Journal of Computer Science and Electrical Engineering
Print ISSN: 2663-1938
Online ISSN: 2663-1946
DOI: https://doi.org/10.61784/jcsee3131

A MULTI-LABEL IMAGE RECOGNITION NETWORK DRIVEN BY LABEL-IMAGE SEMANTIC ALIGNMENT

YuYu Chen*, LinJun Wu, LianZhi Chu
School of Information Science and Engineering, Hunan Institute of Engineering, Xiangtan 411100, Hunan, China.
**Corresponding Author: YuYu Chen*

Abstract: Multi-label image recognition aims to predict a set of semantic labels for an image and has wide applications in other fields. Existing methods have two main problems: the attention regions generated by attention-based methods are insufficiently correlated with label semantics; methods based on label correlation modeling lack dynamic interaction with visual image content, making it difficult to achieve precise alignment between label semantics and image regions. To address these issues, this paper proposes a novel multi-label image recognition algorithm. First, we construct a label graph and leverage a graph convolutional network to learn label semantic priors, modeling the dependencies among labels. Second, we design a semantic decoupling module that adaptively focuses on relevant image regions under the guidance of label semantics to generate label semantic representations. Finally, we introduce a semantic association reasoning module that employs a multi-head self-attention mechanism to dynamically capture semantic correlations among labels, thereby enhancing the discriminative ability of features. Experimental results on the PASCAL VOC 2007 dataset show that our method achieves 95.0% mAP, surpassing existing state-of-the-art methods and improving by 1.6 percentage points over the baseline SSGRL. Ablation studies further validate the effectiveness of each module.
Keywords: Multi-label image recognition; Graph convolutional network; Semantic decoupling; Multi-head self-attention; Label correlation

1 INTRODUCTION

Multi-label image recognition aims to achieve a comprehensive understanding of image content by predicting a set of semantic labels for each image. It is widely applicable in autonomous driving, image retrieval, and other fields.The relationships between the labels are very complex, and the combination of labels and image regions is diverse, which still poses significant challenges. Multi-label image classification has great research significance.

To exploit spatial information in images for enhancing fine-grained object recognition, most existing methods based on object feature extraction [1-5] focus on localizing object regions (e.g., candidate regions, random regions, and image patches) to cover all potential target objects. These methods rely on pre-trained detectors and are difficult to train end-to-end. Attention mechanisms [6-9] guide the model to automatically focus on regions of interest in images and are widely used in multi-label image recognition. However, the attention regions generated by such methods often lack sufficient semantic correlation with the corresponding labels. Label correlations provide important inference evidence for multi-label recognition. Methods based on label correlation modeling typically analyze inter-label relationships from perspectives such as label co-occurrence frequencies. These approaches [10-13] focus on modeling relationships within the label space but lack dynamic interaction with visual image content, making it difficult to achieve precise alignment between label semantics and image regions.

To fully model correlations among labels, dynamically perceive dis-criminative regions in images, and achieve interaction between label semantics and image semantics, this paper proposes a novel multi-label image recognition algorithm (DLSA). Its main innovations are as follows:

(1)A label correlation matrix is constructed based on label co-occurrence relationships, and a graph convolutional network is used to learn label semantic priors, obtaining label embedding vectors with label dependencies.

(2)A semantic decoupling module enables cross-modal interaction between label semantics and image semantics. Guided by label semantics, the model adaptively focuses on image regions relevant to each label, and uses spatial location information to generate semantic representations for the corresponding labels.

(3)A semantic association reasoning module dynamically captures semantic correlations among labels and iteratively updates the label semantic representations, pushing features of correlated labels closer and separating features of uncorrelated labels, thereby enhancing the representational power and discriminative ability of the features.

Experimental results on the PASCAL VOC 2007 dataset validate the effectiveness of the proposed method, and ablation studies further demonstrate the necessity of each module.

