International Journal of Machine Learning and Cybernetics (2023) 14:3427–3435
https://doi.org/10.1007/s13042-023-01841-6

ORIGINAL ARTICLE

# Multi-layered semantic representation network for multi-label image classification

Xiwen Qu[1] · Hao Che[3] · Jun Huang[1,2] · Linchuan Xu[4] · Xiao Zheng[1,2]

Received: 18 May 2021 / Accepted: 17 April 2023 / Published online: 11 May 2023

## Abstract

Multi-label image classification is a fundamental and practical task, which aims to assign multiple possible labels to an image. In recent years, many deep convolutional neural network (CNN) based approaches have been proposed which model label correlations to discover semantics of labels and learn semantic representations of images. This paper advances this research direction by improving both the modeling of label correlations and the learning of semantic representations. On the one hand, besides the local semantics of each label, we propose to further explore global semantics shared by multiple labels. On the other hand, existing approaches mainly learn the semantic representations at the last convolutional layer of a CNN. But it has been noted that the image representations of different layers of CNN capture different levels or scales of features and have different discriminative abilities. We thus propose to learn semantic representations at multiple convolutional layers. To this end, this paper designs a Multi-layered Semantic Representation Network (MSRN) which discovers both local and global semantics of labels through modeling label correlations and utilizes the label semantics to guide the semantic representations learning at multiple layers through an attention mechanism. Extensive experiments on five benchmark datasets including VOC2007, VOC2012, MS-COCO, NUS-WIDE, and Apparel show a competitive performance of the proposed MSRN against state-of-the-art models.

**Keywords** Multi-label image classification · Convolutional neural network · Label embeddings · Multi-layered attention

## 1 Introduction

Multi-label image classification (MLIC) deals with assigning multiple labels to each image, and it has been applied in many fields, including multi-object recognition [17], medical diagnosis recognition [11] and Person re-identification [26]. The recent progress is mainly made by exploiting label correlations and learning semantic representations with deep learning models.

Modeling label correlations has been long studied in multi-label classification and has been demonstrated very effective because correlated labels are highly likely to co-occur [4]. For an image recognition task, convolutional neural networks (CNNs) [10, 12, 24] and unsupervised feature extraction methods [14, 15, 39] have been widely applied to extract image features. Recently, many approaches to MLIC are proposed based on combining CNNs and exploiting label correlations, e.g., [3, 4, 23]. In these approaches, an existing deep learning model is usually employed as a tool to transform an image into a high-level abstract representation. But objects of interest may only be in certain regions of an image. Some recent studies [2, 5, 28, 35, 41] mentioned that semantic label embeddings can make the model generating more likely label combinations in prediction stage. Therefore, these works utilize the correlation between labels and generate semantics to guide the learning of semantic representations of images.

✉ Hao Che
chehao17@gmail.com

✉ Jun Huang
huangjun.cs@ahut.edu.cn

[1] School of Computer Science and Technology, Anhui University of Technology, Maanshan, China

[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China

[3] Australian National University, Canberra, Australia

[4] Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China

This paper advances this research direction by improving both the modeling of label correlations and the learning of semantic representations. On the one hand, in [38], the authors mentioned that higher-order label correlations have stronger modeling ability for multi-label classification. Therefore, unlike existing approaches only learn local semantics of each label, we propose to explore global semantics shared by multiple labels. On the other hand, existing approaches mainly learn the semantic representations at the last convolutional layer of a CNN. But it has been noted that the image representations at different layers of CNN capture different levels or scales of features and have different discriminative abilities [1, 10, 16, 32]. In addition, applying spatial attention can flexibly and adaptively aggregate semantic information and focus on the regions of interest [40]. Therefore, better performance might be achieved by simultaneously exploiting features at multiple layers of a CNN and position-wisely combining the image representations learning with label embeddings.

To realize the proposals mentioned above, we design a novel Multi-layered Semantic Representation Network (MSRN). MSRN generates both label-specific and group embeddings which capture local semantics and global semantics respectively, and then combines them with multiple layers of a CNN by an attention mechanism to learn label and group shared semantic representations of images. To be specific, first, we introduce LGE (Label-Group Embedding) module to capture both local semantics of each label and semantics of a group of labels in embeddings based on the label co-occurrence graph. Second, we propose SGA (Semantic Guided Attention) module to position-wisely guide the CNN to focus on the regions of interest. Third, we design a framework to combine the LGE module with the multiple layers of a CNN through the attention mechanism built in the SGA module. We conduct experiments on five benchmark multi-label image datasets including VOC2007, VOC2012, NUS-WIDE, MC-COCO, and Apparel. The experimental results show that our method outperforms state-of-the-art approaches.

The contributions of this paper are summarized as follows:

- A Multi-layered Semantic Representation Network (MSRN) is designed for multi-label image classification.
- Second-order and high-order label correlations are considered simultaneously to improve the performance of multi-label image classification.
- Semantic representations are learned at multiple layers through the position-wisely attention mechanism by modeling label correlations.

The rest of the paper is organized as follows. Section 2 introduces related work. Section 3 presents the proposed method. Section 4 presents empirical evaluation. Section 5 concludes this paper and introduces future work.

## 2 Related work

In MLIC, images are annotated with multiple labels simultaneously where labels usually have correlations. It has been demonstrated that exploiting label correlations can significantly improve the performance [38]. Recent progress has been made by employing deep learning models, especially convolutional neural networks. Wang et al. [27] extract label semantics and associate it to Recurrent Neural Network (RNN). In addition, Lee et al. [19] apply knowledge graphs to exploit the label dependencies based on the label co-occurrence graph. ML-GCN [4] learns the semantic label embeddings through Graph Convolution Network (GCN), and applies it as inter-dependent object classifiers at the prediction stage. In [29], a label graph superimposing framework is proposed to exploit label correlations. The label graph is constructed by superimposing statistical label graph into knowledge prior oriented graph, which, however, is usually unavailable in real applications.

