Zero-Shot Object Detection with Multi-label Context

Yongxian Wei, Yong Ma
School of Computer Science and Engineering, Nanjing University of Science and Technology
Nanjing, China
{wei_yx,mayong}@njust.edu.cn

ABSTRACT

Zero-shot detection (ZSD) , the problem of object detection when training and test objects are disjoint, i.e. no training examples of the target classes are available. ZSD increasingly gains importance for large scale applications because collecting and labeling sufficient data is extremely hard. In this paper, inspired from human cognitive experience, we propose a simple but effective Multi-label Context (MLC) framework to facilitate the detection ability for both seen and unseen objects by mining contextual cues. We design a multi-label classifier which leverages the holistic image-level context to learn object-level concepts. Then, novel RoI features are generated by exploiting context information beneath both whole images and interested regions. Moreover, background dynamic generator (BDG) can reduce the confusion between background and unseen classes. Our extensive experiments show that MLC outperforms the current state-of-the-art methods on MS-COCO.

Index Terms— Zero-Shot Object Detection, Multi-label Learning, Context Embedding, Computer Vision

1. INTRODUCTION

While object detection methods based on deep learning have achieved great progress over the last few years [1, 2, 3, 4, 5], these gains can be attributed to the availability of the fully supervised training data. Although researchers have struggled to acquire larger datasets with a broader set of categories, the processing procedure is time consuming and tedious. Furthermore, it is hard to collect enough training data for rare categories. Zero-shot learning (ZSL) has been proposed to address the problem for reasoning unseen classes [6, 7, 8], Traditional ZSL researches mainly focus on the classification of unseen objects and achieve high classification accuracy [7]. However, there is still a big gap between ZSL settings and real-world scenes. ZSL only focuses on identifying unseen objects, not detecting them. For example, most of datasets used as ZSL benchmark have only one dominant object per image [9, 10], while in real-world, various objects may appear in a single image without being precisely localized. To close this gap, [11] introduced a new "zero-shot object detection" (ZSD) problem setting method, which aims at detecting objects seen during training as well as detecting unseen classes as and when they appear at test-time.

DOI reference number: 10.18293/SEKE2022-012

(a) Baseline (b) Ours

Fig. 1. Motivation and example results of our MLC framework. By incorporating discriminative context information, the semantic features of "car" are strong evidences for detecting objects that are highly dependent on context information, such as "traffic light".

Existing ZSD approaches mainly focus on learning a visual-semantic correspondence based on intrinsic properties of the target objects by the means of human-defined attributes or distributed representations learned from text corpora. They only focus on local information near an object's region of interest while ignoring rich contextual information within the image, which has been shown to benefit the object detection performance [12, 13, 14].

We therefore propose a novel framework named Multi-label Context (MLC) for ZSD. In this paper, we revisit the RoI features in region-based detectors from the perspective of context information embedding. Our key motivation is that while each RoI in very deep CNNs may have a very large theoretical receptive field which usually spans the whole input image [1]. However, the effective receptive field [15] may only occupy a fraction of the entire theoretical receptive field, making the RoI features insufficient for characterizing objects that are highly dependent on context information. We use a simple but effective process to generate contextual RoI fea-

tures by exploiting embedded multi-label context information beneath both whole images and interested regions, which are also complementary to conventional RoI features.

MLC learns from the cognitive science about how humans reason objects through semantic information. Humans can learn the mapping relationship between vision objects and semantic description from seen objects and transfer it to detect unseen objects. In addition, conventional object detection approaches generally tend to relegate unseen objects into the background leading to missed detection of unseen objects. Previous works [16, 11] used the word-vector of the "background" word to represent background class. Due to the rough word-vector for background class used in detector head is inability to exactly represent the complex background, MLC develops a component denoted as background dynamic generator (BDG) to learn an appropriate word-vector for background class. Our study shows that replacing the rough background word-vector in detector head with the new one learned from BDG can effectively increase the recall rate of unseen classes.

In summary, the contributions of this paper are three-fold: (i) we develop a novel ZSD approach that adaptively exploits the whole image context to learn discriminative features for context-dependent object categories; (ii) to the best of our knowledge, it is the first time to introduce multi-label learning into ZSD task; (iii) extensive experiments on two different MS-COCO splits show significant performance improvement on the existing ZSD benchmarks.

