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

Ziming Liu, Song Guo, Xiaocheng Lu, Jingcai Guo, Jiewei Zhang, Yue Zeng, Fushuo Huo

PEILab, Department of Computing
The Hong Kong Polytechnic University

Post ID: THU-PM-311

Overview

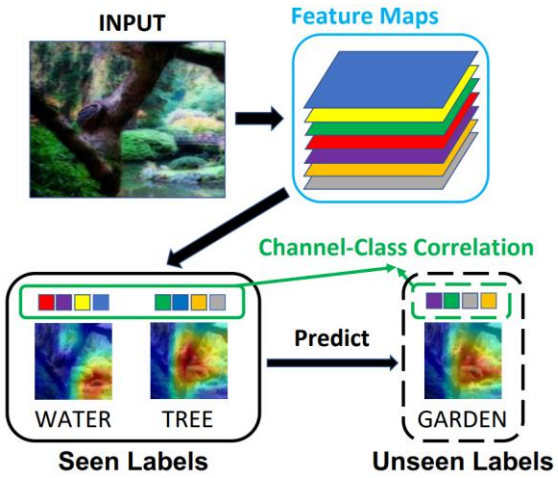
- **What is Multi-label Zero-shot Learning (MLZSL):**

  Extends conventional single-label zero-shot learning to a more realistic scenario.

- **The research difficulties of MLZSL:**

  The complex contents and relationships in a single image across various classes.

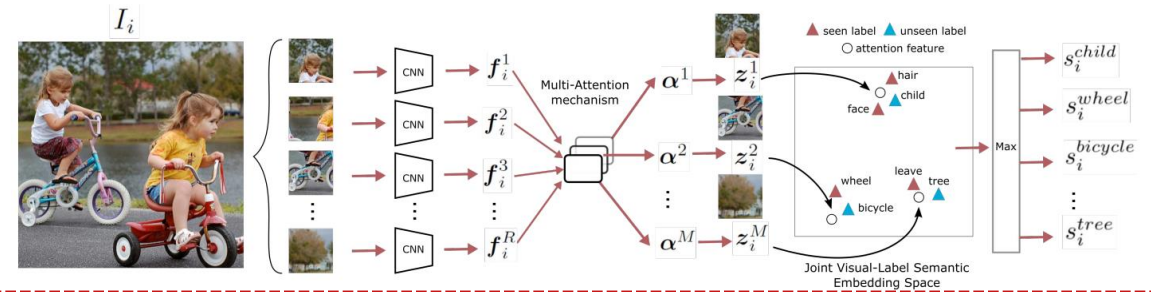
- **Our Methods —— (ML)²P-Encoder:**



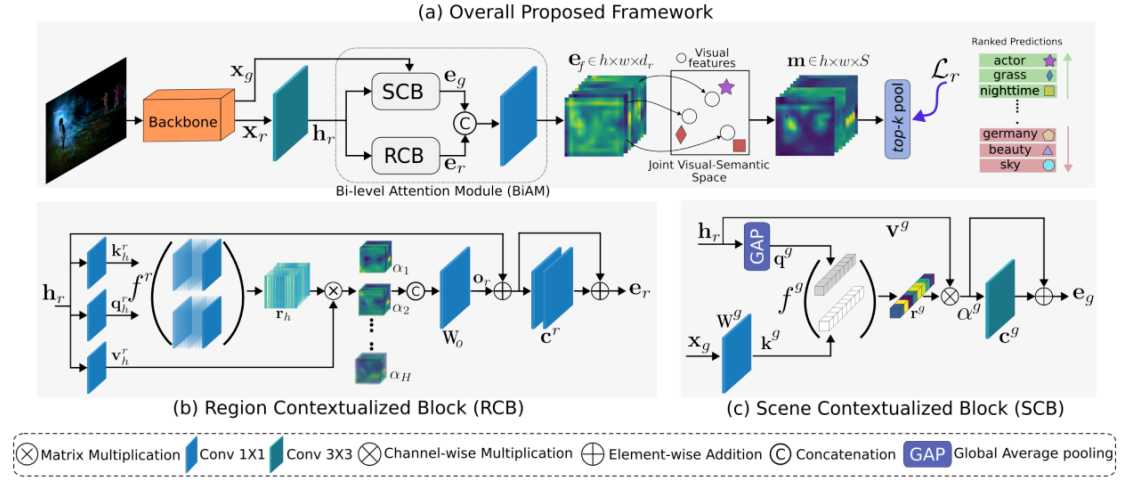
Our interest is to fully explore the power of channel-class correlation as the unique base for MLZSL.

