Structured Semantic Transfer for Multi-Label Recognition with Partial Labels

Tianshui Chen [1], Tao Pu [2], Hefeng Wu [2], Yuan Xie [2], Liang Lin [2] *

[1] Guangdong University of Technology, [2] Sun Yat-Sen University
tianshuichen@gmail.com, putao3@mail2.sysu.edu.cn, wuhefeng@gmail.com, phoenixsysu@gmail.com, linliang@ieee.org

Abstract

Multi-label image recognition is a fundamental yet practical task because real-world images inherently possess multiple semantic labels. However, it is difficult to collect large-scale multi-label annotations due to the complexity of both the input images and output label spaces. To reduce the annotation cost, we propose a structured semantic transfer (SST) framework that enables training multi-label recognition models with partial labels, i.e., merely some labels are known while other labels are missing (also called unknown labels) per image. The framework consists of two complementary transfer modules that explore within-image and cross-image semantic correlations to transfer knowledge of known labels to generate pseudo labels for unknown labels. Specifically, an intra-image semantic transfer module learns image-specific label co-occurrence matrix and maps the known labels to complement unknown labels based on this matrix. Meanwhile, a cross-image transfer module learns category-specific feature similarities and helps complement unknown labels with high similarities. Finally, both known and generated labels are used to train the multi-label recognition models. Extensive experiments on the Microsoft COCO, Visual Genome and Pascal VOC datasets show that the proposed SST framework obtains superior performance over current state-of-the-art algorithms. Codes are available at https://github.com/HCPLab-SYSU/HCP-MLR-PL.

Introduction

Recently, lots of efforts (Chen et al. 2019c,a, 2020) are dedicated to the task of multi-label image recognition as it benefits various applications ranging from content-based image retrieval and recommendation systems to surveillance systems and assistive robots. Despite achieving impressive progress, current leading algorithms (Chen et al. 2019c,a, 2020) introduce data-hungry deep convolutional networks (He et al. 2016; Simonyan and Zisserman 2015) to learn discriminative features, and thus they depend on collecting large-scale clean and complete multi-label datasets. However, it is very time-consuming to collect a consistent and exhaustive list of labels for every image, making collecting clean and complete multi-label annotations more difficult and less scalable. In contrast, it is easy and scalable to annotate partial labels for each image, which can be regarded as an alternative way to address the above problem. In this work, we aim to address the task of learning multi-label recognition models with partial labels (MLR-PL).

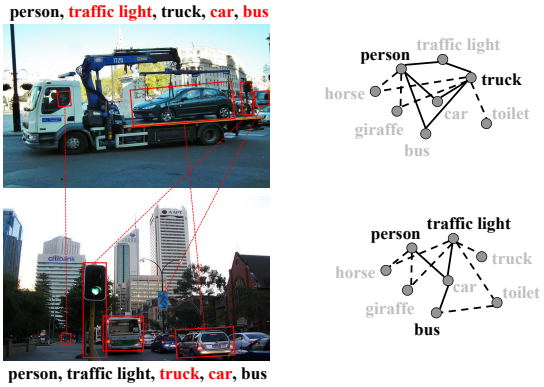

Figure 1: Two examples of images with partial labels (unknown labels are highlighted in red). We can mine the intra-image and cross-image correlations to help complement the unknown labels.

Current algorithms mainly consider multi-label recognition as a multiple binary classification task. Treating the unknown labels as missing or negative labels is an intuitive way to adapt these algorithms to address the MLR-PL task (Sun et al. 2017; Joulin et al. 2016). However, it results in an obvious performance drop as it loses some data or even incurs some noisy labels. Fortunately, strong semantic correlations within each image and cross different images exist, and these correlations can efficiently help to transfer semantic knowledge of known labels to construct the unknown labels: i) Label co-occurrences are widespread in real-world images, e.g., tables tend to co-occur with chairs and cars are likely to co-exist with roads; ii) Objects of the same category in different images may share similar visual appearances, and thus images with similar visual features may have the same labels.

In this work, we explore mining these correlations to help complement the unknown labels by a novel structured semantic transfer (SST) framework. It consists of

*Tianshui Chen and Tao Pu contribute equally to this work and share first authorship. Corresponding author is Liang Lin.

two complementary modules that learn image-specific co-occurrence to help transfer semantic labels within each image and category-specific feature similarities to transfer semantic labels across different images. Although previous work (Huynh and Elhamifar 2020) also takes notice of label/image dependencies, it merely introduces statistical co-occurrence and image-level similarities to regularize training. Instead, the SST framework aims to learn fine-grained image-specific co-occurrence and category-specific feature similarities, which can help construct accurate pseudo labels for the unknown labels to facilitate the MLR-PL task. For example in Figure 1, the feature vectors of *truck* are similar in two different images and we can use the annotated *truck* of the upper image to help complement the unknown *truck* of the lower image. Similarly, *traffic light* has high co-occurrence probability with *car*, and we can complete this unknown label based on the co-occurrence.

The SST framework builds on a semantic-aware representation learning (SARL) module that incorporates category semantic to help learn category-specific feature representation. Then, an intra-image semantic transfer (IST) module is designed to learn a co-occurrence matrix among all categories for each image and map the known labels to complement some unknown labels based on the learned co-occurrences. Meanwhile, a cross-image semantic transfer (CST) module is introduced to measure the similarities of feature representations that belong to the same category and are from different images. It then transfers the semantic known labels to help complement some unknown labels with high similarity. Finally, the known labels and complemented labels are used to supervise training the multi-label recognition model.

