Supplementary Materials of Category-Specific Selective Feature Enhancement for Long-Tailed Multi-Label Image Classification

Ruiqi Du, Xu Tang*, Xiangrong Zhang, Jingjing Ma
Xidian University, China
24171111300@stu.xidian.edu.cn, tangxu128@gmail.com,
xrzhang@mail.xidian.edu.cn, jjma@xidian.edu.cn

This supplementary material contains the following contents. First, we explain the reasons why the proposed category-specific selective feature enhancement (CSSFE) model adopting the asymmetric loss [10] in Section 1. Then, Section 2, as the supplement of ablation analysis, studies the effects of different feature extraction backbones and label embeddings generated by various pre-trained large-scale language models on CSSFE.

1. The Reason for Adopting Ssymmetric Loss

Table 1. mAP (%) of CSSFE with different loss functions.

Datasets	LT-VOC			
Loss	Total	Head	Medium	Tail
DB loss	86.32	82.17	90.34	87.16
PG loss	86.35	82.23	90.68	**87.17**
ASL	**86.44**	**82.33**	**91.10**	87.14

Datasets	LT-COCO			
Loss	Total	Head	Medium	Tail
DB Focal	65.42	64.48	70.14	61.36
PG loss	65.93	65.46	71.08	62.22
ASL	**66.95**	**66.98**	**71.33**	**62.38**

In this paper, we use the basic asymmetric loss (ASL) [10] to optimize the parameters of CSSFE. In fact, there are some alternative options, such as distribution-balanced (DB) loss [13] and probability-guided (PG) loss [6]. However, we find that these losses are unsuitable for CSSFE and have the problem of sacrificing the head categories' performance to improve the tail categories' performance. We use different loss functions to train CSSFEs and count their performance on two datasets to prove this point. As shown in Table 1, the CSSFE trained by DB loss or PG loss only performs well in the tail categories of LT-VOC while having a performance degradation in the head and medium category, decreasing the overall score. For LT-COCO, the CSSFE trained by ASL loss is always the best. This is because DB and CB Focal losses enhance the learning of the tail categories through category-balancing strategies. Although this mechanism alleviates the problem of sparse tail samples, it also produces a gradient suppression effect on the head categories, causing the model to smooth out the features of high-frequency samples during training gradually. The core idea of CSSFE is to use the high sensitivity of deep neural networks to the head categories to improve the confidence of the medium and tail categories. Therefore, DB and PG losses cannot help CSSFE achieve satisfactory results.

2. Supplement of Ablation Analysis

Table 2. mAP (%) of CSSFE with different backbones. The best results are bolded.

Datasets	LT-VOC			
Backbone	Total	Head	Medium	Tail
VGG	86.13	81.04	90.25	86.34
ResNet	86.38	81.93	**91.16**	**87.26**
ViT	**86.44**	**82.33**	91.10	87.14

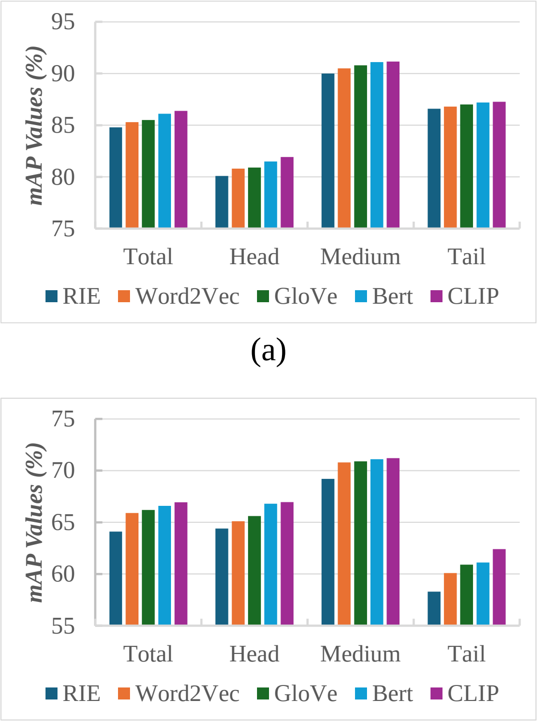
Backbone	LT-COCO			
Methods	Total	Head	Medium	Tail
VGG	65.92	65.13	70.87	60.16
ResNet	66.93	66.96	71.21	**62.41**
ViT	**66.95**	**66.98**	**71.33**	62.38

*:Corresponding author

2.1. Effect of Different Backbones

In this section, we replace the feature extraction backbone of CSSFE with other popular networks to observe the impact of different feature extractors on CSSFE. Specifically, we select VGG16 [12] and ViT-16 [1], which are widely used in the multi-label image classification task [2], as comparisons with ResNet50 [3] that we used. The results of CSSFEs with different backbones counted on various are summarized in Table 2. Through observation, we can find that although the feature extraction capabilities of different backbones vary, the performance of CSSFE is relatively stable, with fluctuations of mAP values are less than 1%. In addition, the parameters of CSSFE with VGG16, ResNet50, and ViT-16 are 142 MB, 151 MB, and 215 MB, respectively. Considering the trade-off between performance and model complexity, we suggest using ResNet50.

