Supplementary Material

In this section, we provide additional results and analysis that could not be included in the main paper owing to space constraints. In particular, we discuss the impact of varying the loss formulation described in Section 3; we consider a different set of unseen classes and show that our results are consistent with observations obtained in Section 5; we also present qualitative results for the MS COCO dataset, including failure cases.

7. Additional Results and Analysis

7.1. Qualitative Results for MS COCO

Figure 6 presents detection results for classes that generally perform well on the AP metric. We can observe from the detections that shortcomings of the individual search space methods MS-Zero-S and MS-Zero-V are rectified by MS-Zero. In Figure 7, we present qualitative results for classes that perform poorly on the AP metric on the new split of COCO. From the qualitative results, we can infer that the model could not recognize some of the unseen objects in the failure cases namely, multiple oranges, laptop, surfboard, clock and bear. A few cases of misclassification are also observed. Misclassification of surfboard to skis and bear to horse in images 7(d) and 7(g) respectively is due to very close occurrence of surfboard to skis and horse to bear in the nearest neighbour search space. A more prominent phenomenon in ZSD is discarding unseen objects as background. We observe that the model misses the detection of unseen class for images 7(a), 7(b), 7(c), 7(e) and 7(f). This can be attributed to the model classifying the unseen objects as background. Since Region Proposal Network is trained using a purely supervised strategy, it suppresses certain unseen class objects in its proposals and considers them to be background. A solution to this problem would be to include semantic information in the region proposal network to avoid misclassification of region proposals containing unseen objects as background class.

7.2. Varying Unseen Classes

We considered a fixed set of unseen classes for both datasets in Section 4 which comprised of the same unseen classes used in earlier work for fairness of comparison. In this section, we present results for a different set of unseen classes for both datasets. We note that earlier work didn't explicitly consider the quality of results on using different splits, hence we don't include their methods for comparison here.

Tables 12 and 13 present the results for the new split of the PASCAL VOC dataset (with table 13 stating the unseen classes in the new split). Tables 7 and 11 present the results for the new split of the MS COCO dataset (with table 7 stating the unseen classes in the new split). Similar to the previous split of the dataset, the MS-Zero and MS-Zero++ models outperform the MS-Zero-S and MS-Zero-V models in all three test settings. Also, MS-Zero++ performs better than MS-Zero in the test-unseen setting whereas MS-Zero performs better than MS-Zero++ in the test-seen and test-mix settings on both PASCAL VOC and MS COCO datasets. This can be attributed to the fact that MS-Zero++, with the help of the correlation loss, learns visual representations which retain semantic properties that helps it to perform well in the unseen setting. But, because of bad separation between seen classes in the semantic space, both MS-Zero++ and MS-Zero-S do not perform well in the test-seen setting when compared to MS-Zero. MS-Zero does not use the correlation loss and hence learns visual representations similar to MS-Zero-V which are relatively less affected by the bad separation between seen classes in semantic space. The performance in test-mix setting is biased towards the test-seen setting.

7.3. Max-Margin Loss

We now discuss the impact of varying the margin in max-margin loss formulation described in Equation 4. We train the visual space branch of MS-Zero on two datasets: MS COCO and PASCAL VOC using a max-margin loss. Max-margin loss enforces a constraint on cosine similarity scores between predicted embedding and ground truth embedding to be at least the value specified by margin m. We consider three values for margin 0.1, 0.7 and 1.0 corresponding to approximately 90, 45 and 0 angle between the unit vectors of the predicted and ground truth embeddings. Tables 8 and 10 present results on PASCAL VOC dataset and tables 9 and 6 present results on MS COCO dataset for different values of margin. It is evident from the results that a margin value of 0.1 performs the best for unseen classes. This can be attributed to the fact that while a margin value of 0.1 enforces a minimum similarity constraint, it also allows the model flexibility over a large range of similarity scores as compared to other margin values. Hard constraints of margin value 0.7 and 1.0 lead to high loss values and model divergence. Results for margin value of 1.0 for MS COCO are not provided in the tables due to model divergence with margin 1.0 leading to very low scores. Thus, we conclude that a margin value of 0.1 is the most appropriate choice. We hence use 0.1 as the margin value for both MS-Zero and MS-Zero++.

