A Multi-Space Approach to Zero-Shot Object Detection

by

Dikshant Gupta, Aditya Anantharaman, Nehal Mamgain, Sowmya Kamath, Vineeth N
Balasubramanian, C V Jawahar

in

2020 IEEE Winter Conference on Applications of Computer Vision (WACV)
: 1
-10

Report No: IIIT/TR/2020/-1

Centre for Visual Information Technology
International Institute of Information Technology
Hyderabad - 500 032, INDIA
March 2020

A Multi-Space Approach to Zero-Shot Object Detection

Dikshant Gupta[1], Aditya Anantharaman[2], Nehal Mamgain[3], Sowmya Kamath S[2], Vineeth N Balasubramanian[3], and C.V. Jawahar[1]

[1]International Institute of Information Technology, Hyderabad
[2]National Institute of Technology Karnataka, Surathkal
[3]Indian Institute of Technology, Hyderabad
dikshant2210@gmail.com, adityanant@gmail.com, nehal.mamgain@gmail.com, sowmyakamath@nitk.edu.in, vineethnb@iith.ac.in, jawahar@iiit.ac.in

Abstract

Object detection has been at the forefront for higher level vision tasks such as scene understanding and contextual reasoning. Therefore, solving object detection for a large number of visual categories is paramount. Zero-Shot Object Detection (ZSD) – where training data is not available for some of the target classes – provides semantic scalability to object detection and reduces dependence on large amount of annotations, thus enabling a large number of applications in real-life scenarios. In this paper, we propose a novel multi-space approach to solve ZSD where we combine predictions obtained in two different search spaces. We learn the projection of visual features of proposals to the semantic embedding space and class labels in the semantic embedding space to visual space. We predict similarity scores in the individual spaces and combine them. We present promising results on two datasets, PASCAL VOC and MS COCO. We further discuss the problem of hubness and show that our approach alleviates hubness with a performance superior to previously proposed methods.

1. Introduction

Object detection consists of both identifying and localizing the objects in an image. It has numerous applications in many domains, including robotics, self-driving cars, medical imaging, and surveillance. Object detection has consequently met with great successes over the last few years [31, 14, 15, 16, 22, 30, 29, 6]. However, the performance of such models is largely limited to the fully supervised domain, viz. detecting classes fully present in the training data. Hence, current state-of-the-art models lack the important property of *semantic scalability*, by virtue of which a model trained on a set of classes can identify classes which are not in the training set, but are semantically related and can occur in the wild.

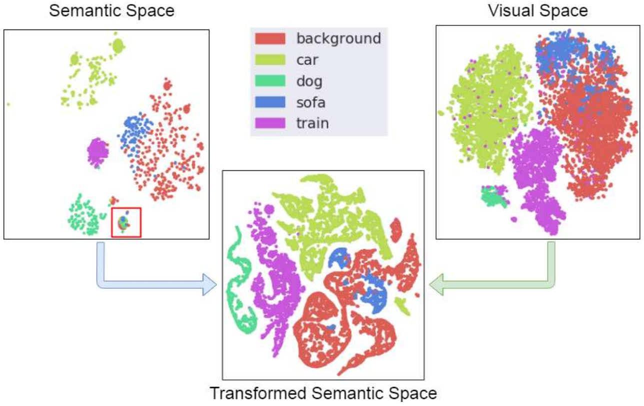

Figure 1. The semantic and visual spaces in a dataset yield complementary information to learn better well-separated embeddings in a transformed semantic space.

For the problem of image classification, there has been a significant amount of effort in recent years to scale up the models semantically [12]. Such efforts, described as zero-shot recognition (ZSR), often use word embeddings learned from large text corpora in an unsupervised manner [1, 2, 11], semantic attributes [19, 18] or concept ontologies [24] as label embeddings to help recognition of unseen classes. While commendable progress has been made, ZSR solves the basic task of image-level classification for unseen classes, where a dominant object is present in an image. It cannot scale to tasks like scene understanding where a sound reasoning of all objects in an image is required. There is an impending need to solve a more complex problem such as Zero-Shot Object Detection (ZSD) where instances of unseen objects are not only recognized but also localized. In this paper, we propose a new solution to ZSD, where very limited work has been done so far, and which can be useful for several mainstream applications that involve scene understanding.

In ZSD, we recognize and localize instances of objects that were not present during training. We refer to the classes present in the training set as seen classes, and the rest as unseen classes. Contemporary deep learning models for object detection [31, 14, 16] employ a background class to discern between foreground and background proposals, which improves the results as it suppresses proposals which contain background elements such as sky, vegetation and roads, and rewards proposals which contain the object(s) of interest. Unlike ZSR, ZSD has an additional task of defining the background class embedding which is non-trivial as background proposals may contain unseen classes. This issue exacerbates other issues inherited from ZSR such as hubness [32] (discussed in Section 5) which occurs when some of the target classes are the nearest neighbours for most of the region proposals.

