Multi Label Loss Correction against Missing and Corrupted Labels

Amirmasoud Ghiassi S.GHIASSI@TUDELFT.NL
Delft University of Technology, Delft, Netherlands

Robert Birke ROBERT.BIRKE@UNITO.IT
Computer Science dept., University of Turin, Turin, Italy

Lydia Y.Chen LYDIAYCHEN@IEEE.ORG
Delft University of Technology, Delft, Netherlands

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

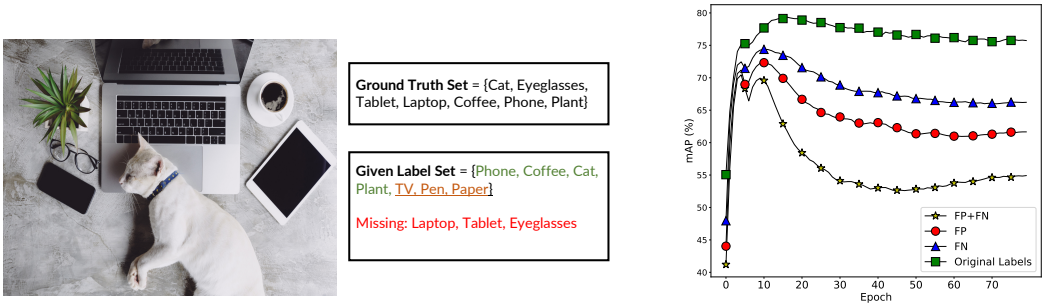
Missing and corrupted labels can significantly ruin the learning process and, consequently, the classifier performance. Multi-label learning where each instance is tagged with variable number of labels is particularly affected. Although missing labels (false-negatives) is a well-studied problem in multi-label learning, it is considerably more challenging to have both false-negatives (missing labels) and false-positives (corrupted labels) simultaneously in multi-label datasets. In this paper, we propose Multi-Label Loss with Self Correction (MLLSC) which is a loss robust against coincident missing and corrupted labels. MLLSC computes the loss based on the true-positive (true-negative) or false-positive (false-negative) labels and deep neural network expertise. To distinguish between false-positive (false-negative) and true-positive (true-negative) labels, we use the output probability of the deep neural network during the learning process. Our method As MLLSC can be combined with different types of multi-label loss functions, we also address the label imbalance problem of multi-label datasets. Empirical evaluation on real-world vision datasets, i.e., MS-COCO, and MIR-FLICKR, shows that our method under medium (0.3) and high (0.6) corrupted and missing label probabilities outperform the state-of-the-art methods by, on average 23.97% and 9.31% mean average precision (mAP) points, respectively.

Keywords: Multi-label learning; Missing labels; Corrupted labels; Loss correction; Robust classifier

1. Introduction

It is challenging to curate error-free training sets for multi-label learning. As opposed to multi-class learning where each sample, e.g. image, is labeled with precisely one correct label, in multi-label learning each sample comes with multiple, varying in number, labels. Web-crawling and crowd-sourcing offer an efficient and inexpensive solution to annotate multi-class data, but multi-label datasets are more prone to curation errors. In addition to erroneous labels, i.e. false-positives, multi-label sets can suffer also from missing labels, i.e., false negatives. Both result in serious degradation of the learning performance (Ghiassi et al., 2021b; Chen et al., 2019).

Multi-label learning is also affected by intrinsic label imbalances. Typically each image only contains a subset of all possible classes. This translates into a number of negative labels



(*a*) Example of image with missing and corrupted labels

(*b*) Impact of false-positive (corrupted) and false-negative (missing) labels on DNN performance with BCE loss

Figure 1: Wrong labels and their impact in multi-label classification.

being higher than the number of positive labels and creates a difference in the distribution of labels which is difficult to avoid or control. Since the number of negative labels is more than positive for each image instance, negative labels contribute more to the loss. In other words, this imbalance significantly impacts the loss (Ridnik et al., 2021b; Lin et al., 2020) and makes it challenging to design loss functions for multi-label learning.

False-negative labels commonly occur because of human annotators missing rare classes. Furthermore, negative label detection, i.e., absence of classes, is more complex than positive label detection, i.e., presence of classes (Wolfe et al., 2005). False-positive labels commonly arise because of confusion of similar classes, e.g., dog and cat, by annotators (Hendrycks et al., 2018; Wu et al., 2021). Despite the fact that label imbalance is a critical issue for training multi-label classifiers, labeling errors mislead the training procedure of deep neural networks (DNNs) even more severely (Cole et al., 2021).

The focus of prior art in multi-label learning is on missing labels (Zhang et al., 2021c; Cole et al., 2021) foregoing the effect of wrong labels. In reality each image can contain some correct labels (true-positives), miss some others (false-negatives), and the annotator can add some wrong labels (false-positives). Hence, different from previous work, we consider annotators that can simultaneously produce missing labels (marked red Figure 1(*a*)) and wrong labels (marked orange in Figure 1(*a*)) for each image. We empirically evaluate the separate and combined impact of false-positives and false-negatives on a multi-label classifier using binary cross-entropy loss by injecting missing and corrupted labels in the MIR-FLICKR dataset. In Figure 1(*b*), one can clearly see that the mean Average Precision (mAP) significantly degrades in the case of false-positives only and false-negatives only, but even more in the presence of both. Due to the memorization effect (Yao et al., 2020; Ghiassi et al., 2022), all curves in Figure 1(*b*) first raise (learn correct labels) then degrade (overfit to wrong labels) in terms of mAP. With both missing and corrupted labels the mAP of Binary Cross Entropy (BCE) is reduced from 75.73% (highly curated dataset) to 54.91%. For real world applications it is thus vital to have loss functions robust to concurrent false-positive and false-negative labels.

To cope with simultaneous missing and corrupted labels we design MLLSC that distinguishes the false-positive (false-negative) from true-positive (true-negative) by using the knowledge of a multi-label classifier. We use the predicted probability for each label of a DNN as a confidence value and as an indicator for false or true positive (negative) detection. After that, we compute the suitable loss based on being true/false positive (negative) since the loss value calculation for positive labels differs from negative labels in multi-label learning to counter label imbalance. Unlike methods that require a small amount of correct labels (Hendrycks et al., 2018; Ghiassi et al., 2021a,b), this efficient and effective loss correction works properly without using any ground truth labels against both missing and corrupted labels. Furthermore, our proposed method can be applied to different kinds of multi-label loss functions, e.g. BCE, Focal (Lin et al., 2020) and ASL (Ridnik et al., 2021b) to make them robust against corrupted and missing labels. MLLSC protects the underlying loss function from the negative impact of missing and corrupted labels with only a slight modification.

