Scene-Aware Label Graph Learning for Multi-Label Image Classification

Xuelin Zhu[1] Jian Liu[3] Weijia Liu[1] Jiawei Ge[1] Bo Liu[2,4] Jiuxin Cao[1,4*]
[1]School of Cyber Science and Engineering, Southeast University
[2]School of Computer Science and Engineering, Southeast University
[3]Ant Group, Hangzhou, China [4]Purple Mountain Laboratories, Nanjing, China
{zhuxuelin,weijia-liu,jiawei_ge,bliu,jx.cao}@seu.edu.cn, rex.lj@antgroup.com

Abstract

Multi-label image classification refers to assigning a set of labels for an image. One of the main challenges of this task is how to effectively capture the correlation among labels. Existing studies on this issue mostly rely on the statistical label co-occurrence or semantic similarity of labels. However, an important fact is ignored that the co-occurrence of labels is closely related with image scenes (indoor, outdoor, etc.), which is a vital characteristic in multi-label image classification. In this paper, a novel scene-aware label graph learning framework is proposed, which is capable of learning visual representations for labels while fully perceiving their co-occurrence relationships under variable scenes. Specifically, our framework is able to detect scene categories of images without relying on manual annotations, and keeps track of the co-occurring labels by maintaining a global co-occurrence matrix for each scene category throughout the whole training phase. These scene-independent co-occurrence matrices are further employed to guide the interactions among label representations in a graph propagation manner towards accurate label prediction. Extensive experiments on public benchmarks demonstrate the superiority of our framework.

1. Introduction

Multi-label image classification is a fundamental task in computer vision, which requires to recognize multiple labels for an image. A main issue for this task is how to fully mine the correlation among these labels. Implicit methods [23, 5] resort to sequential models or graph models to exploit the latent co-occurrence relationship among labels. Instead, explicit methods directly model a co-occurrence probability matrix among labels from data, such as dataset-level label co-occurrence [3] and instance-level label co-occurrence [31]. However, the former is obtained statistically on the entire data, which is coarse and may lead to wrong interactions between labels. As for the latter, it is difficult to learn an accurate co-occurrence matrix for each image, which limits its utility in guiding label interactions.

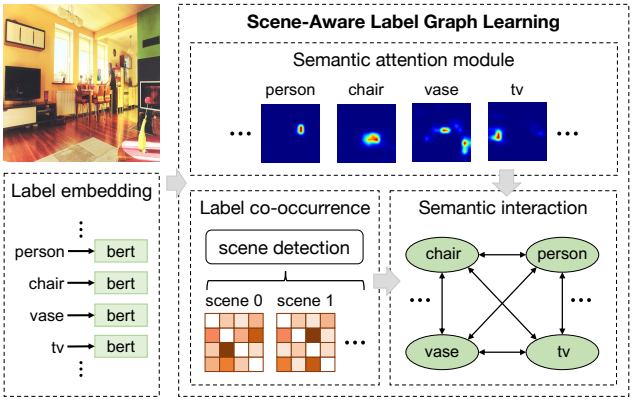

Figure 1. A brief illustration of the proposed scene-aware label graph learning framework. The semantic attention module maps the label embeddings into visual representations, and the label co-occurrence module detects the scene category of the input image and updates the co-occurrence frequency matrix accordingly. The semantic interaction module constructs a label graph for the interaction of label representations under the guidance of the scene-aware co-occurrence matrix for final prediction.

In this work, we consider an important fact that the label co-occurrence heavily relies on the scene categories of images. For example, we expect *chairs* or *tables* co-occurring with *persons* in an *indoor* scene, while *cars* or *buildings* instead in an *outdoor* scene. Therefore, a natural idea is to divide the training images into several independent groups according to their scene categories and count the label co-occurrence matrix for each group separately. Then the images in each group share the same co-occurrence matrix for subsequent feature interactions. Obviously, the obtained group-level label co-occurrence can provide more accurate guidance for feature interaction between labels than the dataset-level one [3] and the instance-

*Corresponding author.

level one [31], However, the scene category of images is unknown and human-expensive to annotate. It is crucial to accurately partition training data into groups with the same scene category without relying on annotations.

To address the aforementioned issue, we propose an effective scene-aware label co-occurrence module that maintains a label co-occurrence frequency matrix for each scene category. Each matrix element represents the number of occurrences of the label pair for the corresponding row and column. They are firstly initialized with zero at the beginning of training, and then continuously counts the co-occurring labels of images by detecting their scene categories throughout the whole training phase. However, due to the lack of scene annotations, we observe the collapse of the scene detection component and the predicted scene distribution is always dominated by a specific scene category in practice, which is known as the winner-take-all phenomenon [21]. To overcome this, we propose an entropy-based loss, which encourages a sharp distribution of scene categories for a single sample but a smoother average distribution over a batch of samples. With this supervision, the scene detection component can not only clearly identify the scene category for an input image, but also provide a diverse prediction on scene categories for the whole dataset.

