Inference Fusion with Associative Semantics for Unseen Object Detection

Yanan Li[1], Pengyang Li[2], Han Cui[3], Donghui Wang[2] *
[1]Zhejiang Lab
[2]Zhejiang University
[3]University of California, Los Angeles
liyn@zhejianglab.com, pyli@zju.edu.cn, elviscuihan@g.ucla.edu, dhwang@zju.edu.cn

Abstract

We study the problem of object detection when training and test objects are disjoint, i.e. no training examples of the target classes are available. Existing unseen object detection approaches usually combine generic detection frameworks with a single-path unseen classifier, by aligning object regions with semantic class embeddings. In this paper, inspired from human cognitive experience, we propose a simple but effective dual-path detection model that further explores associative semantics to supplement the basic visual-semantic knowledge transfer. We use a novel target-centric multiple-association strategy to establish concept associations, to ensure that the predictor generalized to unseen domain can be learned during training. In this way, through a reasonable inference fusion mechanism, those two parallel reasoning paths can strengthen the correlation between seen and unseen objects, thus improving detection performance. Experiments show that our inductive method can significantly boost the performance by 7.42% over inductive models, and even 5.25% over transductive models on MSCOCO dataset.

Introduction

Object detection has shown great success in the deep learning era, relying on a huge amount of training data with accurate bounding box annotations (Ren et al. 2015; Redmon et al. 2016; Lin et al. 2017; Liu et al. 2020). However, detectors can hardly generalize to novel target domain where the labeled data are scarce or none, since some objects are hard to collect, e.g. endangered animals or constantly emerging new products. In contrast, humans exhibit a remarkable ability of learning new concepts quickly, even without seeing any visual instance. To bridge this gap between state-of-the-art detection and human-level intelligence, empowering detectors with the capability of detecting novel classes has become a key area of interest.

Zero-shot object detection (ZSD) is a recently-proposed solution that aims to detect novel (unseen) objects with no annotated samples during training (Bansal et al. 2018; Demirel, Cinbis, and Ikizler-Cinbis 2018; Rahman, Khan, and Porikli 2018; Rahman, Khan, and Barnes 2019; Li et al. 2019; Rahman, Khan, and Porikli 2020; Li, Shao, and Wang 2020). In contrast with zero-shot recognition (Lampert, Nickisch, and Harmeling 2013; Wang et al. 2016; Li et al. 2017; Sung et al. 2018; Zablocki et al. 2019; Luo et al. 2020), it requires the model to not only recognize the object types but also localize the targets among millions of potential regions of interest (RoIs). Current ZSD models generally assume the localization process is class-agnostic (Wang et al. 2020), i.e. capable of proposing high-confidence regions for unseen objects, and mainly focus on attaching a zero-shot classifier within de facto detection networks. They leverage shared semantic information (e.g. attributes, word vectors, text descriptions) to enable the knowledge transfer from seen objects to unseen ones. Specifically, it first learns the visual-semantic mapping from visual features of RoIs to class embeddings during training and then applies the mapping to test RoIs, whose labels are predicted by searching the nearest neighbor class in the semantic embedding space. Several studies (Zhao et al. 2020; Zhu, Wang, and Saligrama 2020) alternatively explore generative models to synthesize unseen RoIs and then retrain a supervised unseen classifier.

Existing methods above generally adopt the basic knowledge transfer mechanism that builds up the feature-label correspondence merely from perceived visual features. For a

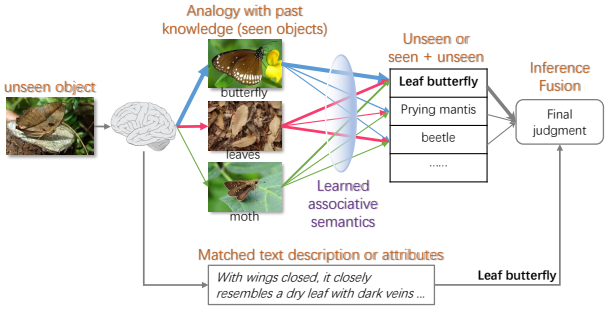


Figure 1: Existing ZSD models generally adopt single-path label prediction for each RoI, realized by aligning RoI with attributes (or text descriptions). Humans learn a previously unseen concept by both inferring through auxiliary knowledge (e.g. attributes) and associating and analogizing with seen objects. Inspired from this, we propose a two-path reasoning mechanism for ZSD and employ inference fusion to obtain the final results.

*Corresponding author

target-domain RoI, a single-path label prediction is made directly from all given concepts, without considering their possible connections in the inference. Intuitively, when humans learn a previously unseen concept, they not only infer directly through auxiliary knowledge (e.g. text descriptions, attributes), but also make an empirical analysis by analogy with past knowledge (e.g. seen objects) to assist the learning process (shown in Fig.1) (Anderson and Bower 2014). For example, the first time we see a "leaf butterfly", we can match with specified attributes, and also concurrently draw an analogy with objects seen in past (e.g. butterflies, dead leaves, moths) to corroborate the former prediction results.

