

Text-Region Matching for Multi-Label Image Recognition with Missing Labels

Anonymous Authors

1 METHODOLOGY

1.1 Why and how does text-region matching help?

To effectively evaluate the performance of text-region image matching, in regard to category c , we begin by simplifying text-image and text-pixel matching¹ as follows:

$$p(g_c, f) = \alpha \sum_{i,j} g_c f_{i,j}^T, \quad (1)$$

where $g_c \in \mathbb{R}^{1 \times d}$ represents the textual representation corresponding to category c , $f_{i,j}$ denotes the pixel-level visual representation of the feature maps $f \in \mathbb{R}^{h \times w \times d}$ at the spatial location (i, j) , and α is a scaling factor. In a similar manner, the text-region matching process can be formulated as follows:

$$p(g_c, f_c) = \alpha \sum_{i,j} g_c f_{i,j}^T r_{c,i,j} = \alpha g_c f_c^T, \quad (2)$$

where $r_c \in \mathbb{R}^{h \times w}$ is visual region for category c , $r_{c,i,j} \in (0, 1)$, and $f_c^T = \sum_{i,j} f_{i,j}^T r_{c,i,j}$ is category-aware region representation.

Assume. In our proposed method, for any text representation g_c embedded in the joint space, the pixel-level representation similarity for relevant categories exceeds 0, while the similarity for irrelevant pixel-level text representations falls below 0. Moreover, for hard samples, the similarity of text representations approaches 0, whereas for easy samples, the similarity tends towards 1 or -1. This assumption can be expressed in the following form:

$$\text{Similarity}(g_c, f_{i,j}) = \begin{cases} g_c f_{i,j}^T \rightarrow 1 & \text{easy positive samples} \\ g_c f_{i,j}^T \rightarrow -1 & \text{easy negative samples} \\ g_c f_{i,j}^T \rightarrow 0 & \text{hard samples} \end{cases} \quad (3)$$

In addition, we compute the cosine similarity between the text embedding and the pixel-level representation. The visualization of the results is depicted in the Figure 1. With the visual region $R = \{r_1, r_2, \dots, r_c\}$ introduced, the aforementioned text-image and

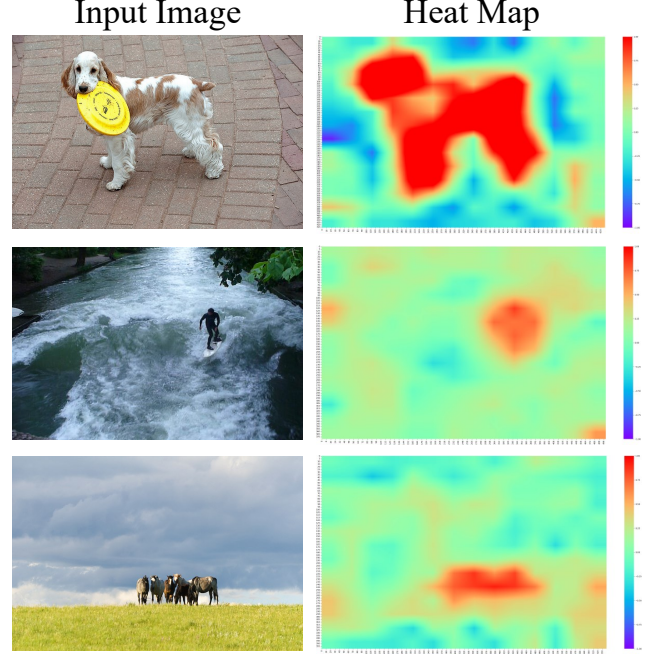


Figure 1: The similarity between text representation and pixel-level representation is visualized using a color gradient, where red indicates higher similarity and blue indicates lower similarity. This is described through three examples: “dog”, “person”, and “horse”.

text-pixel matching method can be expressed as follows:

$$\begin{aligned} & p(g_c, f \cdot (1 - r_c + r_c)) \\ &= \alpha \sum_{i,j} g_c f_{i,j}^T (1 - r_{c,i,j} + r_{c,i,j}) \\ &= \alpha \sum_{i,j} g_c f_{i,j}^T (1 - r_{c,i,j}) + \alpha \sum_{i,j} g_c f_{i,j}^T r_{c,i,j} \\ &= \alpha g_c f_{c-\text{neg}}^T + \alpha g_c f_{c-\text{pos}}^T \end{aligned} \quad (4)$$

where $f_{c-\text{pos}}^T$ represents a regional visual representation associated with category c , and $f_{c-\text{neg}}^T$ represents a regional visual representation unrelated to category c . Given that the regional visual r_c selects the pixel-level representation associated with category c on the feature map, it follows from Eq.(3) that $g_c f_{c-\text{pos}}^T$ is greater than 0, while $g_c f_{c-\text{neg}}^T$ is less than 0. Under the aforementioned assumptions, when category c is present in an image, the prediction scores for both the text representation g_c and the region-level

