Label-Aware Global Consistency for Multi-Label Learning with Single Positive Labels

Anonymous Author(s)
Affiliation
Address
email

Abstract

In single positive multi-label learning (SPML), only one of multiple positive labels is observed for each instance. The previous work trains the model by simply treating unobserved labels as negative ones, and designs the regularization to constrain the number of expected positive labels. However, in many real-world scenarios, the true number of positive labels is unavailable, making such methods less applicable. In this paper, we propose to solve SPML problems by designing a Label-Aware global Consistency (LAC) regularization, which leverages the manifold structure information to enhance the recovery of potential positive labels. On one hand, we first perform pseudo-labeling for each unobserved label based on its prediction probability. The consistency regularization is then imposed on model outputs to balance the fitting of identified labels and exploring of potential positive labels. On the other hand, by enforcing label-wise embedding to maintain global consistency, LAC loss encourages the model to learn a more distinctive representation, which benefits for recovering the information of potential positive labels. Experiments on multiple benchmark datasets validate that the proposed method can achieve state-of-the-art performance for solving SPML tasks.

1 Introduction

Multi-label learning (MLL) is a practical and effective learning framework for tackling objects with complex semantics, where each instance is assumed to be associated with multiple class labels. For example, an image may be annotated with labels *beach*, *sea* and *sky* simultaneously. Multi-label learning aims to train a classifier that can accurately predict all the relevant labels for unseen instances. It has been successfully applied into many real-world applications, such as image annotation [Chen et al., 2019], scene understanding [Shao et al., 2015], facial attribute recognition [Hand et al., 2018].

Given the output space is exponentially larger than that of single-label learning, it often requires a large number of examples with precise annotations to train an effective MLL classifier. However, in many real-world scenarios, it is difficult and costly to collect the precise annotations. To reduce the labeling cost, an alternative solution is to ask the annotators to assign the most obvious label to each instance, leading to only a single positive label available. Such problem has been formulated as a learning framework called single positive multi-label learning (SPML) [Cole et al., 2021]. In this case, except for the annotated positive label, each of other labels could be positive or negative but unknown, thus the learning task becomes much more challenging due to the lack of supervision.

To solve SPML problems, the previous work [Cole et al., 2021] intuitively designs the "assume negative" (AN) loss, which treats all unobserved labels as negative ones, and train an MLL classifier with the conventional binary cross entropy (BCE) loss. Unfortunately, AN loss often introduces many false negative labels, which may lead to a noticeable decrease in performance of the model. To alleviate the negative impact of false negative labels, authors further propose a regularization to

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

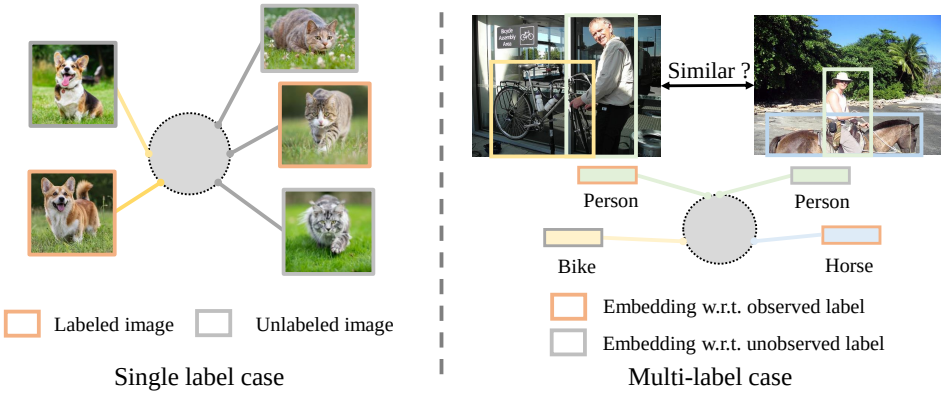


Figure 1: An illustration of the comparison between clustering regularization and our proposed label-aware global consistency regularization.

constrain the number of expected positive labels. However, since the true number of positive labels cannot be accessible in many real-world scenarios, it often requires additional estimation or elaborate hyper-parameter tuning to determine its true value.

Clustering assumption [Zhou et al., 2003] is first proposed to solve semi-supervised learning (SSL) problems. It aims to recover the true labels for unlabeled examples by utilizing the global consistency, *i.e.*, encouraging points on the same structure (typically referred to as a clustering) to be assigned with the same label. The idea of clustering assumption has been successfully used to solve many weakly-supervised learning problems [Snell et al., 2017, Wang et al., 2022]. Unfortunately, applying clustering regularization to solve SPML problems is not as straightforward as in single-label case, *i.e.*, multi-class classification, since it is difficult to measure the underlying similarity between any two examples with multiple labels. For example, in single-label case (see the left side in Figure 1), it is feasible to distinguish examples from different classes by measuring their similarity based on image-level feature representations. While in multi-label case (see the right side in Figure 1), it is hard to measure the similarity between any two images share the same object (*e.g.*, person), while still have individual objects (*e.g.*, bike and horse).