The contributions of this work are summarized into three folds. First, we introduce a structured semantic transfer framework to simultaneously mine intra-image and cross-image correlations to help complement the unknown labels. Second, two complementary modules (i.e., intra-image and cross-image semantic transfer) are incorporated to transfer semantic within each image and cross different images to generate pseudo labels accurately. Finally, we conduct extensive experiments on variant datasets to demonstrate the effectiveness of the proposed SST framework. We also perform ablative studies to analyze the contribution of each module for better understanding.

Related Works

Multi-label image recognition receives increasing attention (Wei et al. 2016; Chen et al. 2020) since it is more practical and necessary than its single-label counterpart. To solve this task, lots of efforts are dedicated to discovering discriminative local regions for feature enhancement by object proposal algorithms (Wei et al. 2016; Yang et al. 2016) or visual attention mechanisms (Ba, Mnih, and Kavukcuoglu 2014; Chen et al. 2018b). Another line of works propose to capture label dependencies to regularize training multi-label recognition models and thus improve their performance (Wang et al. 2016, 2017; Chen et al. 2019c,a). These works either introduce the RNN/LSTM to implicitly capture label dependencies (Wang et al. 2016, 2017) or explicitly model the label dependencies in the form of structured graphs and exploit the graph neural networks (Li et al. 2016) to adaptively capture the label dependencies. Recently, Chen et al. (Chen et al. 2019a) present state-of-the-art results on several multi-label datasets by using semantic decoupling to obtain semantic-aware features for different category labels, and we employ their semantic decoupling module for learning category-specific features in this work. However, despite achieving remarkable progress, all these methods rely on data-hungry deep neural networks (Simonyan and Zisserman 2015; He et al. 2016) to learn discriminative feature representation, and thus require large-scale and clean datasets (e.g., Visual Genome (Krishna et al. 2016), MS-COCO (Lin et al. 2014) and Pascal VOC (Everingham et al. 2010)) to train the deep neural networks. However, it is time-consuming and labor-intensive to annotate a complete list of labels for every image, making collecting large-scale and complete multi-label datasets less practical and scalable.

To reduce the annotation cost, some works propose to learn multi-label recognition models with partial labels, i.e., merely some labels are known (Durand, Mehrasa, and Mori 2019; Huynh and Elhamifar 2020). To deal with this task, some works (Bucak, Jin, and Jain 2011; Wang et al. 2014; Sun et al. 2017) simply regard the unknown labels as negative labels, and train the models with a similar scheme for the fully labeled setting. These methods could suffer from severe performance drop because many positive labels may be wrongly annotated as negative. Some other works (Tsoumakas and Katakis 2007) treat multi-label recognition as multiple independent binary classifications. However, it ignores the label dependencies that play a key role in multi-label recognition. To overcome this issue, some works exploit label dependencies to transfer the known labels to help complement the unknown label (Xu, Jin, and Zhou 2013; Yu et al. 2014). Cabral et al. (Cabral et al. 2011) introduce the low-rank regularization to exploit label correlations and complete unprovided labels, while Wu et al. (Wu, Lyu, and Ghanem 2015) similarly adopts a low rank empirical risk minimization. A mixed graph is also utilized in (Wu, Lyu, and Ghanem 2015) to encode a network of label dependencies. In (Kapoor, Viswanathan, and Jain 2012), missing labels are treated as latent variables in probabilistic models and predicted by posterior inference using Bayesian networks.Most of these works depend on solving an optimization problem that requires loading the whole training set, which cannot be integrated into deep networks for batch-level training. Such limitations result in inferior performance since fine-tuning is critical in transferring pre-trained DNN models. More recently, Durand et al. (Durand, Mehrasa, and Mori 2019) propose a normalized BCE loss to exploit label proportion information and use it to train the model with partial labels. Huynh et al. (Huynh and Elhamifar 2020) introduce statistical label co-occurrence and image-level feature similarity to regularize training networks.

Different from these methods, the proposed framework introduces two complementary modules, in which the first module learns image-specific label co-occurrence correlations to transfer provided labels within the same image to complement unknown labels and the second module learns