Inspired from above, we propose a novel dual-path ZSD model that can automatically explores concept association to supplement the basic visual-semantic knowledge transfer, among which the establishment of analogy between concepts is important. Considering target instances are absent during training, we establish the analogy relationship in advance by using class embeddings, for which a novel target-centric multi-association strategy is proposed. An analogy predictor that is generalizable to unseen domain is then trained on seen instances. Together with the basic visual-semantic knowledge transfer, we have two parallel unseen objects inference paths at the same time. We adopt an inference fusion strategy to make full use of these complementary paths during testing. Obviously, such a dual-path parallel reasoning mechanism strengthens the learning of association relationships between seen and unseen objects, and can effectively improve the model's transferability. The main contribution of this work can be summarized as follows.

- We propose a simple but effective dual-path ZSD method that fuses both visual-semantic transfer and analogy association with previous knowledge to improve the model's generalization ability. It provides a generic, cognitively-plausible solution that can be easily incorporated within one-stage or two-stage detection networks.
- We propose a novel target-centric multi-association strategy to establish the analogy relationships among concepts, through which a generalizable association predictor can be trained on seen objects. Without introducing extra data or complex computation, both the generalization and discrimination ability of our model can be boosted.
- We optimize the proposed model by a two-stage training strategy and test it with inference fusion mechanism. Our experiments show that the proposed inductive model can obtain a large performance gain. We achieve 7.42% and 18.01% absolute boost in mAP and recall over inductive models, even 5.25% and 7.58% over the transductive competitor.

Related Work

Object Detection There have been significant development in object detection over the last few years. The deep detection frameworks are generally categorized into two types, i.e. one-stage detectors (e.g. YOLO (Redmon et al. 2016), SSD (Liu et al. 2016)) and two-stage detectors (e.g. R-CNN series (Girshick 2015; Ren et al. 2015)). These models are roughly composed of three components: proposing bounding boxes, deciding which box contains objects, classifying high-confidence boxes. The first two components are sort of class-agnostic, which means they are capable of proposing high-confidence boxes for previously unseen objects. We mainly work on the classification component for ZSD problem. In this work, we detail the proposed approach based on the two-stage Faster R-CNN framework.

Zero-Shot Recognition Zero-shot recognition (ZSR) mimics the human ability to recognize objects without seeing any visual examples (Lampert, Nickisch, and Harmeling 2013). It uses semantic descriptions that provide relationships among classes, e.g. attributes, word vectors, text descriptions, to transfer knowledge from source domain with abundant training data to target domain. A basic paradigm for ZSR is to learn a direct visual-semantic alignment function. One can map visual feature space to semantic space or vice versa, or map both spaces to a common latent space (Fu et al. 2015; Akata et al. 2013; Wang et al. 2016; Kodirov, Xiang, and Gong 2017; Liu et al. 2018), and then predict class labels by fixed similarity metrics or data-driven metrics (Sung et al. 2018). In our work, we adopt the basic visual-semantic alignment approach in the detection network and use a different target-centric strategy to define the concept association for ZSD, which offers a semantic compensation for the basic alignment.

Zero-Shot Object Detection Zero-shot object detection (ZSD) is a newly-proposed problem and less explored than ZSR. Due to the ill-posed nature and inherent complexity, ZSD is far more challenging. We cannot simply copy the standard problem setup of ZSR into ZSD, since multiple objects instead of one primary object appear in a single image. Contemporary models (Bansal et al. 2018; Rahman, Khan, and Barnes 2020; Li et al. 2019) exploit various zero-shot classifiers in one-stage or two-stage detectors and generally adopt single-path label prediction, e.g. SB and DSES (Bansal et al. 2018) are built on RCNN (Szegedy et al. 2016). (Rahman, Khan, and Barnes 2019) provides a transductive solution to the domain shift problem suffered from above inductive methods by using unlabeled test data during training to iteratively update model parameters. Several works also try to exploit specific associated concepts in the detection process, e.g. (Rahman, Khan, and Barnes 2020) applies related concepts from an external vocabulary to improve class embeddings, (Li, Shao, and Wang 2020) uses context to predict superclasses defined by WordNet and selects the unseen class label from it. Differently, our inductive method uses a dual-path reasoning mechanism that incorporates both visual-semantic knowledge transfer and novel concept association into a detector.

Proposed Approach

The objective of our dual-path approach is to detect novel objects, which have no samples during training. It combines the basic visual-semantic knowledge transfer and analogy association with previous knowledge to learn a new concept,

