Multi-modal Factorized Bilinear (MFB) [37] pooling which enjoys the dual benefits of compact output features of MLB and robust expressive capacity of MCB, thereby efficiently fusing image features and text embeddings as well as greatly speeding up the model convergence. In addition, Huang et al. [12] design a correlational multimodal variational autoencoder (CMVAE) network to learn both the common information in all modalities and private information in each modality to enhance the performance of cross-modal retrieval as well as multi-label classification.

Inspired by the above works, our scheme improves and integrates the MFC to efficiently complete image features and label co-occurrence embeddings fusion process.

## 3 Proposed methodology

In this section, we propose LGLM that mainly contains three modules: an image feature extraction module to learn the visual feature for each input image, a label graph learning module to generate label co-occurrence embeddings, and a cross-modal fusion module to efficiently complete image features and label co-occurrence embeddings fusion process to generate an end-to-end multi-image classification model. Figure 1 presents the overall framework of LGLM and we elaborate the workflow of LGLM below.

### 3.1 Image feature extraction

In this section, we aim to extract the image feature by fine-tuning a state-of-the-art CNN model. As shown in the blue frame of Fig. 1, we adopt ResNet-101 [13] to learn and generate the visual feature for each input image $I$ with 448×448 resolution. Note that this module gets rid of the last fully-connected ($fc$) layer of ResNet-101 and from the "$conv5\_x$"

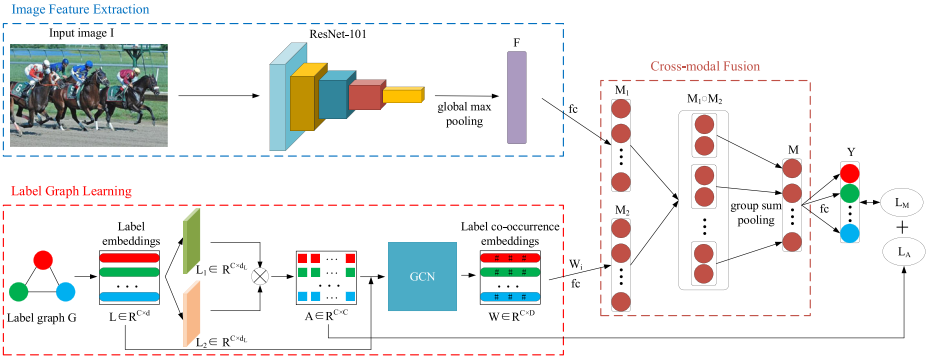


**Fig. 1** The overall framework of LGLM

layer, so we can generate the $2048 \times 14 \times 14$ feature maps for each image. After that, for fair comparisons with the existing mainstream methods, we conduct the global max pooling operation to generate a $D$-dimensional vector $F$ for each image $I$:

$$F = f_{gmp}(f_{cnn}(I; \theta_{cnn})), \tag{2}$$

where $f_{gmp}$ denotes the global max pooling operation, $f_{cnn}$ denotes ResNet-101, $\theta_{cnn}$ denotes the model parameters, and $D = 2048$.

### 3.2 Label Graph Learning

In this part, this module designs GCN with two layers to complete the label graph learning process to obtain the label co-occurrence embeddings. This process aims to capture the label dependencies between different objects.

In the red frame of Fig. 1, we treat each object as a node of our label graph $G$ consisting of $C$ nodes, where $C$ denotes the number of object categories in a dataset. Following ML-GCN [3], we adopt the GloVe [27] model to map each object into a $d$-dimensional word vector. As a result, we will generate a $C \times d$ label embeddings matrix $L$ for all the $C$ object labels. After obtaining the node features, we begin to construct the correlation matrix based on these nodes' features. ML-GCN [3] has to manually count the occurrence times of each object as well as the conditional probability between different objects to model the label dependencies, which becomes inflexible and may bring sub-optimal results for multi-label image recognition. To get rid of this hand-crafted correlation graph, following A-GCN [22], we aim to learn the label correlations and obtain the correlation matrix in an end-to-end manner.

Based on the label embeddings matrix $L$, two $1 \times 1$ convolution layers are used to respectively generate two intermediate $C \times d_L$ matrices $L_1$ and $L_2$, then adopt a DP operation to generate the $C \times C$ correlation matrix $A$. This process can be formulated as:

$$\begin{aligned} L_1 &= f_\alpha(L; \theta_\alpha), \\ L_2 &= f_\beta(L; \theta_\beta), \\ A &= \frac{1}{C} L_1 \otimes L_2^T, \end{aligned} \tag{3}$$

where $f_\alpha$ and $f_\beta$ respectively denote the convolution operations of these two branches, $\theta_\alpha$ and $\theta_\beta$ respectively denote the network parameters of these two branches, $\otimes$ denotes the DP operation. Following the commonly-used normalization trick [19], $A$ will be normalized to $\hat{A}$ as follows:

$$\hat{A} = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}, \tag{4}$$

where $\tilde{A} = A + I_C$ and $\tilde{D}$ is a diagonal matrix with $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ for $i, j \in [1, C]$, $I_C$ is a $C \times C$ identity matrix.

After obtaining both the label embeddings matrix $L$ and normalized label correlation matrix $\hat{A}$, we design GCN with two layers to update the nodes' features as follows:

$$L_{l+1} = f_l(\hat{A} L_l U_l), \quad l \in [0, 1] \tag{5}$$

where $L_l$, $U_l$ and $f_l$ respectively denote the nodes' features, weights and non-linear function (ReLU) in the $l$-layer. It's worth mentioning that the input of this sub-network contains $L$ and $\hat{A}$, and the output is a $C \times D$ label co-occurrence embeddings matrix $W$. Next, we will fuse each row of $W$ with the image feature $F$ in the cross-modal fusion module.

Note that in the GCN propagation process, the feature of each node may become indistinguishable due to the over-smoothing problem. To address this issue, similar to A-GCN [22], we design L1-norm loss to enforce a spare correlation constraint on $\hat{A}$ as follows:

$$L_A = ||\hat{A} - I_C||_2. \tag{6}$$

where $||\cdot||_2$ denotes the 2-norm distance. On the one hand, $L_A$ can avoid over-smoothed features in GCN by highlighting self-correlation weights. On the other hand, $L_A$ will be combined with the multi-label loss function (see (8)) to optimize our model.

