Context-Guided Super-Class Inference for Zero-Shot Detection

Yanan Li
Artificial Intelligence Institute
Zhejiang Lab
liyn@zhejianglab.com

Yilan Shao
Artificial Intelligence Institute
Zhejiang University
ylshao@zju.edu.cn

Donghui Wang
Artificial Intelligence Institute
Zhejiang University
dhwang@zju.edu.cn

Abstract

Zero-shot object detection (ZSD) is a newly proposed research problem, which aims to simultaneously locate and recognize objects of previously unseen classes. Existing algorithms usually formulate it as a simple combination of a typical detection framework and zero-shot classifier, by learning a visual-semantic mapping from the visual features of bounding box proposals to semantic embeddings of class labels. In this paper, we propose a novel ZSD approach that leverages the context information surrounding objects in the image, following the principle that objects tend to be found in certain contexts. It also incorporates the semantic relations between seen and unseen classes to help recognize located instances. Comprehensive experiments on PASCAL VOC and MS COCO datasets show that context and class hierarchy truly improve the performance of detection.

1. Introduction

Object detection is one of the fundamental computer vision problems, which has enjoyed a series of breakthroughs thanks to the advances of deep learning and large-scaled labeled datasets. However, it is difficult to collect and annotate images of all concepts that beyond daily objects. To reduce the limitation, zero-shot detection (ZSD) has emerged as an important research topic recently [11, 2, 1, 12, 7]. Its goal is to simultaneously locate and recognize each individual instance of unseen object class, in the absence of any visual examples of those classes during the training stage.

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 [14, 5, 13]. Also in zero-shot recognition problem, [15] proposed to model the conditional likelihood of an object to appear in a given context by explicitly exploiting all possible objects in the image, to improve the recognition performance. However, since testing categories are completely isolated from the training ones, their connection to context cannot be well-established.

In this paper, we propose a novel context-guided ZSD approach that incorporates into a typical detection framework both context information of an object and high-level semantic relationship among objects, i.e. super-class in the classification branch. We conjecture that semantically analogous objects share similar environment or context. Our ZSD approach consists of two components: (1) A context feature extraction (CFE) module to predict super-class. Differently, we adopt dilated convolution to extract contextual features, consisting of valid information from a lager spatial range. (2) A generic zero-shot detection branch, including object proposal, box coordinates regression and visual-semantic mapping. Final category is determined by both super-class score and semantic vector.

2. Proposed Method

2.1. Problem Definition

We begin by defining the problem of our interest. Let $\mathcal{D}_s = \{(\mathbf{X}_i, \mathbf{b}_i, c_i)\}_{i=1}^{N_s}$ denotes the training dataset, where $\forall \mathbf{X}_i \in \mathcal{I}_s$ is the training image, $\mathbf{b}_i = (x_i, y_i, w_i, h_i)$ is the position and size of target bounding box and $c_i \in \mathcal{S}, \mathcal{S} = \{1, ..., S\}$ is the corresponding category from the set of seen classes $\mathcal{S}$. Let $\mathcal{U} = \{S+1, ..., S+U\}$ denotes the set of unseen classes, where $\mathcal{S} \cap \mathcal{U} = \emptyset$. $\mathcal{K}_s = \{\mathbf{k}_s^1, ..., \mathbf{k}_s^S\}$ and $\mathcal{K}_u = \{\mathbf{k}_u^1, ..., \mathbf{k}_u^U\}$ are the corresponding semantic embeddings of seen and unseen classes, respectively. Given $\mathcal{D}_s$, $\mathcal{K}_s$ and $\mathcal{K}_u$, ZSD requires the algorithm to output the bounding box prediction $\{\mathbf{b}_k, p_k, c_k\}_{k=1}^{M}$ of each image, where $p_k \in (0, 1)$ denotes the probability of object existence.