In this paper, we design a Label-Aware global Consistency (LAC) regularization to leverage the manifold structure information for solving SPML problems. On one hand, we first identify the potential positive labels mislabeled by AN loss based on class probabilities predicted by the model. The pseudo-labeling consistency regularization is then used for training model to boost its ability on identifying the potential positive labels. On the other hand, as illustrated in the right side of Figure 1, the LAC loss is designed to encourage the global consistency among label-wise embeddings, *i.e.*, pulling together intra-class embeddings while pulling apart inter-class embeddings. By leveraging the manifold structure information, the LAC loss enhances the distinctiveness of the learned feature representation, which is beneficial for identifying the potential positive labels. Extensive experimental results on multiple benchmark datasets demonstrate that our method can achieve state-of-the-art performance.

2 Related Work

There are a plenty of studies in multi-label learning literature. Existing methods can be roughly classified into three groups. Considering the traditional binary cross entropy (BCE) loss often suffers from the positive-negative imbalance issue, the first kind of methods aim to design specific loss functions to alleviate this issue. Among them, asymmetric loss (ASL) [Ridnik et al., 2021a] is a representative method that dynamically down-weights and hard-thresholds easy negative examples. The second kind of methods are proposed to capture the label correlations, which are regarded as an essential element for improving the practical performance of MLL. ML-GCN [Chen et al., 2019] and its variant [Wang et al., 2020] utilize the graph convolution network (GCN) to capture the co-occurrence correlations among labels. Considering an image often contains multiple objects,

the last kind of methods are designed to locate areas of interest related to semantic labels by using attention technique.

Traditional multi-label learning assumes that each instance has been annotated with all of its relevant labels. However, in many real-world scenarios, it is costly and difficult to precisely assign an image with multiple labels. To solve this issue, the current effort is mainly dedicated to reduce the labeling cost by training an MLL model with only partial labels. For example, Durand et al. [2019] fine-tune the pre-trained model on ImageNet with partial annotations and perform pseudo-labeling to recover the potential positive labels. In Huynh and Elhamifar [2020], authors propose a two-stage interactive learning framework that performs similarity learning and CNN training interactively to improve the performance of each another. Pu et al. [2022] propose an unified framework to exploit instance-level and prototype-level semantic representation for recovering the potential positive labels.

SPML is the extreme case of multi-label learning with partial labels, where only one of multiple potential positive labels can be observed. The earliest work intuitively treats all unobserved labels as negative ones, and trains an MLL model by introducing the regularization terms that alleviate the negative impact of false negative labels [Cole et al., 2021]. The spatial consistency [Verelst et al., 2022] is designed to avoid the bias toward negative prediction by maintaining the consistency of classification maps between the network and exponential moving average (EMA) model. Unlike aforementioned methods, Zhou et al. [2022] treat unobserved labels as unknown and propose to maximize the entropy of predicted probabilities for unobserved labels. Furthermore, authors combine the asymmetric pseudo-labeling and self-paced strategy to obtain more accurate pseudo labels.

3 The Proposed Method

In the SPML problem, let $\boldsymbol{x} \in \mathcal{X}$ be a feature vector and $\hat{\boldsymbol{y}} \in \mathcal{Y}$ be its corresponding labels, where $\mathcal{X} = \mathbb{R}^d$ is the feature space and $\mathcal{Y} = \{0, 1\}^q$ is the target space with q possible class labels. Here, $\hat{y}_j = 1$ indicates the j-th label is the only observed positive label for instance $\boldsymbol{x}$ while $\hat{y}_j = 0$ indicates the j-th label cannot be observed. Since only a single positive label can be observed, we have $\sum_{j=1}^{q} \hat{y}_{ij} = 1$ for every instance $\boldsymbol{x}_i$. We further denote by $\boldsymbol{y}$ the true label vector for instance $\boldsymbol{x}$. Let $p(\boldsymbol{y}|\boldsymbol{x})$ be the predicted probability distribution over classes and $p(y_j|\boldsymbol{x})$ be the predicted probability of the j-th class for input $\boldsymbol{x}$. We use $[q]$ to denote the integer set $\{1, ..., q\}$.

Label-wise Embedding Model To obtain the label-wise embedding with respect to every label, existing methods can be roughly divided into two groups, *i.e.*, global-average-pooling (GAP) based methods Wang et al. [2017], Guo et al. [2019] and attention-based methods [Liu et al., 2021, Ridnik et al., 2021b, Dao et al., 2021]. For simplicity, we adopt the attention-based method to construct the label-wise embedding model . Assume that the label-wise embedding model is composed of a label-wise embedding encoder f, which can generate a high-dimensional label-wise feature representation $f_j(\boldsymbol{x})$ with respect to the j-th label of instance $\boldsymbol{x}$, and a classification head h, which can output the predicted probability of the j-th class $p(y_j|\boldsymbol{x}) = h(f_j(\boldsymbol{x}))$. We further denote by $g(\cdot)$ a non-linear projection head, which transforms a high-dimensional embedding $f_j(\boldsymbol{x})$ to a low-dimensional embedding $\boldsymbol{z}_j = g(f_j(\boldsymbol{x}))$.