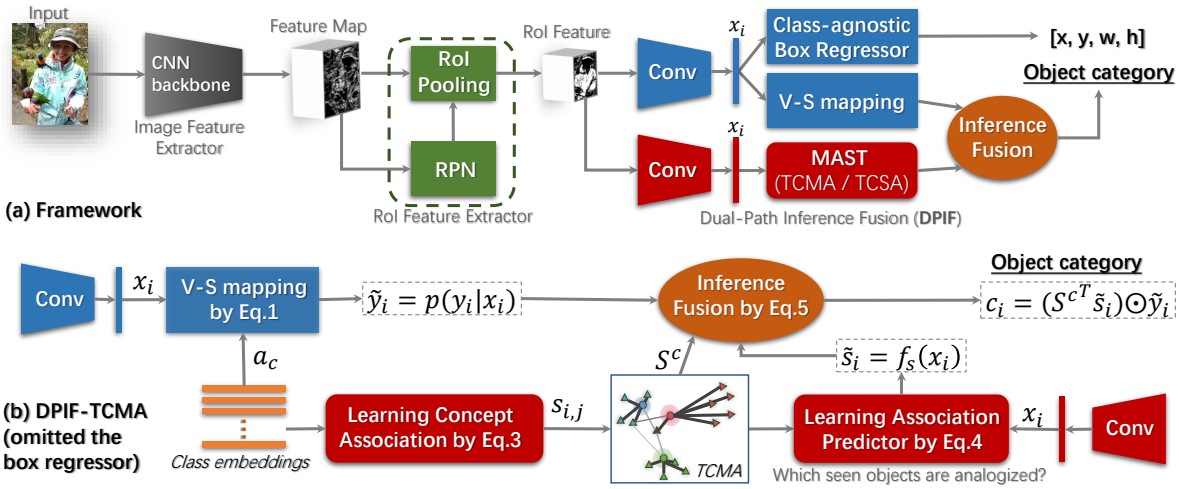


Figure 2: Illustration of our unseen object detection framework, which combines Faster R-CNN with Dual-Path Inference Fusion (DPIF) module, as shown in (a). It contains two parallel inference paths, basic visual-semantic knowledge transfer (V-S mapping) and multi-associative semantic transfer (MAST), whose outputs are fused to make the final prediction.

where the former computes the alignment score between visual features and class embeddings and the latter explicitly match instances with associative semantics. Inference fusion is then performed to obtain the final semantically-enhanced results. See Fig.2. More details will be described below.

Problem Settings

We start by giving a formal definition of zero-shot object detection (ZSD) problem. Let $\mathcal{C} = \mathcal{S} \cup \mathcal{T}$ denote the whole object class set, where $\mathcal{S}$ is the set of C_s source classes, $\mathcal{T}$ is the set of C_t target classes and $\mathcal{S} \cap \mathcal{T} = \emptyset$. Each class is provided in advance with a m-dimensional semantic embedding, acquired in a supervised (e.g. manual attributes) or unsupervised manner (e.g. word2vec). We denote their class embeddings as $\{\boldsymbol{a}_i^s\}_{i=1}^{C_s}$ and $\{\boldsymbol{a}_i^t\}_{i=1}^{C_t}$, respectively. In the training stage, we have N_s labeled images from source classes. Each image $\mathcal{I}_i$ is annotated with N_i bounding boxes and their associated labels, i.e. $(\boldsymbol{b}_j, y_j)_{j=1}^{N_i}$, where $\boldsymbol{b}_j \in \mathbb{R}^4$ and $y_j \in \mathcal{S}$. Similarly, we have N_t images in the test stage, where each image has at least one instance from target classes. In the standard setting, we need to locate and recognize every instance from only unseen objects. While in the more challenging generalized setting, we need to detect all $\mathcal{C}$ classes.

Naive Approach We build the basic ZSD model on Faster R-CNN framework, which contains the feature extraction backbone (e.g. ResNet (He et al. 2016), VGG16 (Simonyan and Zisserman 2014)) to learn image-level features $\boldsymbol{X} \in \mathbb{R}^{H \times W \times d}$, the region proposal network (RPN), RoI pooling and RoI feature extractor to extract proposal-level features $\boldsymbol{x}_i \in \mathbb{R}^{d_v}$, and a box predictor to compute classification score and predict the bounding box coordinates. The key idea is to construct a zero-shot classifier from RoI features and class embeddings, considering that the feature extraction backbone and RPN are class-agnostic (Wang et al. 2020).

We employ the basic visual-semantic alignment strategy to recognize RoIs. It projects the visual features $\boldsymbol{x}_i$ and class embeddings $\boldsymbol{a}_c$ into a common latent space respectively by functions $f_v(\cdot)$ and $f_a(\cdot)$, and compares them using a similarity metric d. The classification score for $\boldsymbol{x}_i$ is defined as:

$$\tilde{\boldsymbol{y}}_{i,c} = p(y_i = c|\boldsymbol{x}_i) = \frac{\exp(d(f_v(\boldsymbol{x}_i), f_a(\boldsymbol{a}_c)))}{\sum_{j=1}^{C_s} \exp(d(f_v(\boldsymbol{x}_i), f_a(\boldsymbol{a}_j)))}, \quad (1)$$

where we utilize cosine similarity for d in this paper, for that it can bound and reduce the variance of neurons and result in models of better generalization (Gidaris and Komodakis 2018). To make the model discriminative enough for classification, we maximize the classification score of ground-truth class and minimize the score of negative classes and use the following cross-entropy loss for optimization,

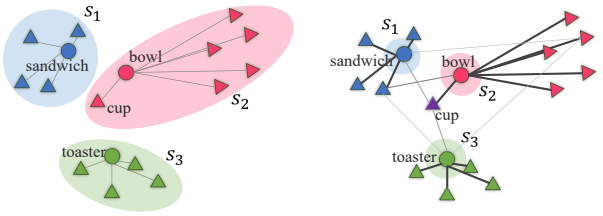
$$L_{cls} = -\sum_{j=1}^{C_s} \boldsymbol{y}_{i,j} \log \tilde{\boldsymbol{y}}_{i,j} + (1 - \boldsymbol{y}_{i,j}) \log(1 - \tilde{\boldsymbol{y}}_{i,j}), \quad (2)$$

where $\boldsymbol{y}_i \in \{0, 1\}^{C_s}$ is the ground-truth label vector of $\boldsymbol{x}_i$. It contains only one element with value 1 indicating the class it belongs to. While for localization, we use the same box regressor as in Faster R-CNN to predict bounding box coordinates, since it is class-agnostic and transferable enough. The localization loss is denoted as L_{loc} (Ren et al. 2015).