Figure 1 illustrates a basic pipeline of the Scene-Aware Label Graph Learning (SALGL) framework. Concretely, for an input image and a candidate label set, label embeddings and visual feature maps are first extracted via a language model and a vision backbone, respectively. They are subsequently fed into the semantic attention module to build alignments with each other, producing semantically related visual representations for each label. Meanwhile, the global pooling feature of the input image is input into the scene-aware label co-occurrence module to detect its scene category and update the label co-occurrence matrix accordingly. Then, in the semantic interaction module, a label graph is constructed with labels as nodes and the co-occurrent relationships as edges. The visual representations of labels are fed into the graph to explore their interactions under the guidance of the scene-aware label co-occurrence. Finally, we train a separate classifier for each label with its visual representation to determine whether the current label exists in the image. Overall, the main contributions of this paper are summarized as follows:

- We are the first to explore the correlation between label co-occurrence and scene categories, and propose an effective approach to dynamically model the label co-occurrence for adapting the variable image scenes.
- We propose an advanced entropy-based auxiliary loss that enables unsupervised learning of scene information of images from dataset and prevents the scene detection component from collapsing.
- We frame an end-to-end label graph learning framework capable of perceiving image scenes and enriching label representations, which achieves state-of-the-art performance on public benchmarks.

2. Related Work

Multi-label image classification task has attracted increasing interest for many years. Early works [26, 29] resort to object detection to generate a set of region proposals for label prediction. Subsequent region-based works focus on spatial dependency among object regions. Wang et al. [25] localized object regions through a spatial transformer layer and utilized a LSTM (Long Short-Term Memory) [13] unit to capture the dependencies among these regions, which are finally employed to predict label confidences sequentially. Chen et al. [2] proposed a recurrent attention based reinforcement learning framework to iteratively discover a sequence of informative regions and explicitly modeled long-term dependency among them for label recognition. Chen et al. [3] explored the semantic interactions between the visual representations of labels under the guidance of statistical label co-occurrence. Wu et al. [27] resorted to the graph matching method to simultaneously explore instance spatial correlation, label semantic dependency and instance-label matching possibility. However, these region-based approaches suffer from the time penalty of object detection, and only provide limited performance in label prediction.

Recently, many studies have been explored to capture the label correlation. Sequence-based methods [23, 30, 1] explored the semantic correlation among label vector representations by RNN. Graph networks also have been used to capture label dependency. Chen *et al.* [5] constructed a directed label graph and utilized GCN to map the graph into a set of independent label classifiers. Wang *et al.* [24] and Chen *et al.* [3] built a label graph with the statistical co-occurrence information from data for label representation learning. With the rise of vision transformer [11], transformer-based architectures have been the primary considerations for multi-label classification. Lanchantin *et al.* [15] proposed a transformer encoder based framework to capture the complex dependencies among visual features and labels. Liu *et al.* [18] proposed to query the existence of a label on the visual features with label embeddings by the transformer decoder. Zhu *et al.* [34] proposed a two-stream transformer network for exploring label correlations and cross-modal textual-visual interactions jointly. TSFormer [34] proposed a two-stream transformer framework to explore the label dependencies and cross-modal correlations as well as spatial correlations simultaneously for robust label prediction. Despite their success, these methods lack fine-grained mining of label associations, ignoring the key factor affecting label co-occurrence, namely scene.

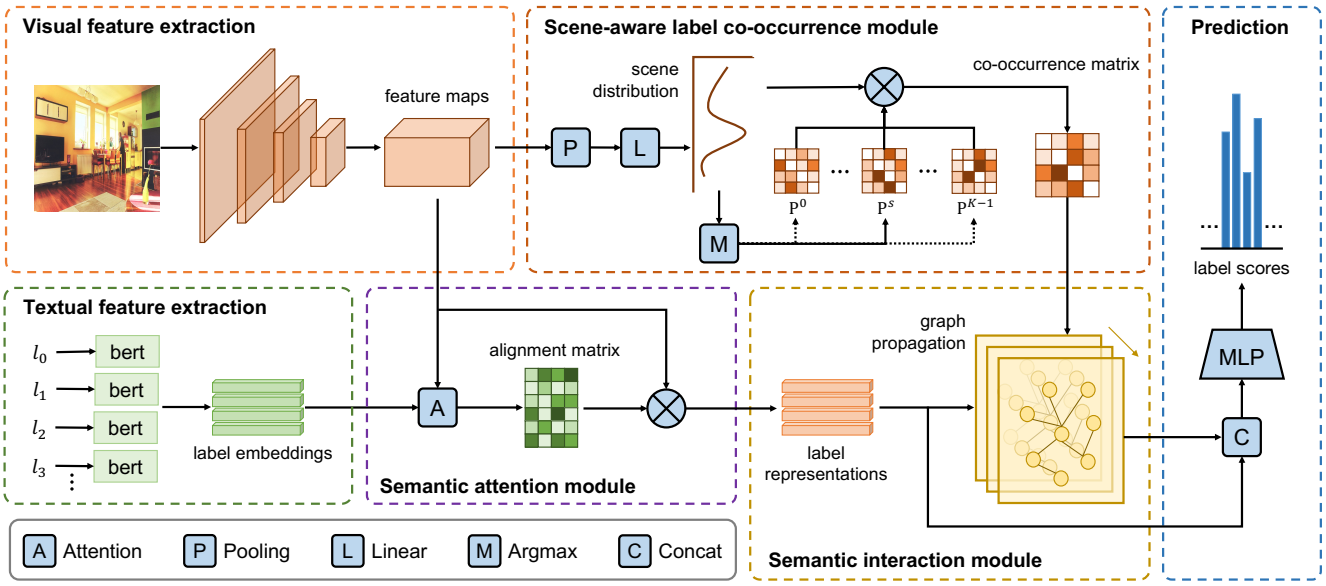

Figure 2. The detailed illustration of the proposed scene-aware label graph learning framework. To build the label graph, the semantic attention module generates visual representations for labels to initialize nodes' features, and the scene-aware label co-occurrence module counts the co-occurrence matrix for each scene category and assigns the weights for edges. The semantic interaction module propagates messages in the graph and produces interacted representations of labels for prediction.