### 3.3 Cross-modal Fusion

Previous works directly adopt the DP operation to fuse image features and label co-occurrence embeddings from two different modalities, but they severely limit the convergence efficiency as well as the performance improvement of the model. Instead, this paper designs MFC to efficiently complete these cross-modal vectors fusion to speed up the model convergence and further promote the classification performance.

Commonly, MFB is implemented by combining several $fc$, element-wise multiplication and pooling layers. As shown in the brown frame of Fig. 1, we improve MFB and design MFC to adapt to our multi-label classification task. Formally, for $\forall i \in [1, C]$, given the feature vector $F \in R^D$ of image $I$, we aim to fuse $F$ and $W_i \in R^D$ to generate the $i$-th element $Y^i$ of the predicted labels $Y \in R^C$:

$$Y^i = f_{MFC}(F, W_i; \theta_{MFC}), \tag{7}$$

where $f_{MFC}$ denotes our cross-modal fusion module, $W_i$ denotes the $i$-th row vector of the label co-occurrence embeddings matrix $W$, and $\theta_{MFC}$ denotes this module parameters. We describe the workflow of this module as follows.

First, $F$ and $W_i$ will be converted into two $m$-dimensional vector $M_1$ and $M_2$ via two $fc$ layers. Next, we adopt Hadamard product ($\circ$) to generate another $m$-dimensional vector $M_1 \circ M_2$ by fusing $M_1$ and $M_2$. Furthermore, we utilize group sum pooling to transform $M_1 \circ M_2$ into a $\frac{m}{g}$-dimensional vector $M$ to greatly speed up the model convergence, where each group that consists of $g$ units will be sequentially converted into one unit. At last, a $fc$ layer will convert $M$ into the $i$-th element $Y^i$ of $Y$. As a result, the complete predicted labels $Y$ of image $I$ will be generated by $C$ times fusion with $F$.

According to the predicted labels $Y$, we adopt a commonly-used multi-label loss function to calculate the loss $L_M$ between $Y$ and the ground truth labels $\hat{Y}$ of image $I$:

$$L_M = \sum_{i=1}^{C} \hat{Y}^i log(Sigmoid(Y^i)) + (1 - \hat{Y}^i) log(1 - Sigmoid(Y^i)), \tag{8}$$

where $\hat{Y}^i$ denotes the $i$-th element of $\hat{Y}$ and $\hat{Y}^i = \{1, 0\}$ denotes whether the label $i$ appears in the image $I$ or not. At last, $L_M$ and $L_A$ (see (6)) will be combined to train the whole network in an end-to-end manner as follows:

$$L_{total} = L_M + \lambda L_A, \tag{9}$$

where $\lambda \in [0, 1]$ is a weighted factor to balance these two losses.

## 4 Experiments

In this section, we evaluate the performance of LGLM and compare it with the state-of-the-art multi-label image recognition methods. We first describe the datasets, then introduce our experimental settings and finally report the experimental results.

### 4.1 Datasets

**MS-COCO** [23] contains 80 semantic label concepts with 82,783 training images, 40,504 validation images and 40,775 test images. We train LGLM on the train set and evaluate its performance on the validation set, owing that the ground truth labels of the test set are not available.

**FLICKR25K** [15] consists of 25,000 multi-label images that belong to 24 labels, and each image is averagely annotated by 4.7 labels. We randomly split the train set and test set with a ratio of 5:5, and then evaluate the performance on the test set.

### 4.2 Experimental settings

**Implementation details**. We employ PyTorch to conduct all the experiments. In the image feature extraction module, we employ ResNet-101 to generate the feature for each image with 448×448 resolution. In the label graph learning module, following ML-GCN, we utilize the GloVe model to obtain a 300-dimensional (*i.e.*, $d = 300$) word embedding for each label word. The output of two convolution branches in (3) are both $C \times 1024$ (*i.e.*, $d_L = 1024$) matrices. In addition, we design GCN with two layers to respectively output 1024 and 2048 dimensional label co-occurrence embeddings after each layer. In the cross-modal fusion module, we set $m = 1024$ and $g = 2$ to respectively complete the cross-modal fusion and group sum pooling process. Our model will be trained using stochastic gradient descent (SGD) with a momentum of 0.9, a batchsize of 32, a weight decay of $10^{-4}$, and an initial learning rate of 0.1 which decays by a factor of 10 every 40 epochs. Note that each experiment has been repeatedly conducted to avoid accidental phenomenon and all results keep stable in each experiment.

**Evaluation metrics**. Following mainstream methods [3, 22, 33, 35], we report two important evaluation metrics, *i.e.*, the average precision (AP) for each category and mean average precision (mAP) over all categories. In addition, we also calculate the per-class precision (CP), per-class recall (CR), per-class F1 (CF1) and the overall precision (OP), overall recall (OR), overall F1 (OF1). For fair comparisons with existing methods, we further list the experimental results (*i.e.*, CP-3, CR-3, CF1-3, OP-3, OR-3, OF1-3) on top-3 labels of the classification scores.

### 4.3 Experimental results

In this section, we first evaluate the convergence efficiency and model complexity of LGLM, then compare the classification results with the state-of-the-art methods, next conduct ablation studies, and finally list the visual results.

#### 4.3.1 Convergence efficiency and model complexity

In this section, we compare the convergence efficiency and model complexity (*i.e.*, FLOPs) [25] of LGLM with ML-GCN [3] and A-GCN [22]. We employ the same configuration settings (*i.e.*, SGD, learning rate, batchsize, *etc*) among these three approaches in the implementation for fair comparisons.

