Supplementary Material for

IGNORE: Information Gap-based False Negative Loss Rejection for Single Positive Multi-Label Learning

GyeongRyeol Song, Noo-ri Kim, Jin-Seop Lee, and Jee-Hyong Lee*

Sungkyunkwan University, Suwon, South Korea
{thd7524, pd99j, wlstjq0602, john}@skku.edu

A CAM Extraction

We adopt two types of backbone models, Q2L [6] and ResNet-50 [3], in our main experiments. Due to differences in their model architectures, the methods for extracting Class-Activation Map (CAM) are implemented differently for these two models. In this section, we discuss the CAM extraction methods for each model. In both cases, when utilizing the extracted CAM in our method, we apply min-max normalization to the CAM and perform interpolation to match the resolution of the original image.

A.1 Extraction of CAM from Q2L

Q2L consists of CNN-based backbone and transformer [7] decoder. For attention mechanism, Q2L utilizes the CNN-based backbone output as key and value. And Q2L utilizes learnable label embeddings as queries. The CAM is extracted by performing an element-wise multiplication between the attention weights of multi-head attention's final layer and the feature from the CNN-based backbone. The attention weights for the j-th class are calculated as follows:

$$AttentionWeights_j = softmax(\frac{Q_j K^T}{\sqrt{d_k}}) \in \mathbb{R}^{1\times HW} \tag{1}$$

$\mathcal{F}_0 \in \mathcal{R}^{H\times W\times d_0}$ represents the CNN-based backbone's output feature for an input image $x \in \mathbb{R}^{H_0\times W_0\times 3}$. For the attention operation, a linear projection layer is used to project the dimension of $\mathcal{F}_0$ from dimension of features, d_0, to d, and the projected features are reshape to be $\mathcal{F} \in \mathbb{R}^{HW\times d}$. This $\mathcal{F}$ is used as the key (K) and value (V) for the attention operation, while the learnable embedding $Q \in \mathbb{R}^{C\times d}$ is utilized as the query in the attention operation, where C represents the number of classes. $Q_j \in \mathbb{R}^{1\times d}$ represents the query corresponding to the j-th class. Subsequently, the CAM feature, A, is caculated through element-wise

* Corresponding author

multiplication of attention weights and $\mathcal{F}^T$. The CAM feature for the j-th class is as follows:

$$A_j = \frac{1}{d}\sum_{i=1}^{d} AttentionWeights_j \odot \mathcal{F}_i^T \in \mathbb{R}^{1\times HW} \tag{2}$$

where, $\mathcal{F}_i$ represents the i-th channel of $\mathcal{F}$. By reshaping A_j from HW to $H \times W$, we obtain the CAM for class j.

$$CAM_j \in \mathbb{R}^{1\times H\times W} \tag{3}$$

A.2 Extraction of CAM from ResNet

In this paper, ResNet consists of residual blocks (image encoder), a 1x1 convolutional layer, and global average pooling for the efficiency of CAM extraction following by BoostLU [4]. In this architecture, because the output feature of the 1×1 convolutional layer becomes the CAM itself, there is no need for additional calculations to obtain the CAM. The CAM, output features of the 1×1 convolutional layer, are as follows:

$$A = g(f(x)) \in \mathbb{R}^{H\times W\times C} \tag{4}$$

$f(x) \in \mathbb{R}^{H\times W\times K}$ denotes the output features of the image encoder, and g represents the 1×1 convolutional layer with weights $w \in \mathbb{R}^{K\times C\times 1\times 1}$, where C represents the number of classes. We utilize $A_j \in \mathbb{R}^{H\times W\times 1}$ as the CAM for the j-th class.

Table 1: Benchmark Dataset Information

Information		VOC	COCO	NUSWIDE	CUB
Dataset Information	Classes	20	80	81	312
	Positives per Train Image	1.46	2.94	1.89	31.4
	Training Images	4,574	65,665	120,000	4,795
	Validation Images	1,143	16,416	30,000	1,199
	Test Images	5,823	40,137	60,260	5,794
Single Positive Multi Label Learning with AN	Single Positives	4,574	65,665	93,156	4,795
	False Negatives	2,091	127,413	133,677	145,756
	True Negatives	84,815	5,060,122	9,493,167	1,345,489
	Assumed Negatives	86,906	5,187,535	9,626,844	1,491,245

B Dataset Information

In this section, we introduce the information about the datasets we used for our experiments. We conduct the experiments using the PASCAL VOC 2012

(VOC) [2], MS COCO 2014 (COCO) [5], NUS-WIDE (NUSWIDE) [1], and CUB-200-2011 (CUB) [8] datasets. As mentioned in the main paper, for all datasets, 80% of the original training dataset is used for training data, and the remaining 20% is used for validation. As shown in Tab. 1, in most of multi-label datasets, it is observed that the number of positive labels per image is very low compared to the number of classes. In other words, the number of negative labels is dominant. This dominance of negative labels also occurs in SPML with AN. In the assumed negative labels, the number of true negative labels is significantly greater than the number of false negative labels. This environment makes it challenging to identify false negative labels among assumed negative labels.

