Supplementary Materials
MambaML: Exploring State Space Models for Multi-Label Image Classification

Xuelin Zhu[1] Jian Liu[2] Jiuxin Cao[3] Bing Wang[1*]
[1]The Hong Kong Polytechnic University [2]Ant Group [3]Southeast University
{xuelin.zhu, bingwang}@polyu.edu.hk, rex.lj@antgroup.com, jx.cao@seu.edu.cn

A. Evaluation Metrics

In this study, we employ the mean average precision (mAP) and the average precision (AP) for each category as standard metrics in the domain of multi-label image classification to assess the performance of our proposed MambaML framework. These metrics are widely recognized for reflecting the overall effectiveness and category-specific accuracies of multi-label classification models. In addition, we follow the previous work [1] to present the precision, recall, and F1-measure score for further comparison. Specifically, top three labels determined by confidence scores in descending order for each image are used to compare with the ground truth labels, including the overall precision, recall, F1-measure (OP, OR, OF1) and per-class precision, recall, F1-measure (CP, CR, CF1), which are formulated as follows:

$$\begin{aligned}
\mathrm{CP} &= \frac{1}{C}\sum_i \frac{N_i^c}{N_i^p}, & \mathrm{OP} &= \frac{\sum_i N_i^c}{\sum_i N_i^p},\\
\mathrm{CR} &= \frac{1}{C}\sum_i \frac{N_i^c}{N_i^g}, & \mathrm{OR} &= \frac{\sum_i N_i^c}{\sum_i N_i^g}, \qquad (1)\\
\mathrm{CF1} &= \frac{2\times\mathrm{CP}\times\mathrm{CR}}{\mathrm{CP}+\mathrm{CR}}, & \mathrm{OF1} &= \frac{2\times\mathrm{OP}\times\mathrm{OR}}{\mathrm{OP}+\mathrm{OR}},
\end{aligned}$$

where C is the number of labels, N_i^c is the number of images that are correctly predicted for the i-th label, N_i^p is the number of predicted images for the i-th label, N_i^g is the number of ground truth images for the i-th label. The above mentioned metrics all require a fixed number of labels, but the label numbers of different images are commonly various. Therefore, we further present the OP, OR, OF1 and CP, CR, CF1 metrics under the setting that a label is predicted as positive if its estimated probability is greater than 0.5. Note that mAP and OF1 as well as CF1 are the most important metrics.

Method	VOC 2007	MS-COCO	NUS-WIDE
SSGRL [1]	93.4	83.8	-
SALGL [5]	95.1	85.8	66.3
MambaML + GGNN	**95.2**	**85.9**	**66.5**
SGRE [6]	**95.4**	85.7	65.7
MambaML + MHA	95.3	**85.9**	**66.4**

Table 1. Comparison of our MambaML framework with more recent methods (mAP in %).

B. Comparison with more state-of-the-arts

In this section, we compare the proposed MambaML framework with more recent approaches, including SALGL [5] and SGRE [6], which aim to enhance multi-label image classification performance by generating label-specific visual representations and incorporating additional network modules. Similar to SSGRL [1], SALGL employs a low-rank bilinear pooling model to produce label-specific visual representations, followed by interaction through a gated graph neural network (GGNN). For a fair comparison, our MambaML framework also integrates the GGNN for the interaction between the produced label-specific visual representations. Experimental results are shown in the upper part in Table 1. Notably, our MambaML + GGNN consistently achieves better performance in mAP across all datasets.

In addition, the mechanism developed in SGRE [6] that utilizes multi-head self-attention (MHA) to augment patch-level representations with object-level representations is applied in our MambaML framework for a fair comparison. Experimental results are reported in the lower part in Table 1. As shown, despite a minor mAP delay on the Pascal VOC 2007 dataset [3], our MambaML + MHA implements better performance than SGRE on both the MS-COCO [4] and NUS-WIDE [2] datasets. Overall, our MambaML framework demonstrates its effectiveness in acquiring label-specific visual representations, thereby exhibiting superior performance in multi-label image classification.

[*] Corresponding author.

References

[1] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 522–531, 2019. 1

[2] 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 *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 1

[3] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1

[4] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[5] Xuelin Zhu, Jian Liu, Weijia Liu, Jiawei Ge, Bo Liu, and Jiuxin Cao. Scene-aware label graph learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1473–1482, 2023. 1

[6] Xuelin Zhu, Jianshu Li, Jiuxin Cao, Dongqi Tang, Jian Liu, and Bo Liu. Semantic-guided representation enhancement for multi-label image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):10036–10049, 2024. 1