On the one hand, we respectively record the change of loss and mAP in Figs. 2 and 3. As shown in Fig. 2, the curve of ML-GCN is shaking violently with the increase of epoch while LGLM and AGCN can converge with a faster and more stable speed. The reason why ML-GCN produces a lower loss is that other two approaches train their respective network with a multi-label classification loss and another L1-norm loss (see (6)). Besides, according to the experimental results in Fig. 3, LGLM has converged within 25 epochs, and obtains higher mAP of 84.2%, and 80.6% on MS-COCO and FLICKR25K respectively. However, ML-GCN and A-GCN never converge at the same time and their mAP results are respectively 39.4% and 37.8% lower on MS-COCO, and 26.7% and 24.2% lower on FLICKR25K compared with LGLM. In addition, we find ML-GCN and A-GCN will respectively take about 200 (more than 8 times of LGLM) and 150 (more than 6 times of LGLM) epochs to train the model. The reasons lie in two parts. One is that Hadmard product and group sum-pooling can increase the interaction of vector elements among different modalities, which promotes the precision. The other is that by means of sum-pooling, it decreases over-fitting and parameter explosion caused by the increasing interaction and thus speeds up the model convergence.

On the other hand, we compare the model complexity including the number of parameters and FLOPs in Table 2. According to our statistics, the parameter number of ML-GCN, A-GCN and LGLM is 42.50M, 42.57M and 44.04M respectively. All these three methods own the same computation complexity, i.e., 31.37 GFLOPs. The reason why our LGLM owns more parameters is that we add another MFC to the model. For each sample, each model owns the same computation complexity, but our LGLM needs much fewer iterations (epochs) to complete the training process via efficient cross-modal fusion, which means the total computation complexity during the model training of LGLM is much less than others. We also talk about how to reduce the model complexity in our future work in Section 5.

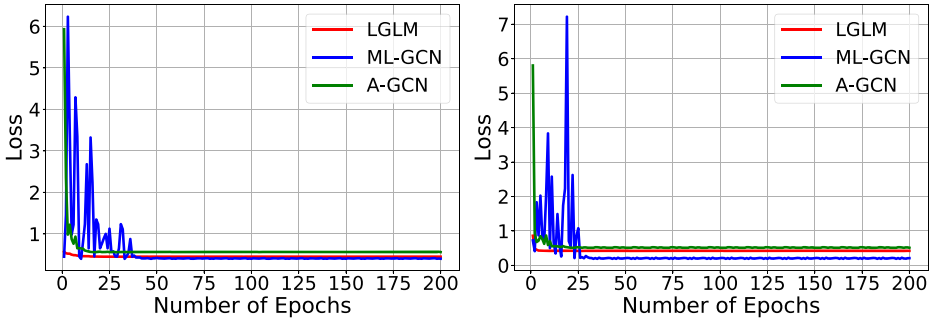


**Fig. 2** The change of loss on the test set with the increase of epoch on the train set

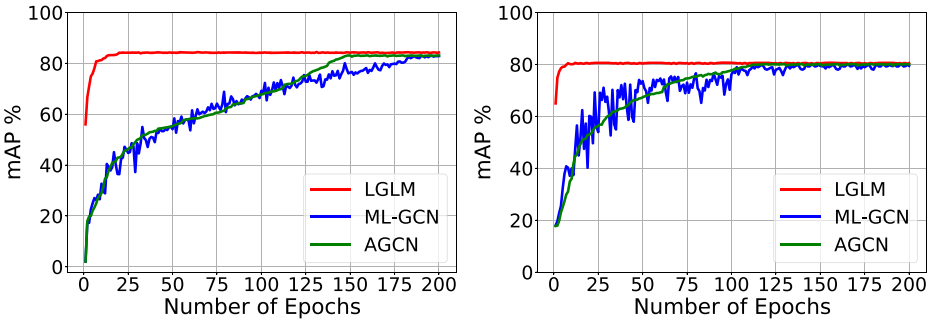


**Fig. 3** The change of mAP on the test set with the increase of epoch on the train set

### 4.3.2 Comparisons with the state-of-the-art methods

In this section, we compare the performance (AP, mAP, CP, CR, CF1, OP, OR, OF1, CP-3, CR-3, CF1-3, OP-3, OR-3, OF1-3) of LGLM with the the state-of-the-art methods on MS-COCO and FLICKR25K.

**Results on MS-COCO** In this part, we compare the performance of LGLM with CNN-RNN [32], RNN-Attention [33], Order-Free RNN [1], ML-ZSL [20], SRN [35], Multi-Evidence [6], ResNet-101 [13], ML-GCN [3] and A-GCN [22]. As shown in Table 3, LGLM obviously outperforms all candidate methods by at least 1.1% mAP improvement and also produces higher results on all other evaluation metrics. This is because most of candidates (such as CNN-RNN, RNN-Attention, Order-Free RNN, ML-ZSL and Multi-Evidence) neglect the label dependencies during the learning process, thus they cannot fuse the label semantic information into image features. Besides, as we see, compared with ResNet-101 baseline, LGLM greatly promotes the performance with 6.9% mAP improvement. The main reason is that we construct a label graph that can effectively capture the relationships between different objects to help generate more accurate image features. In addition, our MFC also efficiently complete the corss-modal embeddings fusion compared with those DP based methods such as ML-GCN and A-GCN, because they failed to well fuse image features and label co-occurrence embeddings. Especially, we list the representative metric (*i.e.*, mAP) between LGLM, ML-GCN and A-GCN in Fig. 4a, which illustrates that LGLM exceeds these two baseline methods by at least 1% mAP on MS-COCO and demonstrates the effectiveness of MFC.