3. Scene-Aware Label Graph Learning

3.1. Overview

Our SALGL framework takes an image and a candidate label set as inputs, and maps them into semantic embeddings and visual feature maps by a language model and a vision backbone, respectively. Next, a semantic attention module is introduced to build alignments between two modalities, producing semantically related visual representations for each label. Meanwhile, the scene-aware label co-occurrence module takes feature maps as input to detect scene category of the input image, and updates corresponding label co-occurrence matrix with its co-occurring labels. Then, in the semantic interaction module, a label graph is constructed to explore the interactions between label representations under the guidance of the scene-aware label co-occurrence. Finally, we train a separate classifier for each label with its visual representation to determine whether the current label exists in the input image. Figure 2 illustrates the detailed pipeline of the proposed SALGL framework.

3.2. Semantic Attention Module

Given a candidate label set $L = \{l_0, l_1, \cdots, l_{n-1}\}$ with n being the number of labels, we use a language model to obtain label embeddings, denoting as $\{\mathbf{t}_i\}_{i=0}^{n-1}$, where $\mathbf{t}_i \in \mathbb{R}^{d_L}$ and d_L is the dimension of label embeddings. Note that label embeddings are fixed and do not require end-to-end training with the whole framework. For an input image I, a vision backbone is used to extract its feature maps, denoting as $\{\mathbf{v}_j\}_{j=0}^{m-1}$, where $\mathbf{v}_j \in \mathbb{R}^{d_I}$ is the feature vector at the j-th spatial region of the input image I and m is the number of spatial regions. Then, we resort to the low-rank bilinear pooling model [14] to build alignment matrix between visual feature maps and label embeddings. Concretely, it first maps the regional feature $\mathbf{v}_j$ and label embedding $\mathbf{t}_i$ into a joint embedding space:

$$\mathbf{x}_{ij} = \mathbf{P}^T \left(\tanh\left(\left(\mathbf{U}^T \mathbf{v}_j \right) \odot \left(\mathbf{V}^T \mathbf{t}_i \right) \right) \right) + \mathbf{b}, \quad (1)$$

where $\tanh(\cdot)$ is the hyperbolic tangent function, $\mathbf{U} \in \mathbb{R}^{d_I \times d_1}$, $\mathbf{V} \in \mathbb{R}^{d_L \times d_1}$, $\mathbf{P} \in \mathbb{R}^{d_1 \times d_2}$ and $\mathbf{b} \in \mathbb{R}^{d_2}$ are all learnable parameters. $\odot$ is the element-wise multiplication operation. d_1 and d_2 are the dimensions of the joint embedding space and the output features, respectively. Then a normalized attention score is calculated as:

$$\alpha_{ij} = \frac{\exp(\Phi_a(\mathbf{x}_{ij}))}{\sum_{j'=0}^{m-1} \exp(\Phi_a(\mathbf{x}_{ij'}))}, \quad (2)$$

where $\Phi_a(\cdot)$ is a learnable feed-forward network that maps input vector to a logit. As a result, the visual representation of label l_i is acquired by the weighted sum of all regional features of the input image I, formulated as:

$$\mathbf{f}_i = \sum_{j=0}^{m-1} \alpha_{ij} \mathbf{v}_j. \quad (3)$$

3.3. Scene-Aware Label Co-occurrence Module

In this section, we introduce an effective approach to model the label co-occurrence relations under the variable scenes without relying on manual annotations.

Scene category detection. Assuming a total of K scene categories in the training data, we detect the scene category of the input image I based on its contextual feature, which is obtained by a global average pooling operation along the spatial dimension, formulated as:

$$\overline{\mathbf{v}} = \sum_{j=0}^{m-1} \mathbf{v}_j / m. \tag{4}$$

Then, the probability that the input image I belongs to the k-th scene category is calculated as follows:

$$\pi_k = \frac{\exp(\mathbf{w}_k^T \overline{\mathbf{v}})}{\sum_{k'=0}^{K-1} \exp(\mathbf{w}_{k'}^T \overline{\mathbf{v}})}, \tag{5}$$

where $\mathbf{w}_k$ for $k \in \{0, 1, ..., K-1\}$ is a parameter vector to be learned and $\mathbf{w}_k \in \mathbb{R}^{d_I}$. It acts as a prototype for the k-th scene category and clusters related images together. With the obtained probability distribution of image scenes, the category with the highest probability is determined as the scene category that the image I belongs to, formulated as:

$$s = \underset{k \in \{0,1,\cdots,K-1\}}{\operatorname{argmax}} \pi_k. \tag{6}$$

In this way, the s-th scene category is assigned to the input image I, whose labels are subsequently used to update the s-th co-occurrence matrix accordingly.

