Two-way Multi-Label Loss

Takumi Kobayashi[†‡]
[†]National Institute of Advanced Industrial Science and Technology, Japan
[‡]University of Tsukuba, Japan
takumi.kobayashi@aist.go.jp

Abstract

A natural image frequently contains multiple classification targets, accordingly providing multiple class labels rather than a single label per image. While the single-label classification is effectively addressed by applying a softmax cross-entropy loss, the multi-label task is tackled mainly in a binary cross-entropy (BCE) framework. In contrast to the softmax loss, the BCE loss involves issues regarding imbalance as multiple classes are decomposed into a bunch of binary classifications; recent works improve the BCE loss to cope with the issue by means of weighting. In this paper, we propose a multi-label loss by bridging a gap between the softmax loss and the multi-label scenario. The proposed loss function is formulated on the basis of relative comparison among classes which also enables us to further improve discriminative power of features by enhancing classification margin. The loss function is so flexible as to be applicable to a multi-label setting in two ways for discriminating classes as well as samples. In the experiments on multi-label classification, the proposed method exhibits competitive performance to the other multi-label losses, and it also provides transferrable features on single-label ImageNet training. Codes are available at https://github.com/tk1980/TwowayMultiLabelLoss.

1. Introduction

Deep neural networks are successfully applied to supervised learning [12, 26, 38] through back-propagation based on a loss function exploiting plenty of annotated samples. In the supervised learning, classification is one of primary tasks to utilize as annotation a class label to which an image sample belongs. As in ImageNet [7], most of image datasets provide a single class label per image, and a softmax loss is widely employed to deal with the single-label annotation, producing promising performance on various tasks.

The single-label setting, however, is a limited scenario from practical viewpoints. An image frequently contains multiple classification targets [39], such as objects, requiring laborious cropping to construct single-label annotations. There are also targets, such as visual attributes [25], which are hard to be disentangled and thereby incapable of producing single-label instances. Those realistic situations pose so-called *multi-label* classification where an image sample is equipped with *multiple* labels beyond a single label.

While a softmax loss works well in a single-label learning, the multi-label tasks are addressed mainly by applying a binary cross-entropy (BCE) loss. Considering multiple labels are drawn from C class categories, the multi-label classification can be decomposed into C binary classification tasks, each of which focuses on discriminating samples in a target class category [28]; the BCE loss is well coupled with the decomposition approach. Such a decomposition, however, involves an imbalance issue. Even in a case of balanced class distribution, the number of positive samples is much smaller than that of negatives, as small portion of whole C-class categories are assigned to each sample as annotation (positive) labels. The biased distribution is problematic in a naive BCE loss. To cope with the imbalance issue in BCE, a simple weighting approach based on class frequencies [25] is commonly applied and in recent years it is further sophisticated by incorporating adaptive weighting scheme such as in Focal loss [18] and its variant [2]. On the other hand, the softmax loss naturally copes with multiple classes without decomposition nor bringing the above-mentioned imbalance issue; it actually works well in the balanced (single-label) class distribution. The softmax loss is intrinsically based on relative comparison among classes (3) which is missed in the BCE-based losses, though being less applicable to multi-label classification.

In this paper, we propose a multi-label loss to effectively deal with multiple labels in a manner similar to the softmax loss. Through analyzing the intrinsic loss function of the softmax loss, we formulate an efficient multi-label loss function to exploit relative comparison between positive and negative classes. The relative comparison is related to classification margin between positive and negative classes, and we propose an approach to enlarge the margin by simply introducing temperature on logits for fur-

ther boosting performance. The proposed loss function is regarded as a generalization of the softmax loss, being reduced into the softmax loss in case of a single-label task. Thus, the general loss formulation enables us to measure losses in *two ways* for computing multi-label classification loss not only at each sample but also for each class to discriminate samples on that class as the BCE focuses on. In summary, our contributions are three-fold as follows.

- We formulate a new loss function to deal with multiple labels per sample while enhancing classification margin.
- By using the loss function, we propose a two-way approach for measuring multi-label loss.
- The loss is thoroughly analyzed from various aspects in the experiments and exhibits competitive performance, compared to the other multi-label losses. It also provides transferrable features on single-label ImageNet training.

1.1. Related works

Multi-label classification: In recent years, the multi-label task is addressed in a deep learning framework by improving models from an architectural viewpoint. Relationships among multiple labels assigned to an image are exploited such as by recurrent neural network [30] and graph neural networks [8] which also leverage external word-related information [4, 32] to extract semantic characteristics of class categories. Regional features are also utilized [39] while transformer-based spatial attention mechanism is incorporated to detect label-related region [33]. In this paper, we focus on a loss function which is an orthogonal direction to the architectural approaches; it would compensate the above-mentioned methods which simply employ a BCE loss.

BCE loss: In the multi-label framework, a binary cross-entropy (BCE) loss plays a key role through decomposing multi-class classification into multiple class-wise binary tasks. For establishing multi-label losses, research effort is mainly devoted to improving the BCE especially in terms of the imbalance issue mentioned above. Frequency-based weighting [25] is widely applied as a naive extension of BCE, and in recent years, sophisticated adaptive weighting schemes are proposed [2, 18]. There are also works to improve BCE so as to cope with particular situations such as class imbalance [35] and partial labeling [1]. In contrast, we derive the proposed method in a framework of a softmax loss, apart from the BCE formulation. Thus, our method bypasses the imbalance issue tackled by [2, 18] while enhancing discriminativity in terms of both classes and samples.