**Results on FLICKR25K** In this part, we further compare LGLM with ML-GCN [3] and A-GCN [22] on FLIKCR25K. In addition, we also compare the classification results with

**Table 2** Model complexity comparisons

| Method | Number of parameters | Float-pointing operations (FLOPs) |
|---|---|---|
| ML-GCN | 42.50M | 31.37G |
| A-GCN | 42.57M | 31.37G |
| LGLM | 44.04M | 31.37G |

**Table 3** Performance comparisons on MS-COCO

| Method | All | | | | | | | Top-3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | CP | CR | CF1 | OP | OR | OF1 | CP | CR | CF1 | OP | OR | OF1 |
| CNN-RNN | 61.2 | - | - | - | - | - | - | 66.0 | 55.6 | 60.4 | 69.2 | 66.4 | 67.8 |
| RNN-Attention | - | - | - | - | - | - | - | 79.1 | 58.7 | 67.4 | 84.0 | 63.0 | 72.0 |
| Order-Free RNN | - | - | - | - | - | - | - | 71.6 | 54.8 | 62.1 | 74.2 | 62.2 | 67.7 |
| ML-ZSL | - | - | - | - | - | - | - | 74.1 | 64.5 | 69.0 | - | - | - |
| SRN | 77.1 | 81.6 | 65.4 | 71.2 | 82.7 | 69.9 | 75.8 | 85.2 | 58.8 | 67.4 | 87.4 | 62.5 | 72.9 |
| Multi-Evidence | - | 80.4 | 70.2 | 74.9 | 85.2 | 72.5 | 78.4 | 84.5 | 62.2 | 70.6 | 89.1 | 64.3 | 74.7 |
| ResNet-101 | 77.3 | 80.2 | 66.7 | 72.8 | 83.9 | 70.8 | 76.8 | 84.1 | 59.4 | 69.7 | 89.1 | 62.8 | 73.6 |
| ML-GCN (DP) | 83.0 | 85.1 | 72.0 | 78.0 | 85.8 | 75.4 | 80.3 | 89.2 | 64.1 | 74.6 | 90.5 | 66.5 | 76.7 |
| A-GCN (DP) | 83.1 | 84.7 | 72.3 | 78.0 | 85.6 | 75.5 | 80.3 | 89.0 | 64.2 | 74.6 | 90.5 | 66.3 | 76.6 |
| LGLM (MFC) | **84.2** | **85.7** | **72.8** | **78.7** | **86.6** | **76.7** | **81.3** | **89.4** | **64.7** | **75.0** | **90.7** | **67.4** | **77.3** |

The bold means the optimal value in the corresponding evaluation metric

the state-of-the-art methods including ADD-GCN [34], VACIT [10] and SS-GRL [2]. We present the experimental results in Table 4. As we see, LGLM achieves a higher mAP value than other candidates by 0.2%-0.7%. Besides, LGLM obviously surpasses ML-GCN and A-GCN on all other evaluation metrics by integrating the cross-modal MFC component. Although ADD-GCN and SS-GRL both construct a graph structure to learn the label embeddings, they also pay more attention to local attention regions but neglect the global label dependencies, which is inferior to LGLM in this aspect. VACIT can effectively recognize different transforms of an multi-label images, but it also failed to integrate label embeddings into image features. ML-GCN and A-GCN treat image features and label embeddings as two independent modalities, but we improve MFC to efficiently fuse these two cross-modal embeddings to further promote the performance. Especially, we list the representative metric (*i.e.*, mAP) between LGLM, ML-GCN and A-GCN in Fig. 4b, which illustrates that LGLM exceeds these two baseline methods by at least 0.6% mAP on FLICKR25K and demonstrates the effectiveness of MFC.

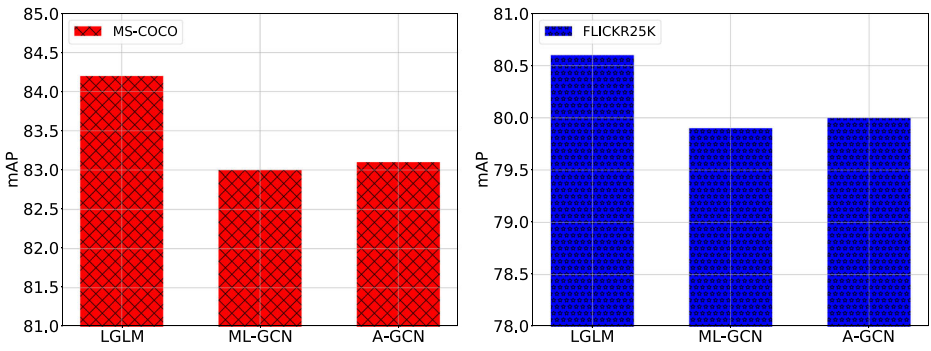


**Fig. 4** mAP comparisons of LGLM with ML-GCN and A-GCN

**Table 4** Performance comparisons on FLICKR25K

| Method | All | | | | | | | Top-3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | CP | CR | CF1 | OP | OR | OF1 | CP | CR | CF1 | OP | OR | OF1 |
| ML-GCN (DP) | 79.9 | 79.2 | 67.9 | 73.1 | 83.5 | 74.8 | 78.9 | 85.1 | 49.3 | 62.4 | 88.6 | 57.1 | 69.4 |
| A-GCN (DP) | 80.0 | 79.1 | 68.0 | 73.1 | 83.3 | 75.1 | 79.0 | 85.1 | 49.4 | 62.5 | 88.7 | 57.2 | 69.5 |
| ADD-GCN | 80.1 | 79.2 | 68.1 | 73.2 | 83.4 | 75.2 | 79.1 | 85.2 | 49.5 | 62.6 | 88.8 | 57.3 | 69.7 |
| VACIT | 80.4 | 83.6 | 64.9 | 72.5 | 86.8 | 73.3 | 79.5 | 85.7 | 48.9 | 62.3 | 89.2 | 57.6 | 70.0 |
| SS-GRL | 80.1 | 79.3 | 68.0 | 73.2 | 83.6 | 75.3 | 79.2 | 85.3 | 49.6 | 62.7 | 89.0 | 57.4 | 69.8 |
| LGLM (MFC) | **80.6** | **79.9** | **68.6** | **73.8** | **84.2** | **75.4** | **79.6** | **85.8** | **50.1** | **63.3** | **89.2** | **57.9** | **70.2** |

The bold means the optimal value in the corresponding evaluation metric

### 4.3.3 Ablation studies

In this part, we conduct ablation studies to analyze how different settings will influence the performance of our model, including the dimension of $d_L$ (see (3)), different GCN layers, the dimension of $m$ in cross-modal fusion, different units $g$ in group sum pooling, weighted factor $\lambda$ in the loss function $L_{total}$, and the effectiveness of MFC.