2.2. CSSFE with Various Label Embeddings



(a)

(b)

Figure 1. Performance of CSSFE with different label embeddings on different datasets. (a) LT-VOC. (b) LT-COCO.

To verify the generalization of CSSFE with different label embeddings, we additionally select random initialization embedding (RIE) [7], Word2Vec [8], GloVe [9], and BERT [5] to generate label embeddings. It is worth noting that different label embeddings are generated by the same prompt, e.g., "a photo of {class}." The results of CSSFEs with different label embeddings counted on LT-COV and LT-COCO are drawn in Figure 1. By observing these bars, we can find the following points. First, the behavior of CSSFE with RIE is the weakest. This shows that the semantic information in the pre-trained language model is beneficial for long-tailed multi-label image classification tasks. Second, CSSFE with CLIP achieves the best results, which can be attributed to CLIP's bidirectional integration of visual and textual information during training. This enhances its semantic representations, making them particularly well-suited for vision-related tasks. Finally, CSSFE maintains competitive performance when using Word2Vec, GloVe, or BERT, demonstrating its robustness in capturing relevant semantic associations from various label embeddings for the progressive attention enhancement mechanism. Based on the above discussion, we select CLIP as the text encoder for CSSFE.

2.3. Expanded Visualization and Analysis

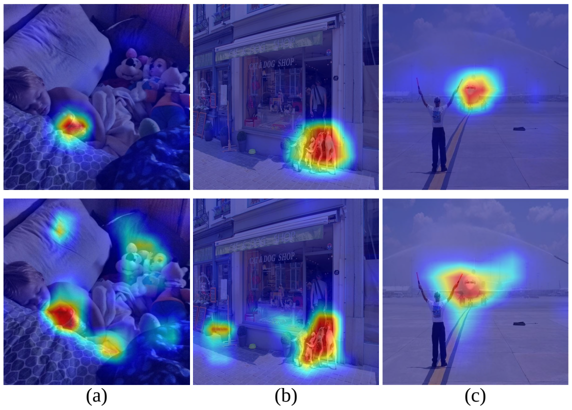


Figure 2. Three small-sized objects and their class activation maps generated by CSSFE with (first row) and without (second row) the distribution-aware binary mask. (a) Head category: cup. (b) Medium category: dog. (c) Tail category: airplane.

To further analyze the impact of the distribution-aware binary mask, we visualize how the class attention scores of different classes change before and after applying the distribution-aware binary mask. As shown in Figure 2, the binary mask effectively suppresses incorrect activations, which are particularly prominent in head and medium categories due to the model's overfitting to these frequently seen classes. Through unidirectional semantic flow, the binary mask not only enhances the model's attention to tail categories but also suppresses interference from other categories. This explains why CSSFE can improve tail category performance without compromising performance on head or medium categories. It is worth noting that we deliberately selected three small-sized objects and used another class activation mapping method Grad-CAM [11] for visualization. The effectiveness of the results highlights CSSFE's localization capability under challenging conditions, while their consistency with the results visualized using CAM in the main paper demonstrates the robustness of CSSFE's visual explanations.

2.4. Expanded Evaluation Metrics and Datasets

In this section, we focus on evaluating the CSSFE's ability to migrate domains. Specifically, we conduct additional experiments on a long-tailed multi-label medical image classification dataset ODIR-5K dataset [14]. It is worth noting that we chose the F1 score [4] to compare from a class-weighted perspective in this section. The results of top-2 comparison methods and CSSFE are shown in Table 3. It is evident that our proposed CSSFE achieves the best performance, which further proves its robustness.

Table 3. F1 scores of different methods across various datasets.

Datasets	LT-VOC	LT-COCO	ODIR-5K
CAE-Net	0.727	0.509	0.872
CPRFL	0.731	0.532	0.893
CSSFE	**0.756**	**0.564**	**0.912**

References

[1] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022. 2

[2] Meng Han, Hongxin Wu, Zhiqiang Chen, Muhang Li, and Xilong Zhang. A survey of multi-label classification based on supervised and semi-supervised learning. *International Journal of Machine Learning and Cybernetics*, 14(3):697–724, 2023. 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[4] Jie Huang, Zhao-Min Chen, Zhao-Min Chen, Xiaoqin Zhang, Xiaoqin Zhang, Yisu Ge, Yisu Ge, Lusi Ye, Lusi Ye, Guodao Zhang, et al. Label decoupling and reconstruction: A two-stage training framework for long-tailed multi-label medical image recognition. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2861–2869, 2024. 3

[5] Mikhail V Koroteev. Bert: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943*, 2021. 2

[6] Dekun Lin. Probability guided loss for long-tailed multi-label image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1577–1585, 2023. 1

[7] 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. 2

[8] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[9] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 2

[10] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 82–91, 2021. 1

[11] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[13] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 162–178. Springer, 2020. 1

[14] Qian Zhou, Hua Zou, and Zhongyuan Wang. Long-tailed multi-label retinal diseases recognition via relational learning and knowledge distillation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 709–718. Springer, 2022. 3