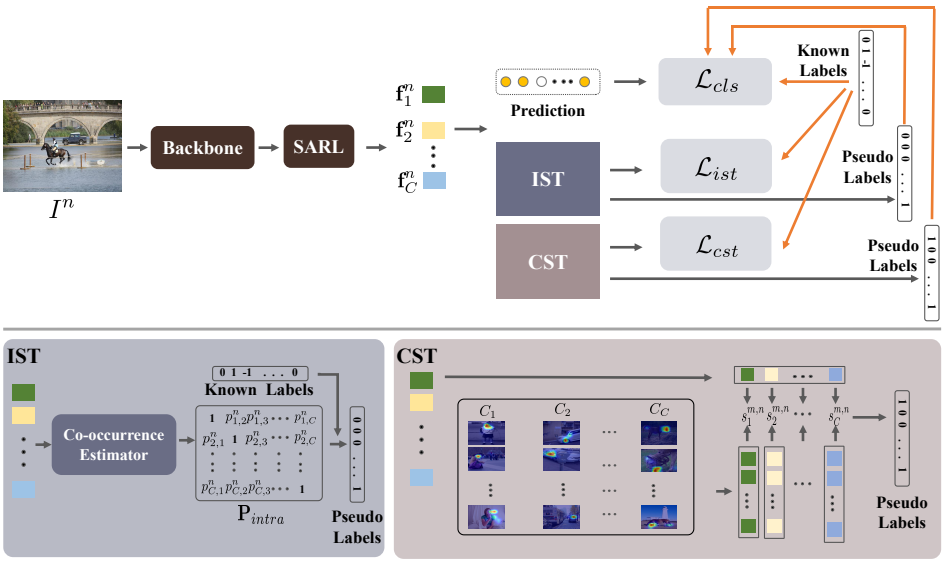

Figure 2: An overall illustration of the proposed structured semantic transfer framework. The upper part is the overall pipeline that consists of the IST and CST modules to generate pseudo labels, which are then fed to supervise training the multi-label recognition model. The lower part is the detailed implementations of the IST and CST modules. The IST module first predicts the label co-occurrence matrix and then maps the known labels to complement the unknown labels. The CST module first learns category-level feature similarities across different images and then also maps to generate the pseudo labels.

category-level feature similarity correlations to transfer provided labels across different images to complement unknown labels. The two modules can be seamlessly incorporated into existing deep neural network models for multi-label recognition and trained in an end-to-end manner.

Structured Semantic Transfer

In this section, we introduce the proposed SST framework that mines intra-image and cross-image correlations to help complement the unknown labels. It adopts a semantic-aware representation learning module to extract category-specific feature vectors for each image. The IST module first learns the co-occurrence probability of each category pair and then constructs a co-occurrence matrix for each image. It then transfers the semantic knowledge of the known labels to complement some unknown labels based on the learned co-occurrence matrix. Meanwhile, the CST module learns the similarities among feature vectors of the same category from different images. Similarly, we can also exploit the known labels to complement some unknown labels based on the learned similarities. In this way, we can obtain pseudo labels for unknown labels accurately and use both known labels and pseudo labels to train the multi-label models. An overall illustration is presented in Figure 2.

Notation. Here, we give an introduction to the notations used in the paper. We denote the training set as $\mathcal{D} = \{(I^1, \mathbf{y}^1), ..., (I^N, \mathbf{y}^N)\}$, in which N is the number of training samples. $\mathbf{y}^n = \{y_1^n, \cdots, y_C^n\} \in \{-1, 0, 1\}^C$ is the label vector of the n-th sample and C is the label number. y_c^n is assigned to 1 if label c exists in the n-th image, assigned to -1 if it does not exist, and assigned to 0 if it is unknown.

Semantic-Aware Representation Learning

Given an input image I^n, we first utilize a backbone network to extract global feature maps $\mathbf{f}^n$, and then follow recent work (Chen et al. 2019a) to adopt the semantic decoupling module to learn semantic-aware representation of each category, denoted as $[\mathbf{f}_1^n, \mathbf{f}_2^n, \cdots, \mathbf{f}_C^n]$. We use a gated graph neural network (Chen et al. 2019b, 2018a, 2021) and linear classifier followed by a sigmoid function to compute the probability score vectors $\mathbf{p}^n$.

Intra-image Semantic Transfer

There exist strong co-occurrence correlations among semantic labels in real-world images, and these correlations can effectively guide transferring semantic knowledge of known labels to generate pseudo labels for unknown labels. Current work (Huynh and Elhamifar 2020) applies dataset-level statistical correlations to achieve this end. However, the statistical correlations are not appropriate for every image, and thus inevitably incur some incorrect labels. To avoid this problem, the IST module is proposed to learn the image-specific co-occurrence matrix and apply this matrix to complement unknown labels for the corresponding image.

Given the semantic feature vectors $[\mathbf{f}_1^n, \mathbf{f}_2^n, \cdots, \mathbf{f}_C^n]$ of input image I^n, we need to compute the co-occurrence probability for each category pair. For categories i and j, we first concatenate the feature vector $\mathbf{f}_i^n$ and $\mathbf{f}_j^n$, and then feed the concatenated features to compute their co-occurrence probability, formulated as

$$p_{i,j}^n = \phi_{intra}([\mathbf{f}_i^n, \mathbf{f}_j^n]), \tag{1}$$

where $\phi_{intra}(\cdot)$ is implemented by several stacked fully connected layers. We compute the probabilities for all pairs and

obtain a co-occurrence matrix $\mathbf{P}^n_{intra} \in \mathcal{R}^{C\times C}$. Then, we estimate pseudo labels for unknown labels based on the co-occurrence matrix and known labels. For category i that is not provided, we can compute its pseudo label by

$$\hat{y}^n_i = \mathbf{1}[(\sum_{\{j|y^n_j=1\}} p^n_{i,j} \cdot y^n_j) \geq \theta_{intra}], \tag{2}$$

where $\mathbf{1}[\cdot]$ is an indicator function whose value is 1 if the argument is positive and is 0 otherwise. θ_{intra} is a threshold that helps to exclude the unlikely labels. We compute the pseudo labels for all unknown labels and combine it with known labels, obtaining $\hat{\mathbf{y}}^n = \{\hat{y}^n_1, \hat{y}^n_2, \cdots, \hat{y}^n_C\}$.