Label co-occurrence modeling. In this work, we aim to mine the label co-occurrence relations under variable scenes. To this end, a label co-occurrence frequency matrix is maintained for each scene category, which tracks co-occurring labels of the input image according to its scene category during the training phase. Specifically, the labels of image I are paired with each other, which are considered as the co-occurrent labels of the s-th scene category. For convenience, they are denoted as a multi-hot vector $\mathbf{y} = [y_0, y_1, \cdots, y_{n-1}]^T$, where $y_i \in \{0, 1\}$ for $i \in \{0, 1, \cdots, n-1\}$ is a binary indicator. $y_i = 1$ if the label l_i presents in the image I and 0 otherwise. Then the label co-occurrence frequency matrix is updated:

$$\mathbf{C}^s = \mathbf{C}^s + \mathbf{y}\mathbf{y}^T, \tag{7}$$

where $\mathbf{C}^s \in \mathbb{R}^{n \times n}$ is a globally maintained frequency matrix of the s-th scene category throughout the whole training phase and initialized with zero. Note that its diagonal element c_{ii}^s counts the number of occurrences of the label l_i, and off-diagonal element c_{ij}^s counts the number of co-occurrences of the label pair l_i and l_j. Hence, the probability that the label l_j appears in an image in the presence of the label l_i in the s-th scene category is computed as:

$$p_{ij}^s = \frac{c_{ij}^s}{c_{ii}^s}. \tag{8}$$

In this way, the co-occurrence probability matrix $\mathbf{P}^s$ of the s-th scene category is obtained. As the training procedure goes, each co-occurrence frequency matrix continuously counts the co-occurring labels for the corresponding scene category, and the co-occurrence probability matrix eventually converges to a steady distribution.

Auxiliary loss function. In practice, due to the lack of effective annotations of scene categories as supervision information, the scene detection component is prone to collapse and dominated by a specific scene category, which is known as the winner-take-all phenomenon [21]. To address this issue, we propose an entropy-based loss to assist the learning of the scene detection component. Concretely, for a batch of samples with $\{\pi_{b0}, \pi_{b1}, \cdots, \pi_{b(K-1)}\}$ being the scene distribution of the b-th sample, the sample-level entropy loss is denoted as:

$$\mathcal{L}_1 = -\frac{1}{B} \sum_{b=0}^{B-1} \sum_{k=0}^{K-1} \pi_{bk} \log \pi_{bk}, \tag{9}$$

where B is the batch size. Notably, the smaller the $\mathcal{L}_1$, the sharper the scene distribution. For the batch-level entropy loss, we first compute the average distribution along the batch dimension, denoting as $\{\overline{\pi}_0, \overline{\pi}_1, \cdots, \overline{\pi}_{K-1}\}$ with $\overline{\pi}_k = \frac{1}{B} \sum_{b=0}^{B-1} \pi_{bk}$, then the loss is formulated as follows:

$$\mathcal{L}_2 = \sum_{k=0}^{K-1} \overline{\pi}_k \log \overline{\pi}_k - \log \frac{1}{K}, \tag{10}$$

where the left item is the negative entropy of the average distribution and the right item is the negative entropy of the uniform distribution on K scene categories. Obviously, the smaller the $\mathcal{L}_2$, the more balanced the average distribution. Finally, auxiliary loss function is defined as:

$$\mathcal{L}_{en} = \mathcal{L}_1 + \mathcal{L}_2. \tag{11}$$

As the loss $\mathcal{L}_{en}$ decreases, the scene detection component can not only clearly identify the scene category for an image, but also avoid collapsing into a single scene model.

3.4. Semantic Interaction Module

In this module, the visual representations of labels interact with each other in a graph propagation mechanism under the guidance of the scene-aware label co-occurrence probability matrix. Firstly, with the maintained co-occurrence probability matrices $\{\mathbf{P}^0, \mathbf{P}^1, \cdots, \mathbf{P}^{K-1}\}$ and the predicted scene probability distribution of the input image I, its co-occurrence probability matrix is calculated in the form of weighted summation:

$$\mathbf{P}^I = \sum_{k=0}^{K-1} \pi_k \mathbf{P}^k. \tag{12}$$

Then, we construct a directed label graph $\mathcal{G} = \{V, E\}$ with the node set V being labels and the edge set E being the co-occurrence relations of the neighboring nodes. Naturally, the weights of edges can be initialized by the label co-occurrence probability matrix $\mathbf{P}^I$. Afterwards, messages are propagated among nodes through the graph $\mathcal{G}$ to learn contextual representations for all nodes. Specifically, the feature vector of node v_i at the t-th time step is denoted as $\mathbf{h}_i^t$ and initialized with the visual representation of label l_i, i.e., $\mathbf{h}_i^0 = \mathbf{f}_i$. Then the message gathered by node v_i from its neighboring nodes at t-th time step is formulated as follows:

$$\mathbf{m}_i^t = \sum_{j=0, j\neq i}^{n-1} p_{ji}^I \mathbf{h}_j^{t-1}. \tag{13}$$

In this way, graph $\mathcal{G}$ encourages message propagation if the node v_i has a high co-occurrent probability with the node v_j, while suppresses message propagation otherwise.