Metric learning: From a viewpoint of formulation, our method is related to the loss functions applied to optimize similarity in the literature of metric learning which is different from the multi-label classification. By regarding a classifier vector as a proxy of the class, one could find resemblance between the similarity learning and the multi-label learning both of which cope with multiple positive and negative pairs. The point is that we derive our multi-label loss through theoretically analyzing a single-label softmax loss, while the similarity losses [27, 31] are formulated based on rather heuristic pair-wise comparison among positive and negative similarities; we discuss the difference in Sec. 3.

2. Method

We first analyze a softmax cross-entropy from a viewpoint of a single-label loss function and then derive a new loss to effectively cope with multiple labels.

2.1. Softmax cross-entropy loss for single label

In *single*-label supervised learning, a softmax cross-entropy provides an effective loss function. Suppose an image sample $\mathcal{I}$ equipped with a class label $y \in \{1, \cdots, C\}$. The image is processed by a (neural network) model $f_{\boldsymbol{\theta}}$ parameterized by $\boldsymbol{\theta}$ to produce a logit vector $\boldsymbol{x} = f_{\boldsymbol{\theta}}(\mathcal{I}) \in \mathbb{R}^C$. The softmax loss is formulated by means of cross-entropy between one-hot label y and softmax of logits $\boldsymbol{x}$ as

$$\ell_{\text{sm}} = -\log \frac{e^{x_y}}{\sum_{c=1}^{C} e^{x_c}}, \tag{1}$$

where x_c indicates the c-th element of a logit vector $\boldsymbol{x}$. Following [15], we reformulate it toward a *loss*-like form akin to hinge loss [6, 29];

$$\ell_{\text{sm}} = \log\left[e^{-x_y}\sum_{c=1}^{C} e^{x_c}\right] = \log\left[1 + e^{-x_y}\sum_{c\neq y} e^{x_c}\right] \tag{2}$$

$$= \mathsf{softplus}\left[\log\left\{\sum_{c\neq y} e^{x_c}\right\} - x_y\right], \tag{3}$$

where $\mathsf{softplus}(\cdot) = \log[1 + \exp(\cdot)]$ is a softplus function, smooth approximation of a hinge function. In this form, the log-sum-exp resembles a maximum operator to provide a *hard* logit over *negative* classes as $\log\sum_{c\neq y}\exp(x_c) \approx \max_{c\neq y} x_c$ [15]. This reformulation (3) reveals that a softmax loss measures difference between the positive x_y and the *hard* negative $\log\sum_{c\neq y}\exp(x_c)$ via a softplus function.

2.2. Multi-label loss

We then consider a multi-label setting where a logit vector $\boldsymbol{x}$ is associated with multiple labels $\mathcal{P}$, a set of positive labels assigned to $\boldsymbol{x}$; the number of positive labels is $1 \leq |\mathcal{P}| < C$ and we denote a set of negative labels by $\mathcal{N}$ such that $|\mathcal{P}\cup\mathcal{N}| = C$ and $\mathcal{P}\cap\mathcal{N} = \varnothing$. For correctly identifying the positive classes $\mathcal{P}$ on multi-label classification, we encourage the logits $\boldsymbol{x}$ to have the relationship of

$$x_p > x_n,\ \forall p \in \mathcal{P}, \forall n \in \mathcal{N} \ \Leftrightarrow\ \min_{p\in\mathcal{P}} x_p > \max_{n\in\mathcal{N}} x_n, \tag{4}$$

where x_p and x_n indicate logits of positive and negative classes, respectively; we refer to them as *positive* and *negative logits*. The analysis in Sec. 2.1 inspires us to formulate a multi-label loss by measuring the difference between the *hard* positive and *hard* negative logits in (4) via softplus.

Since the softmax loss (3) leverages log-sum-exp to represent negative logits, our particular interest is to effectively describe the hard positive of $\min_{p\in\mathcal{P}} x_p$. We follow the approach to replace the min operator by log-sum-exp. The minimum of positive logits can be written by using max operator so that log-sum-exp is applied as follows.

$$\min_{p\in\mathcal{P}} x_p = -\max_{p\in\mathcal{P}}(-x_p) \approx -\log\sum_{p\in\mathcal{P}} e^{-x_p}. \tag{5}$$

We embed this hard positive into a softplus loss function (3) by replacing the single positive logit x_y to construct the multi-label loss of

$$\tilde{\ell} = \mathsf{softplus}\left[\log\left\{\sum_{n\in\mathcal{N}} e^{x_n}\right\} + \log\left\{\sum_{p\in\mathcal{P}} e^{-x_p}\right\}\right] \tag{6}$$

$$= \log\left[1 + \left\{\sum_{n\in\mathcal{N}} e^{x_n}\right\}\cdot\left\{\sum_{p\in\mathcal{P}} e^{-x_p}\right\}\right]. \tag{7}$$

2.3. Classification margin

To enhance the discriminative relationship (4), a margin for multi-label classification is naturally defined as

$$\Delta \triangleq \min_{p\in\mathcal{P}} x_p - \max_{n\in\mathcal{N}} x_n. \tag{8}$$

The multi-label loss (6) implicitly enlarges the margin due to log-sum-exp which has the following property.