Formally, the co-occurrence prediction can be considered as a binary classification task, and we can train it using the binary cross entropy (BCE) loss. However, it is very difficult to train the co-occurrence predictor because positive and negative pairs are extremely imbalanced. To address this task, we introduce the asymmetric loss (Ben-Baruch et al. 2020) that dynamically down-weights the importance of easy negative pair, defined as

$$\mathcal{L}_{ist} = \sum_{n=1}^{N} \sum_{\{i,j\}} \ell^n_{i,j}, \tag{3}$$

where

$$\ell^n_{i,j} = \begin{cases} (1-p^n_{i,j})^{\gamma_1} \log(p^n_{i,j}) & \{i,j\} \in \mathcal{D}^n \\ (p^n_{i,j}-m)^{\gamma_2} \log(1-p^n_{i,j}) & \{i,j\} \notin \mathcal{D}^n. \end{cases} \tag{4}$$

Here, $\mathcal{D}^n$ is the set of label pairs that co-occur in image I^n. γ_1, γ_2, and m are the parameters to balance the loss and they are empirically set to 1, 2, and 0.05.

Cross-image Semantic Transfer

It is intuitive that the objects of the same category in different images share similar visual appearance. In other words, if two images share similar visual features, they tend to have the same labels. In the context of multi-label images, it is difficult to mine label correlation via image-level feature similarities. In this work, we design the CST module to learn category-level feature similarities and transfer known labels of images with high similarities to help complement unknown labels.

For each category c of images I^n and I^m, we use the cosine distance to compute their similarity, formulated as

$$s^{n,m}_c = cosine(\mathbf{f}^n_c, \mathbf{f}^m_c) = \frac{\mathbf{f}^n_c \cdot \mathbf{f}^m_c}{||\mathbf{f}^n_c|| \cdot ||\mathbf{f}^m_c||}. \tag{5}$$

Suppose the label of category c is missing in image I^n, we select image set $\mathcal{D}_c = \{m|y^m_c = 1\}$, in which every image has positive label c. We first compute the average similarities s^n_c between $\mathbf{f}^n_c$ and the correspond feature vectors of the images in $\mathcal{D}_c$, and then estimate the existence of category c by

$$\tilde{y}^n_c = \mathbf{1}[(\frac{1}{|\mathcal{D}_c|} \sum_{\{m\in\mathcal{D}_c\}} s^{n,m}_c \cdot y^m_c) \geq \theta_{cross}]. \tag{6}$$

Similarly, $\mathbf{1}[\cdot]$ is an indicator function and θ_{cross} is a threshold. We also estimate the pseudo labels for all unknown labels and combine it with known labels, obtaining $\tilde{\mathbf{y}}^n = \{\tilde{y}^n_1, \tilde{y}^n_2, \cdots, \tilde{y}^n_C\}$.

It is expected that the similarity between $\mathbf{f}^n_c$ and $\mathbf{f}^m_c$ tends to be high if images I^n and I^m have the same positive label c, and the similarity should be low otherwise. Thus, it can be formulated as a ranking task and we introduce a pair loss for training, formulated as

$$\mathcal{L}_{cst} = \sum_{n=1}^{N} \sum_{m=1}^{N} \sum_{c=1}^{C} \ell^{n,m}_c, \tag{7}$$

where

$$\ell^{n,m}_c = \begin{cases} 1 - s^{n,m}_c & y^n_c = 1, y^m_c = 1 \\ 1 + s^{n,m}_c & otherwise. \end{cases} \tag{8}$$

Optimization

We follow previous work to use the partial binary cross entropy loss as objective function. Specifically, given the predicted probability distribution $\mathbf{p}^n = \{p^n_1, p^n_2, \cdots p^n_C\}$ and the ground truth, the objective function can be defined as

$$\begin{aligned} \ell(\mathbf{p}^n, \mathbf{y}^n) = &\frac{1}{\sum_{c=1}^{C} |y^n_c|} \sum_{c=1}^{C} [\mathbf{1}(y^n_c = 1) \log(p^n_c) \\ &+ \mathbf{1}(y^n_c = -1) \log(1 - p^n_c)]. \end{aligned} \tag{9}$$

We define similar objective functions for the pseudo labels generated by the intra-image and cross-image semantic transfer modules, i.e., $\ell(\mathbf{p}^n, \hat{\mathbf{y}}^n)$ and $\ell(\mathbf{p}^n, \tilde{\mathbf{y}}^n)$. And the final classification loss is defined as summing the three losses over all samples, formulated as

$$\mathcal{L}_{cls} = \sum_{n=1}^{N} (\ell(\mathbf{p}^n, \mathbf{y}^n) + \ell(\mathbf{p}^n, \hat{\mathbf{y}}^n) + \ell(\mathbf{p}^n, \tilde{\mathbf{y}}^n)). \tag{10}$$

The final loss can be defined as summing up the classification loss, the intra-image and cross-image losses

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{ist} + \lambda_2 \mathcal{L}_{cst}. \tag{11}$$

Here, λ_1 and λ_2 are the balance parameters that ensure the three losses have comparable magnitude, so that we set λ_1 and λ_2 to 10.0 and 0.05 in the experiments.

Experiments

Experimental Settings

Datasets. We follow previous works (Durand, Mehrasa, and Mori 2019) to conduct experiments on the MS-COCO (Lin et al. 2014), Visual Genome (Krishna et al. 2016), and Pascal VOC 2007 (Everingham et al. 2010) datasets for evaluation. MS-COCO contains about 120k images that cover 80 daily-life categories. It is further divided into a training set of about 80k images and a validation set of about 40k images. Visual Genome contains 108,249 images and covers 80,138 categories. Since most categories have very few samples, we merely consider the 200 most frequent categories, resulting in a VG-200 subset. We randomly select 10,000 images as the test set and the rest 98,249 images as the training set.