In traditional multi-label classification, the most commonly used loss function is binary cross entropy (BCE) loss, which decomposes the original task into multiple binary classification problems. Formally, given a batch of examples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{b}$, the BCE loss can be defined as follows:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{b} \sum_{i=1}^{b} \sum_{j=1}^{q} y_{ij} \log(p(y_j|\boldsymbol{x}_i)) + (1 - y_{ij}) \log(1 - p(y_j|\boldsymbol{x}_i)) \tag{1}$$

The assume negative (AN) loss However, the conventional BCE loss cannot be directly applied to solve SPML problems, where only a single positive label can be accessible for every instance. Training the neural network with only one positive label often makes model collapse to a trivial solution. To mitigate this issue, an intuitive method is to treat all unobserved labels as negative ones, and then the neural network can be trained with the resulting AN loss [Cole et al., 2021] given a batch of SPML examples $\{(\boldsymbol{x}_i, \hat{\boldsymbol{y}}_i)\}_{i=1}^{b}$:

$$\mathcal{L}_{\text{AN}} = -\frac{1}{b} \sum_{i=1}^{b} \sum_{j=1}^{q} \hat{y}_{ij} \log(p(y_j|\boldsymbol{x}_i))) + (1 - \hat{y}_{ij}) \log(1 - p(y_j|\boldsymbol{x}_i))) \tag{2}$$

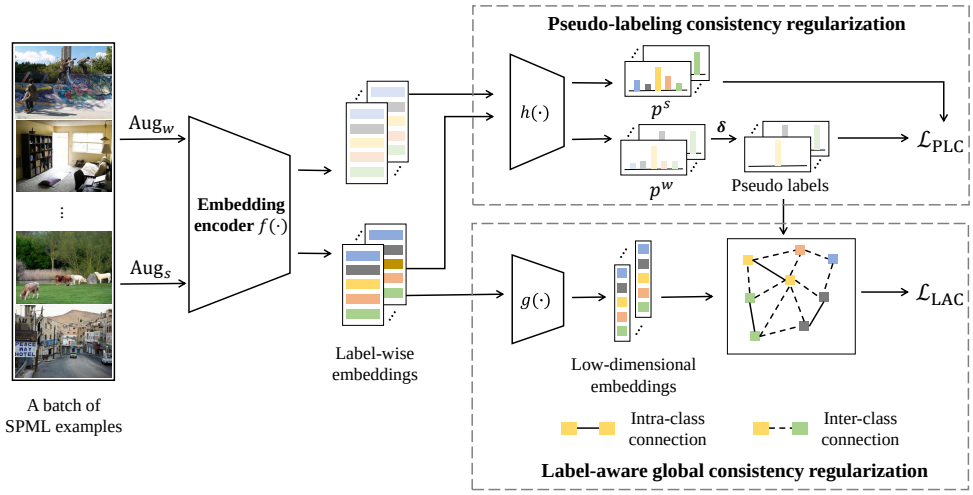


Figure 2: An illustration of the proposed learning framework.

Unfortunately, AN loss often introduces a large number of false negative labels, which leads to a noticeable decrease in generalization performance of the model. To alleviate the negative impact of false negatives, we design the following two regularization losses: 1) the pseudo-labeling consistency regularization $\mathcal{L}_{\text{PLC}}$ to identify the potential positive labels mislabeled by AN loss, 2) the label-aware global consistency loss $\mathcal{L}_{\text{LAC}}$ to leverage the manifold structure information for learning a more distinctive feature representation. The overall objective function can be defined as:

$$\mathcal{L} = \mathcal{L}_{\text{AN}} + \lambda_{\text{PLC}}\mathcal{L}_{\text{PLC}} + \lambda_{\text{LAC}}\mathcal{L}_{\text{LAC}} \tag{3}$$

Figure 2 provides an illustration of the LAC learning framework. For a batch of SPML examples, we generate the pseudo labels based on their weakly-augmented versions. On one hand, the consistency regularization is conducted between the pseudo labels and model outputs on the strongly-augmented versions to prevent from over-fitting to false negative labels. On the other hand, the label-aware global consistency is used to learn a more compact feature representation such that intra-class embeddings are pushed together while inter-class embeddings are pushed away. In the following content, we will introduce these two regularization losses in detail.

3.1 Pseudo-Labeling Consistency Regularization

In this section, we propose the pseudo-labeling consistency regularization for recovering the ground-truth labeling information of unobserved labels.

