$(ML)^2$P-Encoder: On Exploration of Channel-class Correlation for Multi-label Zero-shot Learning

Ziming Liu[1], Song Guo[1,2], Xiaocheng Lu[1], Jingcai Guo[1,2]*, Jiewei Zhang[1], Yue Zeng[1], Fushuo Huo[1]
[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
[2]The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China
{ziming.liu, jiewei.zhang, fushuo.huo}@connect.polyu.hk
{song.guo, xiaoclu, jc-jingcai.guo, zengyue.zeng}@polyu.edu.hk

A. Supplementary Materials

Table 1. **Performance Comparison** using different feature pyramid models.

Pyramid Model	Task	mAP	F1 (K = 3)	F1 (K = 5)
FSSD [2]	ZSL	24.6	29.9	30.7
	GZSL	7.2	12.9	15.8
M2Det [5]	ZSL	29.2	32.1	31.4
	GZSL	9.9	15.4	18.9
Our Approach	ZSL	**29.4**	**32.8**	**32.3**
	GZSL	**10.2**	**15.8**	**19.2**

Table 2. **Inference Time Comparison** using different feature pyramid models.

Pyramid Model	Inference Time (ms)↓
FSSD [2]	2.1
M2Det [5]	3.5
Our Approach	**1.1**

A.1. Different Feature Pyramids

We replace the Forward Pyramid in the C^3-MLZSL model with *FSSD* [2] and *M2Det* [5] in the CNN object detection network, respectively. All experiments use the same VGG19 [4] as the backbone network. Table 1 shows the results produced by our method using different feature pyramids. It can be seen from the experimental results that our proposed Forward Pyramid achieves the best results in both ZSL and GZSL tasks. *M2Det* [5] has also achieved quite good results, but its own structure, each TUM module requires a lot of computing power. The performance of *FSSD* [2] is not satisfactory, mainly because it does not choose to align the larger feature map with the smallest one, but up-samples the smallest feature map, which makes the model lose the judgment of the main information in each image.

Besides, in order to verify the difference in computational power requirements of different feature pyramids, we compare their inference time. Table 2 shows the comparison results. We experiment with three different models, using the same NVIDIA RTX 3090 graphics card, and the results demonstrate that our model has minimal computational overhead. And the longer inference time of *M2Det* [5] proves that its complex structure is the reason.

A.2. Attention Map Comparison

Figure 1 shows the comparison of our method with the region-based bi-attention module BiAM [3] on attention visualization. As can be seen in Figure 1(a), BiAM's attention to the unseen class 'cloud' is almost non-existent, while our method captures the cloud behind 'person'. The same problem also appeared in Figure 1(b). For the recognition of the unseen class 'running', BiAM appears to be shifted to the road, and the attention is not focused on the legs of the person. And our method directly locates the human legs. Beyond that, the focus on the class 'road' is more accurate and comprehensive. Figure 1(c), (d) and (e) show whether the model can still maintain accurate attention to specific unseen classes when faced with a large number of labels for both methods. BiAM's attention has been relatively loose in these two images. We speculate that it is because the model pays too much attention to the spatial region and thus lacks the global perception. Whereas our method pays more attention to the perception of channel responses, which contain both spatial locality and globality, and can better focus on the unseen classes.

*Jingcai Guo is the corresponding author.

A.3. Qualitative Results

Our prediction results for *NUS-WIDE* [1] test images are shown in Figure 2. As can be seen from the figure, we can predict many unseen classes (with asterisks mark) in these images, and the recognition of seen classes in the images is also relatively accurate. In particular, the capture of color is very good, proving that these classes are very dependent on the channel response. In our predicted Top-10 classes, basically every class has a semantic connection with the image. This proves that it is quite correct and sensible to extract class-correlation as semantic information.

References

[1] 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. 2

[2] Zuoxin Li and Fuqiang Zhou. Fssd: feature fusion single shot multibox detector. *arXiv preprint arXiv:1712.00960*, 2017. 1

[3] Sanath Narayan, Akshita Gupta, Salman Khan, Fahad Shahbaz Khan, Ling Shao, and Mubarak Shah. Discriminative region-based multi-label zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8731–8740, 2021. 1

[4] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[5] Qijie Zhao, Tao Sheng, Yongtao Wang, Zhi Tang, Ying Chen, Ling Cai, and Haibin Ling. M2det: A single-shot object detector based on multi-level feature pyramid network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9259–9266, 2019. 1