2.2. Context-Guided Zero-Shot Detector

The proposed method is based on YOLOv3 due to its speediness and efficiency, illustrated in Fig.1. First, we use CNN to extract hierarchical feature maps, in order to be adaptive for objects of different scales. Second, for each

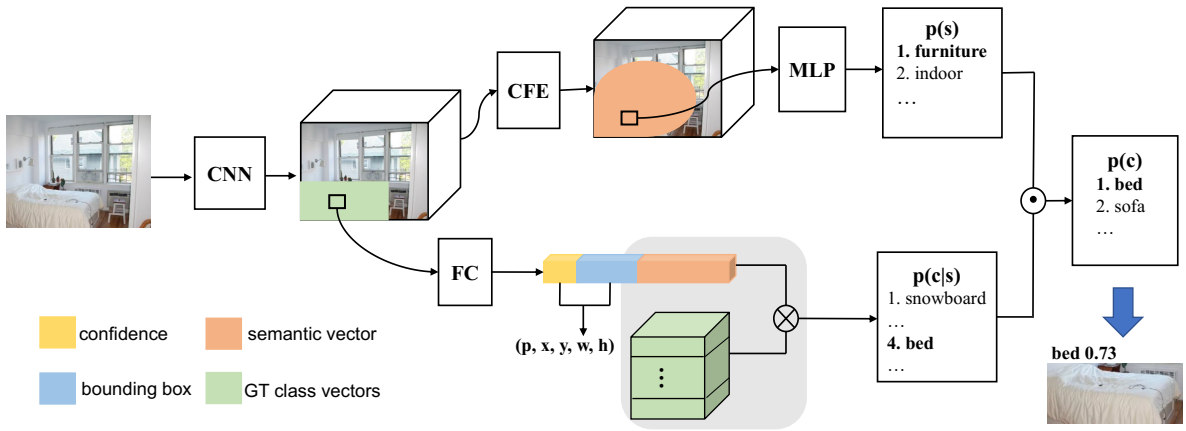

Figure 1. The overview of our proposed model based on YOLOv3. Predictions over the base CNN feature are carried out in 3 scales and one of them are illustrated in the figure. To make category predictions more accurate, super-class probabilities inferred from context feature extracted by context feature extraction module (denoted as CFE) are taken into consideration.

scale, we use CFE module to extract context features for super-class prediction $p(s)$. At the same time, we predict the location and the possible class label $p(c|s)$ for each grid based on their predicted semantic embeddings. The final classification score is obtained by multiplying $p(s)$ and $p(c|s)$.

Context Feature Extraction Modern CNN can draw the valuable parts from feature maps under the supervision of objective function. We use dilated convolution [4] to enlarge the reception fields and harvest context features implicitly, as illustrated in Fig.2. In each context feature extraction blocks (CFEB), we adjust the reception fields of convolution kernels by setting different dilation value, which denotes the multiple of kernel size expansion. We use a 4-way structure, each with different dilations, to extract hierarchical context features. The concatenation of features from each CFEB is then passed through a 3×3 convolution layer to get the final context features.

Unseen Category Prediction For each grid $\boldsymbol{v}_i$, we first predict the semantic embedding $\boldsymbol{k}_i$, compute its similarity with each class j and then multiply them with the super-class probability score $p(s)$ for final prediction, i.e. $\tilde{\boldsymbol{c}}_{i_j} = p(s) \times f(\boldsymbol{k}_i, \boldsymbol{k}_s^j)$. f computes the inner product between two vectors. $\boldsymbol{k}_s^j$ is the provided semantic embedding of j-th class. The product can be viewed as the conditional probability of categories given the super-class prior, i.e. $P(\mathcal{C}|\mathcal{S})$. The training objective is thus formulated as follows.

$$L_{cls} = \sum \mathbb{I}_i^{obj} L_{ce}(\tilde{\boldsymbol{c}}_i, \boldsymbol{c}_i), \tag{1}$$

where L_{ce} denotes cross entropy loss and $\boldsymbol{c}_i$ is the ground-truth one-hot label. $\mathbb{I}_i^{obj}$ is set to 1 if the box is positive.