We empirically demonstrate the performance of MLLSC on MS-COCO (Lin et al., 2014a) and MIR-FLICKR (Huiskes and Lew, 2008) with various injection rates for missing and corrupted labels, against baselines including BCE, Focal (Lin et al., 2020), ASL (Ridnik et al., 2021b), Hill (Zhang et al., 2021c), SPLC (Zhang et al., 2021c), and MPVAE (Bai et al., 2020). MLLSC can improve the mAP compared to the state-of-the-art by on average 23.85%, and 8.88% under severe missing and corruption ratios, i.e., 0.6 probability for each label to be either a false positive or false negative, for MS-COCO, and MIR-FLICKR, respectively.

The contributions of this paper are summarized as follows:

- We design a novel loss function for multi-label classification that uses DNN output probability to distinguish false-positive (false-negative) from true-positive (true-negative) labels.
- We design a robust loss function for multi-label learning called MLLSC that can alleviate the effect of false-negative and false-positive labels. In addition, MLLSC can work with all kinds of multi-label loss functions.
- We improve MLLSC performance compared to state-of-the-art baselines under different ratios of false-negative and false-positive labels.

2. Related Work

Multi-label classification is a well-studied problem (Ridnik et al., 2021b; Liu et al., 2021; Yazici et al., 2020) across a wide range of learning applications, e.g., object detection (Zhang et al., 2021b; Zhao et al., 2020), speech recognition (Zhang et al., 2021a; Cabral et al., 2011), natural language processing (Ishida et al., 2017), and image classification (Lin et al., 2020; Ridnik et al., 2021b). However, all of these methods perform well under the assumption of clean and complete labels for each training sample. Multi-label classifiers robust against noisy labels have recently received attention from literature compared to robust single-label classifiers. We first investigate learning methods addressing the multi-label classification task, and then study the methods robust against missing and noisy label data.

2.1. Multi-Label Loss Functions

In multi-label learning, designing a loss function demands to take into account the label imbalanced label issue because, in practice, the impact of negative (missing) labels is higher than positive (present) ones (Ridnik et al., 2021b). The classic loss function commonly used in multi-label learning is Binary Cross-Entropy (BCE). BCE is agnostic to the imbalance issue and weighs negative and positive labels equally. Focal loss (Lin et al., 2020) tries to solve the imbalance problem by using different weights for negative and positive labels in the loss function. ASL (Ridnik et al., 2021b) computes the weights asymmetrically by shifting the label probability to ensure no loss for negative labels with very low probability. Query2Label (Liu et al., 2021) is a transformer based multi-label classifier that leverages decoder structures to query the presence of certain labels.

2.2. Robust Multi-Label Loss Functions

The standard multi-label methods only consider the label imbalance issue foregoing any corruption in the training data. Hence, they cannot perform well in the presence of data with label noise, i.e. missing or corrupted labels (Zhang et al., 2021c). The label corruption can be categorized into three different scenarios. First, false-negative labels that represent the missing label problem (Yu et al., 2014; Pu et al., 2022; Cole et al., 2021). Second, false positive labels which represent wrongly added labels leading to partial label learning (Xu et al., 2019; Xie and Huang, 2018; Yan and Guo, 2020). Finally, the last and most complex scenario considers the coincident presence of both false negative and false positive labels (Bai et al., 2020).

(Zhang et al., 2021c) proposes a loss re-weighting method for negative labels to solve the missing label problem termed Hill. This method does not take into account the false positive labels. In addition (Zhang et al., 2021c) introduces a self-paced loss correction (SPLC) method using the confidence value of the model under training to regularize the negative and positive labels. Again, SPLC cannot handle false positive labels. Role (Cole et al., 2021) considers the scenario in which only one positive label is provided. Role uses a combination of BCE loss and loss regularization based on the expected positive labels for each sample. The drawback of Role is that the single positive label must be correct, and the average number of positive labels for each data sample must be known. The Multivariate Probit Variational AutoEncoder (MPVAE) (Bai et al., 2020) has been observed to remain robust against a median level of false positive and false negative labels. It is worth mentioning that the primary goal of MPVAE is not robustness against noisy labels.

In contrast to the above, MLLSC is a robust multi-label learning method that leverages the knowledge of the trained model during the learning process to avoid overfitting the loss to false-positive and false-negative labels.

3. Methodology

In this section, we first discuss standard losses in multi-label learning that are not equipped with any mechanism to deal with false-positive and false-negative labels. Next, we introduce our method called MLLSC, which is a robust loss function against missing and corrupted

labels in multi-label learning. Furthermore, we show that our technique can be applied to all multi-label losses and improve their performances compared to the original ones.

3.1. Preliminary

Let $\mathcal{D}$ be a multi-label dataset of pairs of features $\boldsymbol{x} \in \mathbb{R}^d$ and corresponding labels $\tilde{\boldsymbol{y}} \in [0,1]^K$ where K is the number of classes. The presence or absence of label of class k for instance i is represented by $\tilde{y}_k^i = 1$ and $\tilde{y}_k^i = 0$, respectively. Since we consider false-positive and false-negative labels in our problem, we use $\boldsymbol{y}^i$ to represent the ground truth, i.e., the correct label vector for instance i. $\tilde{\boldsymbol{y}}^i$ can contain both correct and corrupted class labels.