2 RELATED WORK

Visual Encoder. Most existing methods employ a unified set of visual features extracted by a CNN or Transformer as the global image representation for all labels. However, since each label typically corresponds to only a small portion of the image, directly using such global features may lead to sub-optimal performance. To address this issue, some approaches [1-2] employ target localization techniques to identify multiple object regions. For instance, Xu et al. [1]

utilize a pre-trained Mask R-CNN as a fixed feature extractor to obtain object-level features, which serve as nodes in a spatial graph. By decoupling detection from classification and separating feature extraction from spatial relationship learning, their method models object spatial relationships while preserving the benefits of pre-trained detection. These approaches can guide the model to focus on semantic regions and suppress background interference, yet they rely on pre -trained detectors and are difficult to train in an end-to-end manner. Other methods [3-9] leverage attention mechanisms to direct the model's focus toward multiple image regions. For example, Chen et al. employ a sliding pooling window to locate the maximum responses in category-specific heatmaps and extract quarter-size candidate regions [5]. Zhou et al. establish connections among all spatial positions via self-attention, enabling precise perception of object locations and feature focusing [9]. Nevertheless, the attention regions generated by these methods often exhibit insufficient semantic correlation with the corresponding labels and relatively coarse localization accuracy.

Label Correlation. Label correlation provides important cues for inferring label sets. Chen et al. employ low-rank bilinear pooling to compute the correlation strength between category semantic vectors and image features [10], generating category-specific attention maps, and then use graph convolutional networks to structurally model and propagate statistical label co-occurrence relationships.Wang et al. adopt block Krylov subspace spectral graph convolution to avoid over-smoothing [11], explore multi-scale label relationships, and design an adaptive module that dynamically constructs a label relationship graph based on image content. Yuan et al. build a label relationship graph using cosine similarity between pre-trained word embeddings [12]. However, these methods focus on modeling label correlations but lack dynamic interaction with the visual content of the input image.

3 PROPOSED METHOD

The overall structure of the model is shown in Figure 1. The following subsections will introduce each component of the model in detail.

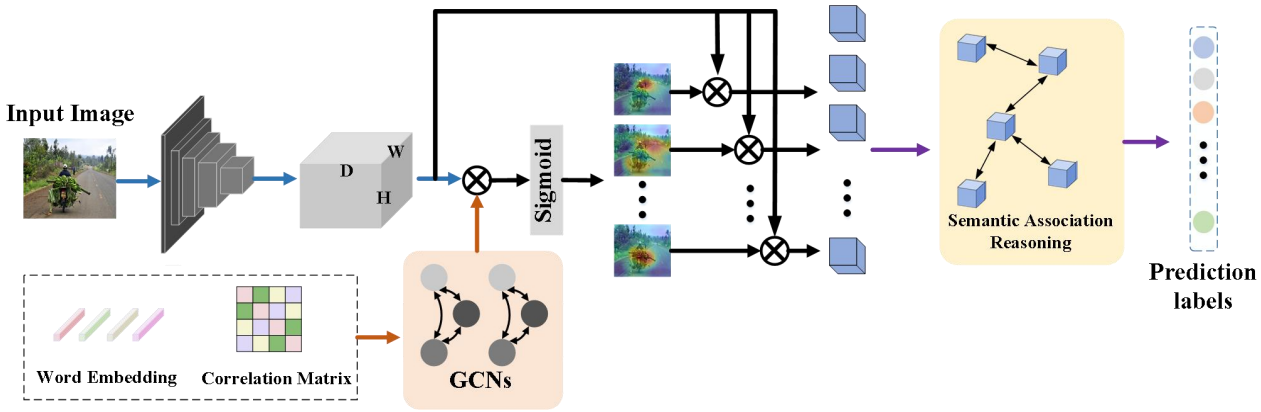

Figure 1 The Overall Architecture of the Proposed Model DLSA

3.1 Image Feature Extraction Module

The image is fed into a convolutional neural network to extract global features for image feature representation. Given an image, the output of the last convolutional layer of ResNet101 is the encoded image feature map $F_{x,\mathrm{i}} \in \mathrm{R}^{\mathrm{D}\times\mathrm{H}\times\mathrm{W}}$, where D, W, and H denote the number of channels, height, and width of the feature map, respectively.