Super-class Inference. The feature map contains the fused knowledge of diverse ranges centered on each grid, which corresponds spatially to the original map. Being a guidance on category prediction, this component is also inferred by the multilayer perceptron under the supervision of ground truth super-class. The objective is formulated as follows.

$$L_{sup} = \sum \mathbb{I}_i^{obj} L_{ce}(P(s), Y_{sup}). \tag{2}$$

We get the ground-truth Y_{sup} between categories and their super-classes from WordNet [9].

Visual-Semantic Projection. We learn a parametric mapping from visual features to semantic embeddings by the following function.

$$L_{emb} = \sum \mathbb{I}_i^{obj} \frac{1}{|\mathcal{Y}_s| - 1} \sum_{j \neq k} \max(0, \Delta - \left\| \hat{\boldsymbol{k}}_i - \boldsymbol{k}_j \right\|_2 + \left\| \hat{\boldsymbol{k}}_i - \boldsymbol{k}_k \right\|_2), \tag{3}$$

where $|\mathcal{Y}_s|$ represents the number of seen classes and k is the index of ground truth class. By using the margin Δ, this projection tries to push away embeddings of other categories while fitting the current one. It is worth mentioning that we give out only one set of super-class scores for each grid, though there are three bounding box outputs based on the anchor priors. This is because these three bounding boxes in the same grid share identical visual features and have similar location.

The overall loss of category prediction is defined as:

$$L = L_{cls} + \alpha L_{sup} + \beta L_{emb} + \gamma L_{reg}. \tag{4}$$

where α, β and γ are adjustable hyper-parameters. L_{reg} is the same regression loss as YOLOv3.

3. Experiments

We conduct experiments on PASCAL VOC and MS COCO datasets [8] for evaluation and use mean average precision (mAP) as a metric under the same settings as [12].

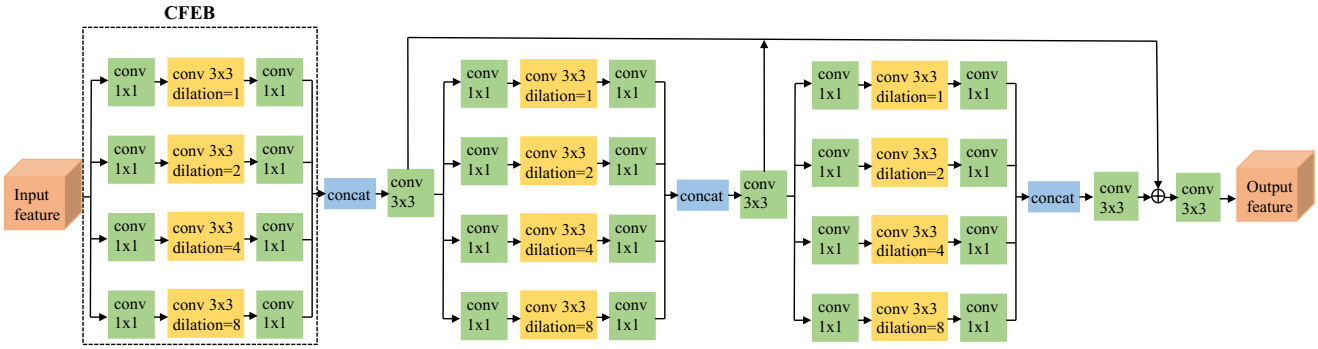

Figure 2. Detailed structure of *Context Feature Extraction* module. The module has 3 context feature extraction blocks (CFEB) where the feature is passed through a 4-way convolution block of different dilations (reception fields) and finally concatenated together.

3.1. Zero-Shot Detection Results

Method	car	dog	sofa	train	mAP(%)
HRE	55.0	82.0	55.0	26.0	54.2
Ours	20.4	92.7	50.1	55.9	**54.8**

Table 1. ZSD results on PASCAL VOC. HRE denotes *hybrid region embedding*, proposed in [2].