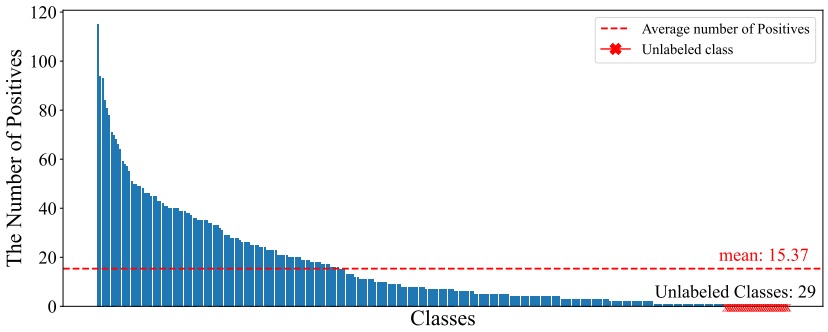


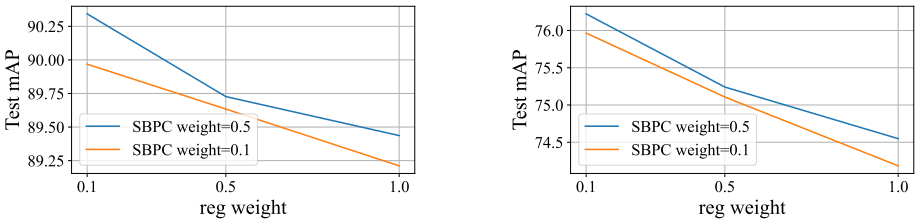
Fig. 1: The number of positive labels per class in the CUB in SPML

In particular, the training images for CUB consist of 4795 images and 312 classes. The number of classes is considerably higher than the number of training images. As shown in Fig. 1, in SPML, the average number of positive labels per class in the CUB training dataset is very low. Additionally, there are classes with no learnable positive labels at all. Therefore, most existing SPML studies have shown low performance on the CUB dataset.

C Analysis on Loss Weights

Our overall loss consists of Single Positive Binary Classification (SPBC), Information Gap-based False Negative Loss Rejection (IGNORE), and Consistency Regularization (reg). In this section, we analyze the performance of the model based on the weight of each loss. For IGNORE, the weight is always fixed at 1, and the CAM threshold τ is fixed at 0.8.

In Fig. 2, it is observed that the performance difference based on the weight of SPBC is slight, and a smaller reg weight results in better performance. For VOC, the performance difference based on the combination of loss weights is insignificant. However, for COCO, the combination of loss weights leads to more variations in model performance compared to VOC.



(a) VOC **(b)** COCO

Fig. 2: Analysis on loss weight for PASCAL and COCO datasets.

D Comparison Through Various Metrics

In this section, we demonstrate the superiority of our method through various metrics. As shown in Tab. 2, we compare the performance of our approach with state-of-the-art methods using not only mAP (mean Average Precision) but also various other metrics (i.e., CP, CR, CF1, OP, OR, and OF1). In the experiments, we utilize Q2L model and PASCAL VOC 2012 dataset. For all metrics except mAP, the confidence threshold is fixed at 0.5. Except for the precision (CP, OP) of PLC + LAC [9], our method outperforms the other methods in all metrics, especially in F1 (CF1, OF1). The reason PLC + LAC [9] has high precision is that it identifies the false negatives through high confidence. As a trade-off for the high confidence, as shown in Tab. 2, PLC + LAC [9] exhibits low recall (CR, OR). In summary, our method demonstrate superior performance across various metrics compared to the other methods.

Table 2: Comparison on PASCAL VOC 2012 with Q2L.

Model	CP	CR	CF1	OP	OR	OF1	mAP
EM+APL [10]	83.7	82.3	82.2	87.3	83.0	85.1	89.3
PLC+LAC [9]	**84.1**	81.1	82.1	**87.4**	80.5	83.8	89.6
OURS	83.7	**84.8**	**83.9**	83.9	**86.7**	**85.3**	**90.1**

References

1. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: Proceedings of the ACM international conference on image and video retrieval. pp. 1–9 (2009) 3
2. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**, 98–136 (2015) 3
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 1

4. Kim, Y., Kim, J.M., Jeong, J., Schmid, C., Akata, Z., Lee, J.: Bridging the gap between model explanations in partially annotated multi-label classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3408–3417 (2023) 2
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 3
6. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834 (2021) 1
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 1
8. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011) 3
9. Xie, M.K., Xiao, J., Huang, S.J.: Label-aware global consistency for multi-label learning with single positive labels. Advances in Neural Information Processing Systems **35**, 18430–18441 (2022) 4
10. Zhou, D., Chen, P., Wang, Q., Chen, G., Heng, P.A.: Acknowledging the unknown for multi-label learning with single positive labels. In: European Conference on Computer Vision. pp. 423–440. Springer (2022) 4