Supplementary Material for 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

A. Bias in logits

We discuss how a bias in logits affects our loss function, which is mentioned in Sec. 3.1. The bias $\epsilon_{\mathcal{P}}$ is added to positive-class logits (positive logits in short) as

$$T_{\mathcal{P}} \log \sum_{p\in\mathcal{P}} e^{-\frac{x_p+\epsilon_{\mathcal{P}}}{T_{\mathcal{P}}}} = T_{\mathcal{P}} \log e^{-\frac{\epsilon_{\mathcal{P}}}{T_{\mathcal{P}}}} \sum_{p\in\mathcal{P}} e^{-\frac{x_p}{T_{\mathcal{P}}}} = -\epsilon_{\mathcal{P}} + T_{\mathcal{P}} \log \sum_{p\in\mathcal{P}} e^{-\frac{x_p}{T_{\mathcal{P}}}}, \tag{i}$$

and similarly the bias $\epsilon_{\mathcal{N}}$ changes the hard negative-class logit (negative logit in short) into

$$T_{\mathcal{N}} \log \sum_{n\in\mathcal{N}} e^{\frac{x_n+\epsilon_{\mathcal{N}}}{T_{\mathcal{N}}}} = \epsilon_{\mathcal{N}} + T_{\mathcal{N}} \log \sum_{n\in\mathcal{N}} e^{\frac{x_n}{T_{\mathcal{N}}}}. \tag{ii}$$

As a result, the loss function is slightly affected by the additive bias $\epsilon_{\mathcal{P}}$ and $\epsilon_{\mathcal{N}}$ as

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

In a case of constant bias, $\epsilon_{\mathcal{P}} = \epsilon_{\mathcal{N}}$, the bias term is canceled out to make the loss invariant against the logit-shift. It also shows that a margin bias presented in [12] is simply reduced into a single parameter $\epsilon = \epsilon_{\mathcal{N}} - \epsilon_{\mathcal{P}}$.

A.1. Comparison to margin bias [12]

Toward large-margin classification, the margin bias ϵ is introduced in [12] to produce a loss ℓ_ϵ by setting $T_{\mathcal{P}} = T_{\mathcal{N}} = 1$ in (iii). The loss gradients with respect to positive logits are given by

$$\frac{\partial \ell_\epsilon}{\partial x_{p'}} = -\texttt{sigmoid}\left(\log \sum_{n\in\mathcal{N}} e^{x_n} + \log \sum_{p\in\mathcal{P}} e^{-x_p} + \epsilon\right) \frac{e^{-x_{p'}}}{\sum_{p\in\mathcal{P}} e^{-x_p}}. \tag{iv}$$

The bias ϵ just modifies the first term to slightly increase the magnitude of the gradients by shifting the logits in the `sigmoid` function.

On the other hand, our temperature-based loss ℓ_T (14) is constructed by setting $T_{\mathcal{P}} = T, T_{\mathcal{N}} = 1$ and $\epsilon = 0$ in (iii). It produces the loss gradient with respect to positive logits as

$$\frac{\partial \ell_T}{\partial x_{p'}} = -\texttt{sigmoid}\left(\log \sum_{n\in\mathcal{N}} e^{x_n} + T \log \sum_{p\in\mathcal{P}} e^{-\frac{x_p}{T}}\right) \frac{e^{-\frac{x_{p'}}{T}}}{\sum_{p\in\mathcal{P}} e^{-\frac{x_p}{T}}}. \tag{v}$$

Compared to a vanilla setting of $T = 1$, the temperature $T > 1$ touches the gradients not only by increasing the first term of `sigmoid` but also by smoothing the second term of softmax on positives to favorably distribute the updating (gradient) across all the positive logits $\{x_p\}_{p\in\mathcal{P}}$. Thus, the proposed method is superior to the margin-bias approach (iv) [12] as shown in Tab. 2.

B. Linear evaluation on transfer learning

In Sec. 4.4 and Tab. 8, loss functions are evaluated in a framework of transfer learning. We first train ResNet-50 on an ImageNet training set by using the loss and then apply it as *frozen* feature extractor to the downstream tasks listed in Tab. A. The linear (FC) classifier is trained on the downstream dataset by using LBFGS with a L_2 regularization weight for which we sampled 45 values equally-spaced on log-scale range between $[10^{-6}, 10^{5}]$; we report the best performance over those regularization parameters.

Table A. Datasets for transfer learning.

Dataset	# class	# training	# test
Aircraft [9]	100	6,667	3,333
Caltech101 [6]	102	3,060	6084
Car [8]	196	8,144	8,041
CUB [13]	200	5,994	5,794
DTD [5]	47	3,760	1,880
Flower [10]	102	2,040	6,149
Food101 [2]	101	75,750	25,250
Pets [11]	37	3,680	3,669
SUN [14]	397	19,850	19,850

C. Additional experimental results

C.1. Comparison to ranking losses

The proposed loss function (14) encourages the positive logits to be larger than the negative ones, $\max_{p\in\mathcal{P}} x_p > \max_{n\in\mathcal{N}} x_n$. In other words, it increases *rank* of the positive logits, thus being connected to ranking losses [3, 7]. We compare the proposed loss with the ranking losses in Tab. B. From a viewpoint of ranking logits, WARP [7] enhances a margin between positive and negative logits at each sample, with a similar motivation to our sample-wise approach (15). Our sample-wise loss discriminates positive and negative classes naturally in a softmax-based formulation, producing superior performance to WARP. While RML [3] incorporates a class-wise loss resembling BCE into the ranking loss, it is inferior to our two-way loss (17).

Table B. Performance comparison to ranking losses on MSCOCO using ResNet-50.

	mAP@class	mAP@sample
WARP [7]	62.54	84.22
Ours (sample-wise) (15)	67.18	86.07
RML [3]	71.37	85.58
Ours (two-way) (17)	74.11	86.66

C.2. NUSWIDE dataset

The methods are also evaluated on NUSWIDE dataset [4] as shown in Tab. Cb, exhibiting favorable performance. We exclude samples which are not equipped with any class labels from provided training/test splits to form the dataset (Tab. Ca).

Table C. Performance results on NUSWIDE dataset [4] using ResNet-50.

(a) Dataset

Dataset	NUSWIDE [4]
# classes	81
# training samples	125,448
# test samples	83,898
# label per sample	2.4

(b) Performance comparison

	mAP@class	mAP@sample
Softmax	49.85	82.77
ASL [1]	58.40	83.73
Ours	59.67	83.60

References

[1] 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. 2

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014. 2

[3] Hakan Cevikalp, Burak Benligiray, Omer Nezih Gerek, and Hasan Saribas. Semi-supervised robust deep neural networks for multi-label classification. In *CVPRW*, pages 9–17, 2019. 2

[4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ACM conference on image and video retrieval*, 2009. 2

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 2

[6] 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. 2

[7] Yunchao Gong, Yangqing Jia, Thomas K. Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv,* 1312.4894, 2014. 2

[8] 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. 2

[9] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv:1306.5151*, 2013. 2

[10] 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. 2

[11] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012. 2

[12] 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. 1

[13] 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. 2

[14] 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. 2