Table 1 shows the comparison between HRE [2] and ours. Training on 16 classes of VOC, our model exceeds HRE on other 4 classes by 0.6% mAP. It is also observed that the average precision of *car* is evidently lower than other unseen classes, which is probably caused by two reasons. First, cars are often heavily occluded at various sizes in street scenes. It is difficult to find them all or separate them completely from complex background, resulting in a low recall. Second, cars and trains usually lie in different environments but they share the same super-class in our setting, where context information cannot be fully leveraged to distinguish between the two. Table 2 shows the performance of zero-shot detectors under two seen and unseen splits on MS COCO. The top part of the table refers to the split in [1] which ignores the category names composed of two words, while the bottom part uses an optimized split [12] where all the classes are involved and distributed across all the super-classes in both training and testing. From the results, the proposed model outperforms the state-of-the-arts approaches in both two cases.

Method	split	mAP(%)
Inductive	65/15	10.8
Ours		**10.9**
SB	48/17	0.70
DSES		0.54
LAB		0.27
Ours		**7.2**

Table 2. ZSD results on MS COCO. The splits of seen/unseen classes are 65/15 and 48/17, respectively.

Figure 3. Visualization of our detection results on MS COCO.

Figure 3 demonstrates the visualized output of testing examples in MS COCO dataset, including various kinds of unseen objects. Most of them are successfully detected, which confirms the effectiveness of the proposed method. It is interesting to note that a gamepad (not belongs to any unseen class) is recognized as fork, probably because it lies beside a tin of coke.

3.2. Ablation Study

We also run a series of ablation studies to further validate the effectiveness of the context information.

Dilated convolution performs better. We introduce dilated convolution for context feature extraction. As a matter of fact, the conventional convolution layer is also able to cover the surrounding areas of the target. To evaluate

their performance, we calculate the precision of super-class prediction of all detected bounding boxes with high confidence on PASCAL VOC dataset. In Table 3, we can see that super-class prediction benefits from dilated convolution for a larger reception field.

Method	animal	vehicle	household	AP(%)
ours-d	62.9	47.3	72.9	61.0
ours	**70.0**	**54.8**	**82.4**	**69.1**

Table 3. Precision of super-class prediction on PASCAL VOC. All the testing categories belong to three super-classes: animal, vehicle and household. ours-d denotes the proposed method without dilated convolution.

Super-class inference leads to better performance. We cut off the context-guided branch of the proposed model to form the baseline method. Table 4 demonstrates the efficacy of using context in unseen class prediction for zero-shot object detection.

Method	Dataset			
	split	VOC(%)	split	COCO(%)
baseline	16/4	48.4	65/15	8.7
ours		**54.8**		**10.9**

Table 4. MAP comparison of baseline model and the proposed model on PASCAL VOC and MS COCO.

Figure 4 lists the average precision of each unseen class in MS COCO. We can know that context-guidance brings an improvement for the majority of testing classes. Taking the category *mouse* with larger increase as an example, the AP goes from 0.6% to 3.9%. By common sense, a mouse would appear in a study scene, accompanied by computer, keyboard, desk, etc. Without its context information, a small and plain object that has never been seen before is much more difficult for the network to recognize.

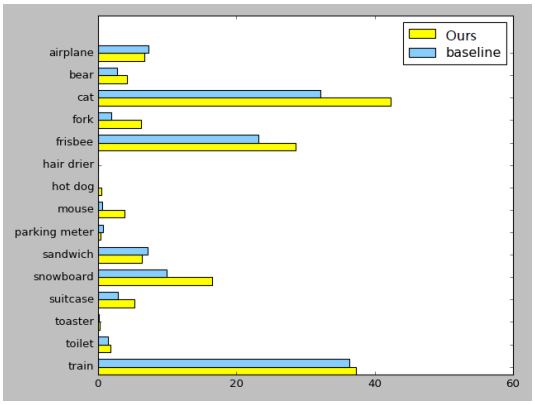

Figure 4. Average precision of unseen classes obtained by baseline and the proposed model on MS COCO.