We define the loss $\mathcal{L}_{\text{PLC}}$ between the pseudo labels and the model predictions. Specifically, for the j-th label of $\boldsymbol{x}_i$, let $p_{ij}^w = p(y_j|\text{Aug}_w(\boldsymbol{x}_i))$ denote the pseudo labels, where $\text{Aug}_w(\cdot)$ is a stochastic weak augmentation. To avoid the trivial solution, let $p_{ij}^s = p(y_j|\text{Aug}_s(\boldsymbol{x}_i))$ denote the model prediction, where $\text{Aug}_s(\cdot)$ is a stochastic strong augmentation. Then, the pseudo-labeling consistency loss can be formulated as:

$$\mathcal{L}_{\text{PLC}} = \frac{1}{b}\sum_{i=1}^{b}\sum_{j=1}^{q}\mathbb{I}(p_{ij}^w \geq \delta|\hat{y}_{ij}=0)\log(p_{ij}^s) + \mathbb{I}(p_{ij}^w \leq \bar{\delta}|\hat{y}_{ij}=0)\log(1-p_{ij}^s) \tag{4}$$

where δ (or $\bar{\delta} = 1-\delta$) is a threshold that controls how many positive (or negative) pseudo labels would be used for model training. Here, $\mathbb{I}(\cdot)$ is the indicator function that output 1 if the condition satisfies while 0, otherwise, and the condition $\hat{y}_{ij} = 0$ is used to ensure the loss $\mathcal{L}_{\text{PLC}}$ is only imposed on the unobserved labels. Generally, the loss $\mathcal{L}_{\text{PLC}}$ imposes a consistency regularization on the model such that encourages it to output the same probability with respect to an unobserved label for an example even after it is strongly augmented. By using a much stronger augmentation, $\mathcal{L}_{\text{PLC}}$ would alleviate the over-fitting issue to false negative labels and improve the ability of identifying potential positive labels.

3.2 Label-Aware Global Consistency Regularization

As discussed in the above section, $\mathcal{L}_{\text{PLC}}$ can be regarded as a consistency regularization on the high-level representation, *i.e.*, the final output of network. With the increase of depth of the neural network, the impact to the network parameters of enforcing the consistency on the high-level representation may be gradually degraded, and thus weakens the effectiveness of the proposed loss $\mathcal{L}_{\text{PLC}}$. To further alleviate the harmfulness caused by the false negative labels, motivated by the clustering assumption [Zhou et al., 2003], we design label-aware global consistency regularization $\mathcal{L}_{\text{LAC}}$ to maintain global consistency among label-wise embeddings. The main idea is to pull the intra-class label-wise embeddings to be close while pull inter-class label-wise embeddings to be separated. The intuition behind the LAC loss is that by maintaining the global consistency among label-wise embeddings, LAC enhances the distinctiveness of learned representation, which benefits for recovering the true labeling information of potential positive labels.

To derive the $\mathcal{L}_{\text{LAC}}$ loss, for a batch of examples, we first use the label-wise embedding encoder f and projection head g to produce a set of embeddings $\mathcal{Z} = \{\boldsymbol{z}_k | 1 \leq k \leq B\}$, where $B = b \cdot q$ is the total number of embeddings for b instances with q classes. Then, we define the mapping $\boldsymbol{i}(\cdot)$ and $\boldsymbol{j}(\cdot)$ to bridge the embedding index k with its original instance index i and class index j. Specifically, for the embedding $\boldsymbol{z}_k$ with respect to j-th class of instance $\boldsymbol{x}_i$, we can locate its original instance index i and class index j by using the mapping, *i.e.*, $i = \boldsymbol{i}(k)$ and $j = \boldsymbol{j}(k)$.

In order to exploit as more as possible supervised information to learn a distinctive representation, we compute the $\mathcal{L}_{\text{LAC}}$ loss not only on unobserved labels but also on observed ones. To obtain a unified formulation, for the single positive label $\hat{y}_{ij} = 1$, we define its pseudo label as $p^w_{ij} = 1$ although it is actually the true label. By denoting $i = \boldsymbol{i}(k), j = \boldsymbol{j}(k)$ and $i' = \boldsymbol{i}(k'), j' = \boldsymbol{j}(k')$, we can construct the intra-class connection matrix I of size $B \times B$ to connect any two intra-class label-wise embeddings:

$$I_{kk'} = \begin{cases} 1 & \text{if } j = j' \text{ and } p^w_{ij} \geq \delta, p^w_{i'j'} \geq \delta \\ 0 & \text{otherwise} \end{cases} . \tag{5}$$

From the equation Eq.(5), every embedding with pseudo labels larger than the threshold δ would be treated as the anchor embedding and be connected with embeddings from the same class that satisfy the same condition. Since a low prediction probability indicates the corresponding label is likely to be negative, its embedding is invalid and would never be used.

Similarly, we can construct the inter-class connection matrix $\bar{I}$ of the same size to connect any two inter-class label-wise embeddings:

$$\bar{I}_{kk'} = \begin{cases} 1 & \text{if } j \neq j' \text{ and } p^w_{ij} \geq \delta, p^w_{i'j'} \geq \delta \\ 0 & \text{otherwise} \end{cases} . \tag{6}$$

From the equation Eq.(6), an anchor embedding is connected with embeddings from other classes with pseudo labels larger than a threshold δ.