The basic visual-semantic knowledge transfer is adopted by current ZSD models, which recognize test RoIs by a single-path label prediction in Eq.1. There is an implicit hypothesis that the relational knowledge between different objects in the feature space is consistent with that in the semantic space. However, visually similar objects may be semantically different, e.g. "leaf butterfly" and "leaves" in Fig.1. This visual-semantic gap will deteriorate the alignment in ZSD.

Multi-Associative Semantic Transfer

Apart from matching with auxiliary knowledge (e.g. attributes), humans can learn a new concept through associations and analogies of similar objects in past experiences. This suggests that the association relationship can be another



(a) Target-Centric Single-Association (b) Target-Centric Multiple-Association

Figure 3: Illustration of target-centric associative methods. Circles and triangles denote target classes and source classes, respectively. The boldness of lines indicates the strength of semantic relationships.

powerful source of information in the context of ZSD. Thus, we supplement the basic ZSD method above with another learning path that explores concept association to better enable transfer knowledge. Then we employ an inference fusion strategy taking advantage of these two paths to obtain the final result.

Establishment of Concept Association The establishment of concept association is important. Apart from being intuitively plausible, it should enable two properties from a model perspective, i.e. transferable and discriminative property. Transferable property indicates that there are at least one seen class related to an unseen one, so that the learning path trained from seen domain can be generalized to the unseen domain. While discriminative property means the predicted results should be discriminative enough to classify unseen classes. Considering unseen objects are absent during training, we use class embeddings to build up the class-level associations between seen and unseen concepts.

A straightforward way is to perform k-means clustering over the class embeddings of all $\mathcal{C}$ classes and use k clusters to form concept association. Seen classes in each cluster are assumed to be associated with the unseen ones that may appear in the same cluster. However, this strategy has two drawbacks. First, a cluster may not contain any unseen object. This means seen classes in this cluster are excluded from learning association predictor, thus weakening the transferability. Second, a cluster may have at least two unseen classes. In this case, more than two different unseen classes will have the exact same association relationships, where discriminative ability may be weakened. In order to enable these two properties at the same time, we propose the following simple but effective target-centric strategy.

Target-Centric Single-Association (TCSA) In TCSA, we use the unseen class embeddings $\{\boldsymbol{a}_j^t\}_{j=1}^{C_t}$ as prototypes and form the associative semantics around them. For each seen object, we assume it is only related to the most similar unseen one. Specifically, for i-th seen class, we compute its similarity with each unseen class and then set the association as a one-hot vector $\boldsymbol{s}_i^c \in \{0,1\}^{C_t}$, where $\boldsymbol{s}^c_{i,\arg\max_j d(\boldsymbol{a}_i^s,\boldsymbol{a}_j^t)} = 1$.

Target-Centric Multi-Association (TCMA) We relax the strict single-association hypothesis made above and assume that each seen object can relate to multiple unseen ones with different similarities, which is intuitively plausible and can encode richer semantic relationships. For example in Fig.3 (b), "cup" is related to "bowl", "toaster" and "sandwich" at the same time. Specifically, for i-th seen class, we first select top-K elements from $\{\boldsymbol{d}_{i,j} : d(\boldsymbol{a}_i^s, \boldsymbol{a}_j^t)\}_{j=1}^{C_t}$, to make sure that there are at most K associated objects. Then, if an element in the top-K list is still less than a threshold α, it will be set to 0 as well. This constraint prevents ambiguous associated concept from being exploited for knowledge transfer in ZSD. The multi-association relationships are thus encoded as follows:

$$\boldsymbol{s}_{i,j}^c = \begin{cases} \boldsymbol{d}_{i,j}, & \text{if } \boldsymbol{d}_{i,j} \in \text{top-}K \cap \boldsymbol{d}_{i,j} > \alpha \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Using $\boldsymbol{s}_i^c$ formed by TCMA as supervision has the following merits. First, it guarantees there is at least one seen object during training that associates with a similar unseen one. In some sense, association predictor can be regarded as learning similar unseen class directly by using training images from a group of related source semantics as their pseudo-instances. It can thus alleviate the domain bias problem caused by lacking of test data and improve model's generalization ability. Second, it ensures that the association relationship for each unseen class is different, meeting the requirement for discriminative property.

