Supplementary Material for Unified Framework for Open-World Compositional Zero-shot Learning

Hirunima Jayasekara Khoi Pham Nirat Saini Abhinav Shrivastava

University of Maryland
College Park

1. Object Attribute Disentanglement

Compositional learning is characterized by the model ability to decompose and compose object primitives and states. In which Object Attribute Disentanglement (OAD) is a vital component to facilitate generalization on unseen pairs. VisProd [6], KG-SP [3] separate attribute and object embedders and TMN [9] utilize word embedding to decompose image features while, HiDC [12] compose novel pairs using word embeddings in order to facilitate disentanglement. Saini *et al.* [10] deploys a visual feature disentanglement of attributes and objects to regularize the common embedding space. Proposed method contains a novel hybrid disentanglement procedure encompassing [10] and [3].

2. Feasibility Score Calculation for Open-World Setting

Feasibility score is determined by assessing the coherence of attribute-object compositions within real-world contexts. For example, while 'small cat' exhibits coherence, 'spilled cat' lacks semantic coherence. This aims to ascertain the viability of a given attribute-object pair within real-world contexts. In order to compute the feasibility scores for each composition, we deploy a similar procedure as KG-SP [3]. First, we aggregate text embedding for each word in the vocabulary by utilizing GloVe [8]. Subsequently, we find similarity between given object and available objects and computing similarity scores for attributes as well. Lastly, for each composition, a feasibility score, $\rho_{Glove}(a, o) \in \mathbb{R}$ is calculated by taking the average of similarity scores between respective attribute and object.

$$\rho_o(a, o) = \max_{\hat{o} \in \mathbb{O}} \frac{\phi(o) \cdot \phi(\hat{o})}{||\phi(o)|| \cdot ||\phi(\hat{o})||} \quad (1)$$

$$\rho_{Glove}(a, o) = \frac{\rho_o(a, o) + \rho_a(a, o)}{2} \quad (2)$$

Where, $\rho_o(a, o)$ is the maximum similarity score between object and other objects in $\mathbb{O}$. $\phi(\cdot)$ is the GloVe embedding function. In order to induce more robustness to the filtering, we compute ConceptNet numberbatch [11] embedding based feasibility scores similar to that of Glove. For each pair, maximum of two feasibility scores set as the final feasibility score $\rho(a, o) \in \mathbb{R}$.

$$\rho(a, o) = \max_{o \in \mathbb{O}, a \in \mathbb{A}} (\rho_{Glove}(a, o), \rho_{Conceptnet}(a, o)) \quad (3)$$

$$f_{pair} = \underset{y, \rho(a,o) > T}{\arg\max}\ p_\theta(y|x) \quad (4)$$

Where, $\rho(a, o)$ is the feasibility score for composition (a, o). Finally, we filter out the infeasible compositions by choosing the compositions with higher scores than a empirically settled threshold value which is calibrated on training set. Resulting a binary mask $f_{pair} \in \mathbb{R}^{|A| \cdot |O|}$.

2.1. Feasibility Results

We examine the least and most feasible attribute-object combinations computed within the MIT-States validation set. Subsequently, we compare the associated labels and corresponding predictions. Column one displays corresponding image, with the ground-truth label positioned at the top (GT), while columns two and three shows attribute and object predictions. Furthermore, fourth column consists of incorrect predictions. Last column indicate the final pair prediction. For each cell, three columns illustrate the top-3 results for attributes, objects, and compositions. We denote the predictions matching the ground truth are highlighted in blue. From Table 1 it is evident that, compositions with lower feasibility scores are susceptible to masking out thereby introducing an induced bias within the network and consequently resulting in incorrect predictions. In contrast, Table 2 illustrates the model's capability to accurately identify labels for compositions with higher feasibility scores. Consequently, this highlights the model's ability to discern such compositions without the risk of masking out. Therefore, above results prompt us to explore novel methodologies to compute more robust and scalable feasibility scores.

Table 1. Qualitative analysis on effect of feasibility mask. Illustrate the five lowest feasibility scores in validation set. Predictions matching the ground truth are highlighted in blue.

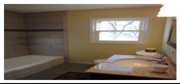
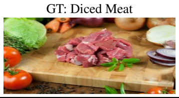
	Feasibility Score for GT	Attribute Predictions	Object Predictions	Pair Prediction Before Binary Mask	Pair Predictions
GT: Dull Bronze 	0.	Brushed Straight Rusty	Bronze Steel Brass	Brushed Bronze Brushed Steel Straight Bronze	Brushed Bronze Brushed Steel Straight Bronze
GT: Full Bathroom 	0.0842	Large Empty Tiny	Bathroom Room Shower	Large Bathroom Empty Bathroom Tiny Bathroom	Large Bathroom Empty Bathroom Tiny Bathroom
GT: Blunt Blade 	0.1257	Large Small Straight	Knife Blade Handle	Large Knife Small Knife Straight Knife	Large Knife Small Knife Straight Knife
GT: Standing Tower 	0.1363	Modern Standing New	Tower Building Church	Standing Tower Modern Tower Ancient Tower	Modern Tower Ancient Tower New Tower
GT: Fallen Tower 	0.1479	Steaming Dry Barren	Lake Mud Farm	Steaming Lake Steaming Water Dry Lake	Steaming Lake Steaming Water Dry Lake

Table 2. Qualitative analysis on effect of feasibility mask. Illustrate the five highest feasibility scores in validation set. Predictions matching the ground truth are highlighted in blue.