There has been limited work so far on ZSD [38, 7, 28, 5], and all of them focus on similarity between embeddings in the semantic space. In this paper, we propose a new methodology for ZSD that leverages the use of the two spaces used in contemporary object detection frameworks: the *visual* and *semantic*. We define the space spanned by semantic label embeddings or word vectors [26] as the *semantic space*, while the space spanned by image features of region proposals as the *visual space*. As seen in Figure 1, 'car' and 'train' classes have poor separation in visual space but are very well separated in semantic space. Similarly, 'dog' and 'background' are poorly separated in semantic space(as shown in the red box) but clearly distinguishable in visual space. We combine both the spaces in a multi-space approach and exploit the complementary information from both spaces, enabling the learning of a better transformed semantic embedding with well separated classes. We carefully leverage this property of complementarity by learning transformed semantic and visual spaces, procuring and combining similarity scores from both spaces while minimizing the correlation between the visual space and the transformed visual space. The resultant multi-space approach discriminates among classes effectively. We propose our methodology within the framework of the widely used region-based object detection networks, in particular the Faster R-CNN [31] framework. We present our results on two datasets: PASCAL VOC [8] and MS COCO [21], and show that the proposed approach yields state-of-the-art results for the ZSD task. We also adapt two common works on ZSR, viz. DeViSE [11] and ConSE [25] to ZSD as baselines, and provide quantitative evaluation of these methods in comparison to ours. We compare them to the multi-space approach and show superior results for both datasets. We also show that using our proposed multi-space model alleviates hubness. In summary, our key contributions in this work are:

- We propose a novel multi-space approach which leverages both visual and semantic spaces for ZSD. We note that a multi-space approach has not been studied for ZSR either, and this is the first such effort to the best of our knowledge.
- In addition to using loss terms on each of the above mentioned spaces, we also introduce a cross-modal consistency loss based on minimizing the correlation between representations from the two spaces.
- We show that the proposed approach leads to state-of-the-art ZSD results, both quantitative and qualitative, on the PASCAL VOC and MS COCO datasets, and study the performance of variants of our methodology on these datasets.
- We show that the proposed method can provide a solution to the hubness problem in ZSD by combining scores from semantic and visual spaces.

The remainder of our paper is organized as follows. Section 2 reviews the related work followed by Section 3 which describes the proposed approach in detail. Experimental results are presented in Section 4, discussions and analysis in Section 5, followed by conclusions in Section 6.

2. Related Work

Object Detection: There have been significant developments in object detection over the last few years. Girshick *et al.* [15] proposed R-CNN (Region-based Convolutional Neural Network) that classifies each region proposal using a deep CNN. Girshick *et al.* [14] proposed Fast R-CNN that extracts convolutional features for region proposals before further processing, and Ren *et al.* [31] improved upon this method by making this approach, including the region proposals, learnable end-to-end. Dai *et al.* [6] proposed R-FCN (Regional Fully Convolutional Network), introducing position-sensitive RoI pooling that shares almost all computations for an image. He *et al.* [16] proposed Mask R-CNN for object detection as well as instance segmentation Redmon *et al.* [30, 29] and Liu *et al.* [22] introduced the YOLO and SSD framework respectively to predict detection and classification probabilities using a single deep neural network. We use Faster R-CNN[31] as the base architecture for our ZSD method, considering its wide use as an accurate method.

Zero-Shot Recognition/Learning: ZSR can be broadly categorized into two approaches: semantic attribute-based and semantic embedding-based. A semantic attribute [9] refers to the characteristics possessed by a class, for e.g., color, shape or annotations such as 'has head'. Lampert *et al.* [19, 18] proposed attribute-based classification that identifies objects based on a high-level description

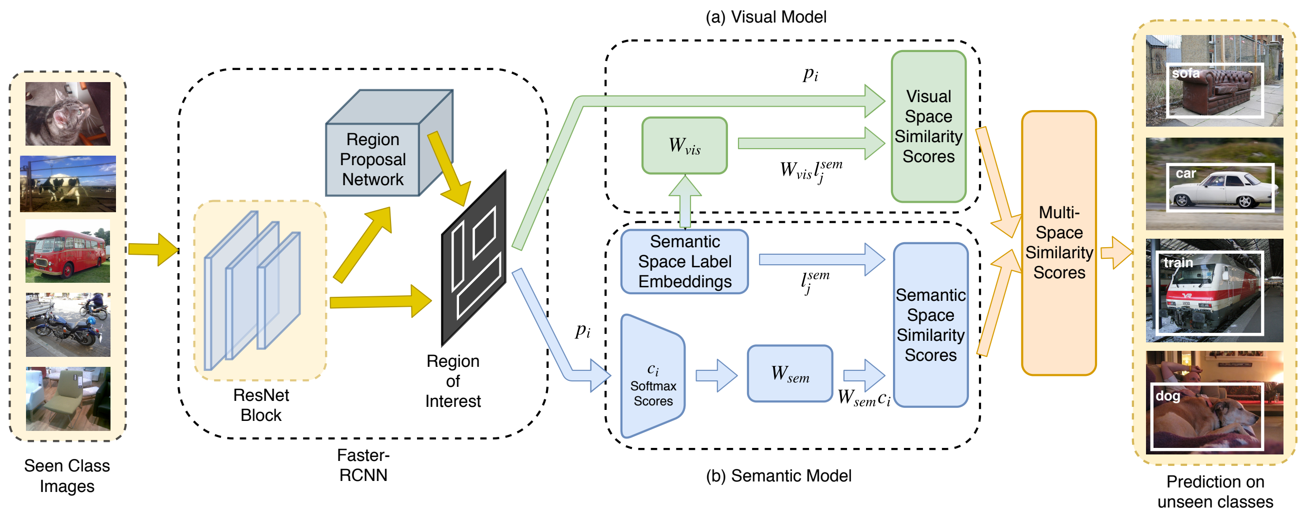

Figure 2. Architecture for multi-space zero-shot object detection. (a) Visual space approach with learnable transformation W_{vis} and distance scores in visual space. (b) Semantic space approach with learnable transformation W_{sem} and distance scores in semantic space.