The aim is to classify an input instance using a deep neural network $f : \mathbb{R}^d \rightarrow [0,1]^K$ which is trained through the learning iteration. We define f to be a classifier with parameters $\boldsymbol{\theta}$ and sigmoid activation function at the last layer which denotes the probability of each label for input instance $\boldsymbol{x}_i$ as $\boldsymbol{p}_i = f(\boldsymbol{x}^i) = \frac{1}{1+e^{-\boldsymbol{x}^i}}$. The common loss function for multi-label classification is binary cross-entropy which is defined as

$$\ell = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}(\tilde{y}_k^i \log\ p_k^i + (1-\tilde{y}_k^i)\log\,(1-p_k^i)) \tag{1}$$

where N denotes the number of samples in the training set. According to Eq. 1, the positive and negative labels losses equal to $\log p_k^i$ and $\log(1-p_k^i)$, respectively. In the rest of the paper for simplicity we ignore the instance superscript i and consider the loss of each instance. Hence we have:

$$\mathcal{L} = -\sum_{k=1}^{K}(\tilde{y}_k \underbrace{\log\ p_k}_{\mathcal{L}_k^+} + (1-\tilde{y}_k)\underbrace{\log\,(1-p_k)}_{\mathcal{L}_k^-}) \tag{2}$$

The BCE loss does not provide any mechanism to handle the imbalance issue of multi-label data. Hence, Focal loss (Lin et al., 2020) is proposed to cope with the imbalance issue of negative and positive labels by weighting differently positive and negative losses. The Focal loss is defined as:

$$\mathcal{L}_{Focal} = -\sum_{k=1}^{K}\tilde{y}_k\underbrace{(\alpha_+(1-p_k)^\gamma \log\ p_k)}_{\mathcal{L}_k^+} + (1-\tilde{y}_k)\underbrace{(\alpha_- p_k^\gamma \log\,(1-p_k))}_{\mathcal{L}_k^-}) \tag{3}$$

where $\alpha_-, \alpha_+ \in [0,1]$ are weighting factors for balancing the impact of positive and negative labels on the training loss. γ represents the focusing parameter that controls the model focus on hard and easy instances during training. For instance, by setting $\gamma > 0$, samples with $p_k << 0.5$ are the easy negative labels, and their loss contributions are reduced.

To weight positive and negative labels differently, ASL (Ridnik et al., 2021b) recently proposed asymmetric multi-label loss to diminish the positive-negative imbalance issue, which is introduced as:

$$\mathcal{L}_{ASL} = -\sum_{k=1}^{K} \tilde{y}_k \underbrace{((1-\mathcal{P'}_{k,m})^{\gamma_+} \log\ \mathcal{P'}_{k,m})}_{\mathcal{L}_k^+} + (1-\tilde{y}_k) \underbrace{(\mathcal{P'}_{k,m}^{\gamma_-} \log\ (1-\mathcal{P'}_{k,m}))}_{\mathcal{L}_k^-} \tag{4}$$

where γ^- and γ^+ are focus parameters for negative and positive labels, respectively, which control the weights of negative and positive labels and $\gamma^- > \gamma^+$. Also, $\mathcal{P'}_{k,m} = \max(p_k - m, 0)$ denotes the shifted probability where m is the probability margin.

3.2. Learning from Noisy Labels

A multi-label classifier when trained with corrupted and missing labels under one of the mentioned losses, e.g., BCE, Focal, and ASL, tends to overfit the false-positive and false-negative labels and model accuracy drops significantly. As shown in $\mathcal{L}_k^+$ and $\mathcal{L}_k^-$ (see Eq. 3), there is no loss correction mechanism to alleviate the impact of false-positive and false-negative labels. We propose MLLSC as a robust multi-label loss correction method that can be applied to all kinds of multi-label loss functions to provide a shield against corrupted and missing labels. MLLSC is inspired by the design of SPLC (Zhang et al., 2021c), which is a loss correction method only for the problem of missing labels. First, we consider the BCE loss as a Maximum Likelihood (ML) estimation problem. The idea of using BCE for multi-label learning (without any corruption and missing labels) can be seen as an ML approximation problem under the Bernoulli distribution:

$$P(\boldsymbol{y}) = \prod_{k=1}^{K} p_k^{y_k} (1-p_k)^{1-y_k} \tag{5}$$

where P is the likelihood for each instance i. By taking log and finding the optimal value of Eq. 5, the BCE loss is inferred. Now, we consider Eq. 5 in presence of false-positive and false-negative labels. Let $q_k \in [0,1]$ be the probability of the corresponding class k being truly positive, and $q'_k \in [0,1]$ is the probability of the corresponding class k being truly negative. In other words, q_k denotes the probability that the ground truth and given labels both assign the same positive labels and q'_k is the probability of a true negative means ground truth and given labels are both negative labels. Hence the probability of being false-positive and false-negative are $1-q'_k$ and $1-q_k$, respectively. Then, we can propose the likelihood of each instance for the new setting as the following:

$$P(\tilde{\boldsymbol{y}}, \boldsymbol{s}, \boldsymbol{t}) = \prod_{k=1}^{K} \{(q_k p_k)^{s_k} ((1-q'_k)(1-p_k))^{(1-s_k)}\}^{\tilde{y}_k} \times \{(q'_k(1-p_k))^{t_k} ((1-q_k)p_k)^{(1-t_k)}\}^{1-\tilde{y}_k} \tag{6}$$

where $s_k \in \{0,1\}$ is the indicating variables of true and false positives, and $t_k \in \{0,1\}$ is the indicating variables of true and false negatives. Besides, $\boldsymbol{s}$ and $\boldsymbol{t}$ are the vectors of indicating variables, i.e., s_k and t_k. Hence we can derive the optimal loss from the likelihood in Eq. 6:

$$\begin{cases} \mathcal{L}_k^+ = s_k\ log(p_k) + (1-s_k)\ log(1-p_k) \\ \mathcal{L}_k^- = t_k\ log(1-p_k) + (1-t_k)\ log(p_k) \end{cases} \tag{7}$$

In Eq. 7, in the case of positive label loss $\mathcal{L}_k^+$, if the label is a true positive $s_k = 1$ and the loss function behaves this sample same as for positive labels. If the label is a false positive, i.e., $s_k = 0$, the loss only considers the negative label term $log\ (1 - p_k)$ because this label is corrupted thus the original label was negative. The same scenario happens the other way around for the case of negative labels loss $\mathcal{L}_k^-$. When the label is true negative $t_k = 1$, the negative BCE loss is calculated, while if the label is false negative $t_k = 0$, then the positive label BCE loss is computed.