Datasets	Methods	10%	20%	30%	40%	50%	60%	70%	80%	90%	Ave. mAP
MS-COCO	SSGRL	62.5	70.5	73.2	74.5	76.3	76.5	77.1	77.9	78.4	74.1
	GCN-ML	63.8	70.9	72.8	74.0	76.7	77.1	77.3	78.3	78.6	74.4
	KGGR	66.6	71.4	73.8	76.7	77.5	77.9	78.4	78.7	79.1	75.6
	Curriculum labeling	26.7	31.8	51.5	65.4	70.0	71.9	74.0	77.4	78.0	60.7
	Partial BCE	61.6	70.5	74.1	76.3	77.2	77.7	78.2	78.4	78.5	74.7
	Ours	**68.1**	**73.5**	**75.9**	**77.3**	**78.1**	**78.9**	**79.2**	**79.6**	**79.9**	**76.7**
VG-200	SSGRL	34.6	37.3	39.2	40.1	40.4	41.0	41.3	41.6	42.1	39.7
	GCN-ML	32.0	37.8	38.8	39.1	39.6	40.0	41.9	42.3	42.5	39.3
	KGGR	36.0	**40.0**	**41.2**	41.5	42.0	42.5	**43.3**	**43.6**	**43.8**	41.5
	Curriculum labeling	12.1	19.1	25.1	26.7	30.0	31.7	35.3	36.8	38.5	28.4
	Partial BCE	27.4	38.1	40.2	40.9	41.5	42.1	42.4	42.7	42.7	39.8
	Ours	**38.8**	39.4	41.1	**41.8**	**42.7**	**42.9**	43.0	43.2	43.5	**41.8**
Pascal VOC 2007	SSGRL	77.7	87.6	89.9	90.7	91.4	91.8	92.0	92.2	92.2	89.5
	GCN-ML	74.5	87.4	89.7	90.7	91.0	91.3	91.5	91.8	92.0	88.9
	KGGR	81.3	88.1	89.9	90.4	91.2	91.3	91.5	91.6	91.8	89.7
	Curriculum labeling	44.7	76.8	88.6	90.2	90.7	91.1	91.6	91.7	91.9	84.1
	Partial BCE	80.7	88.4	89.9	90.7	91.2	91.8	92.3	92.4	92.5	90.0
	Ours	**81.5**	**89.0**	**90.3**	**91.0**	**91.6**	**92.0**	**92.5**	**92.6**	**92.7**	**90.4**

Table 1: Performance of our SST framework and current state-of-the-art competitors for MLR-PL on the MS-COCO, VG-200 and Pascal VOC 2007 datasets. The best results are highlighted in bold.

Pascal VOC 2007 is the most widely used dataset for multi-label evaluation. It contains about 10k images from 20 object categories, which is divided into a trainval set of about 5,011 images and a test set of 4,952 images.

Because the three datasets are fully annotated, we randomly drop some labels to create the training set with partial labels. In this work, the proportion of dropped labels varies from 10% to 90%, resulting in 90% to 10% known labels.

Evaluation Metric. For a fair comparison, we follow current works (Durand, Mehrasa, and Mori 2019; Huynh and Elhamifar 2020) to adopt the mean average precision (mAP) over all categories for evaluation under different proportions of known labels. The proportions are set to 10%, 20%, $\cdots$, 90%. And we also computer average mAP over all proportions. The overall and per-class precision, recall, F1-measure are also widely used to evaluate multi-label image recognition (Chen et al. 2019a) and we also adopt these metrics for more comprehensive evaluation. We present these results in the supplementary material due to the page limit.

Implementation Details. To fairly compare with existing algorithms, we follow previous works (Durand, Mehrasa, and Mori 2019; Chen et al. 2019a) to adopt the 101-layers ResNet (He et al. 2016) as the backbone to extract features $\mathbf{f}^n$. Then, we use exactly the same decoupling module to learn category-specific semantic representation and gated graph neural network to learn contextualized category-specific feature vectors as (Chen et al. 2019a). The co-occurrence estimation function $\phi_{intra}(\cdot)$ is implemented by three fully-connected layers, in which the first layer maps 1024-dimension vector to 512 followed by the ReLU function, the second layer maps 512-dimension vector to 1,024 also followed by ReLU, and the last layer maps to a score that indicates the co-occurrence probability. The proposed framework is trained using the loss $\mathcal{L}$ as shown in Equation 11. The parameters of the ResNet-101 are initialized by those pre-trained on the ImageNet (Deng et al. 2009) dataset and the parameters of all other layers are initialized randomly. The model is trained using the ADAM algorithm (Kingma and Ba 2015) with a batch size of 32, momentums of 0.999 and 0.9, and a weight decay of 5×10^{-4}. The original learning rate is set to 0.00001, and it is divided by 10 for every 10 epochs. It is trained with 20 epochs in total. During training, the input image is resized to 512×512, and we randomly choose a number from {512, 448, 384, 320, 256} as the width and height to crop patch. Finally, the cropped patch is further resized to 448×448. Then we perform randomly horizontal flip and perform normalization. θ_{intra} and θ_{inter} are two crucial parameters that control the accuracy of the generated pseudo labels. In the training process, the parameters are set to 1 during the first 5 epochs to avoid incurring any pseudo labels. Then, they are set to 0.95 at epoch 6 and are decreased by 0.025 for every epoch until they reach the minimum θ_{intra} and θ_{inter}, respectively. Both the minimum θ_{intra} and θ_{inter} are set to 0.75 based on the experimental results. During inference, the intra-image and cross-image semantic transfer modules are removed, and the image is resized to 448×448 for evaluation.