Proposition 1. *The difference between hard positive and negative in* (6) *is larger than the negative margin,* $-\Delta$*:*

$$\log\sum_{n\in\mathcal{N}} e^{x_n} + \log\sum_{p\in\mathcal{P}} e^{-x_p} > \max_{n\in\mathcal{N}} x_n - \min_{p\in\mathcal{P}} x_p. \tag{9}$$

Proof. A log-sum-exp function is rewritten as

$$\log\sum_{n\in\mathcal{N}} e^{x_n} = x_{n^*} + \log\left(1 + \sum_{n\neq n^*} e^{x_n - x_{n^*}}\right) > x_{n^*}, \tag{10}$$

where $n^* = \arg\max_{n\in\mathcal{N}} x_n$ and $\sum_{n\neq n^*} e^{x_n - x_{n^*}} > 0$. Similarly, we have

$$\log\sum_{p\in\mathcal{P}} e^{-x_p} > \max_{p\in\mathcal{P}}(-x_p) = -\min_{p\in\mathcal{P}} x_p. \quad \square \tag{11}$$

Thus, the margin in the multi-label loss (6) is underestimated toward larger margin through minimizing the loss as in large-margin methods [6, 29].

Based on the margin-based analysis, we further enhance large-margin effect in the multi-label loss for improving discrimination. The implicit margin in (6) resorts to the gap

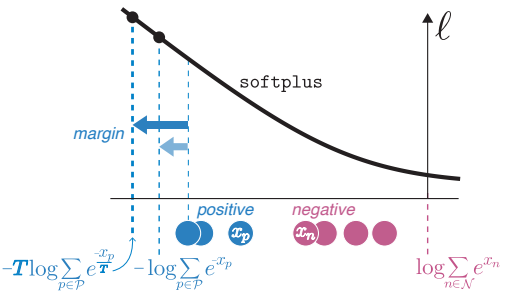

Figure 1. Our multi-label loss function enlarges a margin between positive and negative logits by a temperature parameter T.

shown in (10), that is, $\log\sum_c e^{x_c} - \max_c x_c = \log(1 + \sum_{c\neq c^*} e^{x_c - x_{c^*}})$. We thus introduce *temperature* T into log-sum-exp for enlarging the gap as follows.

Proposition 2. *A temperature parameter* $T > 0$ *reformulates* log-sum-exp *as* $T\log\sum_c e^{\frac{x_c}{T}}$ *which has*

$$\max_c x_c < T\log\sum_c e^{\frac{x_c}{T}} \leqslant \log\sum_c e^{x_c} \quad \textit{for} \quad T \leqslant 1. \tag{12}$$

Proof. The tempered log-sum-exp is written by

$$T\log\sum_c e^{\frac{x_c}{T}} = x_{c^*} + T\log\left(1 + \sum_{c\neq c^*} e^{\frac{x_c - x_{c^*}}{T}}\right), \tag{13}$$

where the second term monotonically increases w.r.t. $T > 0$ as $c^* = \arg\max_c x_c$ and $x_c - x_{c^*} \leq 0, \forall c \neq c^*$. $\square$

Thereby, we propose the following multi-label loss.

$$\ell = \mathsf{softplus}\left[T_{\mathcal{N}}\log\sum_{n\in\mathcal{N}} e^{\frac{x_n}{T_{\mathcal{N}}}} + T_{\mathcal{P}}\log\sum_{p\in\mathcal{P}} e^{-\frac{x_p}{T_{\mathcal{P}}}}\right], \tag{14}$$

where two temperatures $T_{\mathcal{P}}$ and $T_{\mathcal{N}}$ are applied to positive and negative logits, respectively. While such a temperature is incorporated into softmax in the field of knowledge distillation [13], in this study, the temperature controls a margin in the multi-label loss as shown in Fig. 1. Particularly, a margin is further underestimated by $T > 1$ for inducing larger-margin classification. As to negative logits, however, the number of negatives is generally larger than that of positives, $|\mathcal{N}| \gg |\mathcal{P}|$, and thereby sufficiently large gap of $\log(1 + \sum_{n\neq n^*} e^{x_n - x_{n^*}})$ can be given in (10) even for $T_{\mathcal{N}} = 1$, inducing large margin on the side of negatives. Therefore, the temperature parameters can be reduced into only a positive one $T_{\mathcal{P}}$ by $T_{\mathcal{N}} = 1$ in (14) for considering the larger margin only on positive logits via $T_{\mathcal{P}} \geq 1$ as shown in Fig. 1; it is empirically discussed in Sec. 4.2. It should be noted that the proposed loss (14) enhances classification margin without imposing regularization on logits [15] nor manipulating logits [20].

2.4. Two-way multi-label loss

We have analyzed a loss function from a viewpoint of *sample-wise* classification and then formulated the multi-label loss function (14) to work at each sample. On the other hand, a BCE loss commonly applied in the multi-label classification considers rather to discriminate samples at each class by means of a sigmoid function, which can be regarded as *class-wise* classification. These two types of approaches deal with classification in orthogonal directions, and in this study we unify them into *two-way* loss for multi-label classification on the basis of a logit matrix shown in Fig. 2; in the matrix, the sample-wise classification is performed at each *row* while the class-wise one is along each *column*, which are thus regarded as *two-way* classifications.

Notation. Suppose a *bucket*[1] of M image samples. As shown in Fig. 2, it produces a logit matrix $\boldsymbol{X} \in \mathbb{R}^{M\times C}$ in which the (i,c)-th element x_{ic} indicates the logit produced by $f_{\boldsymbol{\theta}}(\mathcal{I}_i)$ for the c-th class. Accordingly, a set of label $y_{ic} \in \{0,1\}$ constitutes a binary label matrix $\boldsymbol{Y} \in \{0,1\}^{M\times C}$.