To compute our proposed loss in Eq. 7, we need to determine $\boldsymbol{s}$ and $\boldsymbol{t}$ for distinguishing false and true positive/negative labels. Since the given label set $\tilde{\boldsymbol{y}}$ is the only available information, and the value of $\boldsymbol{s}$ and $\boldsymbol{t}$ are not known, we leverage the knowledge of the deep neural network itself. We use the predicted label probabilities as a proxy of the model's certainty of each label. Then, we define two thresholds $\tau, \tau' \in (0, 1)$ to detect whether a label should be considered a true positive or false positive and a true negative or false negative, respectively. After training $f(\cdot, \boldsymbol{\theta})$ at each step, we use the model prediction probability $p_k = f_k(\cdot, \boldsymbol{\theta})$ as the confidence value of the DNN for each specific class k based on the input instance. For a given positive label $y_k = 1$, if $p_k > \tau$ it would be a true positive label, otherwise it is a false positive. Also, for the case of $y_k = 0$, the true negative and false negative can be determined by $p_k < \tau'$ and $p_k > \tau'$, respectively. Hence we can write the new robust losses for positive and negative labels based on new thresholds τ, τ' as follows

$$\begin{cases} \mathcal{L}_k^+ = \mathbb{1}(p_k > \tau)\ log(p_k) + (1 - \mathbb{1}(p_k > \tau))\ log(1 - p_k) \\ \mathcal{L}_k^- = \mathbb{1}(p_k < \tau')\ log(1 - p_k) + (1 - \mathbb{1}(p_k < \tau'))\ log(p_k) \end{cases} \tag{8}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. Eq. 8 shows the general form of positive and negative losses when we use BCE as the base multi-label loss function.The complete form of MLLSC for multi-label classification is

$$\mathcal{L}_{\text{MLLSC}} = -\sum_{k=1}^{K} \tilde{y}_k \mathcal{L}_k^+ + (1 - \tilde{y}_k)\mathcal{L}_k^- \tag{9}$$

where $\mathcal{L}_k^+$ and $\mathcal{L}_k^-$ are the proposed losses in Eq. 8 which are equipped with false-negative and false-positive detection mechanisms to select the proper loss term and alleviate the impact of coincident missing and corrupted labels.

The strength of MLLSC is the easy applicability to different multi-label loss functions. Eq. 9 and Eq. 8 are based on the BCE loss. Applying our method to Focal loss gives:

$$\begin{cases} \mathcal{L}_k^+ = \mathbb{1}(p_k > \tau)\ \alpha_+\ (1 - p_k)^\gamma log(p_k) + (1 - \mathbb{1}(p > \tau))\ \alpha_-\ p_k^\gamma log(1 - p_k) \\ \mathcal{L}_k^- = \mathbb{1}(p_k < \tau')\ \alpha_-\ p_k^\gamma log(1 - p_k) + (1 - \mathbb{1}(p_k < \tau'))\ \alpha_+\ (1 - p_k)^\gamma log(p_k) \end{cases} \tag{10}$$

with the same hyper-parameters defined in Eq. 3 and Eq. 8.

Inspired by ASL (Ridnik et al., 2021b) and SPLC (Zhang et al., 2021c), to emphasize more the uncertain classes, i.e., $0.4 < p_k < 0.6$, we make use of Focal margin loss for positive labels instead of Focal loss due to its superior performance in multi-label classification (Lin et al., 2020; Zhang et al., 2021c). More specifically, we define MLLSC loss based on Focal margin loss as

$$\begin{cases} \mathcal{L}_k^+ = \mathbb{1}(p_k > \tau)\ (1 - \mathcal{P}_{k,m})^\gamma log(\mathcal{P}_{k,m}) + (1 - \mathbb{1}(p_k > \tau))\ \mathcal{P}_{k,m}^\gamma log(1 - \mathcal{P}_{k,m}) \\ \mathcal{L}_k^- = \mathbb{1}(p_k < \tau')\ p_k^\gamma log(1 - p_k) + (1 - \mathbb{1}(p_k < \tau'))\ (1 - p_k)^\gamma log(p_k) \end{cases} \tag{11}$$

where $\mathcal{P}_{k,m} = f_k(\boldsymbol{x} - m)$ and m is the margin parameter. For the case of $m = 0$, the Focal margin loss collapses into the Focal loss. It is worth mentioning that throughout this paper we have used Focal margin loss for all experiments. To initialize the knowledge of the DNN, we initially train it for two epochs with uncorrected Focal loss, before switching to the MLLSC corrected loss (Eq. 10). This is because DNNs learn easy (correct) labels first, before overfitting to missing and corrupted labels due to the memorization effect (Xu et al., 2021; Hendrycks et al., 2018; Wang et al., 2019). Switching losses allow us to leverage the easy labels for the initial epochs and afterward actively cope with false-negative and false-positive labels.

4. Evaluation

In this section, we first explain the details of the experiment setup, evaluation metrics, and baselines. Then present the evaluation results of the performance of our proposed MLLSC on two well-known datasets: MS-COCO (Lin et al., 2014a) and MIR-FLICKR (Huiskes and Lew, 2008).

4.1. Experimental Settings

4.1.1. Datasets

MS-COCO. The MS-COCO 2014 dataset is a real-world object detection dataset consisting of 82,021 images for training and 40,137 for testing. MS-COCO is commonly used for multi-label classification. On average, each image has 2.9 labels belonging to 80 classes.
MIR-FLICKR. MIR-FLICKR is a collection of images retrieved from the social photography site Flickr.com through its public API. It includes 23,300 images for training and 1,700 images for testing. The average labels per image is 8.9 belonging to 38 classes.

4.1.2. False Negative and False Positive Labels

Both MS-COCO and MIR-FLICKR are highly curated datasets. We use the original labels as ground truth and synthetically inject corrupted and missing labels. We follow previous work (Patrini et al., 2017) with minor modifications to adapt it to the multi-label scenario and inject both false-positive and false-negatives labels into the training data. We flip one positive class to other classes with uniform probability *eta*. It is worth mentioning that changing to a new label is acceptable only when the new label is not a member of the original label set. Under such noise, we change a positive label to a negative label (missing label/false negative) and create one false positive label (corrupted label). Thus, the corresponding transition matrix $\boldsymbol{C}$ having $c_{i,j}$ elements for the mentioned missing and corrupt labels injection is as follows:

$$c_{ij} = \begin{cases} 1 - \eta & \text{if } i = j \\ \dfrac{\eta}{K - 1} & \text{if } i \neq j \end{cases}$$

In order to evaluate the performance of MLLSC, we test our method against multiple missing and corruption probabilities, i.e., in the range $[0.0, 0.6]$.