Some studies further locate regions of interest because each class label might be determined by some specific regions of an image. Examples include [33, 42] which apply bounding box to focus on the regions of proposal. To learn regions with arbitrary boundaries, more studies propose attention based methods where attention is a spatial weight map representing relative importance among pixels [16]. SRN [41] is an end-to-end CNN model which trains learnable convolutions on the attention maps of labels. In [3], Chen et al. propose an order-free RNN based model for multi-label image classification, which uniquely integrates the learning of visual attention and Long Short Term Memory (LSTM) layers to jointly learn the labels of interest and their co-occurrences.

Recently, some methods [4, 5, 35] apply Graph Neural Network (GNN) techniques to generate semantic label embeddings which can be utilized as visual attention for multi-label image classification. You et al. [35] propose a method of computing cosine similarity between label embeddings to exploit label dependencies. Chen et al. [5] apply a GNN with graph propagation mechanism to exploit the interaction between DNN and label dependencies. Despite having achieved high performance on multi-label image classification, these methods do not explicitly consider high-order label dependencies which may result in semantics shared by a group of labels [38]. Moreover, to the best of our knowledge, no existing approaches utilize image representations extracted from multiple layers of a CNN.

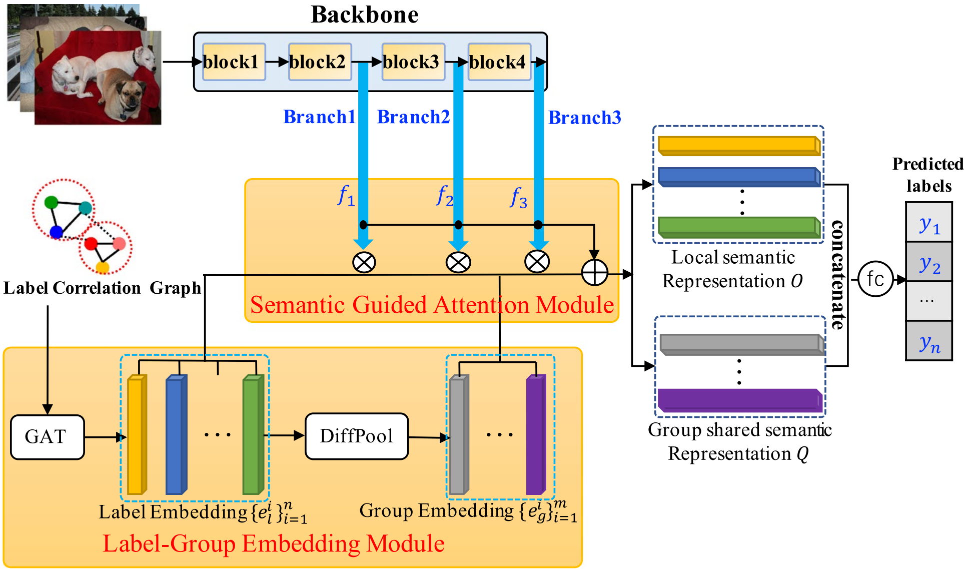
Backbone
block1
block2
block3
block4
Branch1
Branch2
Branch3
$f_1$
$f_2$
$f_3$
Semantic Guided Attention Module
Label Correlation Graph
GAT
DiffPool
Label Embedding $\{e_l^i\}_{i=1}^n$
Group Embedding $\{e_g^i\}_{i=1}^m$
Label-Group Embedding Module
Local semantic Representation $O$
Group shared semantic Representation $Q$
concatenate
fc
Predicted labels
$y_1$
$y_2$
...
$y_n$

**Fig. 1** Overall architecture of our MSRN model. Given an image, a CNN network outputs the image features from different layers to different branches. At the same time, LGE Module generates the the label and the group embeddings. Then, the SGA module produces label-level semantic representations and group-level semantic representations of the image by combining each image feature map from each branch. At last, we concatenate the generated semantic representations and perform the classification

# 3 Method

## 3.1 Architecture

The architecture of the proposed MSRN is shown in Fig. 1. We design an LGE module to generate label and group embeddings with the input of a graph $\mathcal{G} = \{V, A\}$. $V = \{v_i\}_{i=1}^n$ is the feature matrix of labels where $v_i$ is a vector of features and $n$ is the number of labels, and $A = \{a_{ij}\}_{i,j=1}^n$ is the adjacency matrix about label co-occurrence. The outputs of LGE module are label embeddings $E_l = \{e_l^i\}_{i=1}^n \in \mathbb{R}^{n \times d}$ and group embeddings $E_g = \{e_g^i\}_{i=1}^m \in \mathbb{R}^{m \times d}$, where $m$ is the number of groups, and $d$ is the dimension of the embeddings.

The backbone in our framework can be any kind of CNNs, such as VGG [24], ResNet [12] and DenseNet [10]. In this paper, Resnet-101 is chosen for experiment. Given an input image, $\overline{F} = \{\overline{f}_b\}_{i=1}^B$ indicates the output image features of different branches, where $B$ is the total number of branches, and $\overline{f}_b \in \mathbb{R}^{W_b \times H_b \times C_b}$ is image feature for the $b$-th branch with spatial resolution $W_b \times H_b$ and channel $C_b$. The branches marked as blue lines in Fig. 1 are used to receive the image feature map $\overline{f}_b$ from corresponding layers in the CNN. We provide three branches in our work to receive the output image features from the last layer of the last three blocks of Resnet-101. Since the channel $C_b$ of image features from different layers of CNN are distinct, we use a convolutional layer with 1×1 kernel to project image features from $\overline{f}_b$ to $f_b = conv^{1\times1}(\overline{f}_b) \in \mathbb{R}^{W_b \times H_b \times d}$ which has the same dimension $d$ as the label embeddings $E_l$ and group embeddings $E_g$.