7.4. Weighting individual loss terms

In this section, we study the relative importance of the loss terms in Equation 7. We assign a weight, λ, to the following loss terms: softmax loss ($\mathcal{L}_{cls}$ as described in [31] but used in our case only for seen classes), embedding loss ($\mathcal{L}_{sem}$ in Equation 2), max margin loss ($\mathcal{L}_{vis}$ in Equation 4) and correlation loss ($\mathcal{L}_{corr}$ in Equation 5). Similar to cyclegan, λ was set to 10 for all experiments except for weighting

Margin	airplane	bus	cat	dog	cow	elephant	umbrella	tie	snowboard	skateboard	cup	knife	cake	couch	keyboard	sink	scissors
0.1	**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**
0.7	5.3	37.1	**16.7**	5.1	20.1	1.0	0	0	**12.7**	0.9	5.4	2.8	12.7	39.3	13.0	0	4.5

Table 6. Table shows the class-wise Average Precision (AP) (%) for the unseen classes in test-unseen setting on MS COCO dataset for different values of margin.

Model	bicycle	train	horse	sheep	bear	backpack	surfboard	skis	bowl	spoon	sandwich	orange	bed	bench	laptop	refrigerator	clock
MS-Zero-S	11.6	57.6	1.0	21.7	0	**10.5**	1.0	4.8	2.8	0	0	16.1	3.2	44.9	**30.7**	1.8	0
MS-Zero-V	24.1	53.3	4.6	17.2	9.8	1.0	30.7	**39.6**	0	18.6	7.9	6.9	1.0	44.3	2.6	5.0	0
MS-Zero	**27.8**	**61.3**	**9.1**	**28.3**	4.7	2.0	**31.7**	33.7	0	18.6	8.0	6.9	1.0	**54.3**	5.0	7.1	0
MS-Zero++	8.7	58.9	**19.0**	23.0	**42.8**	**24.2**	0	0	**7.1**	17.7	**27.3**	**16.2**	**55.3**	9.9	6.7	**17.5**	**0.7**

Table 7. Table shows the class wise Average Precision (AP) (%) for the unseen classes for all models in test-unseen setting on MS COCO dataset for the new split.

Margin	Seen	Unseen	Mix
0.1	74.49	**62.15**	**60.05**
0.7	**76.77**	44.89	59.07
1.0	76.37	45.64	58.78

Table 8. Table shows the mean average precision (mAP) (%) for PASCAL VOC dataset in three different test settings for different values of margin.

Margin	Seen	Unseen	Mix
0.1	**42.4**	**12.9**	**30.7**
0.7	41.9	10.4	30.1

Table 9. Table shows the mean average precision (mAP) (%) for MS COCO dataset in three different test settings for different values of margin.