**The dimension of $d_L$** As mentioned in (3), we use two convolution layers to first generate the intermediate $C \times d_L$ matrices, then adopt the DP operation to fuse these two matrices to obtain the correlation matrix. In this part, we first vary the dimension of $d_L$, then observe the change of performance on MS-COCO and FLICKR25K. As shown in Fig. 5, we respectively set $d_L = 128, 256, 512, 1024, 2048$ to calculate the mAP on the two datasets. Obviously, we obtain the highest mAP when setting $d_L = 1024$ although other settings will not affect the performance too much. We believe $d_L = 1024$ is a better dimension to learn relationship between different label embeddings to generate a more effective correlation matrix.

**Different number of graph convolution network layers** In this part, we respectively design GCN with two, three and four layers to conduct the experiments and record the change of performance on MS-COCO and FlICKR25K. As shown in Table 5, when using

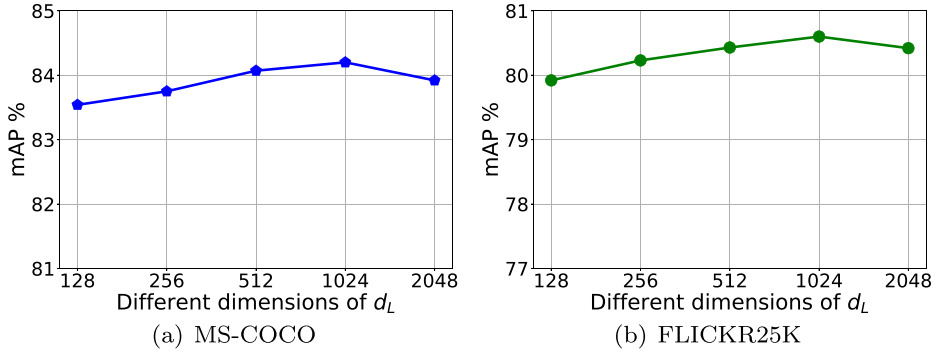


**Fig. 5** The change of mAP using different dimensions of $d_L$

**Table 5** Performance change with different GCN layers

| layers | Datasets | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | MS-COCO | | | | | FLICKR25K | | | | |
| | mAP | CF1 | OF1 | CF1-3 | OF1-3 | mAP | CF1 | OF1 | CF1-3 | OF1-3 |
| 2 | **84.2** | **78.7** | **81.3** | **75.0** | **77.3** | **80.6** | **73.8** | **79.6** | **63.3** | **70.2** |
| 3 | 83.8 | 78.1 | 80.8 | 74.4 | 76.9 | 80.1 | 73.1 | 79.0 | 62.5 | 69.8 |
| 4 | 83.3 | 77.9 | 80.4 | 74.1 | 76.5 | 79.8 | 72.8 | 78.7 | 62.2 | 69.4 |

The bold means the optimal value in the corresponding evaluation metric

two GCN layers, we can achieve the highest mAP, CF1, OF1, CF1-3 and OF1-3 on MS-COCO and FLICKR25K. However, when increasing the layers of GCN, the performance is gradually decreasing on all datasets. This is because the features of nodes may become indistinguishable in the propagation process with multiple accumulated layers, which will reduce the image recognition effect. Therefore, in the experiments, we choose two GCN layers to complete the label graph learning process.

**The dimension of *m* in cross-modal fusion** The output of image feature vector and label co-occurrence embeddings are 2048-dimensional vector pairs, both of which will be converted into $m$-dimensional vectors in the cross-modal fusion module. In this part, we vary $m$ from 128 to 2048, and then observe the change of mAP on MS-COCO and FLICKR25K. As shown in Fig. 6, the performance of LGLM will not change greatly with the increase of $m$, but it will obtain the highest value on these two datasets when setting $m = 1024$. On the one hand, this setting effectively plays an important role in dimension reduction. On the other hand, it also helps effectively complete the cross-modal embeddings fusion process. Therefore, we finally choose to set $m = 1024$ in the experiments.

**Different units of *g* in group sum pooling** In cross-modal fusion module, we first adopt group sum pooling to convert each $m$-dimensional vector into a $\frac{m}{g}$-dimensional vector, which will generate one element of the predicted labels via a $fc$ layer. In this part, we vary the values of $g$ from 1 to 128, and then observe the change of mAP on MS-COCO and FLICKR25K. As shown in Fig. 7, LGLM obtains a better performance when setting $g = 2$

