A Multi-Space Approach to Zero-Shot Object Detection

Dikshant Gupta*
IIIT Hyderabad
dikshant2210@ece.ism.ac.in

Aditya Anantharaman*
NITK Surathkal
15it201.aditya.a@nitk.edu.in

Nehal Mamgain
IIT Hyderabad
cs17mtech11023@iith.ac.in

Sowmya Kamath S.
NITK Surathkal
sowmyakamath@nitk.edu.in

Vineeth N Balasubramanian
IIT Hyderabad
vineethnb@iith.ac.in

C.V. Jawahar
IIIT Hyderabad
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 (ZSOD) - 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 ZSOD 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, as well as in a combined search space. We present promising results on two datasets, PASCAL VOC and MS COCO. We further discuss hubness and the problems associated with background class and its embeddings in ZSOD. We show that our approach alleviates hubness and its performance is superior to previously proposed methods.

Keywords: *Zero-shot Learning, Object detection, Multi-space model.*

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. Although object detection can be complex, the potential impact of successful methods for object detection has made it a research hotspot in recent years. Consequently, object detection has met with great successes over the last few years [1, 2, 3, 4, 5, 6, 7]. However, the performance of such models has been largely limited to the fully supervised domain. While humans can identify thousands of objects, current object detection models are limited to the number of classes present in the training data. This limits the use of currently available models directly for object detection in the wild. 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 should be able to identify classes which are not in the training set, but are semantically related.

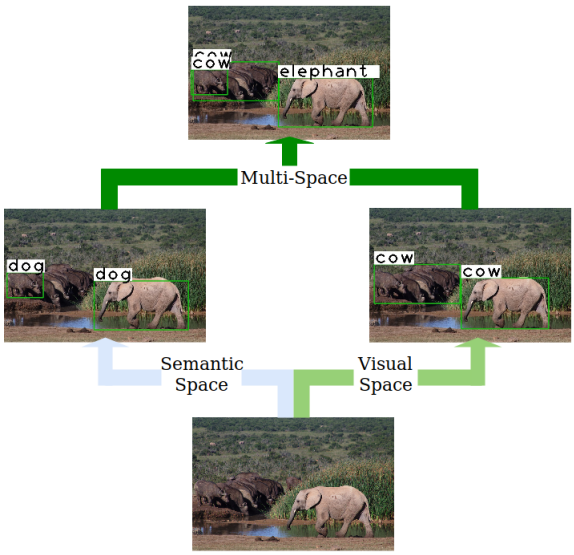

Figure 1: Object detection on a test image from MS COCO. While semantic space and visual space models individually misclassify the detections, a multi-space approach learns from both spaces and provides correct results.

For the problem of image classification, there has been a significant amount of effort to scale up the models seman-

*equal contribution

tically [8] as zero-shot recognition (ZSR). There are works that use word embeddings learned from large text corpora in an unsupervised setting as label embeddings for recognition [9, 10, 11]. Semantic attributes [12, 13] and concept ontology [14] have also been used to generate label embeddings. 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, unfortunately, scale to tasks like scene understanding where a sound reasoning of all objects in an image is required. In this paper, we propose a new solution to ZSOD, where very few efforts exist, and which can be very useful for other high-level vision tasks such as scene and context understanding.

In ZSOD, we recognize as well as 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. In the context of contemporary deep learning models for object detection [1, 2, 3], object detection pipelines 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, ZSOD has an additional and critical task of defining the background class embedding. Handling the background class in ZSOD is non-trivial as background proposals may contain unseen classes. This issue exacerbates the other issues inherited from ZSR such as hubness. In this paper, we propose a new methodology for ZSOD in the widely used Faster R-CNN framework. We define the space spanned by semantic embeddings or word vectors [15] as the *semantic space*, while the space spanned by image features of region proposals as the *visual space*. We propose a multi-space approach, where we combine the predictions of proposals from both semantic and visual spaces. A first set of class scores for each proposal is predicted on the basis of similarity between projection of visual features of proposals to semantic space and the corresponding semantic embeddings for labels; while a second set of class scores is predicted based on similarity between visual features of the proposal and projection of semantic embeddings of class labels onto the visual space. We combine them further to obtain our final prediction. We present our results on two datasets: PASCAL VOC and MS COCO. We evaluate both the sets of similarity scores individually and show that combining them yields superior results. We also adapt two common works on ZSR namely DeVise [11] and ConSE [16] to ZSOD and provide quantitative evaluation of these methods. We compare them to the multi-space approach and show superior results for both datasets.