Existing Models

LESA[1] (CVPR 2020)



BiAM[2] (ICCV 2021)

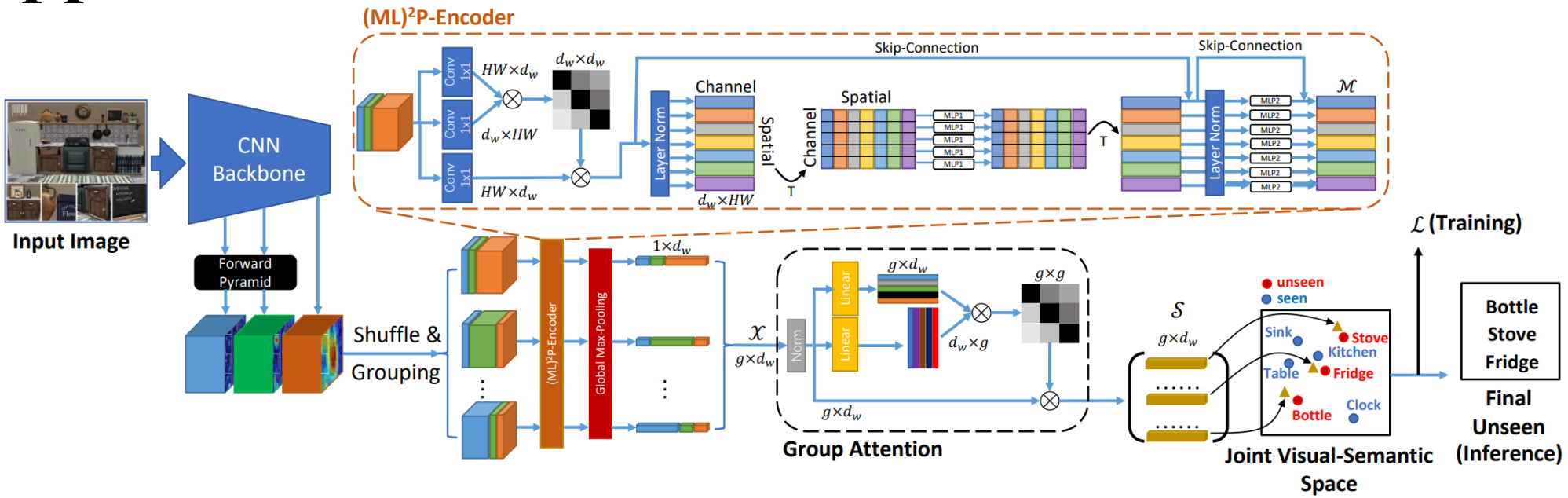


Weaknesses

- The over-reliance on spatial-class correlation fails to capture fine-grained class-specific semantics.
- The additional processing of spatial information greatly increases the computational cost of the model and limits the inference speed.

[1] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[2] 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, 2021.

Our Approach



(ML)²P-Encoder:

$$\mathbf{Q} = W_p^Q \mathcal{F}_a^i \quad \mathbf{K} = W_p^K \mathcal{F}_a^i \quad \mathbf{V} = W_p^V \mathcal{F}_a^i$$