Then we propose an SGA module to position-wisely combine the image feature maps $F = \{f_b\}_{i=1}^B$ with label embeddings $E_l$ and group embeddings $E_g$. The outputs of SGA module are label semantic representations $O = \{o_b\}_{b=1}^B$ and group shared semantic representations $Q = \{q_b\}_{b=1}^B$, where $o_b \in \mathbb{R}^{n \times d}$ and $q_b \in \mathbb{R}^{m \times d}$. Some of the existing approaches utilize the label-specific representation for each label. But since in real applications the labeling results of a dataset usually have noisy or missing labels, the label-specific semantic representation might not be sufficient enough to predict correct labels. Therefore, in the final stage of our framework, we concatenate the generated label and group shared semantic representations into $M = [O||Q]$, and apply fully connected layers to perform the prediction where the cross entropy loss function is adopted as follows:

$$\mathcal{L}_1 = \sum_{i=1}^{n} y^i \log(\sigma(\hat{y}^i)) + (1 - y^i)\log(1 - \sigma(\hat{y}^i)), \tag{1}$$

where $y^i$ is equal to 1 or 0 for image $i$ in terms of a certain label, $\hat{y}^i$ is the output of fully connected layer, and $\sigma(\cdot)$ is the sigmoid function.

## 3.2 Label-group embedding module

Since label correlation is important information in multi-label image classification as we mentioned in Sect. 1, we build a Label-Group Embedding (LGE) module to generate semantic label embeddings $E_l$ and group embeddings $E_g$.

### 3.2.1 Semantic label embeddings

Graph attention Networks (GAT) [25] is a self-attention based model which is most frequently used for learning embeddings of graph-structured data. With GAT algorithm, we can obtain the semantic label embeddings $E_l$ from the label graph $\mathcal{G}$. In our model, GAT first produces the attention coefficient $\alpha_{ij}$ between the $i$-th and $j$-th label as follows:

$$\alpha_{ij} = \frac{\exp(\text{LeakyReLU}(P[Uv_i||Uv_j]))}{\sum_{k\in\mathcal{N}_i}\exp(\text{LeakyReLU}(P[Uv_i||Uv_k]))}, \tag{2}$$

where $P \in \mathbb{R}^{1\times 2w}$ and $U \in \mathbb{R}^{w\times v}$ are two learnable weight matrices, $v$ and $w$ equal to the input and output of feature dimension of the GAT layer respectively, $\mathcal{N}_i$ is the set of neighborhoods of label $i$ in the graph, and $||$ represents the concatenation operation. The negative input slope in LeakyReLU is set to be 0.2 in our work. Then, we can obtain label embeddings $E_l^1 = \{e_l^i\}_{i=1}^n$ from the first GAT layer by linearly combining attention coefficients $\alpha$ with the transformed label features:

$$e_l^i = \sigma\left(\sum_{j\in\mathcal{N}_i}\alpha_{ij}Uv_j + Uv_i\right), \tag{3}$$

where the $\sigma(\cdot)$ is non-linear activation function which is ELU in our method. For simplicity, $\text{GAT}_t(\cdot)$ is used to represent the $t$-th GAT layer that consists of Eqs. (2) and (3), and the semantic label embeddings $E_l^t$ can be generated by the following equation:

$$E_l^t = \text{GAT}_t(E_l^{t-1}, A), \tag{4}$$

where $E_l^0 = V$ is the original feature matrix of labels.

#### 3.2.2 Semantic group embeddings

Differentiable graph pooling (Diffpool) [34] is a graph clustering algorithm that soft map graph nodes to a set of clusters. Once we capture the semantic label embeddings $E_l$, we can apply Diffpool to generate semantic group embeddings $E_g$ as

$$E_g = \text{Diffpool}(E_l, A). \tag{5}$$

Moreover, in order to learn more compact group embeddings, we try to minimize the distance between the group embeddings $E_g$ and the labels embeddings $E_l$ as follow

$$\mathcal{L}_2 = \sum_{k=1}^{m}\sum_{E_l^i\in C_k}||E_g^k - E_l^i||_2^2, \tag{6}$$

where $C_k$ indicates the $k$-th cluster of labels which are highly correlated labels.

### 3.3 Semantic guided attention module

The aim of SGA module is to utilize the semantic embeddings $E_l$ and $E_g$ to guide the learning of semantic representations of images at different branches. As the feature contained in each position (w, h) of an image feature map could be correlated to the semantics of the label embeddings, we propose a position-wise attention mechanism to fully combine the image feature space and the semantic embedding space. Similar to existing studies [5, 18], we adopt the Hadamard product between each position $(w, h)$ of an image feature map from the $b$-th branch and the label, group embeddings to calculate the attention weights as

$$sl^{\,i}_{b_{w,h}} = f_b^{w,h} \odot e_l^i, \quad sg^{\,j}_{b_{w,h}} = f_b^{w,h} \odot e_g^j, \tag{7}$$

where the $\odot$ is Hadamard product, $sl_{b_{w,h}} \in \mathbb{R}^{1\times 1\times n\times d}$ and $sg_{b_{w,h}} \in \mathbb{R}^{1\times 1\times m\times d}$. Then we apply normalization to the computed compatibility scores $al_b \in \mathbb{R}^{W_b\times H_b\times n\times d}$ and $ag_b \in \mathbb{R}^{W_b\times H_b\times m\times d}$

$$al_b^{w,h} = \frac{\exp(sl_{b_{w,h}})}{\sum_{x,y}\exp(sl_{b_{x,y}})}, ag_b^{w,h} = \frac{\exp(sg_{b_{w,h}})}{\sum_{x,y}\exp(sg_{b_{x,y}})}. \tag{8}$$

Once obtained the normalized compatibility scores, we apply the second Hadamard product to generate position-wise attention maps.

$$o_b = \sum_{w,h} al_b^{w,h} \odot f_b^{w,h}, \quad q_b = \sum_{w,h} ag_b^{w,h} \odot f_b^{w,h}. \tag{9}$$

Finally, the total training loss is $\mathcal{L}_1 + \lambda\mathcal{L}_2$, where $\lambda$ is a regularization parameter.