There has been very limited work so far on ZSOD [17, 18, 19, 20], and all of them focus on similarity between embeddings in the semantic space. [21] show that nearest neighbour search on the basis of similarity in semantic space leads to *hubness*. *Hubness* occurs when some of the target classes are the nearest neighbours for most of the region proposals. In this paper, we show that using a multi-space model and combining similarity scores from both semantic and visual space alleviates hubness. In summary, our key contributions in this work are as follows:

- We propose a novel multi-space approach which leverages both visual and semantic spaces for ZSOD. 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.
- We provide a solution to the hubness problem in ZSOD by combining scores from semantic space and visual space.
- We show promising 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.

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. Experimentation and analysis of results is discussed in Section 4 and Section 5 respectively. Ablation studies are provided in Section 6. We conclude the paper and propose future directions in Section 7.

2. Related Work

Object Detection: There has been significant development in object detection over the last few years. Girshick *et al.* [2] proposed an R-CNN (Regions with Convolutional Neural Network features) that extract convolutional features for pre-extracted region proposals. Ren *et al.* [1] improved R-CNN so that it jointly learns to generate, score and classify object proposals. He *et al.* [3] proposed Mask R-CNN that simultaneously predicts object detection as well as instance segmentation and can be generalized to estimate pose as well. Dai *et al.* [7] proposed R-FCN (Regional Fully Convolutional Network), introducing position-sensitive RoI pooling that shares almost all computations for an image. Redmon *et al.* [5, 6] introduced the YOLO (You Only Look Once) framework to predict detection and classification probabilities via a single-stage evaluation. Similar to YOLO, Liu *et al.* [4] proposed the SSD (Single Shot MultiBox Detector) that performs detection and classification using a single deep neural network. In this paper we employ Faster R-CNN as our base architecture with modifications made to the region classification branch of the network to adapt it to ZSOD.

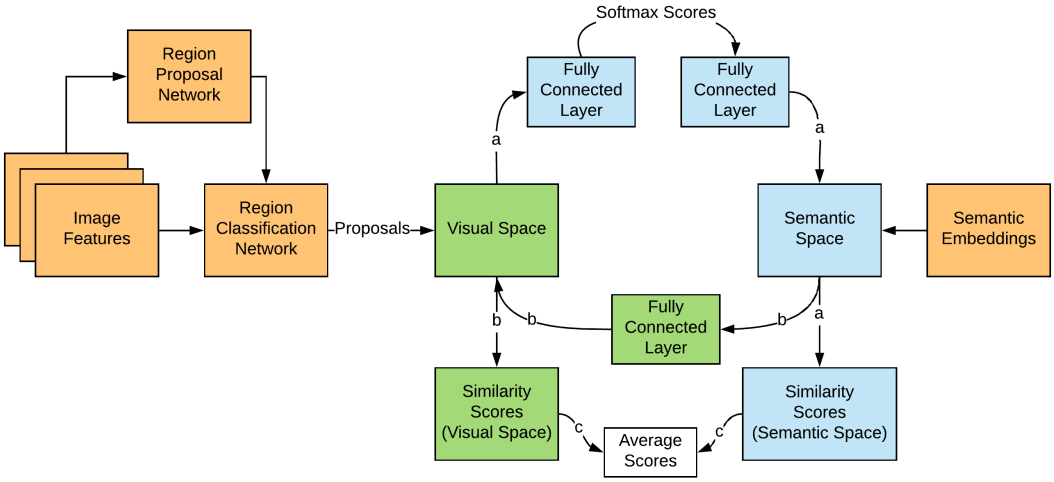

Figure 2: Architecture Diagram of the three approaches. (a) describes the proposed MS-Zero-S approach where softmax probabilities are transformed to semantic embeddings. (b) describes the proposed MS-Zero-V approach where semantic label embeddings are transformed to visual space. (a+b+c) describes the proposed MS-Zero approach which combines both semantic and visual space approaches