Next, we can train the label-wise embedding encoder f and the projection head g by enforcing the global consistency among label-wise embeddings. It is expected to pull together an anchor label-wise embedding and its intra-class label-wise embeddings while separate the anchor label-wise embedding from its inter-class label-wise embeddings. To achieve this goal, given the connection matrices I and $\bar{I}$, we define label-aware global consistency loss as follows:

$$\mathcal{L}_{\text{LAC}} = -\frac{1}{B} \sum_{k=1}^{B} \sum_{k'=1}^{B} I_{kk'} \log \frac{\exp(\boldsymbol{z}_k \cdot \boldsymbol{z}_{k'}/\tau)}{\sum_{t=1}^{B} \bar{I}_{kt} \exp(\boldsymbol{z}_k \cdot \boldsymbol{z}_t/\tau)} \tag{7}$$

It is noteworthy that when $k = k'$, *i.e.*, an anchor embedding and itself, in order to avoid a trivial solution, we produce two different embeddings $\boldsymbol{z}_k$ and $\boldsymbol{z}'_{k'}$ by imposing two different augmentations to its original training image.

Memory queue. Inspired of the previous work [He et al., 2020], we introduce a memory queue to store the past K embeddings and their corresponding pseudo labels $\text{MQ} = \{(\boldsymbol{z}_k, p^w_k) | k \in [K]\}$, where p^w_k is the pseudo label for embedding $\boldsymbol{z}_k$. It is noteworthy that the memory queue consists of embeddings with respect to both unobserved and observed labels. Then, the connection matrices I

and $\bar{I}$ are extended to size of $B \times K$. This means we would impose the label-aware global consistency regularization between an anchor embedding of the current embedding set and every embedding of the memory queue.

Two-stage training strategy. In our experiments, we find that training model with Eq.(3) in an end-to-end fashion often leads to an unfavorable performance that can be worse than that without $\mathcal{L}_{\text{LAC}}$. The main reason is that at the early stage of training, the model suffers from the insufficient training issue, which leads model to produce low-quality label-wise embeddings. Enforcing low-quality label-wise embeddings to maintain global consistency may destroy the feature representation learning and thus significantly degrades the model performance. To solve this issue, we propose a two-stage training strategy, which aims to let model learn a high-quality label-wise embeddings firstly and then perform LAC regularization. Specifically, at the first stage, we train the model by minimizing AN loss and PLC loss, *i.e.*, $\mathcal{L}_{\text{AN}} + \mathcal{L}_{\text{PLC}}$, until it reaches the convergence. The model is expected to produce sufficiently high-quality label-wise embeddings for the second stage training. At the second stage, we fine-tune the model with Eq.3 consisting of all three losses to learn a more distinctive representation.

4 Experiments

In this section, we first perform experiments to compare our method with state-of-the-art methods; then, we conduct ablation studies to examine the effectiveness of each component for the proposed method.

4.1 Experimental Setting

Datasets We perform experiments to evaluate our proposed method on four benchmark datasets: VOC12 [1] [Everingham et al., 2010], COCO [2] [Lin et al., 2014], NUS [3] [Chua et al., 2009], and CUB [4] [Wah et al., 2011]. The detailed characteristics of these datasets are summarized in supplementary material. In order to compare our method with state-of-the-art methods, we use the code shared by Cole et al. [2021] to generate the training, validation and testing sets. For each dataset, we withhold 20% of the training examples for validation. To construct SPML data, we randomly select one positive label for each training example, while keep the validation and testing sets always fully labeled. Follow Verelst et al. [2022], We report the mean average precision (mAP) on the testing set by using predictions of the model with the best validation performance.

Comparing methods Besides the baseline **AN** loss, we compare the proposed method with the following state-of-the-art algorithms: **WAN** [Cole et al., 2021], which introduces a weight parameter to down-weight the losses with respect to negative labels; **EPR** [Cole et al., 2021], which utilizes an expected positive regularization to mitigate the negative impact of false negative labels; **AN-LS**, which incorporates the label smoothing technique to AN loss; **ROLE** [Cole et al., 2021], which adopts regularized online label estimation (ROLE) technique to alleviate the negative impact of false negative labels; **ROLE+LinearInit** [Cole et al., 2021], which combines ROLE and "LinearInit" training strategy; **EN+SCL** [Verelst et al., 2022], which combines expected negative loss with the spatial consistency loss (SPL); **EM+APL** Verelst et al. [2022], which incorporates asymmetric pseudo-labeling (APL) loss into entropy minimization.