¹We forego the normalization step to simplify the expression and facilitate understanding when calculating cosine similarity.

representation f_c surpass those for both image-level representation f matching and pixel-level representation $f_{i,j}$ matching. This process can be summarized as follows:

$$p(g_c, f) = \alpha g_c f_{c\text{-neg}}^\top + \alpha g_c f_{c\text{-pos}}^\top < \alpha g f_{c\text{-pos}}^\top = p(g_c, f_c). \quad (5)$$

$$p(g_c, f) < p(g_c, f_c)$$

Conversely, if a category is absent from a given image, it can be described as follows:

$$p(g_c, f) > p(g_c, f_c). \quad (6)$$

Additionally, we have experimentally confirmed that leveraging high-quality visual regions can enhance the model's performance, thereby validating the efficacy of text-region matching, as illustrated in Table. 1.

Table 1: Ablation study examines the influence of high-quality visual regions on text prompt tuning learning in the MS-COCO and VOC 2007 datasets. The “complete” refers to the category-aware region learning module, trained using complete annotation data. All metrics are in %.

Dataset	Method	10%	20%	30%	40%	50%	Avg.
VOC 2007	w/o complete	92.0	93.9	94.4	94.9	95.0	94.0
	w/ complete	94.8	94.9	95.0	95.1	95.1	95.0
MS-COCO	w/o complete	80.8	82.9	83.7	84.3	84.6	83.3
	w/ complete	83.1	83.8	84.2	84.5	84.7	84.1

1.2 Multiple components' relationships and contributions.

Unlike priors prompt-tuning methods [2, 3, 5, 7, 10], this work's core is optimized for matching between text and visual representations. In the MLR-ML setting, CARL introduces noise and lacks precise supervision, which results in sub-optimal performance. As shown in Figure 2, to mitigate the issue, we introduce KD and MMCP to provide CARL with supervision information and combine these two components with MMCP to enable CARL to learn effective regional-level visual representations. KD, MMCP, and MMCL serve CARL to generate higher-quality regional-level representations. Moreover, MMCP and MMCL help the text encoder get higher-quality text representations. **In summary, MMCP and KD contribute to additional supervisory information that enhances CARL's ability to learn more effective regional-level visual representations, which is meaningful for text and region matching.**

To our knowledge, MMCP and MMCL were first used in MLR-ML tasks and modeling intra- and inter-class relations, which was achieved by cross-image, cross-prototype, and inter-modal interaction.

2 SUPPLEMENTARY EXPERIMENTS

2.1 Multimodal category prototype

Multimodal category prototype estimates unknown-label. To evaluate the efficacy of multimodal category prototypes in

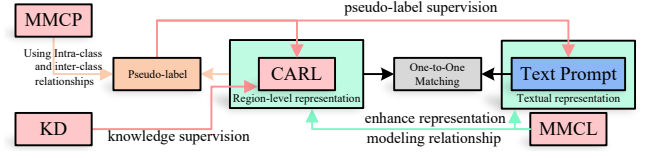


Figure 2: The relationships among multiple components. CARL stands for category-aware region learning. KD refers to the use of knowledge distillation. MMCP represents pseudo-label estimation supported by multimodal category prototypes. MMCL denotes multimodal contrastive learning.

Table 2: The ablation study investigates the effectiveness of multimodal category prototype estimation pseudo-labels in the VOC 2007 dataset at known label proportions of 10% to 50%. The “prototype” refers to the multimodal category prototype module. All metrics are in %.

Dataset	Method	10%	20%	30%	40%	50%	Avg.
VOC 2007	w/o prototype	90.3	92.7	93.8	94.7	94.7	93.2
	w/ prototype	93.6	94.3	94.6	95.0	95.1	94.8

Image	Observed	Estimate	Full label
	N/A	bicycle	bicycle, bottle
	dining table	bottle, dining table, person	bottle, dining table, person
	boat	boat, person	boat, person
	potted plant	chair, dog, potted plant	chair, dog, potted plant, sofa

Figure 3: In the VOC 2007 dataset with 20% known label rate, unknown labels are estimated using multimodal category prototypes.

pseudo-label estimation, we employ two control groups: one without multimodal category prototypes and another with multimodal

category prototypes alone. It is noteworthy that multimodal contrast learning was not employed by either group. As illustrated in the table 2, using category prototypes improves by 1.6% compared to not employing multi-modal category prototypes. Especially at known labels proportion of 10% and 20%, the improvement is more than 2.0%. We further illustrate the accuracy of the proposed method in estimating unknown labels through concrete examples. As illustrated in Figure 3, from left to right, there are the provided image, the observed label, the estimated label, and the complete label. Based on multi-modal prototypes, the method proposed in this paper effectively estimates unknown labels. For instance, in the initial example, although the given image lacks any positive annotations, our method successfully estimates “bicycle” while overlooking “bottle”. This discrepancy may arise from the small size of the “bottle” target relative to the overall image size, rendering it challenging for the model to recognize. In the subsequent examples, even with only one positive sample provided, our method manages to predict all positive annotations. In the third example, while the given image’s positive annotation is “potted plant”, our multimodal prototype model also predicts “chair, dog, potted plant”, while missing “sofa”. This inconsistency could be attributed to the high variability in the appearance of “sofa” and the complexity of image scenes, posing challenges for accurate identification of multi-modal prototypes.