Word/Semantic Embeddings: Mikolov *et al.* [22] and Pennington *et al.* [15] propose word embeddings that map words to the continuous vector space. Word embeddings are trained by exploiting the co-occurrence of words in a large text corpus. They encode semantic and syntactic similarity between words. In this work, we exploit this property of word embeddings to semantically scale up object detection to target classes for which no training data is available.

Zero-Shot Recognition/Learning: ZSR can be broadly categorized into two approaches: semantic attribute-based and semantic embedding based. A semantic attribute [23] refers to the characteristics possessed by a class, for e.g., color, shape or annotations such as 'has head'. Lampert *et al.* [12, 13] proposed attribute-based classification that identifies objects based on a high-level description phrased in terms of semantic attributes. Akata *et al.* [9, 10] 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.* [24] learned a collection of maps with selection instead of learning a single bilinear map, resulting in a piece-wise linear decision boundary. Qiao *et al.* [25] introduced an l2,1-norm based objective function which can simultaneously suppress the noisy signal in the text and learn a function to match the text document and visual features. Ba *et al.* [26] used text features to predict the output weights of both the convolutional and the fully connected layers in a deep convolutional neural network. Recently, Zhang *et al.* [27] propose use of visual space as the embedding space due to nearest neighbour search suffering much less from hubness in visual space when compared to semantic space for the task of ZSR. In a similar approach, Annadani *et al.* [28] also utilize visual space as embedding space and propose an objective function to preserve semantic relations in the visual space.

Hubness: Radovanovic *et. al.* [29] explored hubness as a curse of dimensionality. They investigated the origins of hubness and show that it is inherent to high dimensional data. They also discuss the interaction between hubness and dimensionality reduction. While [29] discussed hubness in general, [30] and [21] studied hubness in the ZSR setting. [30] investigated cross-space mapping properties and proposed the use of max-margin loss to mitigate hubness, while [21] argued that using semantic space as nearest neighbour search space increases hubness and proposed the use of visual space as the nearest neighbour search space. In this paper, we consider both semantic and visual space as our search space in a multi-space approach, as well as extend the analysis to the ZSOD setting, which has not been done before.

Zero-Shot Object Detection: Very few efforts have been proposed on ZSOD so far. Demirel *et al.* [18] proposed a hybrid of convex combination of class embeddings [16] and label embedding-based classification. While the former is a semantic composition method guided by class detection scores, the latter focuses on transforming image features to embedding space. Rahman *et al.* [19] proposed an extension of Faster R-CNN [1] and ConSE [16] with a loss formulation that combines max-margin (separates classes) and semantic clustering (reduces noise in semantic vectors) losses. We use this method for comparison in our experiments. Zhu *et al.* [17] proposed a zero-shot detection framework that fuses semantic attribute prediction with vi-

sual features to predict objectness scores for bounding box proposals. Bansal *et al.* [20] and [17] addressed the problem of confusing background with unseen classes. [20] also proposed using a large open vocabulary for differentiating background regions. In contrast to previous works on ZSOD, we propose a multi-space approach which utilizes both semantic and visual spaces. We show that a multi-space approach outperforms each of the approaches individually on both PASCAL VOC [31] and MS COCO [32] datasets.

3. Approach

In this section we describe the problem statement and our multi-space ZSOD approach. In this appraoch we combine the similarity scores obtained using semantic space and visual space individually as nearest neighbour search space. We further explore the problem of hubness and show how our proposed approach helps alleviate it.