3.2 Label Semantic Prior Learning (LSPL) Module

A graph structure G=(E,M) is employed to model label co-occurrence relationships. The nodes are initialized with GloVe pre-trained word embeddings $\mathrm{E} \in \mathrm{R}^{\mathrm{C}\times\mathrm{K}}$(where C is the number of labels and K is the dimension of the word embeddings). The adjacency matrix M is derived from data-set statistics, where the conditional probability of label *j* co-occurring given label *i* is defined as:

$$P_{i,j}=P(j|i)=O_{ij}/O_i \tag{1}$$

Subsequently, a two-layer graph convolutional network is used for label relationship reasoning. Each node aggregates feature information from its neighboring nodes according to the weights defined by the connecting edges. The computation at each layer is:

$$E^{l+1}=f_g(E^l,\widehat{M})=\sigma(\widehat{M}E^lW^l) \tag{2}$$

where, $\widehat{M}$ is the normalized adjacency matrix, and W^l denotes learnable weights. The output dimensions of the two layers are $E^1 \in R^{C\times d_1}$and $E^2 \in R^{C\times d_2}$, respectively.

3.3 Semantic Decoupling (SD) Module

To adaptively learn label semantic representations from the input feature map $F_{x,\mathrm{i}}$, the image features are first aligned with the label representation E^2, and then *C* class- specific attention maps are generated via Sigmoid normalization. The

attention maps are denoted as $M=[m_1,...,m_c,...,m_C]$. These attention maps guide the model to focus on the spatial regions associated with each label, based on which the feature representation for each label node is aggregated as follows:

$$v_c=(m_c^T)F_{x,i}=\sum_{\mathrm{h}=1}^{H}\sum_{w=1}^{W}(m_{\mathrm{hw}}^c)F_{x,\mathrm{hw}} \tag{3}$$

Where, v_c corresponds to the semantic representation of the c-th label. Repeating this process for all labels yields a complete set of label-related feature vector encodings $V=[v_1,\ldots,v_c,\ldots v_C]\in R^{C\times D}$, where each vector corresponds to a specific label meaning and the labels are partially correlated.

3.4 Semantic Association Reasoning (SAR) Module

To further capture the correlations among label semantic representations, a multi-head self-attention layer is employed to analyze the dependencies between label semantic representations.

For each attention head h, the correlation weight between label semantic features v_{n} and v_{o} is computed as,

$$\alpha_{n,o}^{(\mathrm{h})}=\frac{exp(\langle\ W_q^{\mathrm{h}}v_n,W_k^{\mathrm{h}}v_o\rangle\ )}{\sum exp(\langle\ W_q^{\mathrm{h}}v_n,W_k^{\mathrm{h}}v_o\rangle\ )} \tag{4}$$

where W_q and W_k denotes learnable weights. Then, under the h-th attention head, the label semantic representation is updated by aggregating all relevant features guided by the correlation coefficients α,

$$\bar{v}_n^{\mathrm{h}}=\sum_{n=1}^{C}\alpha_{\mathrm{n,o}}^{(\mathrm{h})}(\mathrm{W}_v^{\mathrm{h}}v_n) \tag{5}$$

The outputs from multiple heads are concatenated, and a standard residual connection is applied to obtain the updated label semantic representation,

$$\bar{v}_n=\bar{v}_n^1||...||\bar{v}_n^{\mathrm{h}}||...||\bar{v}_n^{\mathrm{H}}+v_n \tag{6}$$

where H is the total number of attention heads and $||$ denotes the concatenation operation. The resulting representation is then passed through two fully connected layers (with ReLU) to produce a high-order semantic representation. Finally, the high-order label semantic representation vector $\bar{v}_n$ is processed through mean computation and a fully connected classification layer to generate the label score vector corresponding to the image x_i,

$$Y=[y_i^1,...,y_i^c,...,y_i^C] \tag{7}$$

To enforce that the learned labels approximate the ground-truth labels, the classical binary cross-entropy (BCE) loss is adopted:

$$L=\sum_{i=1}^{N}\sum_{c=1}^{C}(\hat{y}_i^c\log y_i^c+((1-\hat{y}_i^c)\log(1-y_i^c))) \tag{8}$$

where $\hat{Y}_i=[\hat{y}_i^1,\ldots,\hat{y}_i^c,\ldots,\hat{y}_i^C]$ denotes the ground-truth label vector for image x_i.