phrased in terms of semantic attributes. Akata *et al.* [1, 2] and Frome *et al.* [11] developed a bilinear compatibility framework that uses a pairwise ranking formula to learn the parameters of the bilinear model. Xian *et al.* [36] learned a collection of maps with selection instead of learning a single bilinear map, resulting in a piece-wise linear decision boundary. Ba *et al.* [4] used text features to predict output weights of both convolutional and fully connected layers. Recently, Zhang *et al.* [37] proposed the use of visual space as the embedding space due to nearest neighbour search suffering much less from hubness in visual space. Annadani *et al.* [3] also utilized visual space as embedding space and proposed an objective function to preserve semantic relations in the visual space. Mikolov *et al.* [23] and Pennington *et al.* [26] propose word embeddings that map words to the continuous vector space which encodes semantic and syntactic similarity between words. However, all aforementioned efforts are explicitly for classification (not detection, which is the focus of this work). We however adapt popular ZSR methods to detection, and include them as baselines in our experimental studies. We also exploit the property of word embeddings in [23][26] to semantically scale up object detection to target classes for which no training data is available, in this work. Recently, Wang *et al.* [34] proposed a model that learns an intermediate latent space as the embedding space (for classification again). Jiang *et al.*[17] combine visual and semantically aligned intermediate spaces. Intermediate space however suffers from hubness problem as pointed out in [3], and we hence propose a new multi-space model for ZSD. We extend the idea of cross-modal consistency loss proposed in [10] to ZSD by introducing a correlation loss.

Zero-Shot Object Detection: Demirel *et al.* [7] proposed a hybrid of convex combination of class embeddings [25] and label embedding-based classification. Rahman *et al.* [28] proposed an extension of Faster R-CNN [31] and ConSE [25] with a loss formulation that combines max-margin and semantic clustering losses. Zhu *et al.* [38] proposed a zero-shot detection framework that fuses semantic attribute prediction with visual features to predict objectness scores for bounding box proposals. Bansal *et al.* [5] and [38] addressed the problem of confusing background with unseen classes. However, all of these methods rely on the semantic space for the ZSD task. In contrast to previous works on ZSD, we propose a multi-space approach which utilizes both semantic and visual spaces. We show that our multi-space approach outperforms earlier methods on both PASCAL VOC [8] and MS COCO [21] datasets.

3. MS-Zero++: Our Multi-Space Approach

Zero-shot object detection (ZSD) aims at recognition and localization of unseen objects. Our approach is motivated by the observation that there are effectively two spaces, the *visual* and the *semantic*, when performing zero-shot learning. All existing related efforts focus on taking the visual space feature vectors to a semantic space, and comparing with the class label embeddings in the semantic space. This procedure is inherently limiting, and forces the model to rely largely on the discriminative capacity of the semantic space for recognizing classes. However, as evident from the embeddings in the visual space in Fig 1, the visual space has discriminative capacity by itself, which needs to be leveraged to improve the performance of zero-shot learning. In this work, we aim to bring together the capacities of both these spaces to obtain state-of-the-art performance for ZSD.

In particular, we combine the capabilities of the individual spaces in a multi-space (MS-Zero) approach which is summarized in Figure 2. The Visual Space model (MS-Zero-V) captured in part (a) of Figure 2 leverages the similarities between object categories in the visual feature space, while the semantic space model (MS-Zero-S) captured in part (c) of Figure 2 exploits the semantics between category embeddings for object classification. Training the semantic space model using a cross-entropy loss on softmax scores and a mean square error loss on transformation matrix, W_{sem}, ensures that the semantic space not only learns to discriminate categories well, but also the transformation from visual-to-semantic space is learned properly. Similarly, in the visual space model, a max-margin loss on similarity scores enforces the model to discriminate well. We also add a new correlation loss which ensures that the learned transformation matrix in the visual space, W_{vis}, matches the semantics between category embeddings and the visual similarities between the corresponding object categories. We now describe our methodology in detail.

We denote the set of all classes as $\mathcal{C} = \mathcal{S} \cup \mathcal{U}$, where $\mathcal{S}$ denotes the set of seen classes and $\mathcal{U}$ denotes the set of unseen classes, and $\mathcal{S} \cap \mathcal{U} = \emptyset$. Each image is denoted as $I \in \mathbb{R}^{M \times N \times 3}$, with corresponding bounding boxes and ground truth labels denoted as $b_i \in \mathbb{N}^4$ and $y_i \in \mathcal{C}$ respectively. We use region-based CNNs (Faster R-CNN) as the object detector of choice, because of their superior accuracy compared to one-shot methods such as YOLO [30]. The network generates visual features for region proposals denoted by $p_i \in \mathbb{R}^{d_1}$ and softmax scores denoted by $c_i \in \mathbb{R}^{d_2}$ where $d_2 = |\mathcal{S}|$ is the number of seen classes.

Semantic Model (MS-Zero-S): The key idea of the semantic model is to provide reasoning in the semantic space (as defined by the object categories) in an effective manner. The space spanned by semantic embeddings of class labels is denoted by $\mathcal{M}_{sem} \in \mathbb{R}^{d_3}$, where d_3 is the dimensionality of the space. Given $l_j^{sem} \in \mathcal{M}_{sem}$, the embedding for class j in the semantic space (which is obtained using standard methods such as GloVe [26]), we learn a transformation $W_{sem} \in \mathbb{R}^{d_3 \times d_2}$ (defined by a sub-network) that transforms softmax scores $c_i \in \mathbb{R}^{d_2}$ for i^{th} proposal p_i to semantic space. We then use a similarity metric denoted by ϕ (such as a cosine similarity; more implementation details are presented in Section 4), to capture the similarity between the class embeddings and the semantic space embeddings obtained from the softmax scores, i.e. our final similarity scores in semantic space are given by:

$$S_{ij}^{sem} = \phi(l_j^{sem}, W_{sem}c_i) \qquad \forall j \in \{1, ..., \mathcal{C}\} \qquad (1)$$