3.1. Problem Definition

In Zero-Shot Object Detection (ZSOD), we aim at recognition and localization of objects that were previously unseen during training. In this section, we define the problem statement and describe formal notations. We denote 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, given $\mathcal{S} \cap \mathcal{U} = \emptyset$. Each image is denoted as $I \in \mathbb{R}^{M \times N \times 3}$ with corresponding bounding boxes and labels denoted as $b_i \in \mathbb{N}^4$ and $y_i \in \mathcal{S}$ respectively. Our aim is to find all bounding boxes corresponding to unseen classes for each image in the test set. We divide testing in three configurations, testing only on unseen classes ($\mathcal{U}$), testing only on seen classes ($\mathcal{S}$) and testing on both seen and unseen classes which is the generalized zero-shot setting.

3.2. Multi-Space Network (MS-Zero)

We employ Faster R-CNN using ResNet-101 [33] as the backbone architecture. 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 in the number of seen classes. Space spanned by semantic embeddings of class labels is the semantic space and is denoted by $\mathcal{M}_{sem} \in \mathbb{R}^{d_3}$, while the space spanned by visual features p_i is denoted by $\mathcal{M}_{vis} \in \mathbb{R}^{d_1}$, where d_1 and d_3 are the dimension of visual space and semantic space respectively. We generate similarity scores for each proposal with class labels in both semantic as well as visual space. Similarity scores in semantic space are given by,

$$S_i^{sem} = \phi(l^{gt}, W_{sem} c_i) \tag{1}$$

where, ϕ gives cosine similarity, $l^{sem} \in \mathcal{M}_{sem}$ is the class embeddings in semantic space and $W_{sem} \in \mathbb{R}^{d_3 \times d_2}$ is the transformation matrix that transforms softmax scores to semantic space. Softmax layer predicts scores only for seen classes. Classification loss and bounding box regression loss for region proposal as well as region classification network are used as described in [1] 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,

$$\mathcal{L}_{mse}(c_i, y_i^{gt}) = \frac{1}{d_3} \sum_{i=1}^{d_3} (W_{sem} c_i - l_i^{sem})^2 \tag{2}$$

where l_i^{sem} is ground truth class embedding and y_i^{gt} is the ground truth label corresponding to the proposal.

Motivated by [28, 27, 21] we obtain similarity scores for each proposal in visual space. Scores in visual space are given by,

$$S_i^{vis} = \phi(W_{vis} l^{sem}, p_i) \tag{3}$$

where, $W_{vis} \in \mathbb{R}^{d_1 \times d_3}$ is the transformation matrix that transforms semantic class embeddings l^{sem} to visual space. Additional to the loss in semantic space, to learn transformation matrix we introduce max-margin loss given by,

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

where, m refers to the margin. 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 ZSOD background proposals may contain objects that belong to unseen classes, defining a semantic embedding for background class is non-trivial. Similar to [19], in this paper we consider background class embedding in semantic space to be the mean of all semantic class embeddings,

$$l_{bg}^{sem} = \frac{1}{\mathcal{C}} \sum_{j=1}^{\mathcal{C}} l_j^{sem} \tag{5}$$

where, l_{bg}^{sem} represents background class embedding in semantic space. Similarity scores for multi-space model are given by,

$$S_i^{mlt} = \frac{1}{2}(S_i^{sem} + S_i^{vis}) \tag{6}$$

To train the multi-space model we combine all the losses,

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

where, θ represents combined parameters of network and transformation matrices. Figure 2 describes the multi-space model. We also evaluate the visual and semantic parts of the multi-space model as separate models. Arrows marked by 'a' in the diagram describe the semantic space model (MS-Zero-S), while arrows marked by 'b' describe the visual space model (MS-Zero-V). The combination of 'a', 'b' and 'c' together denote the multi space model (MS-Zero).

Model	Seen	Unseen	Mix
DeVise-ZSOD	59.22	50.63	46.57
ConSE-ZSOD	73.19	55.30	56.61
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**
ZSOD-YOLO [18]	65.60	54.20	52.33

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

Model	car	dog	sofa	train
DeVise-ZSOD	44.22	81.51	48.73	28.07
ConSE-ZSOD	62.24	83.70	58.16	17.08
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
ZSOD-YOLO [18]	55.00	82.00	55.00	26.00

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.