4 Experiments and Results

4.1 Datasets and Experimental Setup

The proposed method DLSA is evaluated on the widely used Pascal VOC 2007 dataset. This dataset contains 9,963 images across 20 categories, with 5,011 images for training and 4,952 for testing.

All experiments are implemented using Python 3.10 and PyTorch 2.1.0. ResNet101 is adopted as the backbone network. The dimension D of the extracted image feature map is 2,048, and the dimension K of the GloVe word embeddings is set to 300. The output dimensions after graph convolution, d_1 and d_2, are set to 1,024 and 2,048, respectively. The number of attention heads H in the semantic association reasoning module is set to 4. The model is trained using the Adam optimizer with a batch size of 20, momentum parameters of 0.999 and 0.9, and an initial learning rate of 1×10^{-5}. The learning rate is reduced by a factor of 10 when the validation error plateaus.

4.2 Performance Comparison and Analysis

Following existing work, average precision (AP) and mean average precision (mAP) are used as evaluation metrics on the Pascal VOC 2007 dataset, with mAP serving as the primary performance indicator. Our DLSA model is compared with several state-of-the-art methods, and the results are presented in Table 1.

The experimental results demonstrate that the proposed method achieves the mAP of 95.0%, outperforming all compared methods and attaining the best performance. Compared with the baseline method SSGRL (93.4% mAP), the proposed method improves accuracy in nearly all categories, yielding an increase of 1.6 percentage points in mAP. These results fully validate the effectiveness of the proposed model.

Table 1 Comparison with Various Recognition Algorithms on VOC 2007 Dataset

labels	AP						
	ResNet-101	SSGRL[10]	AMS-GCN[11]	TSGCN[1]	DATran[9]	MSFA[5]	Ours
aero	99.5	99.5	99.7	98.9	**99.9**	<u>99.8</u>	<u>99.8</u>
bike	97.7	97.1	97.6	<u>98.5</u>	**98.7**	98.6	98.4
bird	<u>97.8</u>	97.6	97.6	96.8	**98.6**	98.2	**98.6**
boat	96.4	97.8	98.1	97.3	<u>98.4</u>	**99.1**	98.3

bottle	65.7	82.6	79.6	**87.5**	82.6	84.1	85.2
bus	91.8	94.8	95.4	94.2	96.0	95.5	**96.9**
car	96.1	96.7	97.3	97.4	97.7	98.0	**98.4**
cat	97.6	98.1	97.9	97.7	**98.6**	98.4	98.4
chair	74.2	78.0	81.1	84.1	**85.0**	84.3	83.9
cow	80.9	97.0	96.0	92.6	96.2	**98.4**	95.3
table	85.0	85.6	85.4	**89.3**	84.7	88.3	88.5
dog	98.4	97.8	97.8	98.4	98.5	98.4	**98.7**
horse	96.5	98.3	98.5	98	98.2	98.6	**98.8**
motor	95.9	96.4	96.7	96.1	96.9	**97.3**	96.1
person	98.4	98.8	99.0	98.7	98.8	99.0	**99.3**
plant	70.1	84.9	84.8	84.9	85.0	**86.7**	86.6
sheep	88.3	96.5	96.4	96.6	97.9	**98.3**	96.8
sofa	80.2	79.8	83.7	87.2	86.7	**88.5**	87.7
train	98.9	98.4	98.6	98.4	99.2	**99.8**	99.2
tv	89.2	92.8	94.7	93.7	**95.2**	**95.2**	94.9
mAP	89.0	93.4	93.8	94.3	94.6	94.9	**95.0**

Note: Bold indicates the best result, and underline indicates the second-best result.

4.3 Ablation Study

To validate the effectiveness of each component in the proposed model, a series of ablation experiments are conducted on the VOC 2007 dataset. The results are shown in Table 2. By comparing the first and second rows of the table, it can be observed that after introducing the label semantic prior learning module (graph convolution), the mAP increases from 93.4% to 94.0%, indicating that this module can effectively model the dependencies among labels. Comparing the first and third rows, the semantic decoupling module alone also achieves a mAP of 94.0%, demonstrating its ability to aggregate label-related image region features. Further comparing the second and third rows with the fourth row, when both the label semantic prior learning module and the semantic decoupling module are used together, the mAP improves to 94.6%, suggesting that they have a synergistic effect and can capture richer label semantic information. Finally, comparing the fourth and fifth rows, after adding the semantic association reasoning module, the mAP reaches 95.0%, which is an improvement of 0.4 percentage points over using only the first two modules. This verifies the effectiveness of this module in modeling dynamic correlations among labels.