4.1.3. Baselines

We compare the performance of MLLSC against BCE, Focal (Lin et al., 2020), ASL (Ridnik et al., 2021b), Hill (Zhang et al., 2021c), SPLC (Zhang et al., 2021c), and MPVAE (Bai et al., 2020).

- **Binary Cross Entropy (BCE)** is a standard simple loss function for multi-label classification introduced in Eq. 1.
- **Focal** re-weights the positive and negative terms using the model output probability and weighting factor in the loss to reduce the impact of label imbalance (See Eq. 3).
- **ASL** introduces an asymmetric loss to better address the label imbalance issue for multi-label classification (See Eq. 4).
- **Hill** is mainly designed to address the missing label problem (false-negatives). It re-weights the Mean Squared Error (MSE) loss term for negative labels. The loss term for positive labels remains the same as BCE.
- **SPLC** is a loss correction method for false-negative labels which uses the output probability of the multi-label classifier to distinguish false-negative from true-negative labels.
- **MPVAE** is a multi-label classification model that works based on variational autoencoder. This method encodes the features and labels to Gaussian subspaces and then decodes the samples from the subspaces into a multivariate probit to predict the image labels.

4.1.4. Evaluation Metrics

To evaluate the performance of our proposed method under all aspects against the baselines, we consider and report the following metrics: mean average precision (mAP), average per-class F1 (CF1), and average overall F1 (OF1). These metrics have been commonly used in the related art (Ridnik et al., 2021b; Zhang et al., 2021c) to evaluate the performance of multi-label classification models. We report the average and standard deviation across three runs for Table 1 and Table 2.

4.1.5. DNN Configurations

As a base model we consider a ResNet50 pre-trained on the ImageNet-21K dataset (Ridnik et al., 2021a) which has been widely used in vision classification problems. The DNN architecture is the same for all the baselines according to default values provided in (Zhang et al., 2021c; Ridnik et al., 2021b; Lin et al., 2020) except MPVAE, which is an auto-encoder based method. The encoder and decoder use the structure from (Bai et al., 2020), i.e., 3-layer fully connected neural networks with ReLU activation function. Moreover, we set the hyper-parameters to the default values provided in (Bai et al., 2020) for each dataset.

Table 1: mAP(%) of MIR-FLICKR under different ratios of false-negative and false-positive labels for different multi-label classifiers

Method	$\eta = 0.0$			$\eta = 0.3$			$\eta = 0.6$		
	mAP	***CF1***	***OF1***	***mAP***	***CF1***	***OF1***	***mAP***	***CF1***	***OF1***
BCE	74.79 ± 0.14	74.85 ± 0.07	$\mathbf{78.20} \pm 0.15$	60.56 ± 0.52	62.76 ± 0.20	68.07 ± 0.16	35.43 ± 0.39	33.46 ± 0.07	38.41 ± 0.18
Focal	$\mathbf{75.50} \pm 0.26$	74.34 ± 0.18	77.39 ± 0.13	60.73 ± 0.09	61.53 ± 0.12	68.50 ± 0.32	33.87 ± 0.06	31.34 ± 0.26	36.76 ± 0.23
ASL	74.86 ± 0.56	71.80 ± 0.11	74.99 ± 0.17	51.90 ± 2.02	26.83 ± 1.84	26.78 ± 0.89	34.24 ± 0.29	40.40 ± 0.22	42.68 ± 0.33
Hill	74.79 ± 0.53	74.43 ± 0.06	77.68 ± 0.16	59.23 ± 0.42	62.48 ± 0.18	67.81 ± 0.14	33.04 ± 0.38	35.19 ± 0.21	39.47 ± 0.19
SPLC	73.72 ± 0.04	72.35 ± 0.24	75.98 ± 0.14	63.81 ± 0.05	68.50 ± 0.18	71.55 ± 0.33	43.31 ± 0.23	55.23 ± 0.16	59.95 ± 0.08
MPVAE	41.32 ± 0.27	37.64 ± 0.16	48.85 ± 0.15	39.23 ± 0.19	36.26 ± 0.21	47.39 ± 0.27	35.37 ± 0.29	31.87 ± 0.11	41.72 ± 0.14
MLLSC	75.38 ± 0.24	$\mathbf{75.20} \pm 0.19$	77.72 ± 0.19	$\mathbf{65.09} \pm 0.19$	$\mathbf{68.98} \pm 0.05$	$\mathbf{71.88} \pm 0.12$	$\mathbf{45.33} \pm 0.25$	$\mathbf{56.35} \pm 0.15$	$\mathbf{60.48} \pm 0.24$

We perform all the experiments using PyTorch v1.9.0, and we train all the methods for 60 epochs except MPVAE which trains for 200 epochs. To train with MLLSC, we set the batch size, learning rate, and weight decay to 32, 0.0001, and 10^{-4}, respectively.

4.2. Comparison with Baselines

In this section, we evaluate the performance of our proposed method against baselines on the MIR-FLICKR and MS-COCO datasets under three different ratios of η, i.e., $\{0.0, 0.3, 0.6\}$.

For MIR-FLICKR, we report the mAP, CF1, and OF1 of MLLSC against baselines in Table 1. MLLSC achieves the highest mAP among all the methods except for $\eta = 0.0$. In this case, MLLSC is the second best with only Focal reaching a higher score of 75.50. SPLC is the closest rival to MLLSC with 1.28 and 2.02 percents mAP difference under $\eta = 0.3$ and $\eta = 0.6$, respectively. Besides, the MLLSC achieves the highest CF1 compared to all the baselines for all cases. Since MPVAE uses a Variational Auto Encoder-based method and the network architecture is less deep and complex than other baselines, the MPVAE can not perform well for the case of $\eta = 0.0$ and $\eta = 0.3$. However, it is also a robust method against false-positive and false-negative labels, because the difference between the mAP at $\eta = 0.0$ and $\eta = 0.6$ is the lowest compared to other methods. Our method significantly improves robustness with respect to Focal. Under a severe ratio of missing and corrupted labels, i.e., $\eta = 0.6$, its mAP score is 11.46% points higher.