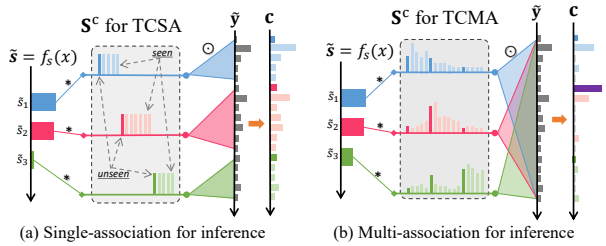
Learning Association Predictor After establishing the concept association, we learn a generalizable association predictor from seen instances. Assume i-th RoI comes from the seen object y_i, whose associative semantics are defined as $\boldsymbol{s}_i = \boldsymbol{s}_{y_i}^c$. We use an extra parallel conv layer on top of RoI pooling to learn its visual features $\boldsymbol{x}_i$. Note the same symbol $\boldsymbol{x}_i$ is used to indicate i-th RoI in both two paths, to avoid confusion. Our goal is to learn a predictor $f_s : \boldsymbol{x}_i \rightarrow \boldsymbol{s}_i$, which can be optimized by minimizing the following cross-entropy loss.

$$L_{ap} = -\sum_{j=1}^{C_t} \boldsymbol{s}_{i,j} \log \tilde{\boldsymbol{s}}_{i,j} + (1 - \boldsymbol{s}_{i,j}) \log(1 - \tilde{\boldsymbol{s}}_{i,j}) \tag{4}$$

where $\tilde{\boldsymbol{s}}_i = f_s(\boldsymbol{x}_i)$. As shown in Fig.2, on the shared backbone, we use two parallel feature extractors to ensure that RoI features in the visual-semantic knowledge transfer are semantic-aligned, while in this branch are appropriate for concept association.

Inference Fusion With the two-path learning, we can generate bounding box coordinates by the box regressor, the classification score $\tilde{\boldsymbol{y}}$ in the basic visual-semantic semantic transfer and the association score $\tilde{\boldsymbol{s}}$ after a forward pass with a test RoI. We fuse $\tilde{\boldsymbol{y}}$ with $\tilde{\boldsymbol{s}}$ to get the semantically-enhanced results, since they provide complementary information. Specially, we first compute the association-guided scores by multiplying $\tilde{\boldsymbol{s}}$ with ground-truth concept associations $\boldsymbol{S}^c$ among concepts. Then, we use these scores to weigh $\tilde{\boldsymbol{y}}$ element-wise to obtain the final result $\boldsymbol{c}$.

$$\boldsymbol{c} = (\boldsymbol{S}^{c^T} \tilde{\boldsymbol{s}}) \odot \tilde{\boldsymbol{y}} \tag{5}$$



(a) Single-association for inference (b) Multi-association for inference

Figure 4: Demonstration of inference fusion using concept association $\boldsymbol{S}^c$. $\tilde{\boldsymbol{s}}$, $\tilde{\boldsymbol{y}}$ and $\boldsymbol{c}$ are the predicted associative semantics, classification score and final results respectively.

Fig.4 demonstrates the difference of inference fusion when $\boldsymbol{s}_i^c$ is formed by TCSA and TCMA. In the ideal ZSD, where test RoIs come from unseen domain for sure, $\tilde{\boldsymbol{y}} \in \mathbb{R}^{C_t}$ is computed using unseen class embeddings and $\boldsymbol{S}^c = [\boldsymbol{s}_1^c, ..., \boldsymbol{s}_{C_t}^c]$ only contain C_t unseen classes. In GZSD, $\tilde{\boldsymbol{y}}$ and $\boldsymbol{S}^c$ contains all $C_s + C_t$ classes. However, we find that in all detection benchmarks, both seen and unseen objects would appear in a single test image in practical. If we merely use unseen class embeddings to obtain the classification scores as in ZSL for ZSD, test instances that actually come from seen domain are inevitably misrecognized. The detection performance is consequently degraded. We cannot simply copy the experimental settings of ZSL into ZSD problem, since the assumption that test samples are only from unseen domain is impractical especially for large-scale detection datasets. In view of this, we use class embeddings of total classes to compute classification scores for both ZSD and GZSD.

Two-Stage Model Training

The final loss objective of the proposed method aggregates the classification loss L_{cls}, regression loss L_{loc} and the association prediction loss L_{ap}, i.e. $L = L_{cls} + L_{loc} + \lambda L_{ap}$. λ is the trade-off parameter, balancing the importance between visual perception module and associative transfer module. Instead of training this model in an end-to-end way, we adopt a two-stage training strategy, inspired from (Wang et al. 2020). In the first stage, we train the standard Faster R-CNN model on all training classes. In the second stage, we modify the classification branch to our proposed two parallel reasoning branches with randomly initialized weights. We fine-tune only the classification and association prediction networks, while keeping the entire feature extractor and box regression network fixed. We find that this two-stage fine-tune strategy outperforms the end-to-end strategy by about 2.5 points in mAP on MSCOCO dataset.

Experiments

Experimental Settings

Datasets. We evaluate the proposed method with three well-known detection benchmarks, Pascal VOC (Everingham et al. 2010), MSCOCO (Lin et al. 2014) and Visual Genome (Krishna et al. 2017). For *Pascal VOC*, we follow the protocol in (Rahman, Khan, and Barnes 2020) and use 16 classes as seen classes and the remaining 4 classes as unseen ones (i.e. car, dog, sofa and train). For *MSCOCO*, we choose the 65/15 source/target split (Rahman, Khan, and Barnes 2019) over the 48/17 source/target split (Bansal et al. 2018), for it considers the desired diverse and rarity nature. *Visual Genome* is a large-scale image dataset, composed of 105K unique object classes, 108K images and 3.8M annotated instances. We follow the protocol in (Bansal et al. 2018), which chooses 478 classes as seen classes and the other 130 classes as unseen ones. For semantic embeddings, we use the 64-dimensional semantic attributes for Pascal VOC and l_2 normalized 300-dimensional word2vec (Mikolov et al. 2013) for MSCOCO and Visual Genome.