2. METHOD

2.1. Problem Formulation

We begin by defining the problem and then present our approach. We denote the set of all classes as $C = C_S \cup C_U$, where C_S denotes the set of seen classes and C_U denotes the set of unseen classes, and $C_S \cap C_U = \phi$. Each image is denoted as $I \in \mathbb{R}^{w \times h \times 3}$, with corresponding bounding boxes and ground truth labels denoted as $b_i \in \mathbb{N}^4$ and $y_i \in C$ respectively. Let D_S denotes the training dataset, which only contains the objects belonging to C_S to train the network and use the unseen classes objects dataset D_U to evaluate the detection performance for unseen classes. For GZSD setting, the test dataset D_T contains objects from both seen and unseen classes ($c \in C = C_S \cup C_U$).

2.2. MLC Framework

The overall framework of our MLC consists of four components: Multi-Label Head for advancing the feature learning of the objects that are highly dependent on larger context clues, contextual RoI features generated by fusing both instance-level and global-level information derived from Multi-Label Head, BDG for generating suitable background word-vector, and Zero-Shot Head for classifying the extracted objects into seen and unseen classes and locating them. The details are indicated in Figure 2.

2.2.1. Multi-Label Head

In parallel with the RPN branch, we exploit Multi-Label Head upon the detection backbone, enabling the backbone to learn object-level concepts adaptively from global-level context. It is worth mentioning that Multi-Label Head does not require additional annotations, as the image-level labels can be conveniently obtained by collecting all instance-level categories in an image. Specifically, we first apply a 3×3 convolution layer on the output of ResNet conv5 to obtain the input feature map, and then follow the practice in [17] to employ both global max-pooling (GMP) and global average-pooling (GAP) for feature aggregation. Formally, let $\mathbf{X} \in \mathbb{R}^{d \times w \times h}$ denote the input feature map, where d is the channel dimensionality, w and h are the width and height, respectively. Then, the multi-label classifier is constructed by N_S binary classifiers for all categories:

$$\hat{\mathbf{y}} = f_{CLS}(f_{GMP}(\mathbf{X}) + f_{GAP}(\mathbf{X})) \in \mathbb{R}^{N_S}, \tag{1}$$

where N_S denotes the number of seen classes, each element of $\hat{\mathbf{y}}$ is a confidence score (logits), and f_{CLS} is binary classifier modeled as one fully-connected layer. We assume that the ground truth label of an image is $\mathbf{y} \in \mathbb{R}^{N_S}$, where $y_i = \{0, 1\}$ denotes whether object of category i appears in the image or not. The multi-label loss can be formulated as follows:

$$\mathcal{L}_{MLL} = -\sum_{i=1}^{N_S} y_i ln(\frac{1}{1+e^{-\hat{y_i}}}) + (1-y_i) ln(\frac{e^{-\hat{y_i}}}{1+e^{-\hat{y_i}}}), \tag{2}$$

2.2.2. Contextual RoI Feature Generation

With the purpose of leveraging larger context, We apply RoIAlign [4] with proposals generated by RPN on the context-embedded feature map X to obtain RoI features:

$$\mathbf{x}_{global} = f_{RoIAlign}(\mathbf{X}; w, h) \in \mathbb{R}^{d \times 7 \times 7}, \tag{3}$$

where $f_{RoIAlign}$ is the RoIAlign operation and w and h are the width and height of the input image, respectively. As the resulting RoI feature $\mathbf{x}_{global}$ absorbs rich context information from the context-embedded image feature $\mathbf{X}$, it is by nature complementary to the conventional RoI feature extracted from the feature pyramid network (FPN) [18]. To integrate our contextual RoI features $\mathbf{x}_{global}$ into the detection pipeline, it is natural to fuse them with the original RoI features extracted from the feature pyramid network (FPN) with element addition. Formally, let $\mathbf{x}_{instance}$ denote the original RoI feature extracted from FPN, and $\mathbf{x}_{fusion}$ denote the fused RoI feature, then we have:

$$\mathbf{x}_{fusion} = \mathbf{x}_{global} + \mathbf{x}_{instance} \in \mathbb{R}^{d \times 7 \times 7}, \tag{4}$$