### 3.4 Model prediction

In the test stage, we concatenate the local semantic representations $O = \{o_b\}_{b=1}^B \in \mathbb{R}^{n\times(Bd)}$ and group shared semantic representations $Q = \{q_b\}_{b=1}^B \in \mathbb{R}^{m\times(Bd)}$ and predict the labels by $\hat{y}^i = fc_2(\text{LeakeyReLU}(fc_1(\tanh(M))))$, where $M = [O||Q]$, $fc_1$ and $fc_2$ are fully connected layers. It should be noted that the label co-occurrence information used in training and testing stage also the same adjacency matrix that computed based on label co-occurrence information of training data set.

## 4 Empirical evaluation

In this section, we will describe the implementation details of our proposed model MSRN and the experimental results.

### 4.1 Implementation details and evaluation metrics

The input label features $V$ are 300-dimensional Glove features pretrained on Wikipedia dataset. The backbone ResNet-101 is pretrained on ImageNet for accelerating training process. We remove the last average pooling layer and classifier of Resnet-101 and apply the MaxPooling with

**Table 1** Comparison of mAP and AP (in %) of our method and state-of-the-art methods on Pascal VOC2007 dataset where numbers in bold indicate the best performance and numbers underlined indicate the second performance

| voc2007 | CNN-RNN | ResNet-101 | AR | ML-GCN | A-GCN | F-GCN | SSGRL | SSGRL (pre) | SIGCN | MSRN | MSRN (pre) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Areo | 96.7 | 99.5 | 98.6 | 99.5 | 99.4 | 99.5 | 99.5 | 99.7 | 99.8 | **100.0** | 99.7 |
| Bike | 83.1 | 97.7 | 97.1 | 98.5 | 98.5 | 98.5 | 97.1 | 98.4 | 98.1 | 98.8 | **98.9** |
| Bird | 94.2 | 97.8 | 97.1 | 98.6 | 98.6 | 98.7 | 97.6 | 98.0 | 97.8 | **98.9** | 98.7 |
| Boat | 92.8 | 96.4 | 95.5 | 98.1 | 98.0 | 98.2 | 97.8 | 97.6 | 98.2 | **99.1** | **99.1** |
| Bottle | 61.2 | 75.7 | 75.6 | 80.8 | 80.8 | 80.9 | 82.6 | 85.7 | 82.7 | 81.6 | **86.6** |
| Bus | 82.1 | 91.8 | 92.8 | 94.6 | 94.7 | 94.8 | 94.8 | 96.2 | 95.3 | 95.5 | **97.9** |
| Car | 89.1 | 96.1 | 96.8 | 97.2 | 97.2 | 97.3 | 96.7 | 98.2 | 97.5 | 98.0 | **98.5** |
| Cat | 94.2 | 97.6 | 97.3 | 98.2 | 98.2 | 98.3 | 98.1 | 98.8 | 97.7 | 98.2 | **98.9** |
| Chair | 64.2 | 74.2 | 78.3 | 82.3 | 82.4 | 82.5 | 78.0 | 82.0 | 84.9 | 84.4 | **86.0** |
| Cow | 83.6 | 80.9 | 92.2 | 95.7 | 95.5 | 95.7 | 97.0 | 98.1 | 96.8 | 96.6 | **98.7** |
| Table | 70.0 | 85.0 | 87.6 | 86.4 | 86.4 | 86.6 | 85.6 | **89.7** | 85.8 | 87.5 | 89.1 |
| Dog | 92.4 | 98.4 | 96.9 | 98.2 | 98.2 | 98.2 | 97.8 | 98.8 | 98.1 | 98.6 | **99.0** |
| Horse | 91.7 | 96.5 | 96.5 | 98.4 | 98.4 | 98.4 | 98.3 | 98.7 | 97.9 | 98.6 | **99.1** |
| Motor | 84.2 | 95.9 | 93.6 | 96.7 | 96.7 | 96.7 | 96.4 | 97.0 | 96.6 | 97.2 | **97.3** |
| Person | 93.7 | 98.4 | 98.5 | 99.0 | 98.9 | 99.0 | 98.8 | 99.0 | 99.6 | 99.1 | **99.2** |
| Plant | 59.8 | 70.1 | 81.6 | 84.7 | 84.8 | 84.8 | 84.9 | 86.9 | 85.9 | 87.0 | **90.2** |
| Sleep | 93.2 | 88.3 | 93.1 | 96.7 | 96.6 | 96.7 | 96.5 | 98.1 | 96.5 | 97.6 | **99.2** |
| Sofa | 75.3 | 80.2 | 83.2 | 84.3 | 84.4 | 84.4 | 79.8 | 85.8 | 86.4 | 86.5 | **89.7** |
| Train | 99.7 | 98.9 | 98.5 | 98.9 | 98.9 | 99.0 | 98.4 | 99.0 | 98.6 | 99.4 | **99.8** |
| Tv | 78.6 | 89.2 | 89.3 | 93.7 | 93.7 | 93.7 | 92.8 | 93.7 | 94.4 | 94.4 | **95.3** |
| mAP | 84.0 | 89.9 | 92.0 | 94.0 | 94.0 | 94.1 | 93.4 | 95.0 | 94.4 | 94.9 | **96.0** |

kernel size 2 × 2 and stride 2 to obtain image features, $\overline{F}$, from last three building blocks of ResNet-101. The output dimension of $fc_1$ is 2048, and the output dimension of $fc_2$ is the same as the number of labels. In addition, to reduce the impact of branches corresponding to lower layers of the backbone on gradients, we add a buffer convolutional layer [1] with kernel size 1 × 1 and stride 1 before we obtain image features from the last two branches. Both the output feature dimension of first GAT layer and input feature dimension of second layer are 300 and the output feature dimension of second GAT layer is 512. The number of groups of labels $m$ is set as 4. The regularization parameter $\lambda$ is set to 0.001. The input image is resized to 448 × 448 for both training and testing. We train our model on one Tesla V100-16GB GPU and set the batch size to 8. For optimization, we apply SGD as optimizer with momentum 0.9 and weight decay $10^{-4}$. The initial learning rate is set to 0.01 and it is decreased by $*$ 0.1 on each 30 epochs in total 90 epochs.[1]