3.3. Extending ZSR methods to ZSOD

We adapt two prior works on zero-shot recognition, namely DeVise [11] and ConSE [16] to ZSOD and use them as our baselines. For DeVise-ZSOD we transform the visual features of each proposal to semantic space and compute the similarity score with semantic embeddings of each class. Scores are penalized using max-margin loss as described in section 3.2. For ConSE-ZSOD we train the model on seen classes in a fully supervised setting. We obtain 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. During testing we use the embedding obtained to predict similarity scores for unseen classes.

3.4. Hubness

Hubness phenomenon is associated with nearest neighbour search and is observed when a few objects in the dataset occur as the nearest neighbour of many objects, therefore becoming universal neighbours or *hubs*. We argue that in ZSOD since, background semantic class embedding is average of all semantic class embeddings, background class by virtue of its definition tends to become a hub and is often assigned to bounding boxes with unseen classes. In contrast, our approach alleviates hubness which we also demonstrate later in our results. Mathematically, we define hubness for each class as,

$$H_k(y) = |\{x, x \in P | y \in NN_k(x, \mathbf{S})\}|$$

where, x denotes features of a region proposal, P denotes the set of all region proposals, $\mathbf{S}$ is the search space and $NN_k(x, \mathbf{S})$ denotes the k nearest neighbour of x in $\mathbf{S}$.

4. Experiments

In this section, we first describe the different datasets that were used to evaluate the proposed approach. We further describe the method used for creating the training and testing splits. Next, we describe the implementation details for different models discussed.

4.1. Datasets

To evaluate the proposed approach we employ two datasets: PASCAL VOC and MS COCO. Since both the datasets have been widely accepted by the research community as standard benchmarks for object detection, using them to evaluate ZSOD is reasonable. It should be noted that, similar to [18], we ignore images that contain both seen and unseen classes in both datasets.

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. 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 [18]. Test-Seen considers images which contain only seen classes, Test-Unseen considers images which contain only unseen objects and Test-Mix is the combination of both seen and unseen classes. Major difference between all three configurations is in the nearest neighbour 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. It is a more challenging dataset as compared to PASCAL VOC as it offers higher number of objects per image, occlusion, clutter, views, etc. and is more similar to real life scenarios. We follow a strategy similar to [20] for creating seen/unseen class splits. We only consider classes that have a synset associated with them in the wordnet 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. We use the 2014 *train* set for training and validation while the 2014 *val* set for testing.

Model	*airplane*	*bus*	*cat*	*dog*	*cow*	*elephant*	*umbrella*	*tie*	*snowboard*	*skateboard*	*cup*	*knife*	*cake*	*couch*	*keyboard*	*sink*	*scissors*
DeVise-ZSOD	**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-ZSOD	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

Table 3: Table shows the class wise Average Precision (AP) (%) for the unseen classes for all models in test-unseen setting on MS COCO dataset.

Model	Seen	Unseen	Mix
DeVise-ZSOD	30.3	10.6	21.5
ConSE-ZSOD	42.4	9.3	30.5
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

Table 4: Table shows the mean average precision (mAP) (%) of all models for MS COCO dataset on three different test settings.

Num of common neighbours	PASCAL VOC	MS COCO
> 1	95.0	96.92
> 2	80.0	87.69
> 3	30.0	36.92
> 4	0.0	1.53

Table 5: Table presents the percentage of classes having a certain number of common neighbours in both semantic and visual space.