We note here that the softmax layer predicts scores only for seen classes. Classification loss and bounding box regression loss for region proposals as well as the region classification network are used as described in [31] and denoted by $\mathcal{L}_{cls}$ and $\mathcal{L}_{reg}$ respectively. The transformation matrix W_{sem} is learned using a Mean Squared Error (MSE) loss given by:

$$\mathcal{L}_{mse}(c_i, y_i^{gt}) = \frac{1}{d_3}\sum_{k=1}^{d_3}((W_{sem}c_i)_k - l_{j,k}^{sem})^2 \qquad (2)$$

where y_i^{gt} is the ground truth label (one-hot vector) for the proposal. $(W_{sem}c_i)_k$ and $l_{j,k}^{sem}$ are the k^{th} elements of the transformed softmax scores and the ground truth semantic embedding l_j^{sem} corresponding to y_i^{gt} respectively.

Visual Model (MS-Zero-V): Motivated by [3, 37, 32] which considered a visual space for ZSR, we propose a module of reasoning directly in the visual space. To this end, we consider the visual features obtained for each region proposal, $p_i \in \mathbb{R}^{d_1}$. In order for us to use these visual features for reasoning w.r.t. class labels, we learn a transformation, W_{vis} (defined by a sub-network) that transforms the class embeddings l_j^{sem} (as used before) to the visual space, $\mathcal{M}_{vis} \in \mathbb{R}^{d_1}$. The similarity score between class categories and region proposals in the visual space is then captured by:

$$S_{ij}^{vis} = \phi(W_{vis}l_j^{sem}, p_i) \qquad \forall j \in \{1, \ldots, \mathcal{C}\} \qquad (3)$$

where ϕ is a similarity metric as before. Since both inputs to similarity measure in (3) are uncertain, we use max-margin loss to learn the transformation matrix. The loss is given by:

$$\mathcal{L}_{margin}(p_i, y_i^{gt}) = \sum_{j \in S, j \neq y_i^{gt}} max(0, m - S_{ii}^{vis} + S_{ij}^{vis}) \qquad (4)$$

where m refers to the margin, S_{ii}^{vis} is the visual space score w.r.t. the ground truth class embedding l_i^{sem}, and S_{ij}^{vis} is the visual space score w.r.t. all class embeddings other than the ground truth class embedding. Max-margin loss enforces a constraint on similarity and separates individual classes.

Modern object detection approaches define an additional background class to differentiate between foreground and background proposals. This eliminates proposals which do not contain any object of interest. Since in ZSD, background proposals may contain objects that belong to unseen classes, defining a semantic embedding for background class is nontrivial. Similar to [28], in this paper we consider background class embedding in semantic space to be the mean of all semantic class embeddings: $l_{bg}^{sem} = \frac{1}{\mathcal{C}}\sum_{j=1}^{\mathcal{C}} l_j^{sem}$, where, l_{bg}^{sem} represents the background class embedding in semantic space. Background embedding is chosen in this manner to not be representative of any particular class in $\mathcal{C}$.

Proposed Multi-Space Model (MS-Zero++): In the proposed multi-space model, we bring together the loss function terms used for both visual and semantic models, and also combine the similarity scores obtained from the semantic and visual space model by averaging them, i.e. $S_{ij}^{mlt} = \frac{1}{2}(S_{ij}^{sem} + S_{ij}^{vis})$. While this simple approach of combining the loss terms worked reasonably well (and is called MS-Zero in our experimental results), we also introduce a cross-modal consistency loss term, which is intended to maintain consistency across the representations

Model	Seen	Unseen	Mix
DeViSE-ZSD	59.22	50.63	46.57
ConSE-ZSD	73.19	55.30	56.61
ZSD-YOLO [7]	65.60	54.20	52.33
MS-Zero-S	75.93	55.55	57.95
MS-Zero-V	70.47	51.39	56.44
MS-Zero	74.49	**62.15**	60.05
MS-Zero++	**76.09**	57.07	**61.08**

Table 1. Table presents the mean average precision (mAP) (%) of all models on PASCAL VOC dataset using different test settings.

Model	car	dog	sofa	train
DeViSE-ZSD	44.22	81.51	48.73	28.07
ConSE-ZSD	62.24	83.70	58.16	17.08
ZSD-YOLO [7]	55.00	82.00	55.00	26.00
MS-Zero-S	60.40	85.89	54.15	21.78
MS-Zero-V	38.42	83.79	54.34	29.04
MS-Zero	**69.00**	86.80	**65.99**	26.81
MS-Zero++	55.24	**89.15**	58.39	25.5

Table 2. Table shows the class wise Average Precision (AP) (%) for the unseen classes for all models in test-unseen setting on PASCAL VOC dataset.

from both modalities. Similar to efforts in style transfer [13, 35, 20], we write this loss term as below, and call it the correlation loss:

$$\mathcal{L}_{corr}(l^{sem}) = \psi(corr(W_{vis}l^{sem}) - corr(l^{sem})) \tag{5}$$

where ψ gives the $smooth\ L1$ loss and $corr$ represents the correlation matrix of the input. Note that the $smooth\ L1$ loss is given by:

$$\psi(x) = \begin{cases} 0.5x^2 & \text{if}|x| < 1 \\ |x| - 0.5 & \text{if}|x| \geq 1 \end{cases} \tag{6}$$

The above term improves the performance of the overall model, as shown in our results. We call the overall model that uses all the loss terms as MS-Zero++. The loss function used to train MS-Zero++ is given by:

$$\mathcal{L}_{mlt}(p_i, y_i^{gt}, \Theta) = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{mse} + \mathcal{L}_{margin} + \mathcal{L}_{corr} \tag{7}$$

where Θ represents the combined parameters of the network, including the transformation matrices introduced.