The evaluation metrics we used in our experiments include mean average precision (mAP) over all categories, precision (CP, OP), recall (CR, OR), and F1 score (CF1, OF1).

### 4.2 Experimental results

*VOC2007* [8] We compare our method with ResNet-101 [12], CNN-RNN [27], AR [2], ML-GCN [4], FGCN [30], SSGRL [5], SIGCN [6]. Following [5], we also pretrain our model on the MS-COCO dataset. The results of all the methods are shown in Table 1. We can see that the result of MSRN(pre) is 2%, 1.9%, and 1% better than ML-GCN [4], FGCN [30], SSGRL [5], SIGCN [6] on mAP respectively. It should be noted that the input image size of SSGRL(pre) is 576×576 which is larger than ours. Our model also achieves the best AP score on 17 categories. The results definitely demonstrate the effectiveness of modeling multi-layered semantic representations.

*VOC2012* [9] We also perform the experiments on VOC2012 dataset and Compare with the RMIC [13], VGG+SVM [24], FeV+LV [33], HCP+AGS [31], SSGRL [5] and SIGCN [6]. The experimental results are shown in Table 2, our method achieves the best AP score on 11 categories. Our method also outperform 0.1%, 0.2% mAP than SSGRL(576 image size) and SIGCN respectively. When applying with the COCO pretrained model, our method is also better than SSGRL(pre) 0.2% on mAP. The results show that our method achieves a competitive performance comparing with other works.

[1] Source codes and pre-trained models of our method are publicly available at https://jiunhwang.github.io/

**Table 2** Comparison of mAP and AP (in %) of our method and state-of-the-art methods on Pascal VOC2012 dataset where numbers in bold indicate the best performance and numbers underlined indicate the second performance

| voc2012 | RMIC | VGG+SVM | FeV+LV | RCP | HCP+AGS | SSGRL | SSGRL(pre) | SIGCN | MSRN | MSRN (pre) |
|---|---|---|---|---|---|---|---|---|---|---|
| Areo | 98.0 | 99.0 | 98.4 | 99.3 | **99.8** | 99.5 | 99.7 | 99.6 | 99.7 | **99.8** |
| Bike | 85.5 | 88.8 | 92.8 | 92.2 | 94.8 | 95.1 | 96.1 | 95.1 | 95.2 | **96.3** |
| Bird | 92.6 | 95.9 | 93.4 | 97.5 | 97.7 | 97.4 | 97.7 | 97.5 | 98.3 | **98.4** |
| Boat | 88.7 | 93.8 | 90.7 | 94.9 | 95.4 | 96.4 | 96.5 | 96.2 | 96.3 | **96.8** |
| Bottle | 64.0 | 73.1 | 74.9 | 82.3 | 81.3 | 85.8 | **86.9** | 84.5 | 84.8 | 85.2 |
| Bus | 86.8 | 92.1 | 93.2 | 94.1 | 96.0 | 84.5 | 95.8 | 95.6 | 96.5 | **97.5** |
| Car | 82.0 | 85.1 | 90.2 | 92.4 | 94.5 | 93.7 | 95.0 | 94.2 | 93.3 | **95.2** |
| Cat | 94.9 | 97.8 | 96.1 | 98.5 | 98.9 | 98.9 | 98.9 | 98.9 | **99.6** | **99.6** |
| Chair | 72.7 | 79.5 | 78.2 | 83.8 | **88.5** | 86.7 | 88.3 | 84.8 | 87.4 | 88.0 |
| Cow | 83.1 | 91.1 | 89.8 | 93.5 | 94.0 | 96.3 | **97.6** | 96.1 | 96.0 | 96.6 |
| Table | 73.4 | 83.3 | 80.6 | 83.1 | 86.0 | 84.6 | 87.4 | 84.3 | 86.3 | **89.8** |
| Dog | 95.2 | 97.2 | 95.7 | 98.1 | 98.1 | 98.9 | **99.1** | 98.6 | 98.9 | 99.0 |
| Horse | 91.7 | 96.3 | 96.1 | 97.3 | 98.3 | 98.6 | **99.2** | 98.5 | 98.3 | 98.8 |
| Motor | 90.8 | 94.5 | 95.3 | 86.0 | 97.3 | 96.2 | **97.3** | 96.2 | 96.9 | 96.8 |
| Person | 95.5 | 96.9 | 97.5 | 98.8 | 97.3 | 98.7 | **99.0** | 98.7 | 98.8 | **99.0** |
| Plant | 58.3 | 63.1 | 73.1 | 77.7 | 76.1 | 82.2 | **84.8** | 83.2 | 80.6 | 84.5 |
| Sleep | 87.6 | 93.4 | 91.2 | 95.1 | 93.9 | 98.2 | **98.3** | 97.7 | 97.7 | 97.1 |
| Sofa | 70.6 | 75.0 | 75.4 | 79.4 | 84.2 | 84.2 | **85.8** | 82.9 | 80.5 | **85.8** |
| Train | 93.8 | 97.1 | 97.0 | 97.7 | 98.2 | 98.1 | 99.2 | 98.5 | **99.4** | 99.3 |
| Tv | 83.0 | 87.1 | 88.2 | 92.4 | 92.7 | 93.5 | 94.1 | 93.3 | 94.7 | **95.9** |
| mAP | 84.4 | 89.0 | 89.4 | 92.2 | 93.2 | 93.9 | 94.8 | 93.7 | 94.0 | **95.0** |