4.2. Implementation Details

As in Faster R-CNN, we use a shallow CNN on top of image features to generate region proposals (figure 2). Image features are extracted from raw image 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 and assigned to the class of the ground-truth box while any proposal with $0 < IoU < 0.2$ with a ground truth box is considered as a background box. We only use boxes belonging to seen classes for training. We use stochastic gradient descent with an initial learning rate of 10^{-3} and momentum 0.9. In case of MS COCO, we train the model for a total of 20 epochs with a learning rate of 10^{-3} for the first ten epochs and 10^{-4} for the remaining ten epochs. 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. For MS-Zero-S, we transform the softmax scores for each proposal to semantic embeddings. The weights for the transformation matrix are initialized with pascal semantic attribute embeddings [34] of seen classes in case of PASCAL VOC dataset and with reduced Word2vec features released by [17] in case of MS COCO. The model is trained end-to-end and results saturate after 6 epochs of training for PASCAL VOC and 10 epochs 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. Results saturate after 15 epochs of training. MS-Zero-S and ConSE-ZSOD utilize 64 dimensional semantic attribute embeddings [34] for PASCAL VOC dataset and 20 dimensional semantic embeddings [17] for MS COCO dataset. Label embeddings for MS-Zero-V and DeVise-ZSOD method for both the datasets are extracted from publicly available GloVe embeddings. We normalize the semantic scores and visual scores using $L2$ norm before combining them for PASCAL VOC dataset in MS-Zero model.

5. Results and Analysis

We use mean average precision (mAP) with an IoU threshold of 0.5 with ground truth boxes as the evaluation metric for the model. We apply greedy non-maximal suppression on all the scored boxes for each test class independently and reject boxes that have an IoU greater than 0.3 with a higher scoring box. Tables 1 and 4 provide quantitative results in terms of mAP for 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. MS-Zero performs best on PASCAL VOC for both test-unseen and test-mix setting. Since we use the same test split and images as [18] in case of PASCAL VOC dataset, we compare our models with their hybrid region embedding model which is a modification of YOLO referred as ZSOD-YOLO in further discussion and table 1 and 2. To our knowledge, ZSOD-YOLO is the best work so far in terms of mAP

Figure 3: Qualitative detection results on test images from PASCAL VOC 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.

on PASCAL VOC. MS-Zero outperforms ZSOD-YOLO by a margin of 14% in test-unseen setting and achieves significant improvements in individual AP for all unseen classes with atleast 10 AP rise for two of the classes.

MS-Zero also performs best on MS COCO in test-unseen setting. MS-Zero-S outperforms MS-Zero in test-seen and test-unseen setting but with a quite small margin. We also observe superior performance by MS-Zero and its variants on 11 out of 17 classes of MS COCO when compared to adaptations of ZSR methods DeVise-ZSOD and ConSE-ZSOD. Extremely low AP for some of the classes in MS COCO dataset is due to a lack of semantically similar classes in the seen set and very small size of objects. The classes 'tie' and 'umbrella' belong to the broad super category 'accessory', therefore almost no semantically similar classes are available in the training set.

We also find 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. A 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 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. Table 5 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. Thus, a higher number of unseen classes in MS COCO helps retain semantic relationships in visual space.

Figure 4 shows KL divergence between region proposal distribution and ground truth bounding box distribution over unseen classes in PASCAL VOC. As described in 3.4, we compute the proportion of region proposals assigned to each

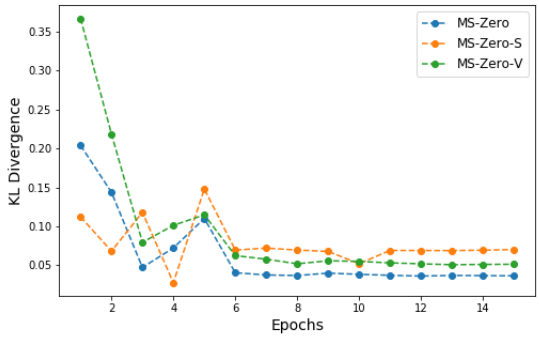

Figure 4: Plot of KL divergence between proposal distribution and ground truth distribution over unseen classes and number of epochs for MS-Zero, MS-Zero-S and MS-Zero-V methods on PASCAL VOC dataset.

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

Table 6: 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.

class. We compare it with the proportion of each class in ground truth and argue that if the proportions are same for ground truth and proposed method, hubness is alleviated. It is evident from the KL divergence plot for PASCAL VOC that as the number of epochs increases, divergence decreases and therefore, hubness decreases. Figure 3 provides qualitative results for PASCAL VOC. MS-Zero performs the best of the three models.