After that, we update node vectors via a gated recurrent update mechanism [16, 3], which is known as an effective approach for graph propagation. We denote it as $\Phi_g(\cdot)$ and the update process is formulated as:

$$\mathbf{h}_i^t = \Phi_g(\mathbf{h}_i^{t-1}, \mathbf{m}_i^t; \Theta), \tag{14}$$

where Θ denotes learnable parameters of $\Phi_g(\cdot)$. The process repeats T times to fully exploit the interactions among label representations with the guidance of the scene-aware label co-occurrence matrix. In this way, labels with higher co-occurring probability in the current scene distribution have more interactions, thus enabling to refine their visual representations. Consequentially, the final vector $\mathbf{h}_i^T$ of the node v_i encodes both the features of label l_i and the contextual message from other labels.

3.5. Label Prediction Module

For label prediction, the feature vectors $\mathbf{h}_i^0$ and $\mathbf{h}_i^T$ of the label l_i are firstly concatenated and input into a feed-forward network to produce a joint representation:

$$\mathbf{o}_i = \Phi_c([\mathbf{h}_i^0, \mathbf{h}_i^T]), \tag{15}$$

where $[\cdot]$ is the concatenation operation and Φ_c is a learnable multi-layer perceptron. Then, a binary classifier is framed to compute the confidence score of the presence of label l_i in the input image I, formulated as:

$$p_i = \sigma(\mathbf{w}_i^T \mathbf{o}_i), \tag{16}$$

where $\mathbf{w}_i \in \mathbb{R}^{d_I}$ is a learnable parameter vector. $\sigma(\cdot)$ is the sigmoid function that maps the input logit into a probability. Then following recent works [20, 18], the asymmetric focal loss is adopted to calculate the loss for the input sample $(I, \mathbf{y})$, formulated as:

$$\mathcal{L}_{cls} = \frac{1}{n} \sum_{i=0}^{n-1} \begin{cases} (1-p_i)^{\gamma^+} \log p_i, & y_i = 1, \\ p_i^{\gamma^-} \log(1-p_i), & y_i = 0, \end{cases} \tag{17}$$

where the γ^+ and γ^- are hyper-parameters and set as 0 and 2, respectively. Together with the auxiliary loss, the final loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{en}, \tag{18}$$

where λ is a hyper-parameter to make a trade-off between the two losses.

4. Experiments

4.1. Implementation Details

In this work, we use pretrained Bert [9] as the language model to initialize label embeddings, which are fixed during training. For a fair comparison, the input images are resized into 448×448 in both training and testing phases throughout all experiments. The whole framework is trained for 80 epochs using AdamW [19] optimizer with a batch size of 128 and 1-cycle policy [22] with a maximum learning rate of 0.0001. Following the previous works [20, 18], we adopt the RandAugment [7] and Cutout [10] for data augmentation, and apply exponential moving average to model parameters with a decay of 0.9997. The graph message propagation times T is empirically set as 3. The hyper-parameter λ is set as 1.0 throughout all experiments.

4.2. Comparisons with State-of-the-Arts

Results on Pascal VOC 2007. Pascal VOC 2007 [12] is a most widely used dataset to evaluate the multi-label image classification task. It has 20 label categories in total and 9,963 images, in which 5,011 images form *train-val* set and remaining 4,952 images are taken as *test* set for evaluation. Following common practice, we train the proposed framework on the *train-val* set and evaluate it on the test set. The number of scene categories on this dataset is set as 3 and experimental results are reported in Table 1. Be aware that the upper part displays the results of methods whose backbone is pretrained on the ImageNet [8], while the lower part is those methods that are further pretrained on the Microsoft COCO dataset [17]. From the table we can see that our SALGL framework achieves best results on both kinds of pre-training settings, reaching 95.1% and 96.7% in mAP, respectively. Particularly, compared with the SSGRL [3] and ADD-GCN [31] that are based on dataset-level and instance-level label co-occurrence respectively, our SALGL framework implements considerable performance gains despite their use of larger input resolution (576×576), arriving 1.7% and 0.7% in mAP respectively, suggesting the superiority of modeling the scene-based label co-occurrence. It is also worth noting that our SALGL shows significant merits in recognizing labels like *bus*, *motorbike* and *sofa*.

Results on NUS-WIDE. The NUS-WIDE [6] is a web dataset with 161,789 images for training and 107,859 for testing. After further manual annotation with 81 concepts,