Comparison with the State-of-the-art Algorithms

To evaluate the effectiveness of the proposed SST framework, we compare it with the following algorithms that can be classified into three folds: 1) **SSGRL** (Chen et al. 2019a), **GCN-ML** (Chen et al. 2019c) and **KGGR** (Chen et al. 2020) introduce graph neural networks to model label dependencies and they achieve state-of-the-art performance on the traditional multi-label image recognition task. We adapt these three methods to address the multi-label recognition with partial labels by replacing the loss with partial BCE loss while keeping other component unchanged. 2) **Curriculum labeling** (Durand, Mehrasa, and Mori 2019) alternately labels the unknown labels with high evidence to update the training set and retrains the model with updated training set.

Methods	10%	20%	30%	40%	50%	60%	70%	80%	90%	Ave. mAP
SSGRL	62.5	70.5	73.2	74.5	76.3	76.5	77.1	77.9	78.4	74.1
Ours IST w/ stat	55.3	62.3	65.9	70.3	71.8	72.7	73.5	74.6	75.2	69.1
Ours IST	64.1	71.3	74.5	75.9	77.2	77.7	78.2	78.8	79.1	75.2
Ours IST w/o L_{ist}	61.9	70.9	73.2	75.0	76.3	76.8	77.6	78.2	78.6	74.3
Ours CST	64.2	72.5	74.4	76.2	77.1	77.9	78.4	78.9	79.3	75.4
Ours CST w/o L_{cst}	63.0	71.7	73.8	74.4	76.3	76.9	77.6	78.3	78.6	74.5
Ours w/ SAM	67.8	73.2	75.3	77.5	78.3	78.6	79.0	79.4	79.7	76.5
Ours	**68.1**	**73.5**	**75.9**	**77.3**	**78.1**	**78.9**	**79.2**	**79.6**	**79.9**	**76.7**

Table 2: Comparison of mAP of the baseline SSGRL, our framework merely using IST with statistical co-occurrence (Ours IST w/ stat), our framework merely using IST (Ours IST), our framework merely using IST without loss L_{ist} (Ours IST w/o L_{ist}), our framework merely using CST (Ours CST), our framework merely using CST without loss L_{cst} (Ours CST w/o L_{cst}), our framework using SAM instead of SD (Ours w/ SAM) and our framework (Ours) on the MS-COCO dataset.

We also treat it as a strong baseline to address this task. 3) **Partial BCE** (Durand, Mehrasa, and Mori 2019) is the most-recent algorithm that is proposed to address this task. It introduces a normalized BCE loss to better exploit partial labels to train the multi-label models. We also include this algorithm for comparison. For fair comparisons, we adopt the same ResNet-101 network as backbone and follow exactly the same train/val split settings.

Performance on MS-COCO We present the comparison results on the MS-COCO dataset as shown in Table 1. We find the traditional multi-label recognition methods SSGRL and GCN-ML can achieve competitive performance when the proportion of known labels is high (e.g., 70%-90%), but suffer from obvious performance drop when the proportion decreases. Partial BCE can achieve competing performance even when the proportion decreases to 30%. By introducing the intra-image and cross-image correlations to generate pseudo labels, our SST framework obtains the best performance for all the settings of different proportions of known labels. Specifically, it obtains the mAPs of 68.1%, 73.5%, 75.9%, 77.3%, 78.1%, 78.9%, 79.2%, 79.6%, 79.9% on the settings of 10%-90% known labels, outperforming the second-best KGGR algorithm by 1.5%, 2.1%, 2.1%, 0.6%, 0.6%, 1.0%, 0.8%, 0.9%, 0.8%, respectively. It is worth noting that SST can achieve more obvious performance improvement when the known labels are small, e.g., mAP improvement of 1.5%, 2.1% when the known label proportions are 10% and 20%.

Performance on VG-200 VG-200 is a more challenging dataset that covers a lot more categories, and we also present the comparison results. As shown in Table 1, our SST framework obtains the best performance over all proportion settings. Specifically, its average mAP is 41.8%, outperforming the second-best KGGR algorithm by 0.3%. Besides, it outperforms leading multi-label methods SSGRL and GCN-ML by 4.2% and by 6.8% when known labels are 10%.

Performance on Pascal VOC 2007 Pascal VOC is the most-widely used dataset for multi-label recognition and we also present the results in Table 1. As this dataset merely covers 20 categories and it is more simple than Visual Genome and MS-COCO, current algorithms achieve comparable results when keeping a certain proportion of known labels (e.g., more than 40%). But their performances drop dramatically when the proportion decreases to 10% and 20%. Our SST framework also suffers performance drop, but it consistently outperforms current methods for all proportion settings. Specifically, it outperforms multi-label methods (i.e., SSGRL and GCN-ML) by 3.8% and by 7.0% and Partial BCE by 0.8% when known labels are merely 10%.