Implementation. Follow Cole et al. [2021], we use the ResNet-50 He et al. [2016] model pretrained on the ImageNet Deng et al. [2009] as the base model. The details for the implementation of the label-wise embedding encoder can be found in the supplementary material. We resize the input images to 448×448, as in Cole et al. [2021]. For training images, we use both a weak augmentation (only containing random horizontal flip) and a strong augmentation (containing Cutout [Devries and Taylor, 2017] and RandAugment [Cubuk et al., 2020]). For training the model, we use the AdamW Loshchilov and Hutter [2019] optimizer with the weight decay of 0.01. The OneCycleLR scheduler is used to change the learning rate with the max learning rate of 0.0001. We train the model for 40 epochs with

[1] `http://host.robots.ox.ac.uk/pascal/VOC/`

[2] `https://cocodataset.org`

[3] `https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html`

[4] `http://www.vision.caltech.edu/visipedia/CUB-200-2011.html`

Table 1: Mean average precision (mAP) of each comparing method on four benchmark datasets. The best performance is highlighted in bold and the sub-optimal performance is underlined.

Loss	Description	mAP (%)			
		VOC	COCO	NUS	CUB
$\mathcal{L}_{\text{WAN}}$	weak assume negative	86.5	64.8	46.3	20.3
$\mathcal{L}_{\text{EPR}}$	expected positive regularization	85.5	63.3	46.0	20.0
$\mathcal{L}_{\text{AN}}$	assume negative	85.1	64.1	42.0	19.1
$\mathcal{L}_{\text{AN-LS}}$ + LI	AN + label smoothing+LinearInit	86.5	69.2	50.5	16.6
$\mathcal{L}_{\text{ROLE}}$ + LI	regularized online label estimation + LinearInit	88.2	69.0	<u>51.0</u>	16.8
$\mathcal{L}_{\text{EN}}$ + $\mathcal{L}_{\text{CL}}$	expected negative + consistency loss	87.6	71.6	-	-
$\mathcal{L}_{\text{EN}}$ + $\mathcal{L}_{\text{SCL}}$	expected negative + spatial consistency loss	88.0	<u>72.1</u>	-	-
$\mathcal{L}_{\text{EM}}$	entropy-maximization	89.1	70.7	47.2	20.9
$\mathcal{L}_{\text{EM}}$ + APL	EM + asymmetric pseudo-labeling	<u>89.2</u>	70.9	47.6	<u>21.8</u>
Ours	assume negative + PLC + LAC	**89.6**	**75.6**	**51.1**	**23.0**

Table 2: The ablation study on COCO.

Method					mAP
Assume negative	PLC	LAC	Two stage	Memory queue	
√					68.5
√	√				69.7
√	√	√			65.6
√	√	√	√		74.7
√	√	√	√	√	75.6

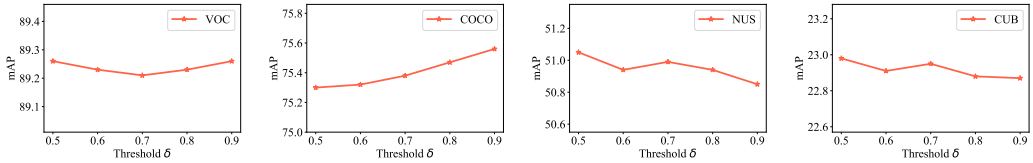
the early stopping. We consider the batch size in the range of $\{8, 16, 32, 64\}$. At the first stage, we set $\lambda_{\text{PCL}} = 1$ and determine the threshold δ from the range of $\{0.5, 0.6, 0.7, 0.8, 0.9\}$. At the second stage, there are two extra parameters, including the balancing parameter λ_{LAC}, the size of memory queue K. We set $\lambda_{\text{LAC}} = 1$, and determine K from the range of $\{512, 1024, 2048, 4096, 8192\}$, respectively. All hyperparameters are determined according to their mAP obtained on the validation set. In Section 4.4, we conduct experiments to analyze the sensitivity for unfixed parameters. Furthermore, we apply exponential moving average (EMA) to model parameters θ with a decay of 0.9997. We perform all experiments on GeForce RTX 3090 GPUs.

4.2 Comparison Results with the State-of-the-Arts

Table 1 reports comparison results between the proposed method and the comparing methods in terms of mAP on four benchmark datasets. From the table, it can be observed that: 1) AN loss achieves the worst performance in almost all cases except for CUB, where it obtains an acceptable mAP of 19.1%. This validates that simply using AN cannot effectively solve SPML problems, since it introduces a large number of false negative labels, which are harmful for model training. 2) ROLE achieves promising performance in most cases except for CUB, where 16.6% achieved by ROLE is significantly worse than 19.1% achieved by AN loss. This validates that by utilizing the regularized online label estimation technique, ROLE can alleviate the negative impact of false negative labels in some extent. However, ROLE requires the true number of positive labels, which can often be unavailable in many realistic tasks and thus limit it practical usefulness. 3) Our proposed method consistently achieves desirable performance and achieves the best performance on all cases. In particular, we improve upon the state-of-the-art performance by 3.5% in terms of mAP on COCO. These results validate that the proposed method can effectively solve SPML problems.