Method	*plane*	*bike*	*bird*	*boat*	*bottle*	*bus*	*car*	*cat*	*chair*	*cow*	*table*	*dog*	*horse*	*moto*	*pers*	*nplant*	*sheep*	*sofa*	*train*	*tv*	mAP
SSGRL† [3]	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	**97.0**	85.6	97.8	98.3	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
ML-GCN [5]	99.5	98.5	**98.6**	98.1	80.8	94.6	97.2	**98.2**	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	**96.7**	84.3	98.9	93.7	94.0
TSGCN [28]	98.9	98.5	96.8	97.3	**87.5**	94.2	97.4	97.7	84.1	92.6	89.3	98.4	98.0	96.1	98.7	84.9	96.6	87.2	98.4	93.7	94.3
ASL [20]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	94.4
CSRA [33]	**99.9**	98.4	98.1	**98.9**	82.2	95.3	97.8	97.9	84.6	94.8	**90.8**	98.1	97.6	96.2	99.1	86.4	95.9	88.3	98.9	94.4	94.7
SALGL	**99.9**	**98.8**	98.3	98.2	81.6	**96.5**	**98.1**	97.8	**85.2**	**97.0**	89.6	**98.5**	**98.7**	**97.1**	**99.2**	**86.9**	96.4	**89.9**	**99.5**	**95.2**	**95.1**
SSGRL† [3]	99.7	98.4	98.0	97.6	85.7	96.2	98.2	98.8	82.0	98.1	89.7	**98.8**	98.7	97.0	99.0	86.9	98.1	85.8	99.0	93.7	95.0
ASL [20]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	94.6
CSRA [33]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	96.0
ADD-GCN† [31]	99.8	99.0	98.4	**99.0**	86.7	**98.1**	98.5	98.3	85.8	98.3	88.9	**98.8**	99.0	97.4	99.2	88.3	98.7	90.7	99.5	97.0	96.0
SALGL	**100.0**	**99.2**	**98.8**	98.6	**87.1**	**98.1**	**99.0**	**99.2**	**87.9**	**98.9**	**92.3**	**98.8**	**99.1**	**98.9**	**99.4**	**89.5**	**99.0**	**93.7**	**99.8**	**97.1**	**96.7**

Table 1. Experiments results on the Pascal VOC 2007 dataset in terms of class-wise precision (AP in %) and mean average precision (mAP in %). The optimal scores are highlighted in bold. Note that the symbol † denotes using a large input resolution (576×576).

Method	mAP	All		Top-3	
		CF1	OF1	CF1	OF1
CNN-RNN [23]	56.1	-	-	34.7	55.2
CMA [32]	61.4	60.5	73.7	55.5	70.0
GM-MLIC [27]	62.2	61.0	74.1	55.3	**72.5**
ICME [4]	62.8	60.7	74.1	56.3	70.6
ASL [20]	63.9	62.7	74.6	-	-
Q2L [18]	65.0	63.1	75.0	-	-
SALGL	**66.3**	**64.1**	**75.4**	**59.5**	71.0

Table 2. Experimental results on the NUS-WIDE dataset under the settings of all and top-3 labels (mAP in %). The optimal scores are highlighted in bold.

125,449 images are left as training set and 83,898 images as test set, respectively. Compared with other benchmarks, NUS-WIDE is more noisy and challenging. The number of scene categories on this dataset is set as 2 and experimental results are reported in Table 2. As shown, our SALGL framework achieves state-of-the-art performance and improves the mAP from 65.0% to 66.3%, a prominent performance gain of 1.3% compared to the sub-optimal method, namely Q2L [18]. Besides, on other important metrics (CF1 and OF1), our SALGL almost implements the best scores. The significant performance boost shows the superior ability of our SALGL in learning from noisy data.

Results on Microsoft COCO. Microsoft COCO (MS-COCO) [17] contains 82,081 images for the training set and 40,137 images for the validation set, and covers 80 label categories with almost 2.9 labels per image. Following previous works [3, 20, 18], we report the precision, recall and F1-measure under the settings of all and top-3 labels. The number of scene categories on this dataset is set as 6. To better evaluate our SALGL framework, we choose two widely used backbone networks in computer vision, namely convolution-based ResNet101 [33] and self-attention-based Vision Transformer (ViT) [11]. Experimental results are presented in Table 3. As shown in the upper part, our SALGL framework accomplishes 85.8% and 87.3% in mAP on the two typical input resolutions (448×448 and 576×576), exceeding the sub-optimal scores by 0.8% in mAP, which is a considerable improvement in terms of multi-label image classification task. Analogously, pronounced performance advantages are also achieved by our SALGL on both pre-training settings when using ViT-L16 [11] as the backbone network, as shown in lower part. Overall, our SALGL implements best scores on all important metrics (mAP, CF1 and OF1), demonstrating its advantageous performance in multi-label classification task.

4.3. Ablation Study

In this section, we investigate the effects of key designs on the performance of the proposed SALGL framework.

The effect of key modules. We first explore the effectiveness of the proposed modules. Experimental results are listed in Table 4. As shown, compared to the backbone network, the semantic attention module (SA) raises the mAP from 83.1% to 84.7% and 63.9% to 64.6% on the MS-COCO and NUS-WIDE datasets respectively. Besides, with the equipment of the semantic interaction module (SI), the performance is further improved, reaching 85.1% and 66.0% in mAP. These remarkable performance improvements demonstrate the effectiveness of these two modules. Most importantly, by considering the scene-based label co-occurrence relationships (K=6 and K=2), extra 0.7% and 0.3% gains in mAP are achieved compared to the dataset-level label co-occurrence (K=1) on the two datasets, suggesting the effectiveness of the proposed scene-aware label co-occurrence module. Overall, our SALGL achieves remarkable performance gains compared to the backbone network, reaching 2.7% and 2.4% in mAP on the MS-COCO and NUS-WIDE datasets respectively, proving its superiority in terms of multi-label image classification task.