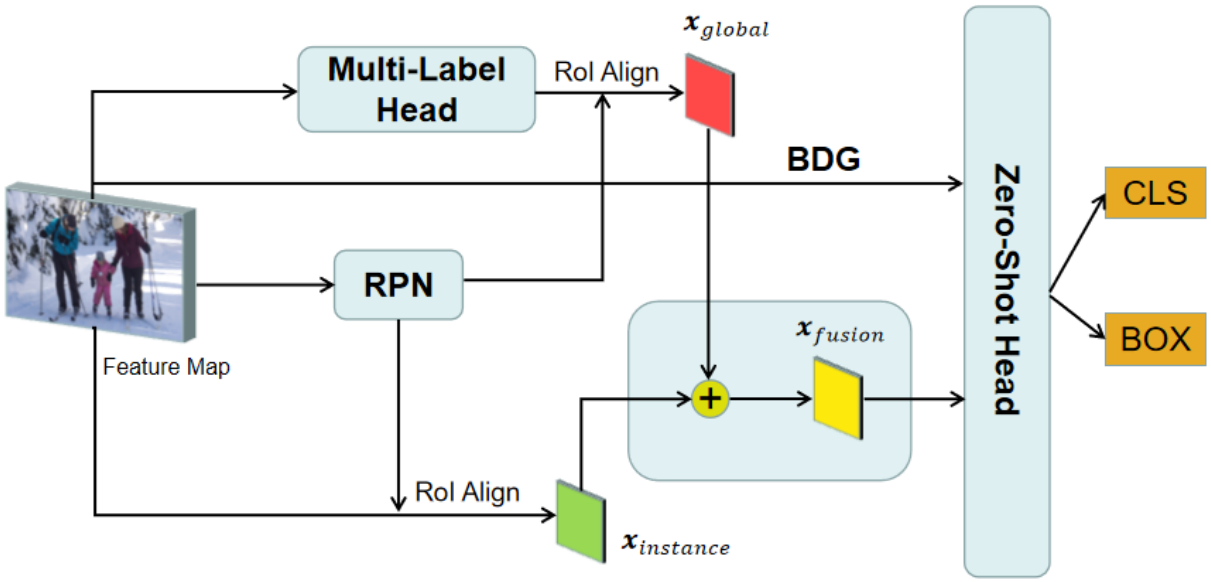

Fig. 2. The architecture for MLC. After acquiring feature map from the backbone, Multi-Label Head enables the network to learn object-level concepts from global-level context. Then, contextual RoI features which are complementary to conventional RoI features, are generated by fusing both instance-level and global-level information. Finally, the Zero-Shot Head uses the $\mathbf{x}_{fusion}$ and BDG to locate and classify the seen and unseen objects, respectively.

As shown in Figure 2, the fused feature map $\mathbf{x}_{fusion}$ is then fed into the Zero-Shot Head to produce refined bounding boxes and classification scores.

2.2.3. *Background Dynamic Generator*

We set a fully connected layer called $\mathbf{v}_b$ without bias and make it trainable. $\mathbf{v}_b$ is used to represent vector for background class, which is initialized with the mean word vectors for all seen classes. BDG will update $\mathbf{v}_b$ during training so that we can learn a new word-vector $\mathbf{v}_b$ for background class. During training, we feed the visual features derived from the backbone network to the BDG branch and get the background binary classification score. The calculation process is formulated as follows:

$$c = \frac{1}{1 + e^{-\mathbf{x}T\mathbf{v}_b}}, \tag{5}$$

specifically, $T \in \mathbb{R}^{N \times d}$ is an FC layer which is used to adjust the dimension of input objective feature to fit d, i.e. the dimension of word vector.