4. Experiments

Datasets: We validated the proposed method on two large commonly used datasets: PASCAL VOC [8] and MS COCO [21]. **PASCAL VOC:** This dataset contains 20 object categories broadly divided in four super-categories namely: 'person', 'vehicle', 'animals' and 'indoor'. We select *car* and *train* from the super-category 'vehicle', *sofa* and *dog* from the super-categories 'indoor' and 'animal' respectively as unseen classes. Since super-category 'person' has only one sub-category, it is excluded from unseen classes.

Model	Seen	Unseen	Mix
	(mAP)	(mAP/recall)	(mAP)
DeViSE-ZSD	30.3	10.6	21.5
ConSE-ZSD	42.4	9.3	30.5
Bansal et al. [5]	-	0.70/27.19	-
MS-Zero-S	**42.6**	9.8	**30.8**
MS-Zero-V	37.7	12.2	26.6
MS-Zero	42.4	12.9	30.7
MS-Zero++	35.0	**13.8/35.0**	26.0

Table 3. Table shows the mean average precision (mAP) (%) and Recall@100 (%) comparison of all models for MS COCO dataset on three different test settings.

In total we consider 16 seen classes and 4 unseen classes. We use *training/validation* sets of the 2007 and 2012 data for training and test the model on the 2007 *test* set. We define three testing configurations, named as *Test-Seen*, *Test-Unseen* and *Test-Mix*. These three settings consist of the same image sets as used in [7]. *Test-Seen* considers images which contain only seen classes, *Test-Unseen* considers images which contain only unseen classes and *Test-Mix* is the combination of both seen and unseen classes. The major difference between all three configurations is in the nearest neighbor search space: *Test-Seen* and *Test-Unseen* contain only seen classes and unseen classes in the search space respectively while *Test-Mix* contains both. **MS COCO:** This dataset contains 80 object categories. We follow a strategy similar to [5] for creating seen/unseen class splits. We only consider classes that have a synset associated with them in the WordNet [24] hierarchy. We split the classes into 10 clusters and select 80% of the classes as seen and 20% as unseen. This gives us 48 seen classes and 17 unseen classes. For testing, we follow the same configurations as described for PASCAL VOC. We use the 2014 *train* set for training and validation while the 2014 *val* set for testing.

Baseline: We present results of [5, 7] as baseline for our model in Tables 3, 2, 3. In order to ensure fair comparison of our proposed method, we also adapted two prior works on zero-shot recognition, namely DeViSE [11] and ConSE [25] to the detection context, and used them as our baselines. Since these methods are considered landmark approaches in ZSR, it is natural to compare against their performance in the ZSD setting. For DeViSE-ZSD, we transformed the visual features of each proposal to the semantic space and learned the transformation using max-margin loss as described in equation 4. For ConSE-ZSD, we trained the model on seen classes in a fully supervised setting. We obtained the embedding for each proposal by a weighted combination of semantic class embeddings where weights for each class are the softmax scores generated by the model.

Implementation Details: As in Faster R-CNN, we use a shallow CNN on top of image features to generate region proposals. In line with all existing efforts [5, 28, 38, 27],

Model	***airplane***	***bus***	***cat***	***dog***	***cow***	***elephant***	***umbrella***	***tie***	***snowboard***	***skateboard***	***cup***	***knife***	***cake***	***couch***	***keyboard***	***sink***	***scissors***
DeViSE-ZSD	**26.1**	28.1	12.1	8.5	23.7	7.1	0	0	0.8	0.3	4.8	0.4	15.6	20.1	13.7	14.1	5.8
ConSE-ZSD	10.7	**51.8**	0	10.1	0	**25.3**	0.1	0	5.5	**5.9**	8.5	2.3	2.7	26.1	1.5	0	8.4
MS-Zero-S	11.8	48.8	0	9.4	0	24.1	0.2	0	**11.5**	2.2	9	**3.6**	3.8	32.9	0.8	0	**8.8**
MS-Zero-V	8.8	36.5	**13.7**	**11.0**	34.2	4.7	**0.3**	0	2.5	0.4	9.3	2.9	11.8	30.1	19	**14.2**	7.7
MS-Zero	10.9	51.0	2.7	8.0	**35.1**	12.9	0	0	10.9	0.7	**10.5**	0.5	5.8	40.6	21.5	1.0	7.0
MS-Zero++	19.6	35.5	4.3	9.3	33.0	4.0	0.1	0	3.5	5.0	10.3	1.1	**20.0**	**41.5**	**26.3**	13.6	7.1

Table 4. Class-wise Average Precision (AP) (%) for unseen classes for all models in test-unseen setting on MS COCO dataset.