Methods	Split	ZSD	GZSD		
			mAP_s	mAP_t	HM
SB	48/17	0.70	–	–	–
DSES	48/17	0.54	–	–	–
VSA-ZSD	48/17	10.01	**35.92**	4.12	7.39
VSA-ZSD	65/15	12.40	34.07	12.40	18.18
TL-ZSD*	65/15	14.57	28.79	14.05	18.80
DPIF-S	65/15	<u>17.39</u>	32.75	<u>16.81</u>	<u>22.22</u>
DPIF-M	65/15	**19.82**	29.82	**19.46**	**23.55**
SB	48/17	24.39	–	–	–
DSES	48/17	27.19	15.02	15.32	15.17
ZSD-Textual	48/17	34.30	–	–	–
VSA-ZSD	48/17	43.56	38.24	26.32	31.18
GTNet	48/17	44.60	42.50	30.40	35.45
VSA-ZSD	65/15	37.72	36.38	37.16	36.76
TL-ZSD*	65/15	48.15	54.14	<u>37.16</u>	<u>44.07</u>
DPIF-S	65/15	**58.19**	**57.59**	32.92	41.90
DPIF-M	65/15	<u>55.73</u>	<u>56.68</u>	**38.70**	**46.00**

Table 1: Comparison with state-of-the-arts on MSCOCO dataset for ZSD and GZSD. We report both mAP (the upper part) and recall@100 (the lower part). $*$ denotes the transductive method, while others are inductive. Numbers in bold and underline denote the best and second best results.

Evaluation Metric and Settings. We use mean Average Precision (mAP) with IoU $= 0.5$ and recall@100 as the main evaluation metric, and conduct experiments under two settings, i.e. standard setting and generalized setting. For GZSD, we report both mAP of source classes (mAP_s), mAP of target classes (mAP_t) and their harmonic-mean value (HM).

Implementation Details. We use Resnet-50 as the feature extraction backbone, for fair comparison. We use a fully-connected (fc) layer as f_v, an identity function as f_a and a fc-layer with sigmoid activation on top for f_s. The RoI feature extractor in both paths share the same architecture. K and α in multi-associative construction are set to 5 and 0.1, respectively. K and α in TCMA are set to 5 and 0.1, respectively.

In the training stage, we first train the standard Faster R-CNN framework on the training set, where the feature extraction backbone is pre-trained on the 1K classes in ILSVRC 2012 (Russakovsky et al. 2015) and Non-Maximum Suppression (NMS) with a threshold of 0.7 is applied to RPN. Second, we freeze the parameters in feature

Figure 5: Qualitative results of MSCOCO under both the standard setting (the top row) and the generalized setting (the bottom row). The left (or right) figure in each pair of images is obtained by our proposed model without (or with) using concept association. Red and green bounding boxes represent seen and unseen objects respectively.

method	mAP_s	mAP_t	car	dog	sofa	train
HRE	57.9	54.5	55.0	82.0	55.0	26.0
VSA-ZSD	63.5	62.1	63.7	87.2	53.2	**44.1**
DPIF	**73.17**	**62.26**	**63.72**	**90.08**	**63.55**	31.7

Table 2: Comparison on Pascal VOC for GZSD problem. Each of the last four columns denotes AP of the specific unseen class.

Methods	mAP	recall
SB	-	4.09
DSES	-	4.75
LAB	-	5.40
ZSD-Textual	-	7.20
GTNet	-	11.30
DPIF-S	1.66	15.89
DPIF-M	**1.81**	**18.25**

Table 3: Comparison with state-of-the-arts on the large-scale Visual Genome. We report both mAP and recall@100.

extraction backbone, RPN and box regressor and then train our proposed model by minimizing $L_{cls} + L_{ap}$. We use the SGD optimizer (Bottou 2010) with a batch size of 14 in the first step and a batch size of 18 in the second step, exponential decay rates of 0.9 and 0.999, weight decay of 0.0001 and a learning rate of 0.01 to train our model. In the inference stage, we apply NMS with a threshold of 0.7 to RPN to generate object proposals and NMS with a threshold of 0.3 on the predicted boxes to obtain the final detection results. We implement our model using Pytorch. For MSCOCO, the training stage takes about 20 hours with 2 TITAN V GPUs. The inference stage takes about 0.14 seconds per test image. The code is available https://github.com/Lppy/DPIF.

Comparison with State-of-the-arts

We compare the proposed method (denoted as DPIF) with inductive state-of-the-arts, i.e. SB, DSES (Bansal et al. 2018), VSA-ZSD (Rahman, Khan, and Barnes 2020), ZSD-Textual (Li et al. 2019) and GTNet (Zhao et al. 2020). To fully evaluate the proposed model, we also compare with a strong transductive competitor TL-ZSD (Rahman, Khan, and Barnes 2019).