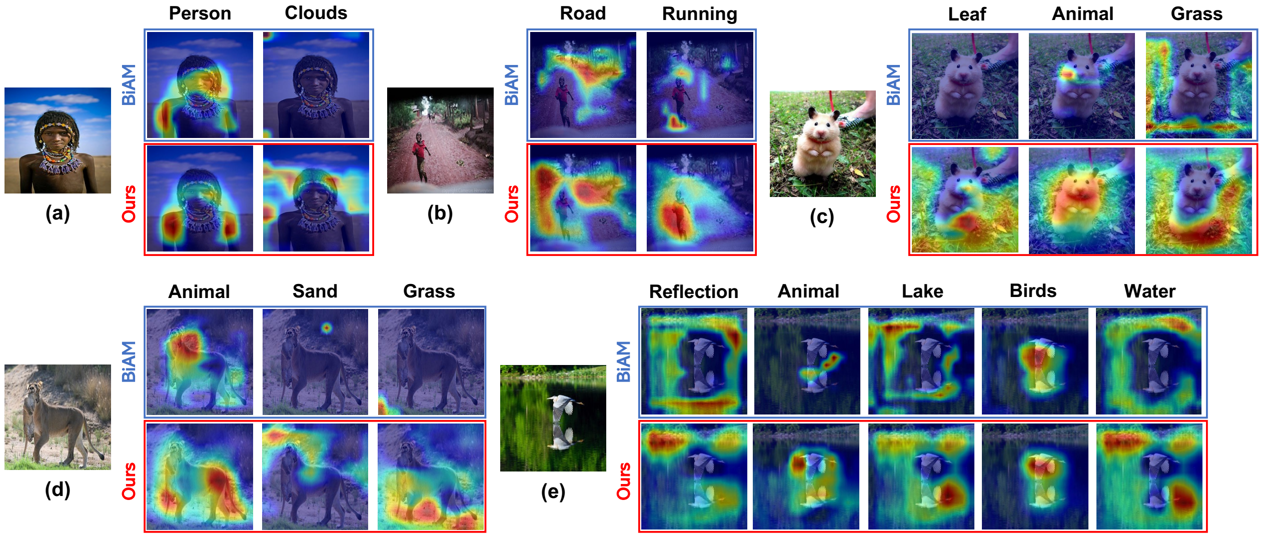

Figure 1. **Attention visualization comparison between our C^3-MLZSL and region-based bi-attention module BiAM.** The images are from **NUS-WIDE** test images.

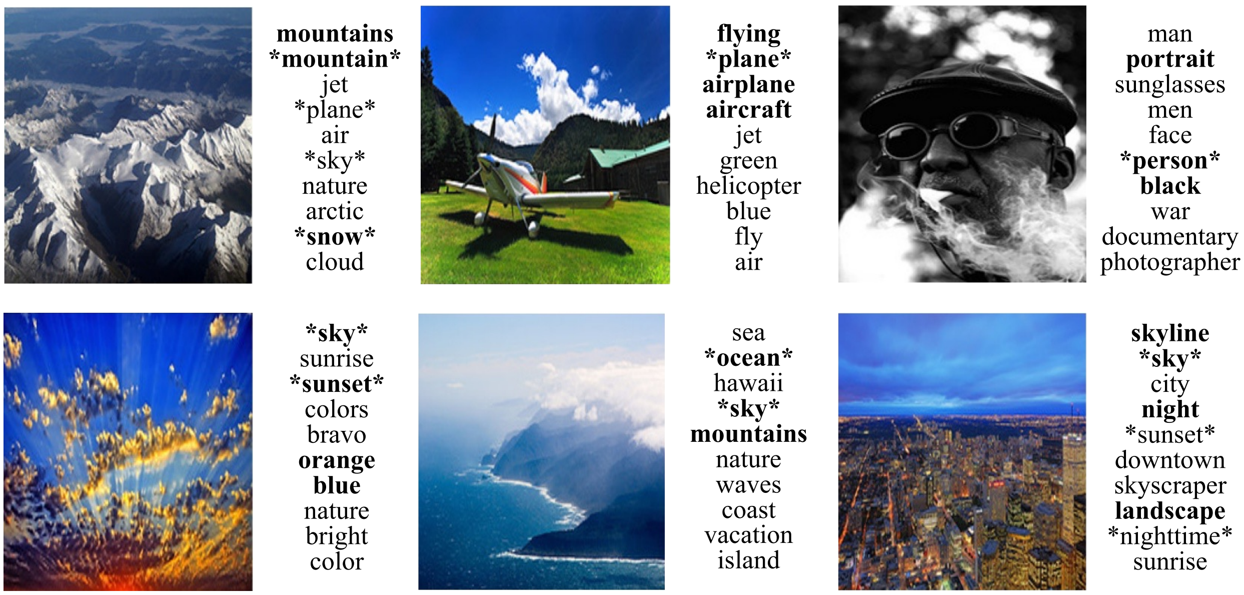

Figure 2. **Qualitative results showing the top-10 labels retrieved using C^3-MLZSL.** Asterisks mark means unseen labels and the bold text indicates that the predicted labels are consistent with the ground truth labels of **NUS-WIDE** test set.