we use ResNet-101 pretrained on the Imagenet-1k dataset to extract image features. The choice can be understood as leveraging common world knowledge. Image features are extracted using a ResNet-101 network. We assign labels to each proposal on the basis of IoU (Intersection over union) threshold. Any proposal with an $IoU > 0.5$ with a ground-truth box is considered a foreground proposal while any proposal with $0 < IoU < 0.2$ with a ground truth box is considered as a background box. We train on images containing only seen classes, therefore none of the unseen categories are misclassified as background during training. We use stochastic gradient descent with an initial learning rate of 10^{-3} and momentum of 0.9. In case of MS COCO, we train the model for a total of 5 epochs with a learning rate of 10^{-3}. In case of PASCAL VOC, we train the model for 8 epochs with a learning rate of 10^{-3} for the first 5 epochs and 10^{-4} for the remaining three epochs after which the results saturate. Margin for ranking loss is set to 0.1. We train the model end-to-end while keeping the semantic embeddings fixed. For MS-Zero-S, the weights for the transformation matrix W_{sem} are initialized with semantic attribute embeddings [9] of seen classes for PASCAL VOC dataset and with GloVe embeddings [26] for MS COCO. For MS-Zero-V we transform the 300 dimensional semantic embedding of each class to 2048 dimensional visual space using a fully connected layer. MS-Zero-S and ConSE-ZSD utilize 64 dimensional semantic attribute embeddings [9] for PASCAL VOC dataset and 20 dimensional semantic embeddings [38] for MS COCO dataset. Label embeddings for MS-Zero-V and DeViSE-ZSD method for both the datasets are extracted from publicly available GloVe [26] embeddings. MS-Zero++ utilizes 64 dimensional semantic attribute embeddings [9] for PASCAL VOC dataset and GloVe [26] embeddings for MS COCO dataset. We normalize the semantic scores and visual scores using $L2$ norm before combining them. In case of the MS-Zero model, similarity metric used for both semantic and visual branches of the model is cosine similarity. In case of the MS-Zero++ model, similarity metric used for the semantic space branch of the model is $L2$ distance for MS COCO and cosine similarity for PASCAL VOC. The visual space branch used cosine similarity for both datasets. Choice of the similarity metric was based on an initial empirical study with different metrics.

Results: We use mean average precision (mAP) with an IoU threshold of 0.5 with ground truth boxes as the evaluation metric for the model. Tables 3 and 3 provide quantitative results in terms of mAP for MS-Zero, MS-Zero++ and its special cases - MS-Zero-S and MS-Zero-V on PASCAL VOC and MS COCO datasets respectively on the three different test settings. Figure 3 provides the qualitative results for the proposed models. We provide an extensive performance comparison of our work with existing work in ZSD. We compare our results on the PASCAL VOC dataset with the hybrid region embedding model proposed by [7] and refer to it as ZSD-YOLO in further discussion, table 3 and 2. Since they do not use MS COCO dataset and do not have publicly available code, we do not compare our results on MS COCO with them. We compare our results on the MS COCO dataset with Bansal et al. [5] which is the best work so far on MS COCO. Since they do not use PASCAL VOC dataset and do not have publicly available code, we do not compare our results on PASCAL VOC with them. We could not compare our work with [38] as they only report class-agnostic average precision thereby relaxing the recognition constraint for them. We also could not compare our work with [28] as they use a different dataset and have no publicly available code.

PASCAL VOC: To our knowledge, ZSD-YOLO [7] is the best work so far in terms of mAP on PASCAL VOC. As seen in Table 3, MS-Zero outperforms ZSD-YOLO by a margin of 14% in test-unseen setting. MS-Zero++ outperforms ZSD-YOLO by a margin of around 16% in both test-seen and test-mix settings. We observe that MS-Zero++ performs best in test-seen and test-mix settings and MS-Zero performs best in test-unseen setting. We attribute this to the fact that the correlation loss (equation 5) helps MS-Zero++ retain the semantic properties in the visual space as well. But the number of seen classes (16) is not enough to learn a generalized transformation from semantic space to visual space. We also note that, as seen in Table 2, MS-Zero and its variants leverage both semantic and visual spaces to outperform other methods on each class.