6. Ablation Studies

In this section, we discuss variations of inference strategy and different methods to combine scores in individual spaces. We also explain the reasoning behind different choice of label embeddings for semantic and visual parts of the multi-space model. We combine scores in both spaces using two approaches: in the first approach, we take the average of the scores denoted by MS-Zero-U and in the second approach, we normalize scores using $L2$ norm before taking average denoted by MS-Zero-N. It is evident from the results in table 6 that MS-Zero-N performs the best on both datasets in test-seen and test-mix setting. While MS-Zero-N gives the best results in test-unseen setting for PASCAL VOC, MS-Zero-U outperforms its normalized counter part by a margin of 15% for MS COCO. 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 scores in semantic space are generally of higher magnitude than scores in visual space causing scale mismatch. Due to the mismatch of scale 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 section 4.2 for the proposed multi-space model, we use separate semantic embeddings for MS-Zero-V and MS-Zero-S methods. We observe 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 observation 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.

7. Conclusion and Future Directions

ZSOD is an exciting avenue and important step towards providing semantic scalability to object detection frameworks and reducing their dependence on annotations. In this work, we propose a novel multi-space approach to solve this problem. We compare both 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 dataset and gives promising results on MS COCO dataset as well. Given the huge relevance of ZSOD, our extensive experimentation indicates that the proposed multi-space approach has the potential to solve ZSOD and is a promising step for ZSOD.

Future directions towards solving ZSOD can be: (1) Multi-space approach can be explored for other object detection pipelines like YOLO and SSD. (2) Incorporating semantic information in the region proposal network is a good problem to investigate. (3) One can also consider extending multi-space approach to open vocabulary detection.

References

[1] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 2, 3, 4

[2] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 1, 2

[3] K. He, G. Gkioxari, P. Dollr, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2

[4] 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. 1, 2

[5] J. Redmon and A. Farhadi. YOLO9000: better, faster, stronger. *CoRR*, abs/1612.08242, 2016. 1, 2

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1, 2

[7] J. Dai, Y. Li, K. He, and J. Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*. 2016. 1, 2

[8] Y. Fu, T. Xiang, Y. Jiang, X. Xue, L. Sigal, and S. Gong. Recent advances in zero-shot recognition. *CoRR*, abs/1710.04837, 2017. 2

[9] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 2, 3

[10] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for image classification. *PAMI*, 2016. 2, 3

[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. 2, 3, 5

[12] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *PAMI*, 2014. 2, 3

[13] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2, 3

[14] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 2

[15] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 2, 3

[16] 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. *CoRR*, abs/1312.5650, 2013. 2, 3, 5

[17] P. Zhu, H. Wang, T. Bolukbasi, and V. Saligrama. Zero-shot detection. *CoRR*, abs/1803.07113, 2018. 2, 3, 4, 6

[18] B. Demirel, R. G. Cinbis, and N. Ikizler-Cinbis. Zero-shot object detection by hybrid region embedding. *CoRR*, abs/1805.06157, 2018. 2, 3, 5, 6

[19] S. Rahman, S. H. Khan, and F. Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. *CoRR*, abs/1803.06049, 2018. 2, 3, 4

[20] A. Bansal, K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran. Zero-shot object detection. In *ECCV*, 2018. 2, 4, 5

[21] 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. 2, 3, 4

[22] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013. 3

[23] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 3

[24] Y. Xian, Z. Akata, G. Sharma, Q. N. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. *CVPR*, 2016. 3

[25] R. Qiao, L. Liu, C. Shen, and A. van den Hengel. Less is more: Zero-shot learning from online textual documents with noise suppression. *CVPR*, 2016. 3

[26] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. *ICCV*, 2015. 3

[27] L. Zhang, T. Xiang, and S. Gong. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2017. 3, 4

[28] Y. Annadani and S. Biswas. Preserving semantic relations for zero-shot learning. In *CVPR*, 2018. 3, 4

[29] M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 2010. 3

[30] A. Lazaridou, G. Dinu, and M. Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015. 3

[31] 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. 4

[32] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 4

[33] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4

[34] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 6