*MS-COCO* [21] The comparison results on MS-COCO dataset are shown in Table 3. We compare our method with ResNet-101 [12], ML-GCN [4], FGCN [30], CMA [35] and SIGCN [6]. Our method can achieve 83.4% mAP score which is in the first rank. Our model achieves comparable performance with the state-of-the-art methods. Although SSGRL and MS-CMA is 0.4% better than our method with 448 input image size in most of metrics, these two methods have a more advantageous experimental setup, such as larger image size and multi-scale training. Therefore, we also test our method by setting inputting image as 576 size. Specifically, MSRN wins the first place in terms of mAP, CR, CF1, OR, and OF1.For the results on top-3 labels, MSRN obtains the best performance in terms of OR and OF1.

*NUS-WIDE* [7] contains 269,648 images and 81 concepts. The dataset is split by following [35]. We compare MSRN with CNN-RNN [27], ML-GCN [4], GATN [36], AT [37], S-CLs [22] and CMA [35]. As shown in Table 4, our model achieves 0.1% better than MS-CMA on mAP. In addition, MSRN achieves the best performance in terms of CF1, OF1 and CF1-3. Both our model and MS-CMA [35] extract image features from lower layers of CNNs, but our model outperforms it by 0.1%, 0.2%, 0.2% and 0.4% in terms of mAP, CF1, OF1 and CF1-3, respectively.

*Apparel*[2] is a clothing dataset for multi-label image classification. We test our model on it and make comparisons with ResNet-101 [12], SSGRL [5] and ML-GCN [4]. In our experiment, we randomly select 50% images from the dataset for training, and other 50% images for testing. The result in Table 5 shows that our model achieves 99.65 mAP score. Our model is 0.18% better than ResNet-101 and 0.06% better than the current best model on mAP. Our model also achieves the best score on all metrics that we employ.

## 4.3 Ablation studies

In this section, we perform ablation studies to evaluate the effectiveness of different components of our framework.

*Label and group embeddings* To verify the effectiveness of label and group embeddings, we conduct experiments with three simplified versions of our proposed method MSRN, i.e., label-E (only using label embedding), group-E (only using group embedding), and no LGE module. The results shown in Table 6 clearly indicate the effectiveness of label and group embeddings.

[2] http://www.kaggle.com/kaiska/apparel-dataset.

**Table 3** Comparison of our method with state-of-the-art methods on MS-COCO dataset where numbers in bold indicate the best performance and numbers underlined indicate the second performance

| Methods | mAP | CP | CR | CF1 | OP | OR | OF1 |
|---|---|---|---|---|---|---|---|
| CNN-RNN [27] | – | – | – | – | – | – | – |
| ResNet-101 [12] | 80.3 | 77.8 | 72.8 | 75.2 | 81.5 | 75.1 | 78.2 |
| ML-GCN [4] | 83.0 | 84.0 | 72.8 | 78.0 | 84.7 | 76.2 | 80.2 |
| A-GCN [20] | 83.1 | 84.7 | 72.3 | 78.0 | 85.6 | 75.5 | 80.3 |
| F-GCN [30] | 83.2 | 85.4 | 72.4 | 78.3 | 86.0 | 75.7 | 80.5 |
| SSGRL(576) [5] | 83.8 | **89.9** | 68.5 | 76.8 | **91.3** | 70.8 | 79.7 |
| CMA [35] | 83.4 | 83.4 | 72.9 | 77.8 | 86.8 | 76.3 | 80.9 |
| MS-CMA [35] | 83.8 | 82.9 | 74.4 | 78.4 | 84.4 | 77.9 | 81.0 |
| SIGCN [6] | 83.5 | 86.8 | 71.8 | 78.6 | 88.1 | 74.5 | 80.7 |
| MSRN | 83.4 | 86.5 | 71.5 | 78.3 | 86.1 | 75.5 | 80.4 |
| MSRN (576) | **84.7** | 84.7 | **74.9** | **79.5** | 85.2 | **79.5** | **81.8** |
| Methods | CP-3 | CR-3 | CF1-3 | OP-3 | OR-3 | OF1-3 | |
| CNN-RNN [27] | 59.3 | 52.5 | 55.7 | 59.8 | 61.4 | 60.7 | |
| ResNet-101 [12] | 84.1 | 59.4 | 69.7 | 89.1 | 62.8 | 73.6 | |
| ML-GCN [4] | 89.2 | 64.1 | 74.6 | 90.5 | 66.5 | 76.7 | |
| A-GCN [20] | 89.0 | 64.2 | 74.6 | 90.5 | 66.3 | 76.6 | |
| F-GCN [30] | 89.3 | 64.3 | 74.7 | 90.5 | 66.6 | 76.7 | |
| SSGRL (576) [5] | **91.9** | 62.5 | 72.7 | **93.8** | 64.1 | 76.2 | |
| CMA [35] | 86.7 | 64.9 | 74.3 | 90.9 | 67.2 | 77.2 | |
| MS-CMA [35] | 88.2 | **65.0** | 74.9 | 90.2 | 67.4 | 77.1 | |
| SIGCN [6] | 90.2 | 64.4 | **75.1** | 92.5 | 66.1 | 77.1 | |
| MSRN | 88.3 | 63.7 | 74.0 | 90.2 | 66.8 | 76.8 | |
| MSRN (576) | 88.8 | 64.5 | 74.7 | 91.0 | **67.5** | **77.6** | |