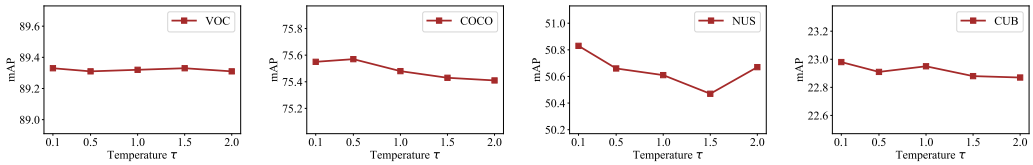
4.3 Ablation study

In this section, to further analyze how the proposed method improves performance for SPML tasks, we conduct a series of ablation studies on COCO and report the results on Table 2. The ablation



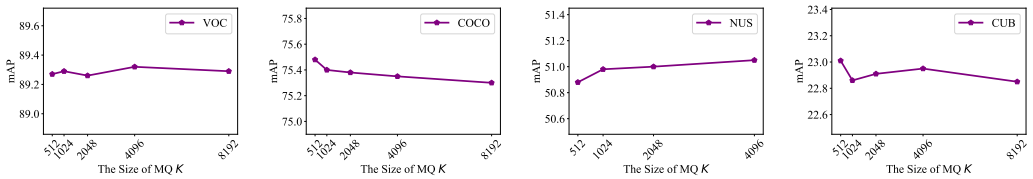
(a) VOC (b) COCO (c) NUS (d) CUB

Figure 3: The performance curve as the threshold δ increases.



(a) VOC (b) COCO (c) NUS (d) CUB

Figure 4: The performance curve as the temperature τ increases.



(a) VOC (b) COCO (c) NUS (d) CUB

Figure 5: The performance curve as the size of memory queue K increases.

experiments for other three datasets are reported in the supplementary material. We first conduct the experiment to validate the effectiveness for the proposed pseudo-labeling consistency regularization $\mathcal{L}_{\text{PLC}}$. The achieved performance is 69.7% by utilizing $\mathcal{L}_{\text{PLC}}$ that is better than 68.5% achieved by only utilizing AN loss $\mathcal{L}_{\text{AN}}$. This discloses that by using $\mathcal{L}_{\text{PLC}}$, the model can recover the true positive labels that benefits for model training. To validate the effectiveness of the proposed label-aware global consistency regularization LAC, we train the model with two strategies, *i.e.*, end-to-end training, which directly optimizes the joint loss Eq.(3) in a pipeline, and two-stage training, which trains the model with the first two losses at the first stage until model converges, and trains the model with Eq.(3) at the second stage. It is can be observed that by adopting the end-to-end training strategy, the performance is degraded from 69.7% to 65.6% by adding LAC into training while is improved to 74.7% by utilizing two-stage training. This is because by adopting the end-to-end training strategy, at the early stage of model training, it is difficult for model to produce high-quality label-wise embeddings for performing LAC regularization. Enforcing low-quality label-wise embeddings to maintain the global consistency may destroy the feature representation learning and thus lead to a noticeable decrease in model performance. Compared to end-to-end training, the two-stage training strategy primarily learn high-quality label embeddings at the first stage, and then utilize the LAC regularization to encourage global consistency among label-wise embeddings, which improves the discrimination ability of model. Finally, we examine the usefulness of memory queue. By using the memory queue, the performance is improved with 0.9% increment from 74.7%. The main reason is that the memory queue provides the model with much more supervision for learning a more distinctive representation, which benefits for model to identify the potential positive labels. These results convincingly validate that each component of the proposed method makes positive contribution to the performance.

4.4 Parameter Sensitivity Analysis

In this section, we study the influence of three parameters, *i.e.*, threshold δ, temperature τ and the size of memory queue K.

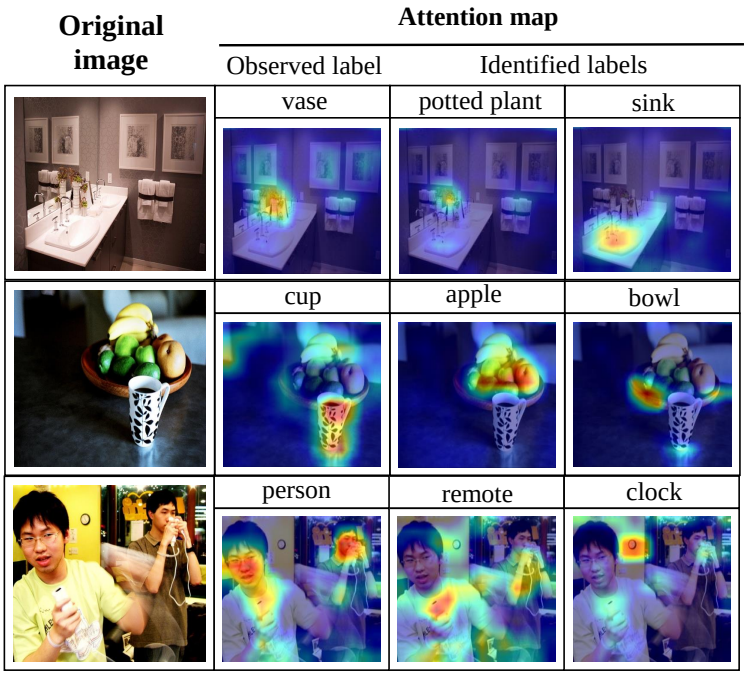


Figure 6: Visualization of attention maps on COCO.