MS-COCO is more challenging than MIR-FLICKR due to the larger number of classes and higher label imbalance. The results are summarized in Table 2. Here MLLSC obtains the highest mAP among all baselines under all three ratios of missing and corrupted labels. For the case of $\eta = 0.0$, our proposed method outperforms all the rivals achieving 68.83% mAP. ASL is the second best with a 0.31% points lower mAP. Under $\eta > 0$, i.e., 0.3 and 0.6, SPLC is the closest competitor since it alleviates the impact of missing labels using a self-paced loss correction method for negative labels. The most considerable difference in mAP between MLLSC and SPLC methods is 5.04% with $\eta = 0.3$. Here MLLSC and SPLC reach 65.69% and 60.65% mAP, respectively. In the case of $\eta = 0.6$, MLLSC achieves the highest mAP, and OF1, whereas the baselines trail far behind. According to the results, not only can the MLLSC withstand missing and corrupted labels, but it also mitigates the impact of imbalance labels even for MS-COCO, which contains a high variation of classes (Lin et al., 2014b). Besides, the performance of all the multi-label losses, i.e., BCE,

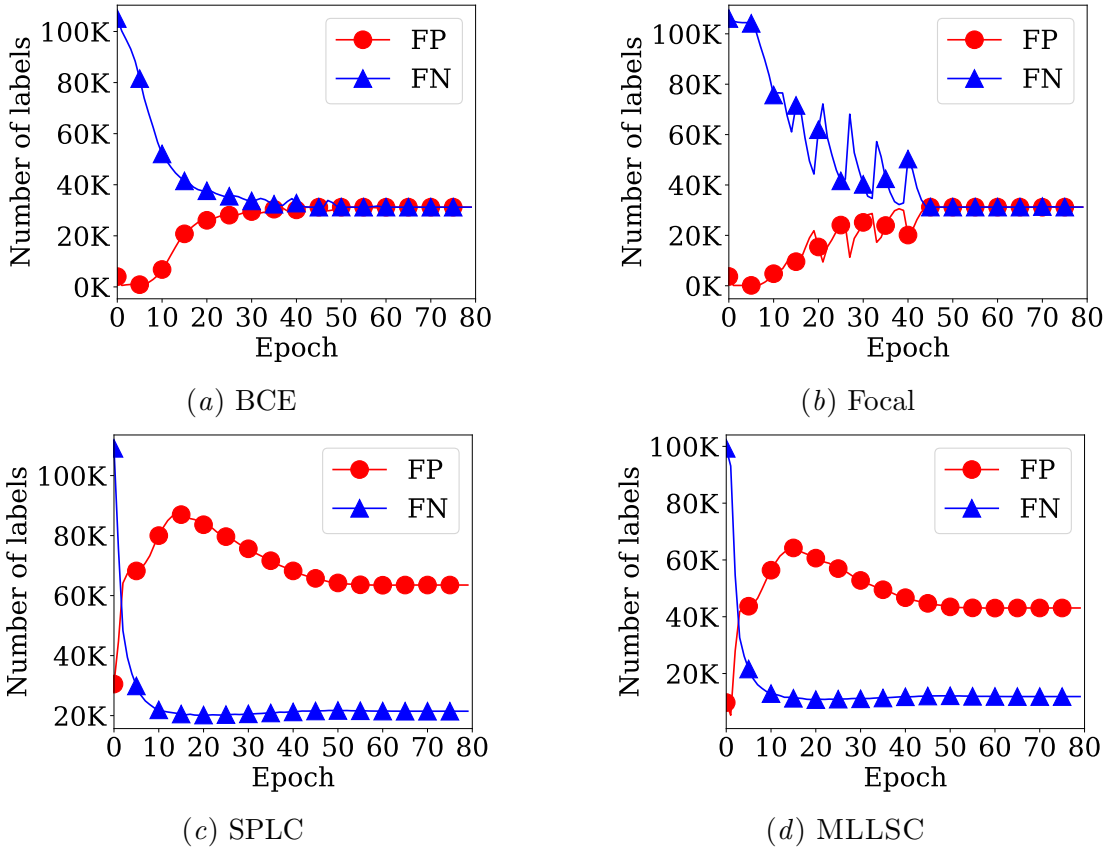
Table 2: mAP(%) of MS-COCO under different ratios of false-negative and false-positive labels for different multi-label classifiers

Method	$\eta = 0.0$			$\eta = 0.3$			$\eta = 0.6$		
	mAP	***CF1***	***OF1***	***mAP***	***CF1***	***OF1***	***mAP***	***CF1***	***OF1***
BCE	65.72 ± 0.15	$\mathbf{64.23} \pm 0.27$	66.29 ± 0.14	43.66 ± 0.24	41.26 ± 0.25	47.31 ± 0.28	19.40 ± 0.31	18.07 ± 0.21	23.41 ± 0.28
Focal	66.77 ± 0.24	63.91 ± 0.17	65.92 ± 0.21	42.73 ± 0.21	36.54 ± 0.29	42.46 ± 0.18	20.67 ± 0.23	20.58 ± 0.14	26.29 ± 0.27
ASL	68.52 ± 0.23	46.72 ± 0.14	60.36 ± 0.27	50.53 ± 0.38	22.74 ± 0.32	51.49 ± 0.38	22.66 ± 0.21	16.70 ± 0.26	40.66 ± 0.22
Hill	63.76 ± 0.14	61.50 ± 0.31	65.78 ± 0.23	48.61 ± 0.24	51.89 ± 0.14	56.50 ± 0.34	27.79 ± 0.20	32.50 ± 0.24	36.42 ± 0.37
SPLC	64.77 ± 0.28	61.71 ± 0.16	65.51 ± 0.33	60.65 ± 0.23	56.68 ± 0.14	61.36 ± 0.42	49.39 ± 0.21	$\mathbf{56.87} \pm 0.15$	57.67 ± 0.16
MPVAE	39.98 ± 0.24	29.71 ± 0.15	46.82 ± 0.23	29.83 ± 0.42	26.58 ± 0.32	46.20 ± 0.16	26.33 ± 0.24	24.72 ± 0.19	43.51 ± 0.24
MLLSC	$\mathbf{68.83} \pm 0.28$	63.58 ± 0.32	$\mathbf{69.83} \pm 0.14$	$\mathbf{65.69} \pm 0.28$	$\mathbf{61.89} \pm 0.18$	$\mathbf{68.75} \pm 0.13$	$\mathbf{51.68} \pm 0.16$	55.57 ± 0.46	$\mathbf{58.60} \pm 0.25$

Focal, and ASL, significantly drops with increasing ratios of false-negative and false-positive labels in the training set.