Quantitative Results Table. 1 reports both mAP and recall for ZSD and GZSD on MSCOCO dataset. From this table, we highlight the following results. (1) The proposed method with multiple-association (DPIF-M) improves the detection performance (mAP) from DPIF-S, our model with single-association and trained by using our two-stage strategy. Although recall is declined slightly from 58.19% to 55.73%, mAP has improved significantly from 17.39% to 19.82% in ZSD. While in GZSD, the improvements are a little higher. This validates the effectiveness of concept association to mitigate the domain bias problem and thus improve the zero-shot generalization capability. (2) The proposed model performs best in both mAP and recall for ZSD among all inductive single-path methods. It improves VSA-ZSD by a large margin of 7.42%, even though the latter uses the best reported detector RetinaNet in this area (Rahman, Khan, and Barnes 2020). Amazingly, our method even beats the transductive competitor by 5.25% improvement, which uses test data during training. These results well verify the efficacy of our proposed model. (3) Our model also achieves the best in unseen prediction and HM value for GZSD. Even though RoIs that come from target classes are distracted by these source ones, their detection performances are still good enough, declined by only 0.36% in terms of mAP. Comparing with other methods that perform better on seen classes while worse on unseen classes, our model can well balance between seen and unseen domain. It achieves the purpose of generalized zero-shot learning.

Fig.5 shows some qualitative detection results by the proposed method on MSCOCO dataset. By using concept association during training, our model is capable of correcting some mis-classification and suppressing false positive results. We can see that the visually-similar objects can be distinguished to some extent. Thus, the association relationships among objects act as a powerful complementary to the basic ZSD model.

Results on Pascal VOC. To compare with current models, we report mAP of both seen and unseen classes, AP of each unseen class for GZSD in Table.2. We can see that our model successfully outperforms the recent inductive models.

training	ZSD	seen	unseen	HM
end-to-end	14.87	31.29	13.76	19.12
	55.67	56.09	27.64	37.03
two-stage	17.39	32.75	16.81	22.22
	58.19	57.59	32.92	41.90

Table 4: Comparison between our method using end-to-end training and the proposed two-stage training strategy. The first (second) row in each part denotes mAP (recall).

Results on Visual Genome. We also report the ZSD results in terms of both recall@100 at IoU = 0.5 and mAP on the large-scale Visual Genome in Table.3, for fair comparison. The proposed method can achieve the best results. Comparing with Table.1 and Table.2, we can see that the detection performance is much lower than that on Pascal VOC and MSCOCO. Because of the large number of object categories and dense labeling, unseen objects are actually regarded as background during training. Unseen instances in the test images thus tend to be neglected. This issue inevitably restricts the development towards large-scale applications. Distinguishing between unseen objects and background remains one of the biggest challenges in practical applications.

Ablation Studies

To study the effects of different components in our model, we conduct a series of experiments on MSCOCO and report mAP and recall@100.

Effect of Two-Stage Training Strategy We propose to train our model using a two-stage training strategy, inspired from (Wang et al. 2020). To evaluate its effectiveness, we give a performance comparison in Table.4 between our method (DPIF-S) using two-stage training and end-to-end training. End-to-end training denotes that after the feature extraction backbone is pre-trained on the 1K classes in ILSVRC 2012 (Russakovsky et al. 2015), the whole network is trained by minimizing the overall L. We can see that two-stage training strategy can actually improve the detection performance.

Effect of Associative Semantic Transfer Associations among different concepts in the training stage plays an important role in our framework. It can enhance the model's transferability, when fused with the basic visual-semantic knowledge transfer for ZSD. Because of the lack of labeled data, we use class embeddings to build up the relationship in advance and learn a generalizable predictor on training data. To better enable the generalizable and discriminative properties, we employ target-centric single-associative (TCSA) or multi-associative (TCMA) methods, rather than simply using k-clustering over class embeddings. Fig.6 (right) reports the detection performance comparison for ZSD, when using different semantic transfer respectively. For fair comparison, k is set to $\mathcal{C}_t$ in k-clustering method. We observe that the multi-association approach achieves the best, which encodes richer semantic relationships, benefiting explicit knowledge transfer. Although k-means offers comparable results with single-association method when k is set to $\mathcal{C}_t$, it needs to manually select a proper k.

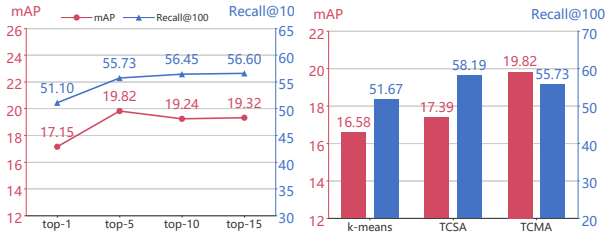


Figure 6: Left: mAP and recall on MSCOCO for ZSD with different value of K in constructing multi-associative semantics. Right: Comparison with the baseline and ours that employs associative semantic transfer on MSCOCO for ZSD. "k-means", "TCSA" and "TCMA" denote that concept associations are formed by k-means, TCSA and TCMA, respectively.

Methods	metric	ZSD	seen	unsen	HM
FL-80	mAP	10.36	**36.69**	10.33	16.12
	recall	34.29	39.53	36.62	36.34
DPIF-S	mAP	14.75	32.72	13.95	19.56
	recall	54.43	57.33	28.76	38.30
DPIF-M	mAP	**16.89**	29.33	**16.36**	**21.00**
	recall	52.25	56.34	34.84	43.03

Table 5: mAP and recall with GloVe embeddings for both ZSD and GZSD.