Table 2 Comparison of Module Ablation Experiments

LSPL Module	SD Module	SAR Module	mAP
			93.4
√			94.0
	√		94.0
√	√		94.6
√	√	√	95.0

5 CONCLUSION

This paper addresses the problems of inaccurate alignment between label semantics and image regions, as well as the lack of dynamic interaction in label correlation modeling for multi-label image recognition. To this end, a multi-label image recognition algorithm oriented to label-image semantic alignment is proposed. The algorithm learns label co-occurrence relationships via a graph convolutional network and achieves dynamic interaction between label semantics and image features using a semantic decoupling module along with an association reasoning module. Experimental results on the VOC2007 dataset demonstrate that the proposed method achieves 95.0% mAP, verifying its effectiveness. Ablation studies further confirm the synergistic enhancement of the three modules.

However, existing methods, including the one proposed in this paper, primarily focus on positive correlations among labels, while the modeling of label irrelevance (e.g., mutual exclusion relationships) remains insufficient. Moreover, large-scale pre-trained models have not been introduced. Future research will explore the incorporation of label irrelevance constraints and leverage large-model techniques (e.g., CLIP) to further enhance the semantic understanding and generalization capability of the model.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This article is funded by the Talent Research Fund Project of Hunan Institute of Engineering (No. 09001003).

REFERENCES

[1] Xu Jiahao, Tian Hongda, Wang Zhiyong, et al. Joint input and output space learning for multi-label image classification. IEEE Transactions on Multimedia, 2020(23): 1696-1707.
[2] Wu Yanan, Feng Songhe, Yang Wang. Semantic-aware graph matching mechanism for multi-label image recognition. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(11): 6788-6803.
[3] Hu Yunqing, Chen Qianglong, Zhang Yin. Semantic perception enhancement region pyramid model for multi-label image recognition. Journal of Computer-Aided Design & Graphics, 2025, 37(10): 1770-1786.
[4] Gao Bin-Bin, Zhou Hongyu. Learning to discover multi-class attentional regions for multi-label image recognition. IEEE Transactions on Image Processing, 2021, 30: 5920-5932.
[5] Chen Jiale, Feng Xu, Tao Zeng, et al. MSFA: Multi-stage feature aggregation network for multi-label image recognition. IET Image Processing, 2024, 18(7): 1862-1877.
[6] Li Liang, Wang Shuhui, Jiang Shuqiang, et al. Attentive recurrent neural network for weak-supervised multi-label image classification. In Proceedings of the 26th ACM international conference on Multimedia, 2018: 1092-1100.
[7] Zhou Wei, Xia Zhiwu, Dou Peng, et al. Aligning image semantics and label concepts for image multi-label classification. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(2): 1-23.
[8] Ye Qingwen, Zhang Qiuju. Multi-label image recognition using channel pixel attention. Computer Science and Exploration, 2024, 18(08): 2109-2117.
[9] Zhou Wei, Zheng Zhijie, Su Tao, et al. DATran: Dual attention transformer for multi-label image classification. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(1): 342-356.
[10] Chen Tianshui, Xu Muxin, Hui Xiaolu, et al. Learning semantic-specific graph representation for multi-label image recognition. In Proceedings of the IEEE/CVF international conference on computer vision, 2019: 522-531.
[11] Wang Xuesong, Rong Xiaolong, Cheng Yuhu, et al. Multi-label image recognition based on adaptive multi-scale graph convolutional network. Control and Decision, 2022, 37(07): 1737-1744. DOI: 10.13195/j.kzyjc.2021.0179.
[12] Yuan Jin, Chen Shikai, Zhang Yao, et al. Graph attention transformer network for multi-label image classification. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(4): 1-16.
[13] Qu Xiwen, Che Hao, Huang Jun, et al. Multi-layered semantic representation network for multi-label image classification. International journal of machine learning and cybernetics, 2023, 14(10): 3427-3435.