The effect of auxiliary loss. To explore the effect of auxiliary loss, we remove it from the final loss function and observe the change in the performance of the proposed

Method	Backbone	Resolution	mAP	All						Top-3					
				CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
ML-GCN [5]	ResNet101	448×448	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
CMA [32]	ResNet101	448×448	83.4	82.1	73.1	77.3	83.7	76.3	79.9	87.2	64.6	74.2	89.1	66.7	76.3
TSGCN [28]	ResNet101	448×448	83.5	81.5	72.3	76.7	84.9	75.3	79.8	84.1	67.1	74.6	89.5	69.3	69.3
CSRA [33]	ResNet101	448×448	83.5	84.1	72.5	77.9	85.6	75.7	80.3	88.5	64.2	74.4	90.4	66.4	76.5
ASL [20]	ResNet101	448×448	85.0	-	-	80.3	-	-	82.3	-	-	-	-	-	-
Q2L-R101 [18]	ResNet101	448×448	84.9	84.8	**74.5**	79.3	86.6	76.9	81.5	78.0	**69.1**	73.3	80.7	**70.6**	75.4
SALGL	ResNet101	448×448	**85.8**	**87.2**	**74.5**	**80.4**	**87.8**	**77.6**	**82.4**	**90.4**	65.7	**76.1**	**91.9**	67.9	**78.1**
SSGRL [3]	ResNet101	576×576	83.6	**89.5**	68.3	76.9	**91.2**	70.7	79.3	**91.9**	62.1	73.0	**93.6**	64.2	76.0
C-Tran [15]	ResNet101	576×576	85.1	86.3	74.3	79.9	87.7	76.5	81.7	90.1	65.7	76.0	92.1	**71.4**	77.6
ADD-GCN [31]	ResNet101	576×576	85.2	84.7	75.9	80.1	84.9	79.4	82.0	88.8	66.2	75.8	90.3	68.5	77.9
Q2L-R101 [18]	ResNet101	576×576	86.5	85.8	76.7	81.0	87.0	78.9	82.8	90.4	66.3	76.5	92.4	67.9	78.3
SALGL	ResNet101	576×576	**87.3**	87.8	**76.8**	**81.9**	88.1	**79.5**	**83.6**	91.1	**66.9**	**77.2**	92.4	69.0	**79.0**
ViT-L16 [11]	ViT-L16	448×448	80.4	83.8	67.0	74.5	86.6	72.0	78.6	86.8	60.0	70.1	90.3	64.7	75.4
CSRA [33]	ViT-L16	448×448	86.9	**89.1**	74.2	81.0	**89.6**	77.1	82.9	**92.5**	65.8	76.9	**93.4**	68.1	78.8
SALGL	ViT-L16	448×448	**87.6**	88.4	**77.6**	**82.6**	88.3	**80.3**	**84.1**	92.1	**67.6**	**78.0**	92.6	**69.9**	**79.7**
ViT-L16 [11]	ViT-L16(22k)	448×448	89.4	88.4	**81.4**	84.8	88.5	**83.4**	85.9	92.5	**70.4**	79.9	93.4	**71.5**	81.0
SALGL	ViT-L16(22k)	448×448	**90.1**	**90.4**	80.8	**85.3**	**89.9**	83.0	**86.3**	**93.8**	69.8	**80.0**	**94.0**	**71.5**	**81.2**

Table 3. Experimental results on the Microsoft COCO dataset under the settings of all and top-3 labels (mAP in %). The optimal scores are highlighted in bold. The backbones noted with 22k are pretrained on the ImageNet 22k dataset.

Method	MS-COCO		NUS-WIDE	
	K=1	K=6	K=1	K=2
ResNet-101	83.1	-	63.9	-
ResNet-101 + SA	84.7	-	64.6	-
ResNet-101 + SA + SI	85.1	85.8	66.0	66.3

Table 4. The effect of key modules on the performance of the proposed SALGL framework (mAP in %). "SA" and "SI" are the abbreviations of semantic attention module and semantic interaction module, respectively. K is the number of scene categories.

Method	MS-COCO		NUS-WIDE	
	Entropy	mAP	Entropy	mAP
SALGL w/o $\mathcal{L}_{en}$	0.001	85.3	0.037	66.1
SALGL	1.759	85.8	0.686	66.3

Table 5. The effect of auxiliary loss on the performance of the proposed SALGL framework (mAP in %).

SALGL. Experimental results are reported in Table 5. The entropy is calculated from the number distribution of scene categories on the test set; the higher the entropy, the more balanced the predicted scene distribution, and vice versa. The table shows that the entropy of scene distribution in the absence of the auxiliary loss is very low on both datasets. This is, the scene distribution is very sharp and dominated by a specific scene category. In contrast, with the help of auxiliary loss, the issue is alleviated and the scene distribution is more balanced, suggesting its effectiveness in assisting the SALGL to detect the scene of the input image.

The effect of graph propagation. We also explore the effect of graph propagation. Experimental results are displayed in Figure 3. As shown, the performance of SALGL

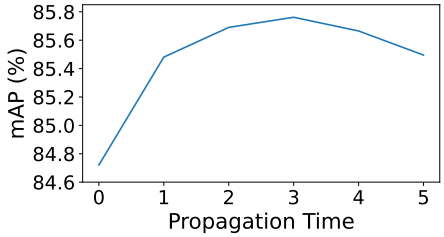

Figure 3. The effect of graph propagation time T on the performance of the proposed SALGL framework.

rises rapidly at the initial propagation stage, and reaches a maximum score in mAP when T=3. After that, the curve shows a downward trend due to the over-smoothing problem. Notably, SALGL benefits a lot from the graph propagation, reaching a gain of 1.9% in mAP, showing the effectiveness of the introduced graph propagation mechanism.