Sample-wise way: In Sec. 2.2, we have considered a sample-wise loss by applying the multi-label loss function to row-wise logits $\{x_{ic}\}_{c=1}^{C}$ as

$$\ell_i = \mathsf{softplus}\left[\log \sum_{n|y_{in}=0} e^{x_{in}} + T \log \sum_{p|y_{ip}=1} e^{-\frac{x_{ip}}{T}}\right], \quad (15)$$

which enhances a discrimination among classes enforcing (4) to improve classification accuracy of top-K prediction at the i-th sample. On the other hand, it pays less attention to the relationship among samples such as x_{ic} and x_{jc} at the c-th class, due to which a learnt model might fail to discriminate samples on the c-th class. It leads to inferior performance on the metric of mAP across classes, a popular performance metric in multi-label classification. This is the main reason why the popular softmax loss is less frequently applied to multi-label classification.

Class-wise way: To enhance the discrimination among samples, we apply the multi-label loss function (14) in a *class-wise* way, i.e., column-wise manner in Fig. 2, to produce the loss of

$$\ell^c = \mathsf{softplus}\left[\log \sum_{i|y_{ic}=0} e^{x_{ic}} + T \log \sum_{j|y_{jc}=1} e^{-\frac{x_{jc}}{T}}\right]. \quad (16)$$

It measures separation between positive and negative *samples* in the c-th class, promoting the relationship of $\min_{j|y_{jc}=1} x_{jc} > \max_{i|y_{ic}=0} x_{ic}$ to improve precision at each class. It is noteworthy that we apply the same loss function as (15) just by feeding column-wise samples $\{x_{ic}\}_{i=1}^{M}$. While a pair-wise comparison $x_{jc} > x_{ic}$ is considered in a SVM framework for shallow models [36], we leverage our multi-label function (14) to effectively formulate the column-wise classification loss with a proper margin induced by a temperature T, working in end-to-end learning of deep models.

[1]We use a term of *bucket* to distinguish from *batch* in deep learning. B samples in a batch can be divided into k buckets of $M = B/k$ samples.

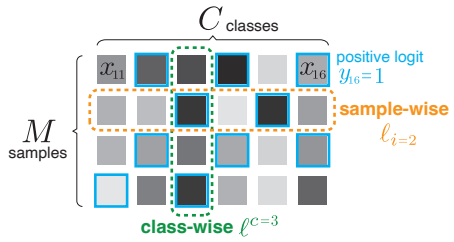

Figure 2. Logit matrix on a bucket of M images.

We finally formulate our *two-way* multi-label loss $\bar{\ell}$ by

$$\bar{\ell} = \frac{1}{M}\sum_{i=1}^{M} \ell_i(\{x_{ic}, y_{ic}\}_{c=1}^{C}; T) + \frac{1}{C}\sum_{c=1}^{C} \ell^c(\{x_{ic}, y_{ic}\}_{i=1}^{M}; T). \quad (17)$$

The proposed two-way loss contributes to improving both sample-wise and class-wise discrimination. It should be noted that even in a case of single (one-hot) label $\sum_c y_{ic} = 1, \forall i$, the class-wise samples demand multi-label classification (column in Fig. 2). Thus, this two-way formulation is intrinsically coupled with the multi-label loss function (14).

3. Discussion

3.1. Loss function

Softmax loss: It is possible to apply a softmax loss in a multi-label scenario through describing labels in a probabilistic way,

$$\ell_{\mathrm{ce}} = -\sum_c \tilde{y}_c \log \frac{e^{x_c}}{\sum_{c'} e^{x_{c'}}}, \quad (18)$$

where $\tilde{y}_c = \frac{y_c}{\sum_{c'} y_{c'}}$ is a normalized label. This enforces the positive logits $\{x_p\}_{p\in\mathcal{P}}$ to exhibit the same value so as to produce $\frac{e^{x_p}}{\sum_c e^{x_c}} = \frac{1}{|\mathcal{P}|}, \forall p \in \mathcal{P}$. It, however, is less relevant to the multi-label classification (4), being a meaningless constraint on feature representation. In contrast, the proposed loss (14) lets the positive logits take any higher values $x_p > x_{p^*}$ by paying attention to the hard positive logit. It also coincides with the softmax loss in case of single-label classification, i.e., $|\mathcal{P}| = 1$.

BCE loss: The multi-label classification is addressed mainly by a binary cross-entropy (BCE) loss of

$$\ell_{\mathrm{bce}} = -\sum_c y_c \log \frac{1}{1+e^{-x_c}} - (1-y_c)\log \frac{1}{1+e^{x_c}}, \quad (19)$$

which can be contrasted with our multi-label loss in the following three points.

First, the BCE loss is faced with imbalance issue due to $|\mathcal{P}| \ll |\mathcal{N}|$. To mitigate it, some weighting schemes are proposed in a simple statistical manner [25] and a sophisticated adaptive manner [2, 18]. On the other hand, the proposed loss (17) is based on comparison of logits among classes and samples in two ways through excavating the hard negatives and positives by means of log-sum-exp, which naturally alleviates the imbalance issue.

Second, BCE (19) is based on a point-wise loss without paying much attention to relationships among classes and samples, which thus leads to the above-mentioned issue. The proposed loss effectively exploits the discriminative characteristics through relative comparison among classes and samples.

Third, the point-wise loss in BCE stems from a sigmoid function which provides classification on the basis of *zero*, i.e., $x \lessgtr 0$, which is regarded as a constraint in the classification model. The proposed loss based on relative comparison is actually invariant to logit shift[2], $\bar{\ell}(\boldsymbol{X}, \boldsymbol{Y}) = \bar{\ell}(\boldsymbol{X} + \epsilon, \boldsymbol{Y})$, without imposing constraints on logits, which thereby enables flexible feature representation learning.