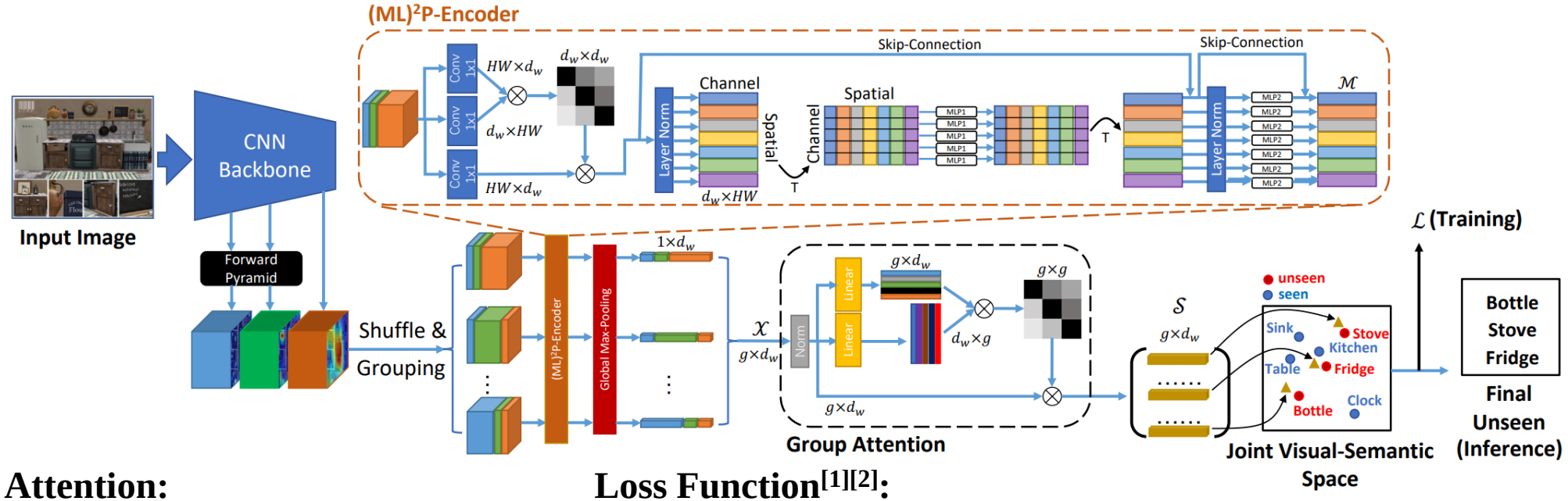
$$\mathrm{Att}(\widehat{\mathbf{Q}}, \widehat{\mathbf{K}}, \widehat{\mathbf{V}}) = \widehat{\mathbf{V}} \cdot \mathrm{softmax}(\underbrace{\widehat{\mathbf{K}} \cdot \widehat{\mathbf{Q}}}_{\mathcal{R}})$$

$$\widehat{\mathcal{F}}_a^i = \mathcal{F}_a^i + \mathrm{Att}(\widehat{\mathbf{Q}}, \widehat{\mathbf{K}}, \widehat{\mathbf{V}})$$

$$\mathcal{F}_{mlp1}^i = \mathcal{F}_{mlp}^i + \mathbf{W}_2 \sigma(\mathbf{W}_1 \mathrm{LayerNorm}(\mathcal{F}_{mlp}^i))$$

$$\mathcal{M} = \mathcal{F}_{mlp1}^i + \mathbf{W}_4 \sigma(\mathbf{W}_3 \mathrm{LayerNorm}(\mathcal{F}_{mlp1}^i))$$

Our Approach



Group Attention:

$$\mathbf{Q_x} = W_x^Q \mathcal{X} \quad \mathbf{K_x} = W_x^K \mathcal{X}$$

$$\mathcal{S} = (\mathbf{Q_x} \cdot \mathbf{K_x}) \cdot \mathcal{X}$$

Loss Function[1][2]:

$$\mathcal{L}_{rank} = \frac{1}{(|P|\,|\bar{P}|)} \sum_i \sum_j \log\left(1 + e^{\mu_{ij}}\right) \text{ where } \mu_{ij} = \max(\mathcal{S} \cdot n_i) - \max(\mathcal{S} \cdot p_j)$$

$$\mathcal{L}_{reg} = \left\| \sum_n var(\mathcal{S}_n) \right\|_1$$

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} ((1-\lambda) \cdot \omega \mathcal{L}_{rank}(\mathcal{S}_i, Y_i) + \lambda \mathcal{L}_{reg}(\mathcal{S}_i)) \text{ where } \omega = 1 + \sum_i var(P^i)$$

[1] Yang Zhang, Boqing Gong, and Mubarak Shah. Fast zero-shot image tagging. In 2016 IEEE Conference on Computer Vision and Pattern Recognition.
[2] Avi Ben-Cohen, Nadav Zamir, Emanuel Ben Baruch, Itamar Friedman, and Lihi Zelnik-Manor. Semantic diversity learning for zero-shot multi-label classification. In 2021 IEEE/CVF International Conference on Computer Vision.