Ablative Studies

As discussed above, SSGRL can be treated as the baseline method and we stress the comparison with SSGRL to verify the contribution of the structured semantic transfer module. As shown in Table 2, the SSGRL obtains an average mAP of 74.1%. By introducing the structured semantic transfer module to complement the unknown labels, SST boosts the average mAP to 76.7%, with an improvement of 2.6%. And SST also performs consistently better than the baseline SSGRL method on different proportion settings.

The SST framework depends on the semantic-aware representation learning (SARL) module to learn semantic-aware representation. In this work, we use the semantic decoupling algorithm proposed in (Chen et al. 2019a) to implement this module, because it achieves the state-of-the-art performance for the multi-label recognition task. It is noteworthy that we can also use other algorithms for learning semantic-aware representation. To verify this point, we replace the semantic decoupling algorithm with the semantic attention module (SAM) proposed in (Ye et al. 2020) to generate category-specific representation. As shown in Table 2, "Ours w/ SAM" can also achieve comparable performance, suggesting the universality of the proposed SST framework.

Since SST consists of two complementary modules, i.e., intra-image semantic transfer (IST) and cross-image semantic transfer (CST) modules, in the following we will conduct more ablative experiments to analyze the separate contributions of these two modules in detail.

Analysis of Intra-image Semantic Transfer (IST) Contribution of the IST module. We then evaluate the actual contribution of the IST module by comparing the performance with and without this module. As shown in Table 2, "Ours IST" means we merely use the IST module to generate pseudo labels. We find it achieves obvious mAP improvement compared with the baseline SSGRL, i.e., an aver-

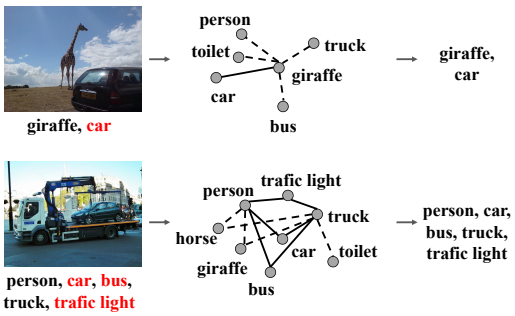

Figure 3: Examples of the image-specific co-occurrence matrices and the complemented labels: input images and labels (left), partial graphs of within-image co-occurrence (middle), and pseudo positive labels (right). The missing labels are highlighted in red. Two categories that have high co-occurrence probability are connected by solid line, otherwise, connected by a dotted line.

age mAP improvement of 1.1%. Besides, the loss L_{ist} helps to learn accurate co-occurrence matrix. To evaluate its effectivenss, we conduct experiments that remove this loss for comparison (namely Ours IST w/o L_{ist}). As shown in Table 2, it further decreases the average mAP by 0.9%.

Here, we learn image-specific co-occurrence matrix to generate pseudo labels. To demonstrate its effectiveness, we perform experiments that use statistical co-occurrence matrix computed on the training dataset to generate the pseudo labels, namely "Ours IST w/ stat". As shown in Table 2, it suffers from dramatic performance drop. Specifically, the average mAP is merely 69.1%, worse than that using image-specific co-occurrence matrix by 6.1% in average mAP. One reason for this phenomenon is that statistical co-occurrence is not suitable for every image and thus it may incur many false positive labels for the unsuitable images.

To delve deep into the IST module, we visualize some examples of the image-specific co-occurrence matrices and how these matrices generate the pseudo labels in Figure 3. As shown, it can capture the category pairs that frequently co-occur, like *car* and *person* in the second example. It can also assign a high co-occurrence probability to the pairs that rarely co-exist, e.g., *giraffe* and *car* in the first example. This also suggests that learning image-specific co-occurrences can better capture label correlations for each image and thus facilitate generating more accurate pseudo labels.

Analysis of Cross-image Semantic Transfer (CST) Contribution of the CST module. In this section, we add the CST module to the baseline SSGRL, namely "Ours CST", and compare it with baseline method to verify the contribution of CST. As shown in Table 2, it shows that adding the CST module improves the average mAP from 74.1% to 75.4%, an improvement of 1.3%. In this module, the loss L_{cst} plays an important role in learning category-specific feature similarity. Here, we also evaluate its contribution by conducting an experiment that removes this loss (namely Ours CST w/o L_{cst}). It is observed that the average mAP decreases from 75.4% to 74.5%.

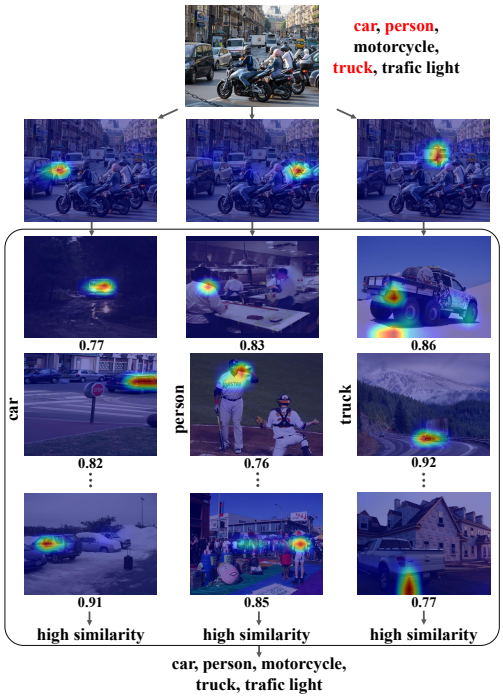

Figure 4: An example of category-specific feature similarities and the complemented labels: input image and labels (top), category-specific feature vectors (middle top), category-specific feature vectors of other images with known labels that are missing for the given image (middle in box), and the generated pseudo labels (bottom). The missing labels are highlighted in red.