Metric learning loss: In the literature of metric learning, a multi-similarity loss [31] is formulated through relative similarity weighting as

$$\ell_{\mathrm{ms}} = \log\left(1 + \sum_{p\in\mathcal{P}} e^{-x_p}\right) + \log\left(1 + \sum_{n\in\mathcal{N}} e^{x_n}\right), \quad (20)$$

where a logit x_c is regarded as a similarity to the c-th class. It is rewritten by using log-sum-exp with a BCE loss (19) to

$$\ell_{\mathrm{ms}} = \ell_{\mathrm{bce}}\left(\left\{-\log\sum_{p\in\mathcal{P}} e^{-x_p}, \log\sum_{n\in\mathcal{N}} e^{x_n}\right\}, \{+1, -1\}\right). \quad (21)$$

This reformulation clarifies that the multi-similarity loss [31] considers logit relationships within the respective groups of positives and negatives to partially mitigate the above-mentioned issues of BCE, though inheriting the zero-basis constraint.

In the metric learning, a circle loss [27] to learn similarities is also presented in a similar form to our loss (7). The circle loss is derived from maximizing pair-wise discrepancy of $\sum_{p\in\mathcal{P}}\sum_{n\in\mathcal{N}} e^{x_n - x_p}$ in a rather heuristic manner. In contrast, we theoretically formulate the multi-label loss (14) through analyzing a softmax loss from the viewpoint of a margin-aware loss function. Thus, the tempered log-sum-exp is introduced into the loss to adaptively control margin based on the logit distribution for enhancing classification margin. In [27], large-margin discrimination is realized by means of logit bias as $x_p - \epsilon_{\mathcal{P}}$ and/or $x_n + \epsilon_{\mathcal{N}}$. In the multi-label scenario, however, due to the above-mentioned invariance against logit shift, the logit biases are reduced into a constant bias[2] $\epsilon = \epsilon_{\mathcal{P}} + \epsilon_{\mathcal{N}}$ in a softplus function (6); such a bias simply increases the loss in disregard of logits, thus less contributing to larger-margin classification.

[2]Details are shown in the supplementary material.

3.2. Two-way formulation

Class-wise way: In the proposed two-way loss, the class-wise loss is especially distinctive in comparison to a standard softmax loss (1) which considers only sample-wise relationships. It resembles contrastive learning in self-supervised learning [11]; from that view, the target c-th class vector is trained so as to be close enough to the samples associated with the c-th class while being away from the other samples. Thus, the proposed two-way loss would contribute to favorable feature representation learning.

Joint way: The proposed loss (17) compares logits in respective two ways of class- and sample-wise directions (Fig. 2). It is also conceivable to contrast positive logits with negatives in a *joint* manner for encouraging $\min_{j,p|y_{jp}=1} x_{jp} > \max_{i,n|y_{in}=0} x_{in}$, which leads to

$$\ell_{\mathrm{j}} = \mathsf{softplus}\left[\log\sum_{i,n|y_{in}=0} e^{x_{in}} + T\log\sum_{j,p|y_{jp}=1} e^{-\frac{x_{jp}}{T}}\right]. \quad (22)$$

The joint loss is different from (15, 16) in that summation indexes run over all the elements in a logit matrix. As the joint loss measures two-way discrimination at once by using only a single loss function, it would be vulnerable to irregular logit values, such as too large negative and/or small positive logits, which dominate the loss via log-sum-exp function. We empirically compare the proposed two-way loss with the joint loss (22) in Sec. 4.2.

4. Results

We apply the proposed loss to train CNNs on multi-label image classification. Performance of multi-label classification is measured based on the following two metrics.

mAP@class: Average precision over N samples is computed at each class and then is aggregated across classes by

$$\frac{1}{C}\sum_{c=1}^{C} \mathtt{AP}(\{x_{ic}, y_{ic}\}_{i=1}^{N}), \quad (23)$$

where $\mathtt{AP}(\cdot)$ computes average precision of input samples. This is a popular metric in the multi-label classification.

mAP@sample: In contrast to mAP@class, we measure mean average precision over *samples* by

$$\frac{1}{N}\sum_{i=1}^{N} \mathtt{AP}(\{x_{ic}, y_{ic}\}_{c=1}^{C}). \quad (24)$$

It is connected to classification accuracy, a standard metric for *single*-label classification; AP at each sample summarizes top-K accuracies over Ks.

Table 1. Performance results with various temperatures. The performances of vanilla setting, $T_{\mathcal{P}} = 1$ and $T_{\mathcal{N}} = 1$, are underlined.

		mAP@class				mAP@sample			
	$T_{\mathcal{N}}$	0.5	1	2	4	0.5	1	2	4
	$\to 0$	71.03	71.58	71.88	71.86	85.01	85.37	85.55	85.47
	0.5	72.20	72.30	72.17	71.95	85.72	85.81	85.68	85.51
$T_{\mathcal{P}}$	1	72.85	<u>72.93</u>	72.56	72.20	86.07	<u>86.14</u>	85.94	85.66
	2	73.42	73.66	73.23	72.53	86.27	86.48	86.34	85.85
	4	73.56	**74.11**	73.90	73.24	86.26	**86.66**	86.66	86.23

Table 2. Performance results by applying logit bias ϵ [27].

ϵ	0.5	1	2	4	8
mAP@class	73.08	73.16	73.28	**73.44**	73.18
mAP@sample	86.19	86.22	**86.25**	86.21	86.02

4.1. Training procedure

We finetune ImageNet-pretrained CNN models on a target dataset with a multi-label loss. SGD with momentum of 0.9 and weight decay of 10^{-4} is applied to a batch of 512 samples over 40 training epochs with a cosine-scheduled learning rate starting from 0.01 for the FC classifier and 0.001 for the other layers. We apply distributed training across 4 GPUs of NVIDIA V100, which produces 4 buckets of $M = 512/4 = 128$ samples to construct a logit matrix (Fig. 2); the bucket-level losses are averaged in the batch.