We conducted a quantitative analysis of the accuracy of multi-modal prototype estimation for unknown labels, with specific results presented in Table 2. This analysis was performed on the VOC2007 dataset across a range of label proportions from 10% to 50%. Three key metrics were evaluated: Accuracy, Precision, and Recall. As indicated in Table 2, the effectiveness and accuracy of the multi-modal prototype pseudo-labels improve with increasing label proportions. This may be attributed to the observation that a larger volume of labeled data enables the model to more effectively learn the distribution of the data, resulting in more reliable pseudo-labels.

Visualization of multimodal category prototypes. To offer further insights into the multimodal category prototypes, we conducted visualization analyses of these prototypes within the latent feature space. As shown in Figure 5, t-SNE [6] is used to visualize multi-modal category prototypes. Within the embedding space, categories are separated to the greatest extent feasible, whereas categories that belong to the same superclass are clustered more closely. Textual prototypes exhibit a more uniform data distribution compared to visual prototypes, which forge stronger connections based on category relevance. For instance, in the vicinity of the “traffic signals” category, there is often a cluster of “transportation” categories (e.g., bus, truck, and bicycle), and some “outdoor” categories (e.g., “parking meters” and “fire hydrants”), which exhibit higher co-occurrence among themselves.

2.2 How to establish a one-to-one match

As shown in Figure 5, we visualize one-to-one matching. To achieve one-to-one matching between text representations and corresponding visual representations for each category, we first learn semantic-aware region representations. To this end, the proposed method first identifies regions of interest (ROIs) on the spatial

Table 3: The ablation study investigates the effectiveness of multimodal category prototype estimation pseudo-labels in the VOC 2007 dataset. The “prototype” refers to the multi-modal category prototype module. All metrics are in %.

Dataset	Label Proportions	Accuracy	Precision	Recall
VOC 2007	10%	74.4	98.5	75.7
	20%	77.9	98.8	79.7
	30%	80.8	99.1	82.1
	40%	83.6	99.4	83.9
	50%	86.3	99.5	87.1

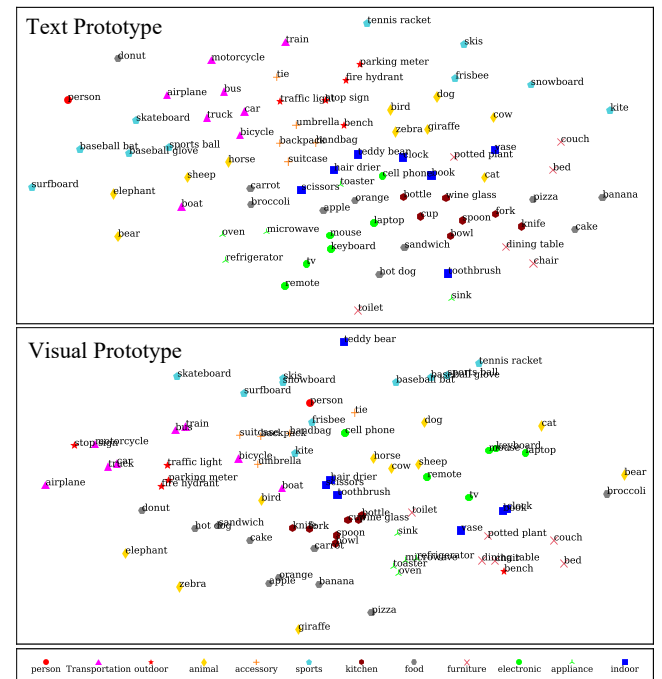


Figure 4: Visualize multimodal prototypes on MS-COCO with known labels proportion of 10%, including text prototypes and visual prototypes. Different colors and shapes mean different superclass .

representation of the image using semantic-aware region learning. Then, it aggregates the spatial visual representations based on the predicted probability (i.e., energy) of each position within the ROI to obtain semantic-aware region representations. Next, similar to the previous multi-label prompt tuning methods [3, 5, 9], text representation can be easily obtained through flexible parameter fine-tuning. Finally, one-to-one matching is achieved by calculating the cosine similarity between the text representation and the visual representation.