Experiments

- **Datasets:**

NUS-Wide dataset: About 270K images for the whole dataset. During the experiment, 925 labels were used as 'seen labels', and 81 labels were used as 'unseen labels'.

Open-Images-V4 dataset: Contains nearly 9 million training images, 125,456 images as test images. The training set contains 7,186 'seen labels'. While 400 most frequent labels are used as 'unseen labels'.

- **Evaluation Metrics:**

We use the two most common evaluation metrics, the **mean Average Precision (mAP)** and **F1-Score**.

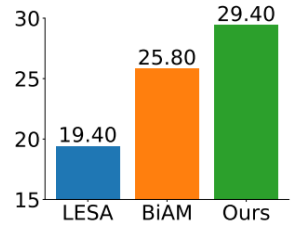
Performance Evaluation

Method	Task	mAP	F1 (K = 3)	F1 (K = 5)
CONSE [37]	ZSL	9.4	21.6	20.2
	GZSL	2.1	7.0	8.1
LabelEM [2]	ZSL	7.1	19.2	19.5
	GZSL	2.2	9.5	11.3
Fast0Tag [50]	ZSL	15.1	27.8	26.4
	GZSL	3.7	11.5	13.5
Kim *et al.* [23]	ZSL	10.4	25.8	23.6
	GZSL	3.7	10.9	13.2
Attention per Cluster [22]	ZSL	12.9	24.6	22.9
	GZSL	2.6	6.4	7.7
LESA [22]	ZSL	19.4	31.6	28.7
	GZSL	5.6	14.4	16.8
BiAM [36]	ZSL	25.8	32.0	29.4
	GZSL	8.9	15.5	18.5
Our Approach	ZSL	**29.4**	**32.8**	**32.3**
	GZSL	**10.2**	**15.8**	**19.2**

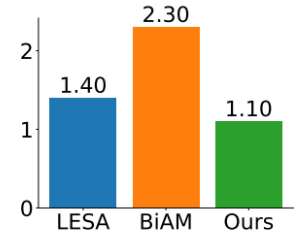
NUS-Wide Dataset

Method	Task	mAP	F1 (K = 10)	F1 (K = 20)
CONSE [37]	ZSL	40.4	0.4	0.3
	GZSL	43.5	2.6	2.4
LabelEM [2]	ZSL	40.5	0.5	0.4
	GZSL	45.2	5.2	5.1
Fast0Tag [50]	ZSL	41.2	0.7	0.6
	GZSL	45.2	16.0	13.0
Attention per Cluster [22]	ZSL	40.7	1.2	0.9
	GZSL	44.9	16.9	13.5
LESA [22]	ZSL	41.7	1.4	1.0
	GZSL	45.4	17.4	14.3
BiAM [36]	ZSL	62.8	4.1	3.7
	GZSL	79.6	17.6	15.1
Our Approach	ZSL	**65.7**	**7.5**	**6.5**
	GZSL	**79.9**	**27.6**	**24.1**