4.2. Ablation study

We first analyze the method from various aspects through the following ablation studies, and then in Sec. 4.3 compare the method to the other loss functions on various multi-label tasks. The ablation study is conducted on MS-COCO [19] using ResNet-50 [12].

Temperature T: In the loss (14), temperature parameters $T_{\mathcal{P}}$ and $T_{\mathcal{N}}$ are applied to enhance classification margin on positive and negative logits, respectively. Tab. 1 shows performance results across various temperatures. As discussed in Sec. 2.3, performance is improved by increasing the positive temperature $T_{\mathcal{P}}$ to enlarge margin on the positive side. On the other hand, the negative temperature provides favorable performance by $T_{\mathcal{N}} = 1$, as the number of negative logits is larger than that of positives, in this case $\times 30$ (Tab. 6), to provide a sufficient margin even by $T_{\mathcal{N}} = 1$. Thus, we apply $T_{\mathcal{N}} = 1$ and $T_{\mathcal{P}} = T = 4$ in (15, 16).

We also evaluate the naive hard positive logit of $\min_{p \in \mathcal{P}} x_p$ by simply setting $T_{\mathcal{P}} \to 0$. It significantly degrades performance as shown in the top row of Tab. 1. Such an extreme operator produces sparse backward update as well as provides no margin in the loss. This comparison clarifies that log-sum-exp function effectively produces hard positive logit with a proper margin via T.

Table 3. Performance regarding ways to compute multi-label loss.

	(a) Ours			(b) Softmax		
way	both	class	sample	both	class	sample
mAP@class	**74.11**	<u>73.06</u>	67.18	**69.19**	<u>68.13</u>	58.00
mAP@sample	**86.66**	82.75	<u>86.07</u>	**84.33**	69.15	<u>83.60</u>

Table 4. Performance results by other related loss functions.

	(a) Multi-sim loss (20)			(b) Joint-way (22)		
$T_{\mathcal{P}}$	1	2	4	1	2	4
mAP@class	70.36	70.94	**71.27**	70.19	71.71	**71.82**
mAP@sample	85.42	85.75	**85.95**	84.82	85.57	**85.64**

Table 5. Performance about bucket size M in a batch size of 512.

M	16	32	64	128	256	512
mAP@class	70.76	72.84	73.68	74.11	**74.15**	74.01
mAP@sample	85.79	86.33	86.60	**86.66**	86.58	86.47

Logit bias: In [27], the margin is also discussed through adding bias to logits, which is eventually described by a single bias parameter ϵ to shift the logit difference in a softplus function as described in Sec. 3.1. Tab. 2 shows performance results of various bias ϵ. The logit bias slightly improves performance by $\epsilon = 4$, though being inferior to our temperature approach of $T_{\mathcal{P}} = 4$ (Tab. 1). As shown in (13), the temperature T adaptively controls margin based on the logit distribution while a bias ϵ constantly affects logits in the softplus function, less contributing to large-margin classification during end-to-end learning.

Ways: We then analyze the ways to apply the multi-label loss function (14) to a logit matrix. The proposed *two*-way approach (17) is compared to class-wise (16) and sample-wise (15) losses as shown in Tab. 3a. Those two approaches improve class-wise and sample-wise performances, respectively. By combining those two ways into the loss (17), performance is significantly improved to outperform the respective approaches. The feature representation is effectively learned from these two perspectives to discriminate both samples and classes.

Loss function: The proposed loss is compared to the related loss functions of softmax loss (18), multi-similarity loss (20) [31] and joint-way loss (22), which are mentioned in Sec. 3. As in our loss, the softmax loss is applicable in two ways to improve performance as shown in Tab. 3b, though being inferior to ours (Tab. 3a). Particularly, one-way softmax loss significantly degrades performance on the counter-metric; e.g., performance of class-way loss deteriorates at the metric of mAP@sample. This result indicates that the implicit constraint of uniform positive logits (Sec. 3.1) in the softmax loss would make the feature repre-

Table 6. Multi-label image datasets.

Dataset	MSCOCO [19]	VISPR [23]	VAW [25]	WIDER [17]	VOC2007 [9]	VOC2012 [9]
# classes	80	68	620	14	20	20
# training samples	82,081	14,167	229,076	28,340	5,011	5,717
# test samples	40,137	8,000	31,819	29,117	4,952	5,823
# label per sample	2.9	5.2	1.8	2.9	1.4	1.4

sentation overly fit the target performance metric while impeding generalization. In contrast, the proposed multi-label loss is suitable for the two-way approach without imposing any constraints, to improve performance.

As discussed in Sec. 3.1, the multi-similarity loss (20) [31] evaluates logits on the basis of zero point and thus is directly applicable to multi-label classification as in a BCE loss. For fair comparison, we apply the proposed temperature scaling to the loss as shown in Tab. 4a. The multi-similarity loss works well on both the performance metrics due to the zero-basis measurement. It, however, is significantly inferior to our two-way loss. This comparison clarifies the efficacy of our two-way approach to relatively compare logits without introducing a fixed zero-basis.

The joint loss (22), a variant of our loss, produces poor performance as shown in Tab. 4b. It would be hard to compare diverse logits including inconsistent ones on different samples for different classes, x_{ic} and x_{jd} $(i \neq j, c \neq d)$, in a single loss function at once. On the other hand, the proposed two-way approach measures losses along either of class-wise and sample-wise ways by fixing the other way, which contributes to consistent loss measurement with robustness against irregular logits.