2.2.4. *Zero-Shot Head*

The main idea for our Zero-Shot Head is learning the relationship between visual and semantic concepts from seen classes data and transferring it to detect unseen objects. To this end, we replace the classification branch in Faster R-CNN with a new semantic-classification branch. Keeping the non-trainable seen class word vectors W_S, we allow projection of the visual feature $\mathbf{x}_{fusion}$ to the word embedding space to calculate classification scores P_S. In inference, we follow [11] to use an additional procedure to calculate the classification scores for unseen classes. The process can be briefly demonstrated as follows:

$$P_U = (P_S W_S^T) W_U, P_S = T_e(\mathbf{x}_{fusion}) W_S, \tag{6}$$

where, W_U contains unseen class word vectors. So we can get the scores by performing the matrix multiplication of the semantic feature and W_U.

2.2.5. *Loss Function*

The whole loss function $\mathcal{L}_{MLC}$ for our end-to-end network has four components:

$$\begin{aligned} \mathcal{L}_{MLC} &= \mathcal{L}_{MLL} + \mathcal{L}_{RPN} + \mathcal{L}_{BDG} + \mathcal{L}_{ZSH}, \\ \mathcal{L}_{BDG} &= -(c \log(\hat{c}) + (1-c) \log(1-\hat{c})), \\ \mathcal{L}_{ZSH} &= -\sum_{i=1}^{N_S} P_{S,i} \log(P_{S,i}) + l_1(\mathbf{r}, \hat{\mathbf{r}}), \end{aligned} \tag{7}$$

where $\mathcal{L}_{ZSH}$ is the losses for Zero-Shot Head and it contains smooth l_1 regression loss. All loss terms are considered equally important, without extra hyper-parameters to characterize the trade-off between them, which reveals MLC is generalized and not trick.

Table 1. ZSD performance of Recall@100 and mAP with different IoU thresholds on MS COCO dataset.

Method	Seen/Unseen	Recall@100			mAP
		0.4	0.5	0.6	0.5
SB [16]	48/17	34.46	22.14	11.31	0.32
DSES [16]	48/17	40.23	27.19	13.63	0.54
TD [11]	48/17	45.50	34.30	18.10	-
PL [19]	48/17	-	43.59	-	10.10
Gtnet [20]	48/17	47.30	44.60	35.50	-
BLC [21]	48/17	51.33	48.87	45.03	10.60
Ours	48/17	**56.03**	**52.52**	**47.73**	**11.30**
PL [19]	65/15	-	37.72	-	12.40
BLC [21]	65/15	57.23	54.68	51.22	14.70
Ours	65/15	**60.11**	**57.81**	**52.49**	**15.70**

Table 2. Comparison of Recall@100 and mAP at IoU=0.5 under GZSD setting on MS COCO dataset. HM denotes the harmonic average for seen and unseen classes.

Method	Seen/Unseen	seen		unseen		HM	
		mAP	Recall	mAP	Recall	mAP	Recall
DSES [16]	48/17	-	15.02	-	15.32	-	15.17
PL [19]	48/17	35.92	38.24	4.12	26.32	7.39	31.18
BLC [21]	48/17	42.10	57.56	4.50	46.36	8.20	51.37
Ours	48/17	**47.26**	**71.46**	**5.39**	**50.92**	**9.68**	**59.46**
PL [19]	65/15	34.07	36.38	12.40	37.16	18.18	36.76
BLC [21]	65/15	36.00	56.39	13.10	51.65	19.20	53.92
Ours	65/15	**40.95**	**67.83**	**14.86**	**59.64**	**21.81**	**63.47**

3. EXPERIMENT

3.1. Dataset and Setting

We validate our proposed method on the widely used object detection dataset MSCOCO. This dataset is more challenging than Pascal VOC as it has 80 object classes, more small objects, and more complex background. Following the dataset splits of MSCOCO proposed in [16] and [19], we use both two splits of the dataset in experiments: (1) 48 seen classes and 17 unseen classes; (2) 65 seen classes and 15 unseen classes. Note that the seen classes and unseen classes are disjoint.

We use mAP and Recall@100 as the evaluation metrics, in which 100 means that only the top 100 detections are valid for evaluation. The experimental results are reported under ZSD (zero shot detection) and GZSD (generalized zero shot detection) benchmarks. The ZSD setting only requires the detection results of unseen objects, while for GZSD setting, it requires the model predict both the seen and unseen objects. GZSD is more challenging than ZSD, and more suitable for practical application.