4.3. Study the Number of False-Positive and False-Negative Labels During Training MLLSC

To provide insights on MLLSC and evaluate the ability of our proposed method to distinguish false-positive (FP) and false-negative (FN) labels, we monitor the number of FN and FP during the training process. We compute FP and FN labels by considering the whole label vector of each predicted image and comparing it to the ground truth. Note that the ground truth is only used to compute these statistics but not for training. Figure 2 plots the number of FP and FN labels during training for BCE, Focal, SPLC, and MLLSC over 80 epochs on MIR-FLICKR under $\eta = 0.3$. At the beginning of training, the number of FN labels is high because there is no knowledge to detect and correct labels. On the contrary, the number of FP labels is low because the model refrains from predicting any labels due to low confidence. With increasing training epochs, the number of FN labels decreases, and the number of FP labels increases for both BCE and Focal until they reach a steady-state (see Figure 2(a) and Figure 2(b)). Since BCE and Focal are not equipped with any mechanism to make them robust against FP and FP, they overfit to missing and corrupted labels and reach a steady-state in which the ratio of FP and FP labels are both equal to the η in use. In contrast SPLC and MLLSC use resilient losses against FP and FN. Thus they reduce the impact of label errors and reach different steady-state values. Comparing SPLC and MLLSC to BCE and focal shows that the number of FN labels significantly decreases with training epochs. Moreover, the end values for MLLSC and SPLC are approximately $87.5K$ and $80K$ less, respectively. MLLSC and SPLC share the same trend for the number of FP labels, but our proposed method incurs about $43K$ FP labels which is approximately $7K$ FP labels less than SPLC (see Figure 2(c) and Figure 2(d)). The loss term for positive labels in SPLC is margin Focal loss which is not equipped with a robust mechanism to alleviate the impact of false-positive labels. Hence, the number of false-positive is higher than for MLLSC. Although the number of FP labels in BCE and Focal are slightly less than SPLC and MLLSC, handling both FP and FN labels simultaneously in MLLSC improves the mAP significantly compared to the baselines (see Table 2 and Table 1).



(a) BCE (b) Focal

(c) SPLC (d) MLLSC

Figure 2: Number of false-positive and false-negative labels during training under $\eta = 0.3$ on MIR-FLICKR dataset.

4.4. Study the Performance of MLLSC with Different Loss Functions

In this part, we complement our method and SPLC with three different multi-label losses, i.e., ASL, Focal, and BCE, to evaluate their improvements under $\eta = 0.4$ on the MIR-FLICKR dataset. The detailed formula of MLLSC with ASL loss is given in the supplementary material due to space reasons. Focal loss with margin shows the best performance improvement combined with both SPLC and MLLSC compared to ASL and BCE (see Table 3). Moreover, the mAP of MLLSC outperforms SPLC when using BCE, ASL and Focal loss functions by 9.12%, 10.55% and 1.37% point difference, respectively. Although ASL can deal well with label imbalance compared to BCE, it can not handle missing and corrupted labels appropriately.

4.5. Ablation Study

4.5.1. Impact of the hyper-parameters (τ,τ')

To distinguish false-positive and false-negative labels correctly, we need to find the best threshold values τ and τ' used to compute the correct loss term. We empirically evaluate

Table 3: mAP(%) of multi-label classifiers under $\eta = 0.4$ for different loss functions on MIR-FLICKR

Method	***mAP***	***CF1***	***OF1***
SPLC + BCE	50.10	60.49	61.94
SPLC + ASL	48.60	42.91	47.59
SPLC + margin Focal	59.24	66.17	69.14
MLLSC + BCE	59.22	67.23	70.31
MLLSC + ASL	59.15	66.30	69.39
MLLSC + margin Focal	**60.61**	**66.92**	**70.87**

Table 4: mAP(%) of MLLSC under $\eta = 0.4$ for different thresholds (τ, τ') on MIR-FLICKR

Different Threshold (τ)							
τ	0.4	0.45	0.5	0.55	0.6	0.65	0.7
mAP	60.05	60.30	60.53	**60.61**	60.32	60.29	60.26
Different Threshold (τ')							
τ'	0.4	0.45	0.5	0.55	0.6	0.65	0.7
mAP	20.69	26.76	50.26	59.74	**60.61**	58.74	56.14
Different Thresholds (τ,τ')							
(τ, τ') = (0.4, 0.7) 55.91					(τ, τ') = (0.7, 0.4) 20.55		

the best combination of (τ, τ'). Table 4 presents the mAP of MLLSC under different $\tau, \tau' \in [0.4, 0.7]$ using Focal margin loss under a ratio of missing and corrupted labels of $\eta = 0.4$ on MIR-FLICKR. Here we keep one of the two parameters fixed while changing the other one to determine the effect on mAP. All other parameters use their default values, i.e., $\tau' = 0.6$ when varying τ, and $\tau = 0.55$ when varying τ'. From Table 4, one can see that the mAP of MLLSC is more sensitive to τ' than τ because the difference between the highest and lowest mAPis 39.92% and 0.56%, respectively. This sensitivity is due to the predominant number of negative labels compared to positive labels in each image. A wrong threshold for negative labels, i.e., τ', creates a more significant number of false label identifications. Thus it influences the performance of MLLSC more. Higher thresholds (τ, τ') make substantial restrictions for true-positive and true-negative label detection, and consequently, in the case of uncertain labels, the number of false-negative and false-positive labels increases. Overall we identify as best values for τ and τ' to detect false-positive and false-negative labels via the model output confidence to be 0.55 and 0.6, respectively. We also consider two extremes for (τ, τ') when we change both values at the same time. We can see that the mAP degrades significantly, i.e., 35.36% points, when we increase τ from 0.4 to 0.7 and decrease τ' from 0.7 to 0.4.