Bucket size M: The class-wise loss (16) is computed on a bucket of M samples which are randomly drawn trhough random mini-batch sampling. We investigate performances across various bucket sizes. In this case, 512 samples in a batch are grouped into k buckets each of which contains $M = 512/k$ samples where $k \in \{1, 2, 4, 8, 16, 32\}$ produces $M \in \{512, 256, 128, 64, 32, 16\}$; it could be naturally and efficiently implemented by distributed training on k GPUs. It should be noted that we fix the batch size to 512 so that the bucket size M affects only the class-wise loss (16) in the two-way loss (17). As shown in Tab. 5, the score of mAP@sample is less sensitive to M since the sample-wise loss (15) responsible for the metric is irrelevant to the bucket size. On the other hand, mAP@class is improved by larger M, implying that it would be somehow difficult to explore discriminative features from a smaller number of samples; as $M \rightarrow 1$, the proposed two-way loss is reduced to the sample-wise loss disregarding discrimination over samples. Discriminativity among samples is enhanced by moderate size of bucket to boost performance on *both* metrics in our two-way framework. We use $M = 128$ throughout the experiments while $M \geq 64$ produces favorable performance.

4.3. Performance comparison

We compare the proposed loss with the others by using various CNN models on diverse datasets. We employ ResNet-50 [12], ResNeXt-50 [38], DenseNet-169 [14] and RegNetY-32gf [26]. Performances are reported on datasets of MS-COCO [19], VISPR [23], VAW [25], WIDER Attribute [17] and VOC-2007/2012 [9], which are summarized in Tab. 6. The comparison methods include softmax cross-entropy loss (18), binary cross-entropy (BCE) loss weighted by class frequencies [25], Focal loss [18] and asymmetric loss (ASL) [2]; we set the hyper-parameters to $\gamma = 2$ in FocalLoss and $\{\gamma_+ = 0, \gamma_- = 4, m = 0.05\}$ in ASL as suggested in [2, 18]. For fair comparison, we apply the same training protocol as shown in Sec. 4.1; only the loss functions are compared by fixing the other components.

The performance results are shown in Tab. 7, demonstrating efficacy of our loss on diverse datasets with various CNN models. While the softmax loss works well only on the metric of mAP@sample, the proposed two-way loss is superior on both metrics, outperforming the others. Some other experimental results are shown in the supplementary material.

4.4. ImageNet of single-label dataset

Finally, we apply the loss to train ResNet-50 [12] on ImageNet [7] posing single-label classification; the training protocol is the same as Sec. 4.1 except that the model is trained from scratch over 90 epochs with an initial learning rate of 0.1. In the ImageNet, the proposed two-way loss differs from a softmax loss in that the class-wise loss (16) is incorporated, while the sample-wise loss (15) is reduced to the softmax loss (1) in this single-label learning.

To evaluate the feature representation learnt by the losses, we measure classification accuracy not only on ImageNet validation set but also on the other datasets by following the transfer learning scheme; pretrained ResNet-50 backbone is applied as frozen feature extractor and only an FC classifier is tuned on the downstream tasks, details of which are shown in supplementary material. Tab. 8 reports classification accuracies of ASL [2], softmax (1) and our loss (17) on various datasets of single-label classification.

While the three losses produce similar performances on ImageNet, one can find difference among the transfer learning performances. ASL is inferior to the other two losses

Table 7. Classification accuracies (%) on diverse datasets with various CNNs.

		mAP@class				mAP@sample			
	CNN	ResNet50	ResNeXt50	DenseNet169	RegNetY32gf	ResNet50	ResNeXt50	DenseNet169	RegNetY32gf
MSCOCO [19]	Softmax	58.00	59.53	58.21	64.14	83.60	84.46	83.13	86.92
	BCE	67.71	69.68	64.04	73.38	79.65	80.62	76.14	83.21
	Focal [18]	69.42	71.33	67.18	74.99	84.38	85.22	83.33	87.51
	ASL [2]	70.92	73.04	69.25	76.70	85.05	86.06	84.40	88.29
	Ours	**74.11**	**75.44**	**73.51**	**79.57**	**86.66**	**87.11**	**86.62**	**89.54**
VISPR [23]	Softmax	36.61	36.97	28.64	36.79	85.23	**85.43**	83.75	85.90
	BCE	44.22	45.34	39.73	46.11	72.39	73.14	69.31	73.75
	Focal [18]	46.89	47.76	40.78	48.75	84.35	84.26	82.91	85.29
	ASL [2]	48.53	49.53	42.61	51.03	84.81	84.99	83.99	86.15
	Ours	**51.89**	**52.79**	**48.57**	**53.75**	**85.64**	85.40	**85.88**	**86.67**
VAW [25]	Softmax	52.59	53.33	47.30	55.02	77.68	78.09	75.97	78.99
	BCE	51.21	51.31	44.53	52.25	72.43	72.29	66.50	72.43
	Focal [18]	54.38	54.50	48.94	56.77	77.66	77.70	75.81	78.71
	ASL [2]	55.39	55.72	48.17	57.88	78.05	78.32	75.91	79.03
	Ours	**56.42**	**57.00**	**54.28**	**59.33**	**78.81**	**78.95**	**78.36**	**80.07**
WIDER [17]	Softmax	63.91	65.14	63.61	66.47	83.09	83.74	83.02	84.65
	BCE	70.16	71.40	70.16	73.26	77.62	78.56	77.36	79.94
	Focal [18]	65.88	67.29	64.49	68.72	82.27	82.92	81.89	83.66
	ASL [2]	67.99	69.71	67.34	71.11	83.44	84.12	83.38	85.00
	Ours	**72.28**	**72.77**	**73.03**	**74.92**	**85.43**	**85.43**	**85.87**	**86.97**
VOC2007 [9]	Softmax	83.49	84.31	82.53	86.63	93.23	93.66	92.61	94.99
	BCE	85.58	86.65	81.96	88.25	91.44	91.91	87.55	92.62
	Focal [18]	85.59	86.27	78.33	87.87	93.04	93.24	89.83	94.56
	ASL [2]	86.70	87.53	81.44	89.26	93.42	93.80	91.05	95.08
	Ours	**89.04**	**89.57**	**88.67**	**91.44**	**94.44**	**94.53**	**94.13**	**95.72**
VOC2012 [9]	Softmax	82.46	83.37	81.56	86.48	93.65	93.81	92.72	95.23
	BCE	85.56	86.56	81.94	88.27	92.60	93.01	89.63	94.05
	Focal [18]	85.59	86.04	79.41	87.99	93.56	93.74	90.65	95.17
	ASL [2]	86.56	87.09	81.83	89.06	94.01	94.19	92.18	95.46
	Ours	**88.12**	**88.72**	**87.95**	**91.01**	**94.38**	**94.73**	**94.28**	**96.04**