**Table 4** Comparion with state-of-the-art methods on NUS-WIDE dataset where numbers in bold indicate the best performance and numbers underlined indicate the second performance

| Methods | mAP | CF1 | OF1 | CF1-3 | OF1-3 |
|---|---|---|---|---|---|
| ML-GCN [4] | 54.9 | 51.6 | 68.1 | – | – |
| GATN [36] | 59.8 | 56.9 | 70.7 | – | – |
| CNN-RNN [27] | 56.1 | – | – | 34.7 | 55.2 |
| AT [37] | 57.6 | 55.2 | 70.3 | 51.7 | 68.8 |
| S-CLs [22] | 60.1 | 58.7 | 73.7 | 53.8 | **71.1** |
| CMA [35] | 60.8 | 60.4 | 73.3 | 55.5 | 70.0 |
| MS-CMA [35] | 61.4 | 60.5 | 73.8 | 55.7 | 69.5 |
| MSRN | **61.5** | **60.7** | **74.0** | **56.1** | 69.5 |

**Table 5** Comparison with state-of-the-art methods on Apparel dataset where numbers in bold indicate the best performance and numbers underlined indicate the second performance

| Methods | mAP | CF1 | OF1 | CF1-3 | OF1-3 |
|---|---|---|---|---|---|
| ResNet-101 [12] | 99.47 | 97.51 | 97.78 | 97.51 | 97.78 |
| SSGRL [5] | 99.57 | 97.77 | 98.01 | 97.77 | 98.01 |
| ML-GCN [4] | 99.56 | 97.68 | 97.88 | 97.68 | 97.87 |
| MSRN | **99.65** | **98.21** | **98.36** | **98.21** | **98.36** |

**Table 6** Comparison among different versions of MSRN

| Setting | No LGE | Label-E | Group-E | MSRN |
|---|---|---|---|---|
| mAP | 91.75 | 94.42 | 94.20 | **94.85** |

Numbers in bold face indicate the best performance

**Table 7** Comparison among different number of branches

| Number of branches | Last 1 | Last 2 | Last 3 | All 4 |
|---|---|---|---|---|
| mAP | 94.53 | 94.66 | **94.85** | 94.65 |

Numbers in bold face indicate the best performance and numbers underlined indicate the second performance

*Number of branches* As ResNet-101 contains four blocks, we conduct experiments to validate whether the multi-branch architecture is better than the single-branch architecture and whether the model performs better with all branches. The experimental results are shown in Table 7. We can find that the multi-branch architecture can improve at least 0.12% compared to the single-branch architecture, and achieves the best performance with the last 3 branches.

**Table 8** Comparison among different values of $m$ and $\lambda$

| $m$ | 2 | 4 | 6 | 8 | 10 | 20 |
|---|---|---|---|---|---|---|
| mAP | 94.59 | **94.85** | 94.66 | 94.56 | 94.63 | 94.48 |
| $\lambda$ | $10^{-1}$ | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | $10^{-6}$ |
| mAP | 94.43 | 94.74 | **94.85** | 94.69 | 94.66 | 94.55 |

Numbers underlined indicate the second performance

### 4.4 Parameter sensitivity

In this section, we study the sensitivity of MSRN to two hyper-parameters, i.e., the number of groups of labels $m$ and the regularization parameter $\lambda$. Due to the limitation of space, we only present the analyses on VOC2007 dataset. For the number of groups $m$, we conduct experiments of six different cases corresponding to 2, 4, 6, 8, 10 and 20, respectively, with $\lambda$ fixed as $10^{-3}$. The experimental results in Table 8 show that the performance in terms of mAP is not much sensitive to $m$. For $\lambda$, we study the values from $\{10^{-1}, 10^{-2}, \ldots, 10^{-6}\}$. The results with different values of $\lambda$ are shown in Table 8, which shows the performance is not much sensitive to $\lambda$.

## 5 Conclusion and future work

This paper proposes a novel Multi-layered Semantic Representation Network (MSRN) for multi-label image classification. MSRN for the first time considers both local semantics and global semantics of labels through modeling label correlations, and learns semantic representations of images at multiple layers of a convolutional neural network through an attention mechanism. Extensive experiments show that MSRN outperforms many state-of-art methods on VOC2007, VOC2012, MS-COCO, NUS-WIDE and Apparel datasets. In the future, we will improve our method to explicitly utilize labels which exist but are unobservable due to lack of labeling efforts.

**Acknowledgements** This work is supported by NSFC: 61806005 and 61906003, the University Synergy Innovation Program of Anhui Province: GXXT-2019-025, GXXT-2020-012, GXXT-2022-052, the Outstanding Young Talents Support project of Anhui Province: gxyqZD2022032, and the Natural Science Foundation of the Educational Commission of Anhui Province of China: KJ2021A0373.