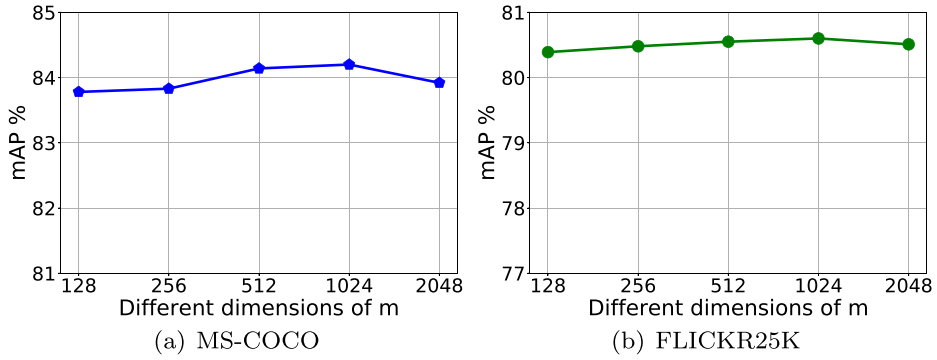


**Fig. 6** The change of mAP using different values of $m$

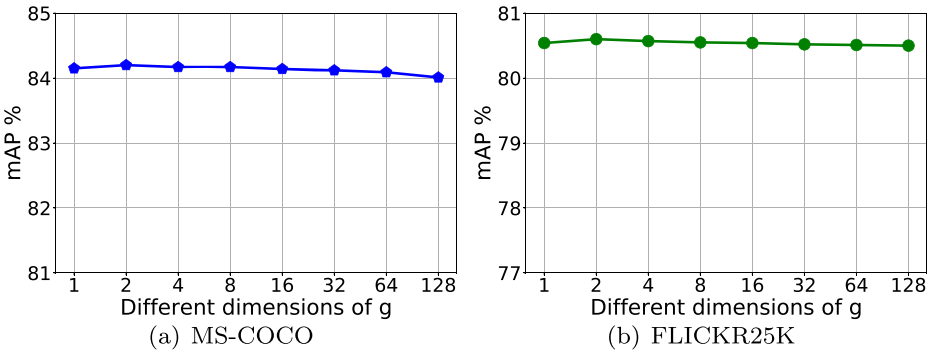


**Fig. 7** The change of mAP using different values of $g$

although the change of mAP is very slight on these two datasets. According to the experimental results, we believe $g = 2$ is a better pooling parameter that not only generates a light-weight fusion vector but also retains the original semantic information. Therefore, in the experiments, we choose $g = 2$ to complete the group sum pooling operation.

**Weighted factor λ in the loss function $L_{total}$** When training our model, we combine the multi-label loss and L1-norm loss to update the network. The weighted factor λ indicates the contribution of $L_A$ in the whole loss $L_{total}$. In this part, we vary λ from 0 to 1 and record the change of mAP on MS-COCO and FLICKR25K in Fig. 8. As we see, when we set $\lambda = 1$, the performance on all the two datasets will achieve the highest values. We believe this regularization helps optimize the model via avoiding the over-smoothing problem in GCN. Therefore, we choose $\lambda = 1$ as the default setting of LGLM.

**The effectiveness of multi-modal fusion component** In this part, we verify the effectiveness of MFC by recording the number of convergence epochs, mAP, CF1 and OF1 of LGLM with MFC, LGLM without MFC. As shown in Table 6, LGLM will converge within 25 epochs on both MS-COCO and FLICKR25K if we keep the MFC component, while it will respectively take 150 epochs to converge on these two datasets when removing MFC from LGLM. This reflects MFC greatly speeds up the model convergence. Besides, LGLM obviously achieves higher classification performance (including mAP, CF1 and

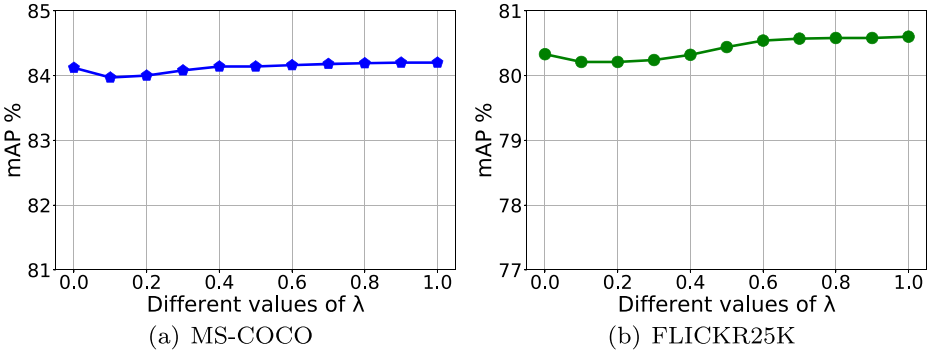


**Fig. 8** The change of mAP using different values of λ

**Table 6** Performance of LGLM with/without MFC

| Metric | MS-COCO | | FLICKR25K | |
|---|---|---|---|---|
| | LGLM with MFC | LGLM without MFC | LGLM with MFC | LGLM without MFC |
| Epoch | **25** | 150 | **25** | 150 |
| mAP | **84.2** | 83.1 | **80.6** | 80.0 |
| CF1 | **78.7** | 78.0 | **73.8** | 73.1 |
| OF1 | **81.3** | 80.3 | **79.6** | 79.0 |

The bold means the optimal value in the corresponding evaluation metric

OF1) when integrating MFC into our model to help effectively fuse image features and label co-occurrence embeddings. These results illustrate the effectiveness of our designed MFC component.

### 4.3.4 Visual results

In this section, we evaluate the multi-label image recognition performance of LGLM by presenting the visual results including the retrieval results and visualization of activation maps.

**Retrieval results** In this part, we evaluate LGLM by illustrating the visual retrieval results on FLICKR25K. We return the top-5 images by the KNN algorithm for the given query image. As shown in Fig. 9, we first randomly choose an input image that contains three objects: "sky", "clouds" and "structures". Obviously, the returned 5 images also contain these three objects. In addition, we randomly choose another image that contain two objects: "animal" and "plant_life". Similarly, the returned top-5 images also contain these two objects. These visual results illustrates that our approach owns a good classification ability to recognize multi-label images. Note that our model can be applied to fast large-scale image retrieval, which means it can be deployed on real systems to help users to retrieve similar images according to a query image. For each given query image and database images, we first transform these images into feature vectors via our model, then compute the similarity between feature vectors to find the most similar images. We help others in this community can benefit from our method.

**Visualization of activation maps** In this part, we compare the performance of our LGLM with ResNet-101 baseline by visualizing the activation maps of 5 images. As shown in


**Fig. 9** Visual retrieval results on FLICKR25K

Fig. 10, for a single-label image that only contains "animals" object, our LGLM can activate all "animals" regions, but ResNet-101 focuses more on one region. Similarly, for those images that contain two objects, LGLM can also highlight more valid regions corresponding to objects, but ResNet-101 performs not well in this aspect. At last, we input an image that contains three objects. As expected, LGLM can activate three valid regions but ResNet-101 only focuses on "people" object. These results reflect that LGLM can effectively capture the global label dependencies to activate more valid regions corresponding to multiple objects, which greatly boosts the multi-label image classification performance.