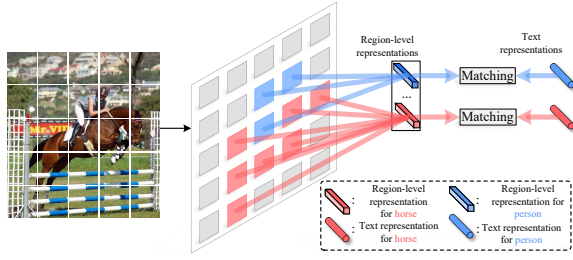


Figure 5: Visualize one-to-one matching process.

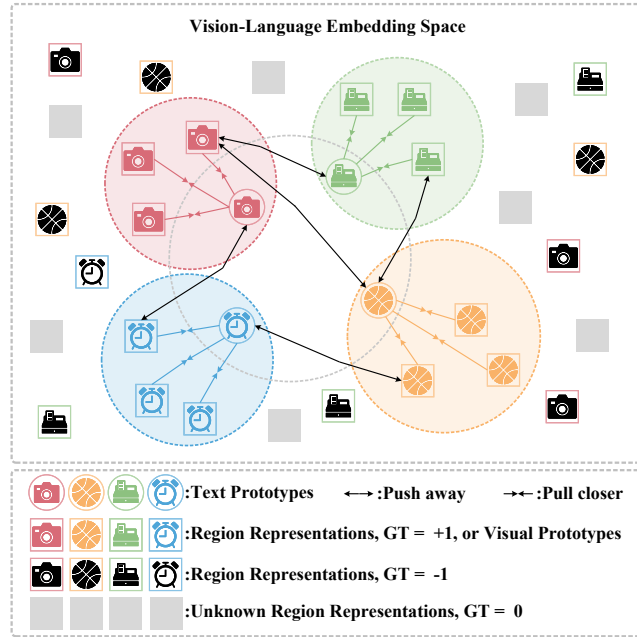


Figure 6: Visualization of illustrates how the proposed multimodal contrastive learning is implemented in the visual-language embedding space.

2.3 How multimodal contrastive learning works

At its core, multimodal contrastive learning revolves around the principle of contrasting similar and dissimilar data pairs. By carefully crafting these pairs and leveraging appropriate loss functions, the learning process drives the model to embed similar multimodal instances into close proximity in a latent representation space, while pushing apart those that are dissimilar. As depicted in the Figure 6, MMCL selects text representations as anchors, category-aware region representations of the same category as positive samples, and category-aware region representations of irrelevant categories as negative samples. It is worth noting that all visual region representations are sourced from the current batch or memory bank queue. In the visual language embedding space, the category-aware area representation is drawn nearer to the text representation of the corresponding category and distanced from non-related text representations. Furthermore, the relationship between intra-class and inter-class distances is also modeled, aiming to minimize intra-class

distances and maximize inter-class distances. In the visual language embedding space, the category-aware region representation is drawn nearer to the text representation of the corresponding category and distanced from non-related text representations. Furthermore, the relationship between intra-class and inter-class distances is also modeled, aiming to minimize intra-class distances and maximize inter-class distances, which is different from modeling correlation methods on graph in multi-label recognition tasks [1, 4, 8].

REFERENCES

- [1] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-Label Image Recognition With Graph Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5177–5186.
- [2] Zixuan Ding, Ao Wang, Hui Chen, Qiang Zhang, Pengzhang Liu, Yongjun Bao, Weipeng Yan, and Jungong Han. 2023. Exploring Structured Semantic Prior for Multi Label Recognition With Incomplete Labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3398–3407.
- [3] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, and Wangmeng Zuo. 2023. Texts as Images in Prompt Tuning for Multi-Label Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2808–2817.
- [4] Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. 2022. Semantic-Aware Representation Blending for Multi-Label Image Recognition with Partial Labels. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 2 (Jun. 2022), 2091–2098.
- [5] Ximeng Sun, Ping Hu, and Kate Saenko. 2022. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems* 35 (2022), 30569–30582.
- [6] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
- [7] Shuo Yang, Zirui Shang, Yongqi Wang, Derong Deng, Hongwei Chen, Qiyuan Cheng, and Xinxiao Wu. 2024. Data-free Multi-label Image Recognition via LLM-powered Prompt Tuning. *arXiv preprint arXiv:2403.01209* (2024).
- [8] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. 2020. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI* 16. Springer, 649–665.
- [9] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.
- [10] Xuelin Zhu, Jiuxin Cao, Dongqi Tang, Furong Xu, Weijia Liu, Jiawei Ge, Bo Liu, Qingpei Guo, Tianyi Zhang, et al. 2023. Text as Image: Learning Transferable Adapter for Multi-Label Classification. *arXiv preprint arXiv:2312.04160* (2023).