## References

1. Cai Z, Fan Q, Feris RS, Vasconcelos N (2016) A unified multi-scale deep convolutional neural network for fast object detection. In: ECCV, pp 354–370
2. Chen T, Wang Z, Li G, Lin L (2017) Recurrent attentional reinforcement learning for multi-label image recognition. In: AAAI
3. Chen S, Chen Y, Yeh C, Wang YF (2018) Order-free rnn with visual attention for multi-label classification. In: AAAI
4. Chen Z, Wei X, Wang P, Guo Y (2019) Multi-label image recognition with graph convolutional networks. In: CVPR, pp 5172–5181. https://doi.org/10.1109/CVPR.2019.00532
5. Chen T, Xu M, Hui X, Wu H, Lin L (2019) Learning semantic-specific graph representation for multi-label image recognition. In: ICCV, pp 522–531
6. Chen B, Zhang ZLY, Chen F, Lu G, Zhang D (2021) Semantic-interactive graph convolutional network for multilabel image recognition. IEEE Trans Syst Man Cybernet Syst. https://doi.org/10.1109/TSMC.2021.3103842
7. Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) Nus-wide: a real-world web image database from national university of Singapore. In: ICIVR, ACM
8. Everingham M, Gool L, Williams CK, Winn J, Zisserman A (2010) The pascal visual object classes (voc) challenge. IJCV 88(2):303–338
9. Everingham M, Gool LV, Williams CK, Winn J, Zisserman A (2012) The PASCAL visual object classes challenge (VOC2012) results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
10. Gao H, Z, Liu M, Laurens W, Kilian Q (2017) Densely connected convolutional networks. In: CVPR, pp 4700–4708
11. Ge ZY, Mahapatra D, Sedai S, Garnavi R, Chakravorty R (2018) Chest x-rays classification: a multi-label and fine-grained problem. arXiv
12. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
13. He S, Xu C, Guo T, Xu C, Tao D (2018) Reinforced multi-label image classification by exploring curriculum. In: Proceedings of the AAAI conference on artificial intelligence, vol 32(1). https://ojs.aaai.org/index.php/AAAI/article/view/11770
14. Hou C, Zhou Z (2018) One-pass learning with incremental and decremental features. IEEE Trans Pattern Anal Mach Intell 40(11):2776–2792. https://doi.org/10.1109/TPAMI.2017.2769047
15. Hou C, Zeng L, Hu D (2019) Safe classification with augmented features. IEEE Trans Pattern Anal Mach Intell 41(9):2176–2192. https://doi.org/10.1109/TPAMI.2018.2849378
16. Jetley S, Lord NA, Lee N, Torr PHS (2018) Learn to pay attention. In: ICLR
17. Kang K, Ouyang W, Li H, Wang X (2016) Object detection from video tubelets with convolutional neural networks. In: CVPR, pp 817–825. https://doi.org/10.1109/CVPR.2016.95
18. Kim J, On K, Kim J, Ha J, Zhang B (2016) Hadamard product for low-rank bilinear pooling (10)
19. Lee C, Fang W, Yeh C, Wang FY (2018) Multi-label zero-shot learning with structured knowledge graphs. In: CVPR, pp 1576–1585. https://doi.org/10.1109/CVPR.2018.00170
20. Li Q, Peng X, Qiao Y, Peng Q (2020) Learning label correlations for multi-label image recognition with graph networks. PRL 138:378–384

21. Lin T, Maire M, Belongie S, Bourdev L, Girshick R, Hays J, Perona P, Ramanan D, Zitnick CL, Dollr P (2014) Microsoft coco: common objects in context
22. Liu Y, Sheng L, Shao J, Yan J, Xiang S, Pan C (2018) Multi-label image classification via knowledge distillation from weakly supervised detection. In: ACM MM, pp 700–708
23. Ma C, Chen Z, Lu J, Zhou J (2018) Rank-consistency multi-label deep hashing. In: ICME, pp 1–6
24. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: ICLR
25. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y (2018) Graph Attention Networks. In: ICLR
26. Wang D, Zhang S (2020) Unsupervised person re-identification via multi-label classification. In: CVPR, pp 10978–10987
27. Wang J, Yang Y, Mao J, Huang Z, Huang C, Xu W (2016) CNN-RNN: a unified framework for multi-label image classification. In: CVPR, pp 2285–2294. https://doi.org/10.1109/CVPR.2016.251
28. Wang Z, Chen T, Li G, Xu R, Lin L (2017) Multi-label image recognition by recurrently discovering attentional regions. In: ICCV, pp 464–472
29. Wang Y, He D, Li F, Long X, Zhou Z, Ma J, Wen S (2020) Multi-label classification with label graph superimposing. In: AAAI, pp 12265–12272
30. Wang Y, Xie Y, Liu Y, Zhou K, Li X (2020) Fast graph convolution network based multi-label image recognition via cross-modal fusion. In: CIKM, pp 1575–1584
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. https://doi.org/10.1109/TPAMI.2015.2491929
32. Yang S, Ramanan D (2015) Multi-scale recognition with dag-cnns. In: ICCV, pp 1215–1223
33. Yang H, Zhou T, Zhang Y, Gao B, Wu J, Cai J (2016) Exploit bounding box annotations for multi-label object recognition. In: CVPR, pp 280–288
34. Yin R, You J, Morris C, Ren X, Hamilton WL, Leskovec J (2018) Hierarchical graph representation learning with differentiable pooling. In: NIPS, pp 4805–4815
35. You R, Guo Z, Cui L, Long X, Bao Y, Wen S (2020) Cross-modality attention with semantic graph embedding for multi-label classification. AAAI 34:12709–12716. https://doi.org/10.1609/aaai.v34i07.6964
36. Yuan J, Chen S, Zhang Y, Shi Z, Geng X, Fan J, Rui Y (2022) Graph attention transformer network for multi-label image classification
37. Zagoruyko S, Komodakis N (2017) Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: ICLR
38. Zhang ML, Zhou ZH (2014) A review on multi-label learning algorithms. IEEE Trans Knowl Data Eng 26(8):1819–1837
39. Zhang Z, Xu Y, Shao L, Yang J (2018) Discriminative block-diagonal representation learning for image recognition. IEEE Trans Neural Netw Learn Syst 29(7):3111–3125. https://doi.org/10.1109/TNNLS.2017.2712801
40. Zhao H, Zhang Y, Liu S, Shi J, Loy C, Lin D, Jia J (2018) Psanet: point-wise spatial attention network for scene parsing. In: Computer vision—ECCV 2018. . Springer, Cham, pp 270–286
41. Zhu F, Li H, Ouyang W, Yu N, Wang X (2017) Learning spatial regularization with image-level supervisions for multi-label image classification. In: CVPR, pp 2027–2036. https://doi.org/10.1109/CVPR.2017.219
42. Zitnick CL, Dollár P (2014) Edge boxes: locating object proposals from edges. In: ECCV, pp 391–405

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