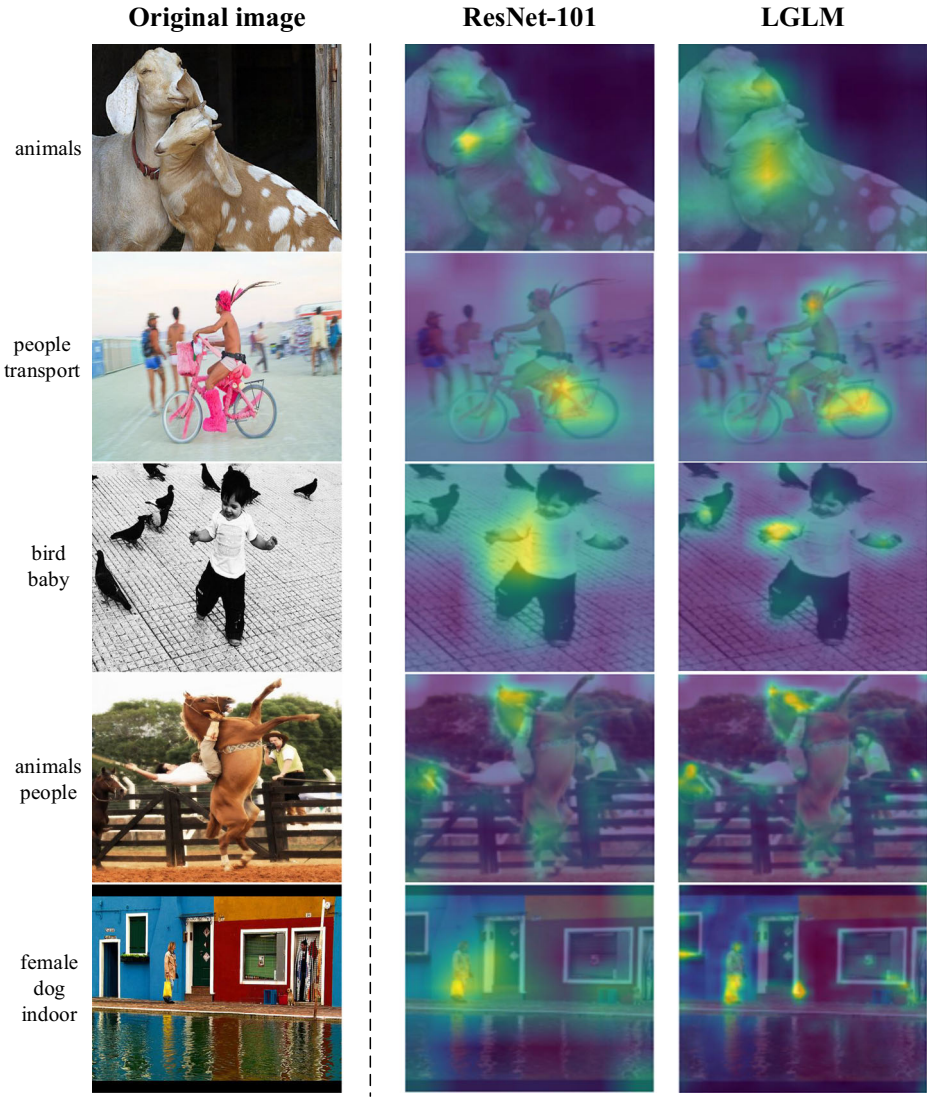


**Fig. 10** Visualization of activation maps

## 5 Conclusion and future work

In this paper, we propose a label graph learning model (LGLM) for multi-label image recognition, which integrates a MFC to efficiently complete the cross-modal embeddings fusion process. First, LGLM uses CNN to learn the feature for each image. Second, we first construct a label graph according to the word vector of each object and then adopt GCN to learn the label correlations to generate label co-occurrence embeddings. Finally, our MFC efficiently fuses image features and label co-occurrence embeddings to generate an end-to-end image recognition model. We conduct extensive experiments on MS-COCO and FLICKR25K and the experimental results demonstrate the superiority of LGLM compared with the state-of-the-art image recognition methods. Nevertheless, the MFC may lead to a complex model with too many components. Therefore, it becomes necessary to explore how to reduce the complexity of such model. Luckily, transformers, which come from the nature language processing field, have been proved to be effective in various computer vision tasks. As a result, in the next step of our study, we would like to combine transformer and GCN to overcome the low fusion efficiency to reduce the number of model components.

**Acknowledgements** Thanks for the support of the Innovation Group Project of the National Natural Science Foundation of China No.61821003 and the National Natural Science Foundation of China No.61902135.

### Declarations

**Conflict of Interests** The authors declare that they have no conflict of interest.

## References

1. Chen S-F, Chen Y-C, Yeh C-K, Wang Y-CF (2018) Order-free RNN with visual attention for multi-label classification, proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 6714–6721 AAAI Press
2. Chen T, Xu M, Hui X, Wu H, Lin L (2019) Learning semantic-specific graph representation for multi-label image recognition, 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), october 27 - november 2, 522–531. IEEE
3. Chen Z-M, Wei X-S, Wang P, Guo Y (2019) Multi-Label image recognition with graph convolutional networks, IEEE Conference on computer vision and pattern recognition, CVPR, Long beach, CA, USA, June 16-20, 5177–5186. IEEE Computer Vision Foundation
4. Defferrard M, Bresson X, Vandergheynst P (2016) Convolutional neural networks on graphs with fast localized spectral filtering, advances in neural information processing systems 29: annual conference on neural information processing systems 2016, December 5-10, Barcelona, Spain, 3837–3845
5. Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding, proceedings of the 2016 conference on empirical methods in natural language processing, EMNLP 2016, austin, texas, USA, November 1-4, 457–468. The Association for Computational Linguistics
6. Ge W, Yang S, Yizhou Y (2018) Multi-Evidence Filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning, 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 1277–1286. IEEE Computer Society
7. Ghamrawi N, McCallum A (2005) Collective multi-label classification, Proceedings of the 2005 ACM CIKM International conference on information and knowledge management, Bremen, Germany, October 31 - November 5, 195–200. ACM