As discussed above, the CST module measures category-level feature similarity of the same category from different images to help complement the unknown labels. Here, we also visualize an example that loses labels of *car*, *person*, and *truck* (Figure 4). We can see that the features belonging to the same category but from different image share very high similarities, which help to recall the missing labels.

Conclusion

In this work, we propose a novel structured semantic transfer framework, which consists of an intra-image semantic transfer module that mines image-specific label co-occurrences and a cross-image semantic transfer module that mines category-level feature similarities, to transfer semantics of known labels to complement unknown labels for model training. We conduct extensive experiments on various multi-label datasets (e.g., MS-COCO, VG-200, and Pascal VOC) to demonstrate its superiority.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61876045, 61836012 and 62002069), the Natural Science Foundation of Guangdong Province (No. 2017A030312006) and Guangdong Provincial Basic Research Program (No. 102020369).

References

Ba, J.; Mnih, V.; and Kavukcuoglu, K. 2014. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Ben-Baruch, E.; Ridnik, T.; Zamir, N.; Noy, A.; Friedman, I.; Protter, M.; and Zelnik-Manor, L. 2020. Asymmetric Loss For Multi-Label Classification. *arXiv preprint arXiv:2009.14119*.

Bucak, S. S.; Jin, R.; and Jain, A. K. 2011. Multi-label learning with incomplete class assignments. In *CVPR 2011*, 2801–2808.

Cabral, R. S.; Torre, F.; Costeira, J. P.; and Bernardino, A. 2011. Matrix completion for multi-label image classification. In *Advances in neural information processing systems*, 190–198.

Chen, T.; Lin, L.; Chen, R.; Wu, Y.; and Luo, X. 2018a. Knowledge-Embedded Representation Learning for Fine-Grained Image Recognition. In *IJCAI*, 627–634.

Chen, T.; Lin, L.; Hui, X.; Chen, R.; and Wu, H. 2020. Knowledge-Guided Multi-Label Few-Shot Learning for General Image Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, T.; Pu, T.; Wu, H.; Xie, Y.; Liu, L.; and Lin, L. 2021. Cross-domain facial expression recognition: A unified evaluation benchmark and adversarial graph learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, T.; Wang, Z.; Li, G.; and Lin, L. 2018b. Recurrent Attentional Reinforcement Learning for Multi-label Image Recognition. In *Proc. of AAAI Conference on Artificial Intelligence*, 6730–6737.

Chen, T.; Xu, M.; Hui, X.; Wu, H.; and Lin, L. 2019a. Learning Semantic-Specific Graph Representation for Multi-Label Image Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, 522–531.

Chen, T.; Yu, W.; Chen, R.; and Lin, L. 2019b. Knowledge-Embedded Routing Network for Scene Graph Generation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6163–6171.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019c. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5177–5186.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Durand, T.; Mehrasa, N.; and Mori, G. 2019. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 647–657.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huynh, D.; and Elhamifar, E. 2020. Interactive multi-label CNN learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9423–9432.

Joulin, A.; Van Der Maaten, L.; Jabri, A.; and Vasilache, N. 2016. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, 67–84. Springer.

Kapoor, A.; Viswanathan, R.; and Jain, P. 2012. Multilabel Classification using Bayesian Compressed Sensing. In *Proceedings of Advances in Neural Information Processing Systems*, 2654–2662.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; Bernstein, M.; and Fei-Fei, L. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations.

Li, Y.; Tarlow, D.; Brockschmidt, M.; and Zemel, R. S. 2016. Gated Graph Sequence Neural Networks. In *International Conference on Learning Representations*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 740–755.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (LCLR)*.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, 843–852.

Tsoumakas, G.; and Katakis, I. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3): 1–13.

Wang, J.; Yang, Y.; Mao, J.; Huang, Z.; Huang, C.; and Xu, W. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, 2285–2294.

Wang, Q.; Shen, B.; Wang, S.; Li, L.; and Si, L. 2014. Binary codes embedding for fast image tagging with incomplete labels. In *European Conference on Computer Vision*, 425–439. Springer.

Wang, Z.; Chen, T.; Li, G.; Xu, R.; and Lin, L. 2017. Multi-label Image Recognition by Recurrently Discovering Attentional Regions. In *ICCV*, 464–472.

Wei, Y.; Xia, W.; Lin, M.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; and Yan, S. 2016. HCP: A flexible CNN framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9): 1901–1907.

Wu, B.; Lyu, S.; and Ghanem, B. 2015. ML-MG: Multi-label Learning with Missing Labels Using a Mixed Graph. In *Proceedings of IEEE International Conference on Computer Vision*, 4157–4165.

Xu, M.; Jin, R.; and Zhou, Z.-H. 2013. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in neural information processing systems*, 2301–2309.

Yang, H.; Tianyi Zhou, J.; Zhang, Y.; Gao, B.-B.; Wu, J.; and Cai, J. 2016. Exploit bounding box annotations for multi-label object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 280–288.

Ye, J.; He, J.; Peng, X.; Wu, W.; and Qiao, Y. 2020. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *ECCV*, 649–665.

Yu, H.-F.; Jain, P.; Kar, P.; and Dhillon, I. 2014. Large-scale multi-label learning with missing labels. In *International conference on machine learning*, 593–601.