Table 3. Ablation study of our method in different splits. ZSH means Zero-Shot Head and MLH means Multi-Label Head.

Seen/Unseen	ZSH	MLH	BDG	Recall/mAP		
				seen	unseen	HM
48/17	✓			65.1/40.7	43.1/4.5	51.8/7.7
	✓	✓		70.9/47.1	45.3/5.5	55.3/9.8
	✓	✓	✓	71.4/47.2	50.9/5.3	59.4/9.6
65/15	✓			60.6/32.0	54.3/12.7	57.3/18.2
	✓	✓		65.7/39.2	55.0/14.6	59.9/21.3
	✓	✓	✓	67.8/40.9	59.6/14.8	63.4/21.8

3.2. Comparison with Other Methods

We compare the performance for MLC with the state-of-the-art zero-shot detection approaches on both 48/17 and 65/15 splits of MSCOCO under ZSD and GZSD settings. For ZSD setting, we show the results in Table 1. Our method outperforms all other work and improves up to 30.38% and 20.09% of the Recall@100 metric over the 48/17 and 65/15 splits, respectively. Moreover, the improvement in mAP also shows that the contextual RoI feature has an effective discrimination ability to unseen class. For GZSD setting, we report the results in Table 2. MLC surpasses all previous works in terms of mAP and Recall@100 on both seen and unseen classes. The "HM" performance gain reveals that our method maintains a good balance between seen and unseen classes.

3.3. Ablation Study

We conduct a controlled study of our proposed method on GZSD evaluation. As shown in Table 3, the baseline method with Zero-Shot Head gives a foundation and achieves comparable mAP and recall at IoU = 0.5. Our method is able to consistently bring improvement on both seen and unseen categories. From these results, we can learn the significant effectiveness of the Multi-Label Head. We can also observe that BDG brings an improvement of 4.6% in terms of Recall@100 for unseen classes.

4. CONCLUSION AND FUTURE WORK

In this paper, we find that contextual information is quite important in zero-shot detection, hence we propose a novel framework to embed global-level context to advance the learning of context-dependent categories with the help of multi-label learning. In the experiment part, we described the extensive experiments which were conducted to demonstrate

the superiority of the proposed model, and investigated the effectiveness of different components. In the future, we would like to find a better approach to obtain the semantic feature since traditional word vectors like word2vec are noisy.

5. REFERENCES

[1] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[2] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.

[3] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[5] Joseph Redmon and Ali Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[6] Abhijit Bendale and Terrance E Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.

[7] Ziming Zhang and Venkatesh Saligrama, "Zero-shot learning via joint latent similarity embedding," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 6034–6042.

[8] Yong Ma, Huaqi Mao, Haofeng Zhang, and Wenbo Wang, "Prototype relaxation with robust principal component analysis for zero shot learning," *IEEE Access*, vol. 8, pp. 170140–170152, 2020.

[9] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[10] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona, "Caltech-ucsd birds 200," 2010.

[11] Shafin Rahman, Salman Khan, and Fatih Porikli, "Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 547–563.

[12] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng, "Direction-aware spatial context features for shadow detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7454–7462.

[13] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta, "Iterative visual reasoning beyond convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7239–7248.

[14] Haowen Deng, Tolga Birdal, and Slobodan Ilic, "Ppfnet: Global context aware local features for robust 3d point matching," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 195–205.

[15] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 4905–4913.

[16] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran, "Zero-shot object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 384–400.

[17] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[19] Shafin Rahman, Salman Khan, and Nick Barnes, "Improved visual-semantic alignment for zero-shot object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 11932–11939.

[20] Shizhen Zhao, Changxin Gao, Yuanjie Shao, Lerenhan Li, Changqian Yu, Zhong Ji, and Nong Sang, "Gtnet: Generative transfer network for zero-shot object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, vol. 34, pp. 12967–12974.

[21] Ye Zheng, Ruoran Huang, Chuanqi Han, Xi Huang, and Li Cui, "Background learnable cascade for zero-shot object detection," in *Proceedings of the Asian Conference on Computer Vision*, 2020.