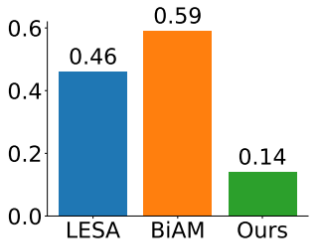
Open-Images V4 Dataset



(a) mAP

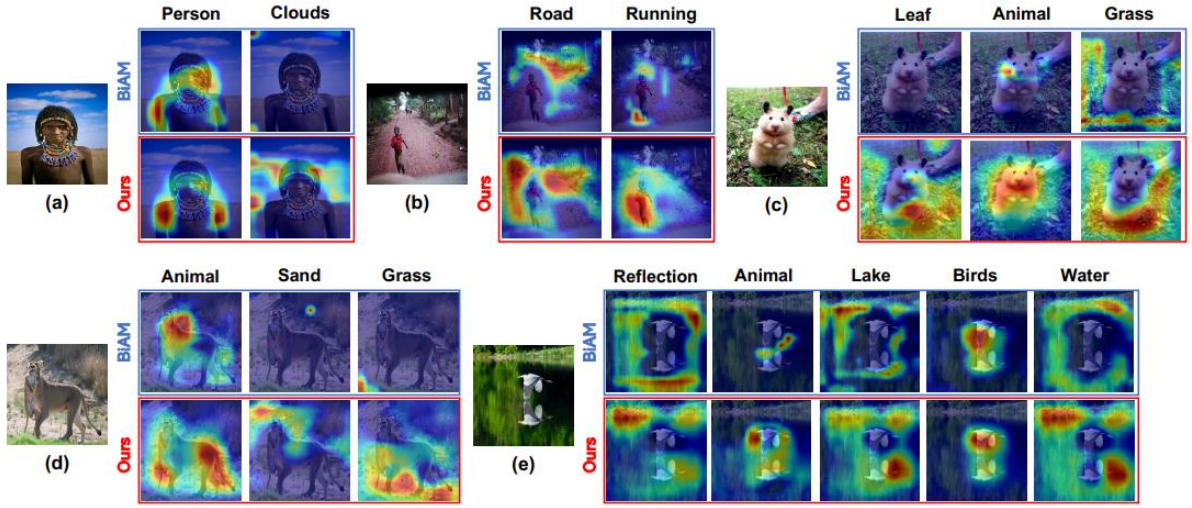


(b) Inference time (ms)

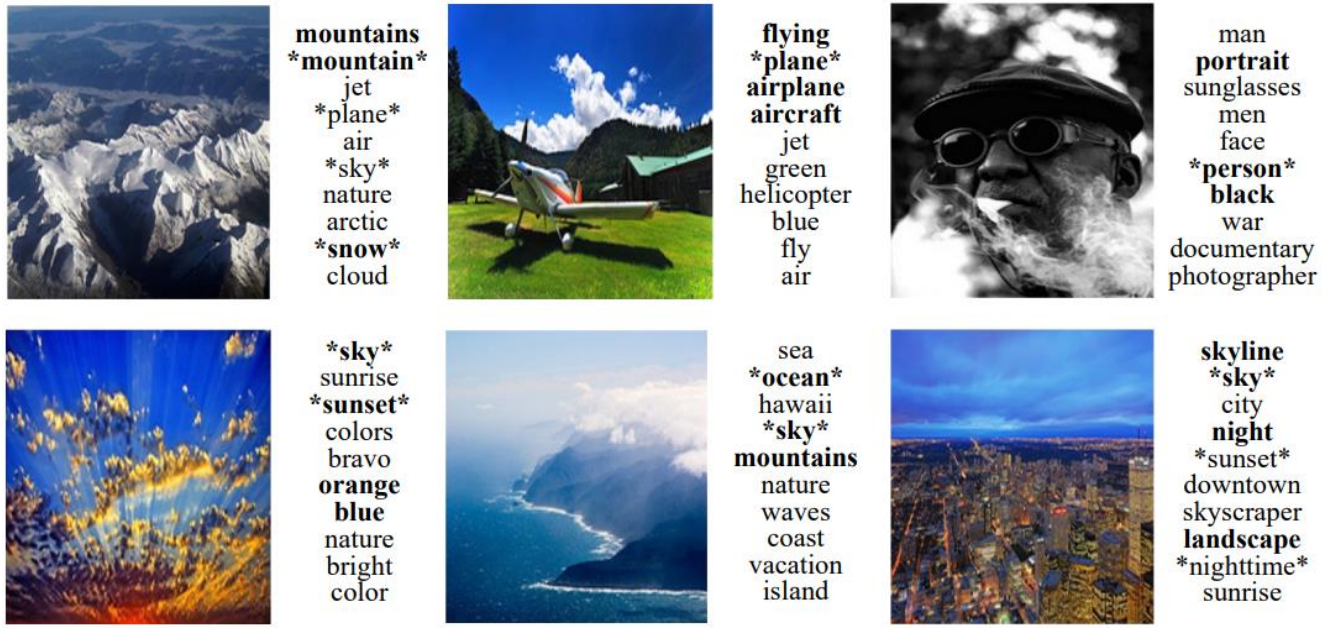


(c) GFLOPs

Qualitative & Visualization Results



Attention Visualization Comparison



Qualitative Results

Thanks!