Impact of Top-K. In multi-association semantic transfer, we associate each source class with top-K unseen objects based on their cosine similarities. Fig.6 (left) shows the impact of different K on the detection performance of unseen classes for ZSD in terms of mAP and recall. We can observe that when $K = 5$, our model achieves the best. The performance drops down if a source class is associated with too many unseen objects, which is reasonable.

GloVe Embedding Our method can work equally with other semantic embeddings apart from word2vec. Table.5 compares the results when using GloVe embeddings in the model. We can observe the same trend as in Table.1. Our model still can achieve the best.

Conclusion

In this paper, we proposed a simple but effective dual-path zero-shot object detection (ZSD) model. It explores both association relationships among concepts and basic visual-semantic knowledge transfer to recognize RoIs. We developed a novel target-centric multi-association method to establish concept associations in advance and used seen instances to learn a generalizable association predictor. These two parallel reasoning paths can strengthen the model's transferability by a reasonable inference fusion mechanism. The dual-path reasoning method is generic and can be possibly applied to recognition and segmentation of unseen objects. Extensive experiments show that our model yields state-of-the-art performance for both standard and generalized settings on various benchmarks.

Acknowledgments

We thank Prof. Zhouchen Lin for helpful discussions. This work was supported by Zhejiang Natural Science Foundation (LQ20F030007), Research Project of Zhejiang Lab (2020KD0AA03, 2019KB0AC01) and the National Natural Science Foundation of China (61473256).

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 819–826.

Anderson, J. R.; and Bower, G. H. 2014. *Human associative memory*. Psychology press.

Bansal, A.; Sikka, K.; Sharma, G.; Chellappa, R.; and Divakaran, A. 2018. Zero-shot object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 384–400.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, 177–186. Springer.

Demirel, B.; Cinbis, R. G.; and Ikizler-Cinbis, N. 2018. Zero-shot object detection by hybrid region embedding. *arXiv preprint arXiv:1805.06157* .

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision* 88(2): 303–338.

Fu, Y.; Hospedales, T. M.; Xiang, T.; and Gong, S. 2015. Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence* 37(11): 2332–2345.

Gidaris, S.; and Komodakis, N. 2018. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4367–4375.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kodirov, E.; Xiang, T.; and Gong, S. 2017. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3174–3183.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123(1): 32–73.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence* 36(3): 453–465.

Li, Y.; Shao, Y.; and Wang, D. 2020. Context-Guided Super-Class Inference for Zero-Shot Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 944–945.

Li, Y.; Wang, D.; Hu, H.; Lin, Y.; and Zhuang, Y. 2017. Zero-shot recognition using dual visual-semantic mapping paths. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3279–3287.

Li, Z.; Yao, L.; Zhang, X.; Wang, X.; Kanhere, S.; and Zhang, H. 2019. Zero-Shot Object Detection with Textual Descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8690–8697.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; and Pietikäinen, M. 2020. Deep learning for generic object detection: A survey. *International journal of computer vision* 128(2): 261–318.

Liu, S.; Long, M.; Wang, J.; and Jordan, M. I. 2018. Generalized zero-shot learning with deep calibration network. In *Advances in Neural Information Processing Systems*, 2005–2015.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *European conference on computer vision*, 21–37. Springer.

Luo, R.; Zhang, N.; Han, B.; and Yang, L. 2020. Context-Aware Zero-Shot Recognition. In *AAAI*, 11709–11716.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

Rahman, S.; Khan, S.; and Barnes, N. 2019. Transductive Learning for Zero-Shot Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 6082–6091.

Rahman, S.; Khan, S.; and Barnes, N. 2020. Improved Visual-Semantic Alignment for Zero-Shot Object Detection. In *AAAI*, 11932–11939.

Rahman, S.; Khan, S.; and Porikli, F. 2018. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Asian Conference on Computer Vision*, 547–563. Springer.

Rahman, S.; Khan, S. H.; and Porikli, F. 2020. Zero-Shot Object Detection: Joint Recognition and Localization of Novel Concepts. *International Journal of Computer Vision* 1–21.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 779–788.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3): 211–252.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* .

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1199–1208.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261* .

Wang, D.; Li, Y.; Lin, Y.; and Zhuang, Y. 2016. Relational knowledge transfer for zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Wang, X.; Huang, T. E.; Darrell, T.; Gonzalez, J. E.; and Yu, F. 2020. Frustratingly Simple Few-Shot Object Detection. *arXiv preprint arXiv:2003.06957* .

Zablocki, E.; Bordes, P.; Piwowarski, B.; Soulier, L.; and Gallinari, P. 2019. Context-Aware Zero-Shot Learning for Object Recognition. *arXiv preprint arXiv:1904.12638* .

Zhao, S.; Gao, C.; Shao, Y.; Li, L.; Yu, C.; Ji, Z.; and Sang, N. 2020. GTNet: Generative Transfer Network for Zero-Shot Object Detection. *arXiv preprint arXiv:2001.06812* .

Zhu, P.; Wang, H.; and Saligrama, V. 2020. Don't Even Look Once: Synthesizing Features for Zero-Shot Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11693–11702.