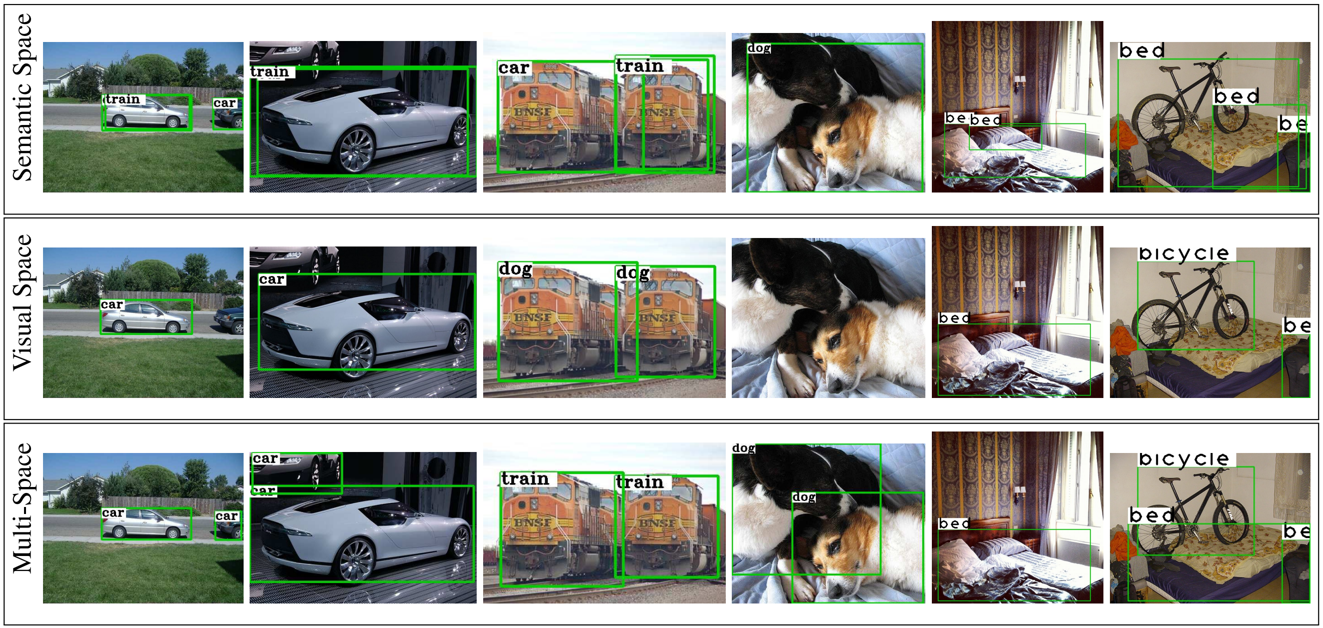

Figure 3. Qualitative detection results on test images from PASCAL VOC (first four images) and MS COCO (last two images) dataset. Topmost row shows detection and classification results using MS-Zero-S. Middle row shows detection and classification results using MS-Zero-V. Results from both models are prone to misclassification. Bottom row shows detection results using MS-Zero. It is evident from results that MS-Zero outperforms individual search space methods.

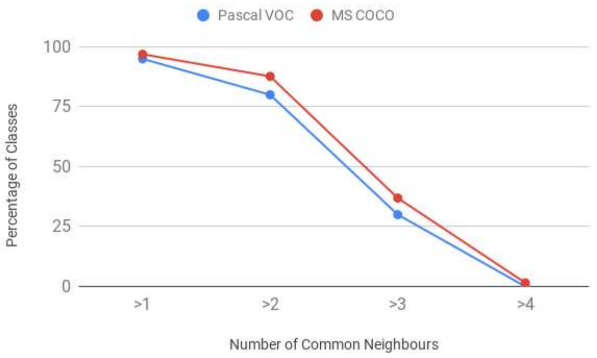

Figure 4. Percentage of classes having a certain number of common neighbours in both semantic and visual space.

MS COCO: MS-Zero++ performs the best on MS COCO in the test-unseen setting. MS-Zero-S outperforms MS-Zero in test-seen and test-unseen setting by a small margin. We observe that our MS-Zero++ model achieves a huge 13 mAP increase and a 28% increase in recall@100 value when compared to the previous best, Bansal et al. [5] as seen in Table 3. We observe that MS-Zero-S performs better than all other models in test-seen and test-mix settings. This can be attributed to good separation for the seen classes in the semantic space. The performance in the test-mix setting is biased towards seen classes therefore the results for test-mix are higher for MS-Zero-S as well. A low AP for some of the classes in MS COCO dataset as seen in Table 4 occurs when there is a lack of semantically similar classes in the seen set and very small size of objects. For e.g. the classes ‘tie’ and ‘umbrella’ belong to the broad super category ‘accessory’, for which almost no semantically similar classes are available in the training set. Addressing this challenge, which is an issue across existing methods [27, 7, 28], will be an important direction of our future work.

We observe that MS-Zero-S performs better compared to MS-Zero-V on PASCAL VOC while, MS-Zero-V outperforms MS-Zero-S on MS COCO in test-unseen setting. We posit that since the number of seen classes is much lesser in PASCAL VOC (16) than MS COCO (48), it is difficult for MS-Zero-V to learn the high dimensional visual space embeddings from semantic embeddings, that can retain semantic properties for PASCAL VOC dataset. Higher number of seen classes offered by MS COCO dataset enables the model to learn visual space embeddings that can retain semantic properties. To justify this, we compare the number of common neighbours for each class in semantic space and visual space. We find top-five nearest neighbours for each class in semantic space and visual space and calculate the number of neighbours shared in both space. Figure 4 presents the percentage of classes having a certain number of common neighbours. It is evident from the results that even though PASCAL VOC has a smaller search space, number of shared neighbours in the semantic and visual space are lower as compared to MS COCO dataset.

We recently came across another work in ZSD based on

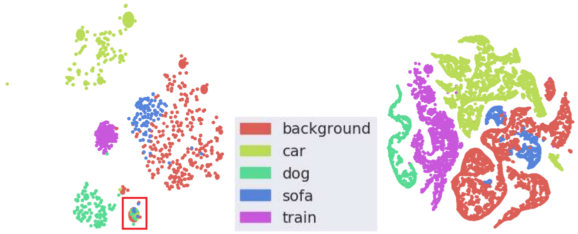

Figure 5. 2-D t-SNE [33] plot of the semantic attribute embeddings [9] (human-annotated) (left) and the transformed semantic embeddings (right) on PASCAL VOC dataset.

Model	Seen	Unseen	Mix
PASCAL VOC			
MS-Zero-Avg-U	68.15	55.25	54.57
MS-Zero-Max	73.45	57.35	58.75
MS-Zero-Avg-N	**74.49**	**62.15**	**60.05**
MS COCO			
MS-Zero-Avg-U	42.4	**12.9**	30.7
MS-Zero-Max	43.3	10.0	31.3
MS-Zero-Avg-N	**43.5**	11.2	**31.4**

Table 5. Table shows mean average precision (mAP) (%) for the un-normalized, max and average models on PASCAL VOC and MS COCO datasets on different test settings.

polarity loss [27] which we refer to as ZSD-POLARITY. On the PASCAL VOC dataset, our MS-Zero model with an mAP of 62.15% outperforms ZSD-POLARITY by a small margin which has an mAP of 62.1%. MS-Zero also outperforms ZSD-POLARITY in the test-seen setting with a 17% improvement in mAP (74.49% vs 63.5%). On the MS COCO dataset, our MS-Zero++ model with an mAP of 13.8% outperforms ZSD-POLARITY which has an mAP of 10.01%. Our proposed models outperform ZSD-POLARITY on both datasets in terms of mAP.