8. Gong Y, Jia Y, Leung T, Toshev A, Ioffe S (2014) Deep convolutional ranking for multilabel image annotation, 2nd International conference on learning representations, ICLR 2014, Banff, AB, Canada, April 14-16. Conference Track Proceedings
9. Guo Y, Suicheng G (2011) Multi-label classification using conditional dependency networks, IJCAI 2011, Proceedings of the 22nd International joint conference on artificial intelligence, Barcelona, Catalonia, Spain, July 16-22, 1300-1305. IJCAI/AAAI
10. Guo H, Zheng K, Fan X, Hongkai Y, Wang S (2019) Visual attention consistency under image transforms for multi-label image classification, IEEE conference on computer vision and pattern recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 729–739. IEEE Computer Society
11. Huang F, Zhang X, Zhao Z, Jie X, Li Z (2019) Image-text sentiment analysis via deep multimodal attentive fusion. Knowl Based Syst 167:26–37
12. Huang F, Zhang X, Jie X, Zhao Z, Li Z (2021) Multimodal learning of social image representation by exploiting social relations. IEEE Trans Cybern 51(3):1506–1518
13. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition, 2016 IEEE conference on computer vision and pattern recognition, CVPR las vegas, NV, USA, June 27-30, 770–778. IEEE Computer Society
14. He T, Jin X (2019) Image emotion distribution learning with graph convolutional networks, Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 392–390. ACM
15. Huiskes MJ, Lew MS (2008) The MIR flickr retrieval evaluation, Proceedings of the 1st ACM SIGMM International conference on multimedia information retrieval, MIR 2008, Vancouver, British Columbia, Canada, October 30-31, 39–43. ACM
16. Inoue N, Simo-Serra E, Yamasaki T, Ishikawa H (2017) Multi-label fashion image classification with minimal human supervision, 2017 IEEE International conference on computer vision workshops, ICCV Workshops. Venice, italy, october 22-29, 2261–2267. IEEE Computer Society
17. Johnson J, Gupta A, Li F-F (2018) image generation from scene graphs, 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 1219–1228. IEEE Computer Society
18. Kim J-H, On KW, Lim W, Kim J, Ha J-W, Zhang B-T (2017) Hadamard product for low-rank bilinear pooling, 5th International Conference on Learning Representations, ICLR 2017, Toulon, france, april 24-26, conference track proceedings. OpenReview.net
19. Kipf TN, Welling M (2017) Semi-supervised classification with graph convolutional networks, 5th International conference on learning representations, ICLR 2017, Toulon, france, april 24-26, conference track proceedings. OpenReview.net
20. Lee C-W, Fang W, Yeh C-K, Wang Y-CF (2018) Multi-label zero-shot learning with structured knowledge graphs, 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 1576–1585. IEEE Computer Society
21. Li J, Huang C, Loy CC, Tang X (2016) Human attribute recognition by deep hierarchical contexts, computer vision - ECCV 2016 - 14th European Conference, Amsterdam, The netherlands, october 11-14, proceedings, Part VI 684–700. Springer
22. Li Q, Peng X, Qiao Y, Peng Q (2020) Learning label correlations for multi-label image recognition with graph networks. Pattern Recognit Lett 138:378–384
23. Lin T-Y, Maire M, Belongie SJ, Hays J, Perona P, Ramanan D, Dollàr P, Lawrence Zitnick C (2014) Microsoft coco: common objects in context, computer vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part v, 740–755. Springer
24. Malinowski M, Fritz M (2014) A multi-world approach to question answering about real-world scenes based on uncertain input, advances in neural information processing systems 27: annual conference on neural information processing systems 2014, december 8-13, montreal, quebec, Canada, 1682–1690
25. Molchanov P, Tyree S, Karras T, Aila T, Kautz J (2017) Pruning Convolutional Neural Networks for Resource Efficient Inference, 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference track proceedings. OpenReview.net
26. Monti F, Boscaini D, Masci J, Rodolà E, Svoboda J, Bronstein MM (2017) Geometric deep learning on graphs and manifolds using mixture model CNNs, 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 5425–5434. IEEE Computer Society
27. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation, Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, October 25-29, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, 1532–1543 ACL
28. Razavian AS, Azizpour H, Sullivan J, Carlsson S (2014) CNN features off-the-shelf: an astounding baseline for recognition, IEEE conference on computer vision and pattern recognition, CVPR Workshops 2014, Columbus, OH, USA, June 23-28, 512–519 IEEE Computer Society

29. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein MS, Berg AC, Li F-F (2015) ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis 115(3):211–252
30. Shao J, Kang K, Loy CC, Wang X (2015) Deeply learned attributes for crowded scene understanding, IEEE conference on computer vision and pattern recognition, CVPR Boston, MA, USA, June 7-12, 4657–4666. IEEE Computer Society
31. Wei Y, Xia W, Lin M, Huang J, Ni B, Dong J, Zhao Y, Yan S (2016) HCP: A flexible CNN framework for Multi-Label image classification. IEEE Trans Pattern Anal Mach Intell 38(9):1901–1907
32. Wang J, Yi Y, Mao J, Huang Z, Huang C, Wei X (2016) CNN-RNN: a unified framework for multi-label image classification, 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2285–2294 IEEE Computer Society
33. Wang Z, Chen T, Li G, Xu R, Lin L (2017) Multi-label image recognition by recurrently discovering attentional regions, IEEE International conference on computer vision, ICCV, Venice, italy, october 22-29, 464–474. IEEE Computer Society
34. Ye J, He J, Peng X, Wu W, Qiao Y (2020) Attention-driven dynamic graph convolutional network for multi-label image recognition, computer vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, part XXI, 649–665. Springer
35. Zhu F, Li H, Ouyang W, Nenghai Y, Wang X (2017) Learning spatial regularization with image-level supervisions for multi-label image classification, 2017 IEEE conference on computer vision and pattern recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2027–2036. IEEE Computer Society
36. Zhou B, Tian Y, Sukhbaatar S, Szlam A, Fergus R (2015) Simple baseline for visual question answering. arXiv:1512.02167
37. Zhou Y, Jun Y, Xiang C, Fan J, Tao D (2018) Beyond bilinear: generalized multimodal factorized High-Order pooling for visual question answering. IEEE Trans Neural Networks Learn Syst 29(12):5947–5959