Figure 3, Figure 4 and Figure 5 illustrate the performance curves of the proposed method as the threshold δ, the temperature τ and the size of memory queue K change in the range of $\{0.5, 0.6, 0.7, 0.8, 0.9\}$, $\{0.1, 0.5, 1.0, 1.5, 2.0\}$ and $\{512, 1024, 2048, 4096, 8192\}$, respectively. As shown in the figures, the performance of the proposed method is generally insensitive to these three parameters. Specifically, the largest performance gap with respect to δ, τ and K are respectively about 0.3% on COCO, 0.2% on NUS and 0.2% on COCO. The results indicate that we can safely set these parameters in a large range in practice.

4.5 Case Study

To disclose the mechanism behind the effectiveness of the proposed method for identifying the potential positive labels, Figure 6 visualizes some cases of attention maps on COCO. For every original image in the first column, we illustrate the attention maps of the single observed positive label and identified positive labels in the second column and last two columns, respectively. From the figures, it is interesting to observe that some small objects without annotations can be captured by the proposed method precisely, such as *potted plant* in the first case, *bowl* in the second case as well as *clock* and *remote* in the last case. The observations disclose that the proposed method can capture small objects by maintaining the global consistency of embeddings between the object and its intra-class objects. This further indicates LAC can significantly enhance the ability of model for identifying the potential positive labels.

5 Conclusion

The paper studies the problem of multi-label learning with single positive labels. To solve SPML problems, we design the PLC regularization that exploits the model outputs predicted on weakly-augmented and strongly-augmented images, and LAC regularization that leverages the manifold structure information to recover the true labeling information of the potential positive labels. By utilizing these two regularizations, the model is expected to learn a more distinctive representation, which benefits for identifying the potential positive labels. Experiments on multiple benchmark datasets validate the proposed method can achieve state-of-the-art performance. In the future, we plan to improve the performance of SPML by using other structure information, such as label correlations.

References

Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5177–5186. Computer Vision Foundation / IEEE, 2019.

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 Stéphane Marchand-Maillet and Yiannis Kompatsiaris, editors, *Proceedings of the 8th ACM International Conference on Image and Video Retrieval, CIVR 2009, Santorini Island, Greece, July 8-10, 2009*. ACM, 2009.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 933–942. Computer Vision Foundation / IEEE, 2021.

Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Son D. Dao, Ethan Zhao, Dinh Phung, and Jianfei Cai. Multi-label image classification with contrastive learning. *CoRR*, abs/2107.11626, 2021. URL `https://arxiv.org/abs/2107.11626`.

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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206848. URL `https://doi.org/10.1109/CVPR.2009.5206848`.

Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017. URL `http://arxiv.org/abs/1708.04552`.

Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 647–657. Computer Vision Foundation / IEEE, 2019.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010.

Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 729–739. Computer Vision Foundation / IEEE, 2019.

Emily M. Hand, Carlos Domingo Castillo, and Rama Chellappa. Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 6878–6885. AAAI Press, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020.

Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer, 2014.

Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *CoRR*, abs/2107.10834, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL `https://openreview.net/forum?id=Bkg6RiCqY7`.

Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. Semantic-aware representation blending for multi-label image recognition with partial labels. *CoRR*, abs/2203.02172, 2022.

Tal Ridnik, Emanuel Ben Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 82–91. IEEE, 2021a.

Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. *CoRR*, abs/2111.12933, 2021b.

Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4657–4666. IEEE Computer Society, 2015.

Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4077–4087, 2017.

Thomas Verelst, Paul K. Rubenstein, Marcin Eichner, Tinne Tuytelaars, and Maxim Berman. Spatial consistency loss for training multi-label classifiers from single-label annotations. *CoRR*, abs/2203.06127, 2022.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Haobo Wang, Ruixuan Xiao, Yixuan Li, Lei Feng, Gang Niu, Gang Chen, and Junbo Zhao. Pico: Contrastive label disambiguation for partial label learning. *CoRR*, abs/2201.08984, 2022.

Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 12265–12272. AAAI Press, 2020.

Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 464–472. IEEE Computer Society, 2017.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 321–328. MIT Press, 2003.

Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Acknowledging the unknown for multi-label learning with single positive labels. *CoRR*, abs/2203.16219, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3 and Section 4.
   (b) Did you describe the limitations of your work? [Yes] See Section 4.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 4.1.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We report the random seed used for performing experiments.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 4.1.
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] The code is attached in the supplementary material.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]