5. Discussion and Analysis

Hubness: Hubness phenomenon is associated with nearest neighbor search and is observed when a few objects in the dataset occur as the nearest neighbor of many objects, therefore becoming universal neighbours or *hubs*. We argue that in ZSD since, the background semantic class embedding is the average of all semantic class embeddings, the background class by virtue of its definition tends to become a hub and is often assigned to bounding boxes with unseen classes. We define hubness for each class as, $H_k(y) = |\{x, x \in P | y \in NN_k(x, \mathbf{S})\}|$ where x denotes a region proposal, P denotes the set of all region proposals, **S** is the search space and $NN_k(x, \mathbf{S})$ denotes the k nearest neighbor of x in **S**. We show how our proposed multi-space approach alleviates this problem of hubness in ZSD. Figure 5 shows 2-D t-SNE [33] plots of the semantic attribute embeddings (human-annotated) obtained from [9] and the transformed semantic embeddings on PASCAL VOC dataset. It can be seen that the semantic embeddings of MS-Zero learn a better representation for the 'dog' class than the human-annotated embeddings for 'dog' class which overlap with 'background', 'car' and 'sofa' classes (as shown in the red box). Furthermore, the semantic embeddings of MS-Zero learn a more closely associated representation for 'car' and 'background' classes separately and overcome the hubness caused by the 'background' class (for 'sofa' and 'dog' classes) in the semantic attribute embeddings. This reinforces our claim that hubness decreases as MS-Zero combines information from both the semantic and visual spaces.

Combining scores through other methods: We combined scores in both spaces using two approaches: (1) MS-Zero-Avg-U where we take the average of the scores. (2) MS-Zero-Avg-N where we normalize scores using $L2$ norm before taking average. Table 5 shows that MS-Zero-N performs the best on both datasets in test-seen and test-mix setting. MS-Zero-Avg-N and MS-Zero-Avg-U give the best results for PASCAL VOC and COCO datasets respectively. This can be attributed to the fact that, since the unnormalized scores are a weighted combination of normalized scores, some unseen categories perform well on normalized scores while some on unnormalized scores. We also consider combining the scores by taking max scores from each space denoted by MS-Zero-Max. We observe that the scores in semantic space are generally of higher magnitude than scores in visual space causing scale mismatch. Due to this scale mismatch, MS-Zero-Max is biased towards semantic space and does not fully exploit the benefits of multi-space model leading to lower results.

As described in Sec 4, we use separate semantic embeddings for MS-Zero-S and MS-Zero-V methods. We observed in our experiments that low-dimensional semantic embeddings perform better for MS-Zero-S whereas high dimensional semantic embeddings perform better for MS-Zero-V. We attribute this to the fact that since MS-Zero-S transforms visual features to semantic embedding, it is easier for the model to learn a transformation from visual features to low dimensional (20-d and 64-d) semantic embeddings as compared to high dimensional (300-d) GloVe embeddings. High dimensional GloVe embeddings makes it easier for MS-Zero-V to learn transformations of embeddings to visual space that can retain semantic properties.

6. Conclusion

ZSD is an exciting avenue and an important step towards providing semantic scalability to object detection. In this work, we propose a novel multi-space approach to solve this problem. We compare semantic and visual space scores of the model separately and show that multi-space model improves on individual scores and mitigates the hubness problem. Our multi-space model outperforms the previous best on PASCAL VOC as well as MS COCO dataset. Our extensive experimentation indicates that the proposed multi-space approach is a promising step towards solving ZSD.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *PAMI*, 2016.

[3] Y. Annadani and S. Biswas. Preserving semantic relations for zero-shot learning. In *CVPR*, 2018.

[4] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2015.

[5] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran. Zero-shot object detection. In *ECCV*, 2018.

[6] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*. 2016.

[7] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis. Zero-shot object detection by hybrid region embedding. In *BMVC*, 2018.

[8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015.

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[10] R. Felix, V. B. G. Kumar, I. Reid, and G. Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 2018.

[11] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*. 2013.

[12] Y. Fu, T. Xiang, Y.-G. Jiang, X. Xue, L. Sigal, and S. Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, 2018.

[13] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015.

[14] R. Girshick. Fast R-CNN. In *ICCV*, 2015.

[15] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[16] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[17] H. Jiang, R. Wang, S. Shan, and X. Chen. Learning class prototypes via structure alignment for zero-shot recognition. In *ECCV*, 2018.

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[19] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *PAMI*, 2014.

[20] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *CVPR*, 2016.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013.

[24] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

[25] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations*, 2014.

[26] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[27] S. Rahman, S. Khan, and N. Barnes. Polarity loss for zero-shot object detection. *arXiv preprint arXiv:1811.08982*, 2018.

[28] S. Rahman, S. H. Khan, and F. Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *ACCV*, 2018.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[30] J. Redmon and A. Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, pages 7263–7271, 2017.

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[32] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2015.

[33] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[34] Q. Wang and K. Chen. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision*, 07 2016.

[35] Z. Wang, L. Zhao, W. Xing, and D. Lu. Glstylenet: Higher quality style transfer combining global and local pyramid features. *CoRR*, abs/1811.07260, 2018.

[36] Y. Xian, Z. Akata, G. Sharma, Q. N. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[37] L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017.

[38] P. Zhu, H. Wang, T. Bolukbasi, and V. Saligrama. Zero-shot detection. *CoRR*, abs/1803.07113, 2018.