Table 8. Classification accuracies (%) on various datasets by transferring ImageNet-pretrained ResNet-50 features.

Dataset	ImageNet	Aircraft [21]	Caltech101 [10]	Car [16]	CUB [34]	DTD [5]	Flower [22]	Food101 [3]	Pets [24]	SUN [37]
ASL [2]	**76.76**	27.42	85.22	32.99	55.26	64.89	75.88	57.08	91.33	53.38
Softmax	76.32	39.18	88.17	45.13	63.15	70.85	85.75	65.44	92.12	58.85
Ours	76.29	**44.01**	**88.36**	**46.50**	**65.96**	**72.71**	**87.72**	**66.83**	**92.18**	**59.49**

of softmax and ours; it degrades performance especially on Aircraft dataset[3]. As discussed in Sec. 3.1, ASL which is a variant of BCE loss imposes the zero-basis constraint on the logits in a sigmoid function. It might lead to over-fitting toward the primary (ImageNet) task, hampering generalization of the learnt features. On the other hand, the losses of softmax and ours based on relative comparison among logits let features be flexibly learned to enhance generalization performance. The proposed loss further enhances the discriminative power of feature representation through comparison along two directions of classes and samples in the two-way formulation, which leads to better performance, such as on Aircraft dataset, as shown in Tab. 8.

[3]Aircraft dataset [21] poses fine-grained discrimination of aircraft appearances. Since the ImageNet pre-training task pays less attention to those visual features, the performance comparison on that dataset might highlight difference in general discriminative power of feature representations.

5. Conclusion

We have proposed a novel loss to cope with multiple labels. The multi-label loss function is theoretically formulated in a margin-aware form through analyzing the softmax loss. Then, it is effectively applied in the two-way manner to finally construct multi-label loss for improving both class-wise and sample-wise performance. The experimental results show that the proposed loss is effective not only for improving performance on multi-label classification but also for providing transferrable features on single-label ImageNet pre-training.

References

[1] Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label classification with partial annotations using class-aware selective loss. In *CVPR*, pages 4764–4772, 2022. 2

[2] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, pages 82–91, 2021. 1, 2, 5, 7, 8

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014. 8

[4] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, pages 5177–5186, 2019. 2

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 8

[6] Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001. 2, 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 7

[8] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *CVPR*, pages 647–657, 2019. 2

[9] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 7, 8

[10] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Computer Vision and Pattern Recognition Workshop*, 2004. 8

[11] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 5

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 6, 7

[13] Geoffrey E. Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning Workshop*, 2014. 3

[14] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 2261–2269, 2017. 7

[15] Takumi Kobayashi. Large margin in softmax cross-entropy loss. In *BMVC*, 2019. 2, 3

[16] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Workshop on 3D Representation and Recognition*, 2013. 8

[17] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *ECCV*, 2016. 7, 8

[18] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 1, 2, 5, 7, 8

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. *arXiv,* 1405.0312, 2014. 6, 7, 8

[20] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang. Large-margin softmax loss for convolutional neural networks. In *ICML*, pages 507–516, 2016. 3

[21] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 8

[22] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 8

[23] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *ICCV*, pages 3706–3715, 2017. 7, 8

[24] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012. 8

[25] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *CVPR*, pages 13018–13028, 2021. 1, 2, 5, 7, 8

[26] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *CVPR*, pages 10428–10436, 2020. 1, 7

[27] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *CVPR*, pages 6398–6407, 2020. 2, 5, 6

[28] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007. 1

[29] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, 1998. 2, 3

[30] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*, pages 2285–2294, 2016. 2

[31] Xun Wang, Xintong Han, Weiling Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair-weighting for deep metric learning. In *CVPR*, pages 5022–5030, 2019. 2, 5, 6, 7

[32] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *AAAI*, pages 12265–12272, 2020. 2

[33] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *ICCV*, pages 464–472, 2017. 2

[34] Peter Welinder, Steve Branson, Takashi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010. 8

[35] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *ECCV*, 2020. 2

[36] Xi-Zhu Wu and Zhi-Hua Zhou. A unified view of multi-label performance measures. In *ICML*, 2017. 4

[37] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 8

[38] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995, 2017. 1, 7

[39] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations formulti-label object recognition. In *CVPR*, pages 280–288, 2016. 1, 2