4.5.2. Impact of m on MLLSC with Focal margin loss

To assess the impact of the margin parameter (m) of the Focal loss on MLLSC, we vary its value from 0.0 to 2.0 in Table 5. Since the margin manages the attention of the loss on positive labels, a small value of m concentrate more on hard positive labels. As shown in

Table 5: mAP(%) of MLLSC under $\eta = 0.4$ for different margins (m) on MIR-FLICKR

Different margins m					
m	0.0	0.5	1.0	1.5	2.0
mAP	55.69	59.87	**60.61**	59.38	55.73

Table 6: mAP(%) of MLLSC under $\eta = 0.4$ for different γ on MIR-FLICKR

Different γ				
γ	0.0	1.0	2.0	4.0
mAP	51.47	59.95	**60.61**	59.03

Table 5, for the case of $m = 0$, i.e., standard Focal loss, MLLSC achieves 55.69% mAP, while when increasing the margin, MLLSC can achieve 60.61% mAP when $m = 1$. Increasing the value beyond this shifts the focus of the loss function from hard positives to semi-hard positive labels, which leads again to an mAP reduction. Hence, we set $m = 1$ in all the experiments.

4.5.3. Impact of γ on MLLSC with Focal margin loss

Here, we evaluate the sensitivity of MLLSC to the focus parameter (γ) of the Focal margin loss (see Eq. 11). γ controls the weights of positive and negative labels. With γ below 2.0, the loss can not bring down the weight of easy negative labels, and this degrades the performance of MLLSC by 9.14% and 0.66% points when is $\gamma = 0.0$ and $\gamma = 1.0$, respectively (see Table 6). A large value of γ causes a significant weight reduction of positive labels which are rarely seen in the training data. According to the empirical study, we set $\gamma = 2.0$ through the experiments for best mAP performance.

5. Conclusion

We have investigated the performance impact on multi-label classifiers of combined false-negative (missing) and false-positive (corrupted) labels in multi-label datasets. We have shown that achieving high mAP for a multi-label classifier which is competitive to training on a complete and correct multi-label training set is still possible. To do this, we need to correctly optimize the loss function by computing proper loss terms based on the input label being a true positive or true negative. This paper introduces MLLSC that enhances multi-label loss robustness against missing and corrupted labels by detecting false-positive (false-negative) and true-positive (true-negative) to calculate the corresponding correct loss values. MLLSC distinguishes false-positive (false-negative) and true-positive (true-negative) labels through the model prediction probability, which proxies the confidence of the classifier for each label.We evaluate MLLSC on two real-world datasets subject to different degrees of false-negative and false-positive labels. Under noise ratios of 0.3 and 0.6, MLLSC improves the mAP compared to six baselines drawn from the state-of-the-art by between 9.33-19.48% and 8.88-23.85% points, respectively.

References

Junwen Bai, Shufeng Kong, and Carla Gomes. Disentangled variational autoencoder based multi-label classification with covariance-aware multivariate probit model. In *(IJCAI)*,

2020.

Ricardo Cabral, Fernando Torre, Joao P Costeira, and Alexandre Bernardino. Matrix completion for multi-label image classification. *NIPS*, 24, 2011.

Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *ICML*, volume 97, pages 1062–1070, 2019.

Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *CVPR*, pages 933–942, 2021.

Amirmasoud Ghiassi, Robert Birke, Rui Han, and Lydia Y Chen. Labelnet: Recovering noisy labels. In *IJCNN*, pages 1–8. IEEE, 2021a.

Amirmasoud Ghiassi, Robert Birke, and Lydia Y.Chen. Trustnet: Learning from trusted data against (a)symmetric label noise. In *BDCAT*, page 52–62, 2021b.

Amirmasoud Ghiassi, Robert Birke, and Lydia Y Chen. Labnet: A collaborative method for dnn training and label aggregation. In *ICAART*, pages 56–66, 2022.

Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. *NIPS*, 31, 2018.

Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *International Conference on Multimedia Information Retrieval*, pages 39–43, 2008.

Takashi Ishida, Gang Niu, Weihua Hu, and Masashi Sugiyama. Learning from complementary labels. *NIPS*, 30, 2017.

Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014a.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014b.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, 2020.

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

Giorgio Patrini, A. Rozza, A. Menon, R. Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. *CVPR*, pages 2233–2241, 2017.

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

Tal Ridnik, Emanuel Ben Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *CoRR*, abs/2104.10972, 2021a.

Tal Ridnik, Emanuel Ben Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, pages 82–91. IEEE, 2021b.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330, 2019.

Jeremy M Wolfe, Todd S Horowitz, and Naomi M Kenner. Rare items often missed in visual searches. *Nature*, 435(7041):439–440, 2005.

Songhua Wu, Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Nannan Wang, Haifeng Liu, and Gang Niu. Class2simi: A noise reduction perspective on learning with noisy labels. In *ICML*, pages 11285–11295. PMLR, 2021.

Ming-Kun Xie and Sheng-Jun Huang. Partial multi-label learning. In *AAAI*, volume 32, 2018.

Han Xu, Xiaorui Liu, Wentao Wang, Wenbiao Ding, Zhongqin Wu, Zitao Liu, Anil Jain, and Jiliang Tang. Towards the memorization effect of neural networks in adversarial training. *arXiv preprint arXiv:2106.04794*, 2021.

Ning Xu, Jiaqi Lv, and Xin Geng. Partial label learning via label enhancement. In *AAAI*, volume 33, pages 5557–5564, 2019.

Yan Yan and Yuhong Guo. Partial label learning with batch label correction. In *AAAI*, volume 34, pages 6575–6582, 2020.

Quanming Yao, Hansi Yang, Bo Han, Gang Niu, and James Tin-Yau Kwok. Searching to exploit memorization effect in learning with noisy labels. In *ICML*, pages 10789–10798, 2020.

Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *CVPR*, 2020.

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *ICML*, pages 593–601. PMLR, 2014.

Dong Zhang, Xincheng Ju, Wei Zhang, Junhui Li, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *AAAI*, volume 1, 2021a.

Yixin Zhang, Zilei Wang, and Yushi Mao. Rpn prototype alignment for domain adaptive object detector. In *CVPR*, pages 12425–12434, 2021b.

Youcai Zhang, Yuhao Cheng, Xinyu Huang, Fei Wen, Rui Feng, Yaqian Li, and Yandong Guo. Simple and robust loss design for multi-label learning with missing labels. *arXiv preprint arXiv:2112.07368*, 2021c.

Zhen Zhao, Yuhong Guo, Haifeng Shen, and Jieping Ye. Adaptive object detection with dual multi-label prediction. In *ECCV*, pages 54–69. Springer, 2020.