4.4. Visualization and Analysis

In this section, we first look at the six scene categories detected by the scene detection component on the Microsoft COCO dataset to investigate whether it works as expected. Note that the diagonal elements of the label co-occurrence frequency matrices $\{\mathbf{C}^0, \mathbf{C}^1, \cdots, \mathbf{C}^{K-1}\}$ count the number of labels whose images are classified into the correspond scene category during the whole training phase. Therefore, we create word cloud figures based on these diagonal elements to better understand the label preferences of these scene categories. As shown in Figure 5, it is clear that the six scene categories are traffic, animals, furniture, food, sports and transport. Their distinct contrasts demonstrate the efficacy of the scene detection component.

Figure 4. Visualization of label co-occurrence probabilities of two scene categories and their images with top confidence scores. Some labels in heat maps have been shortened, such as *mcycle* for *motorcycle*, *tlight* for *traffic light*, *ssign* for *stop sign*, *pmeter* for *parking meter*, *pplant* for *potted plant*, *dtable* for *dining table* and *kboard* for *keyboard*.

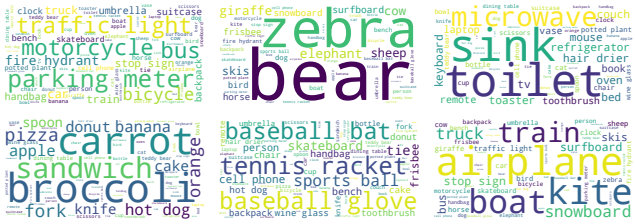

Figure 5. The word cloud visualization on the frequency distribution of labels under different scene categories.

Then, we visualize the label co-occurrence probabilities of two typical scene categories as well as their images with top confidence scores. As shown in Figure 4, we can see from the images exhibited on the left part that two scene categories favors outdoor images of traffic and indoor images of furniture, respectively. With this observation, we carefully choose 10 related labels from the candidate set and visualize their co-occurrence probabilities with each other. As shown in the heat maps on the right, label co-occurrence probabilities between the two scene categories have a large difference, proving the capacity of our SALGL in modeling scene-aware label co-occurrence relationships.

Furthermore, by focusing on the rows of *person* on the heat maps (highlighted in dashed boxes), we investigate how scene category affects the probabilities that other labels appear in the presence of *person*. In the middle column, the traffic-related labels (*car*, *bus* and *traffic light*, etc.) are more likely to co-occur with *person* in the traffic scene than in the furniture scene, where the co-occurrence probabilities are nearly zero. As a comparison, in the right column, the labels associated to furniture (*couch*, *dining table* and *chair*, etc.) have higher chances of co-occurring with *person* in the furniture scene than in the traffic scene. Notably, our SALGL is competent to detect the scene categories of images and further determine reasonable co-occurrence probabilities of their labels, thus providing precise guidance for subsequent graph-based semantic interaction.

5. Conclusion

In this paper, we propose a novel scene-aware label graph learning framework for multi-label image classification. Concretely, the scene-aware label co-occurrence module maintains a label co-occurrence matrix for each scene category and tracks co-occurring labels during the training phase, which is used to guide the interactions of label representations via graph propagation. An advanced entropy-based auxiliary loss is proposed to prevent it from collapsing. Experimental results on public benchmarks demonstrate the superiority of our SALGL framework.

Acknowledgments

This work is supported by National Key R&D Project of China under Grants No.2021QY2102, National Natural Science Foundation of China under Grants No.62172089, No.61972087, No.62172090, No.62106045, Natural Science Foundation of Jiangsu Province under Grants No.BK20191258, Purple Mountain Laboratories, Jiangsu Provincial Key Laboratory of Computer Networking Technology, Jiangsu Provincial Key Laboratory of Network and Information Security under Grants No.BM2003201, and Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grants No.93K-9.

References

[1] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Frank Wang. Order-free rnn with visual attention for multi-label classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[2] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. Recurrent attentional reinforcement learning for multi-label image recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[3] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 522–531, 2019.

[4] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 622–627. IEEE, 2019.

[5] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019.

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling, 2017.

[15] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021.

[16] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. Gated graph sequence neural networks, 2017.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[20] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021.

[21] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[22] Leslie N Smith. A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay. *arXiv preprint arXiv:1803.09820*, 2018.

[23] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016.

[24] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12265–12272, 2020.

[25] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision*, pages 464–472, 2017.

[26] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1901–1907, 2015.

[27] Yanan Wu, He Liu, Songhe Feng, Yi Jin, Gengyu Lyu, and Zizhang Wu. Gm-mlic: Graph matching based multi-label image classification, 2021.

[28] Jiahao Xu, Hongda Tian, Zhiyong Wang, Yang Wang, Wenxiong Kang, and Fang Chen. Joint input and output space

learning for multi-label image classification. *IEEE Transactions on Multimedia*, 23:1696–1707, 2020.

[29] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–288, 2016.

[30] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2020.

[31] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020.

[32] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12709–12716, 2020.

[33] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 184–193, 2021.

[34] Xuelin Zhu, Jiuxin Cao, Jiawei Ge, Weijia Liu, and Bo Liu. Two-stream transformer for multi-label image classification. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3598–3607, 2022.