In search of higher precision, we compare the results when using different semantic knowledge for visual-semantic mapping. We report both U2U and U2T performance on two datasets in Table 5. U2U denotes that the label search space during testing consists of only unseen classes. U2T denotes that the label search space during testing includes both seen and unseen object classes. It is a more generalized and challenging setting in zero-shot learning.

Knowldge	dim	VOC		COCO	
		U2U	U2T	U2U	U2T
attributes	64	50.2	**7.4**	-	-
GloVe	300	49.8	1.1	10.3	0.06
BERT	768	**54.8**	4.6	**10.9**	**0.55**

Table 5. Results obtained by using different semantic knowledge. Attributes are from [6], consisting of 64 specific manual attributes. GloVe [10] and BERT [3] embeddings are generated by their corresponding language models with raw class names.

As illustrated in Table 5, embeddings generated by BERT, currently acknowledged as the most powerful language model, outperform others on the whole, especially under the U2U circumstance. Attributes receive decent results as well with 7.4% mAP (U2T) on PASCAL VOC dataset because clear attribute definition avoids ambiguity when large distances exist among classes. The more accurately semantic knowledge describes the categories, the stronger transferability of the model.

4. Conclusion

We present a novel context-guided approach for zero-shot object detection. It mainly uses context information to predict super-class in unseen class prediction and exceeds the state-of-the-art methods on PASCAL VOC and MSCOCO datasets. However, there are also some limitations in the proposed method when contexts are counter-intuitive, e.g. a fork lies near the bathroom sink instead of being in the outdoors. When two object classes share the same super-class but are very different in appearance, e.g. the unseen umbrella and the seen suitcase (seen), the proposed method would largely fail in unseen prediction. Moreover, the context in the image is not always in accordance with the super-class taxonomy defined in WordNet, which can deteriorate the performance to some extent. The above issues will be mainly investigated in the future work.

5. Acknowledgments

This work was supported by the Natural Science Foundation of Zhejiang Province (No.LQ20F030007), the National Natural Science Foundation of China (No.61473256) and the Youth Science Foundation of Zhejiang Lab (No.2020KD0AA03).

References

[1] Bansal Ankan, Sikka Karan, Sharma Gaurav, Chellappa Rama, and Divakaran Ajay. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 384–400, 2018. 1, 3

[2] Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. Zero-shot object detection by hybrid region embedding. In *British Machine Vision Conference*, page 56, 2018. 1, 3

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019. 4

[4] Yu Fisher, Koltun Vladlen, and Funkhouser Thomas. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 2

[5] Deng Haowen, Birdal Tolga, and Ilic Slobodan. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 195–205, 2018. 1

[6] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 951–958. IEEE, 2009. 4

[7] Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. Zero-shot object detection with textual descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8690–8697, 2019. 1

[8] 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. 2

[9] Ted Pedersen, Siddharth Patwardhan, Jason Michelizzi, et al. Wordnet:: Similarity-measuring the relatedness of concepts. In *AAAI*, volume 4, pages 25–29, 2004. 2

[10] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543, 2014. 4

[11] Rahman Shafin, Khan Salman, and Porikli Fatih. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Asian Conference on Computer Vision*, pages 547–563. Springer, 2018. 1

[12] Rahman Shafin, Khan Salman, and Barnes Nick. Transductive learning for zero-shot object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6082–6091, 2019. 1, 2, 3

[13] Hu Xiaowei, Zhu Lei, Fu Chi-Wing, Qin Jing, and Heng Pheng-Ann. Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7454–7462, 2018. 1

[14] Chen Xinlei, Li Li-Jia, Fei-Fei Li, and Gupta Abhinav. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248, 2018. 1

[15] Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. Context-aware zero-shot learning for object recognition. In *Proceedings of the 36th International Conference on Machine Learning*, pages 7292–7303, 2019. 1