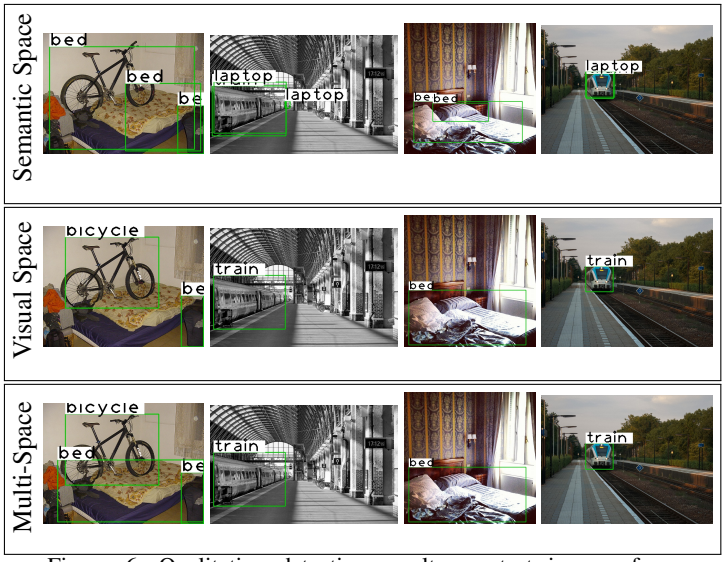

Figure 6. Qualitative detection results on test images from MS COCO dataset. The topmost row shows detections using MS-Zero-S. The middle row shows detections using MS-Zero-V. The bottom row shows detections using MS-Zero.

$\mathcal{L}_{cls}$: for which λ was set to 0.1. Further, this choice of λ was made as L_{sem}, L_{vis} and L_{corr} are of significantly lower order than $\mathcal{L}_{cls}$. On weighting $\mathcal{L}_{cls}$ with $\lambda = 0.1$, we obtain a marginally better mAP of 57.95 than MS-Zero++ in test-unseen setting on PASCAL VOC. This is because of reducing the order of magnitude for $\mathcal{L}_{cls}$ to the same order as L_{sem}, L_{vis} and L_{corr} which allows better contribution of each loss term. On weighting $\mathcal{L}_{sem}$ and L_{vis} with $\lambda = 10$, we get an improved mAP of 58.46 and 58.85 respectively as due to an increase in loss values, a higher gradient is obtained which improves the training of MS-Zero++. On weighting the correlation loss with $\lambda = 10$, we observe an even higher mAP of 61.49 as it enforces correlation amongst classes to be the same in semantic and visual spaces thus enabling the model to learn good representations in both spaces. We also note that weighting these losses further improves upon the mAP obtained by the unweighted loss objective for MS-Zero++ on the PASCAL VOC dataset in test-unseen setting. We cannot compare these results to MS-Zero results in table 3 as MS-Zero uses GloVe embeddings [26] for the visual component of the model and semantic attribute embeddings [9] for the semantic component of the model whereas MS-Zero++ uses only semantic attribute embeddings [9] on PASCAL VOC.

Figure 7. Images showing some failure cases and potential areas of improvement for MS-Zero as described in section 7.1. Incorporating semantic information in Region Proposal Network is a possible step towards improving these detections.

Margin	car	dog	sofa	train
0.1	**69.00**	**86.80**	**65.99**	**26.81**
0.7	21.45	85.66	61.49	10.96
1.0	21.89	86.6	60.21	13.87

Table 10. Table shows the class-wise average precision (AP) (%) for the unseen classes in test-unseen setting on PASCAL VOC dataset for different values of margin.

Model	Seen	Unseen	Mix
MS-Zero-S	44.3	12.2	31.9
MS-Zero-V	44.8	15.7	32.4
MS-Zero	**46.0**	17.6	**33.1**
MS-Zero++	39.6	**19.7**	30.1

Table 11. Table shows the mean average precision (mAP) (%) of all models for MS COCO dataset in three different test settings for the new split.

Model	Seen	Unseen	Mix
MS-Zero-S	70.39	48.14	52.74
MS-Zero-V	79.46	47.71	63.15
MS-Zero	**80.49**	49.18	**63.54**
MS-Zero++	76.94	**54.04**	54.2

Table 12. Table shows the mean average precision (mAP) (%) of all models for PASCAL VOC dataset in three different test settings for the new split.

Model	aeroplane	chair	motorbike	sheep
MS-Zero-S	21.95	20.61	73.47	76.53
MS-Zero-V	25.4	17.86	77.00	70.60
MS-Zero	24.98	22.85	77.95	70.90
MS-Zero++	**31.5**	**23.28**	**83.01**	**78.38**

Table 13. Table shows the class wise average precision (AP) (%) of all models for PASCAL VOC dataset in test-unseen setting for the new split.