	Feasibility Score for GT	Attribute Predictions	Object Predictions	Pair Prediction Before Binary Mask	Pair Predictions
GT: Small Bathroom 	0.9968	Small Tiny Clean	Bathroom Shower Tile	Small Bathroom Tiny Bathroom Clean Bathroom	Small Bathroom Tiny Bathroom Clean Bathroom
GT: Small Kitchen 	0.9968	Small Tiny Large	Kitchen Cabinet Room	Small Kitchen Tiny Kitchen Large Kitchen	Small Kitchen Tiny Kitchen Large Kitchen
GT: Diced Meat 	1.	Diced Sliced Raw	Meat Beef Chicken	Diced Meat Diced Beef Sliced Meat	Diced Meat Diced Beef Sliced Meat
GT: Frozen Beef 	1.	Frozen Thawed Raw	Beef Meat Chicken	Frozen Beef Frozen Meat Thawed Beef	Frozen Beef Frozen Meat Thawed Beef
GT: Sliced Beef 	1.	Sliced Chipped Diced	Beef Plant Leaf	Sliced Beef Chipped Beef Cooked Beef	Sliced Beef Chipped Beef Cooked Beef

Such that it would encompass a right-skewed distribution to accurately represent viable compositions.

3. Hyper Parameter Tuning

For experiments, we found the best hyperparameters by random search and choose the best hyperparameters for each dataset based on best AUC on the validation split. We reduce the number of epochs for C-GQA [5] and MIT-States [2] since the model tend to converge earlier due the low number of training samples. We increase the number of epochs for VAW-CZSL [10] to compensate for high number of training samples. Table 6 shows hyperparameters used to train the proposed model on all datasets.

Table 5. The effect of number of frozen layers in transformer encoder for MIT-States.

Number of frozen layers	S	U	HM	AUC
0 frozen layers	36.3	12.5	12.4	3.1
2 frozen layers	35.3	12.4	12.3	3.0
4 frozen layers	35.0	12.5	12.2	2.9
6 frozen layers	35.4	12.0	11.8	2.9
8 frozen layers	35.0	12.3	12.1	2.9
11 frozen layers	31.6	11.8	11.0	2.4

Furthermore, we experiment the effect of number of frozen layers in transformer encoder. As shown in Table 5,

Table 6. Hyperparameter tuning on MIT-States, C-GQA and VAW-CZSL

	MIT-States [2]	C-GQA [5]	VAW-CZSL [10]
LR for TopK Selection	1e-6	1e-6	1e-6
LR for Transformer Encoder	3.5e-5	3.5e-5	3.5e-5
LR for Sparse Linear Compositor	3.6e-5	3.6e-5	3.5e-5
Weight Decay	0.001	0.001	0.001
K for TopK Selection	3	3	3
Batch Size	32	64	64
Epochs	20	30	85
GPU	1080Ti	RTXA4000	RTXA4000

Table 7. Closed world performance on MIT-States, C-GQA and VAW-CZSL. As evaluation matrices we refer to AUC with seen and unseen accuracies and HM. X_{vit} denots the networks with transformer based image encoders.

Method	MIT-States				C-GQA				VAW-CZSL			
	S	U	HM	AUC@1	S	U	HM	AUC@1	S	U	HM	AUC@3
CompCos [5]	26.9	24.5	16.9	4.8	28.1	11.8	12.1	2.6	23.9	18.0	14.2	3.2
CGE [7]	28.9	25.0	18.1	5.3	27.5	11.7	11.9	2.5	23.4	16.8	13.0	2.9
OADIs [10]	31.1	25.6	18.9	5.9	-	-	-	-	24.9	18.7	15.2	3.6
CoT [4]	30.8	26.8	19.6	6.2	33.1	16.6	16.6	4.5	24.6	19.1	15.7	3.8
CGE$_{vit}$ [7]	39.7	31.6	24.8	9.7	38.0	17.1	18.5	5.4	30.1	25.7	20.1	6.2
OADIs$_{vit}$ [10]	39.2	32.1	25.2	10.1	38.3	19.8	20.1	7.0	31.3	26.1	20.4	6.5
CoT$_{vit}$ [4]	39.5	33.0	25.8	10.5	39.2	22.7	22.1	7.4	32.9	28.2	21.7	7.2
Ours$_{vit}$	36.5	30.9	22.1	8.2	36.1	18.7	19.3	5.8	30.0	26.4	20.2	6.2

we can observe a decrement in all four evaluation matrices when the number of frozen layers increases. This may be attributed to the inherent limitation of the ViT [1] encoder, which was not pre-trained to process multi modal inputs.

4. Closed World Testing

In order to measure the flexibility of proposed model, we conduct experiments on closed world setting. During the evaluation we adjust the feasibility mask to represent total number seen and unseen pairs present in the dataset. This transitions the proposed model from open world setting to closed world setting.

We compute seen, unseen accuracy and HM for all three datasets and similar to CoT [4] and OADIs [10], we compute the AUC with top 1 for MIT-States, C-GQA and AUC with top 3 for VAW-CZSL. In particular, despite an increase in the total number of compositions in output space, proposed model was able to attain a narrower the gap between itself and closed-world models, thereby showcasing the model's inherent flexibility.

5. Identifying Multiple Object Instances

In Table 8, we examine the model's capability in managing multiple object instances. For column 1, model successfully recognized both 'bear' and 'forest' while giving priority for 'bear' during predictions. However in column 2, predictions are predominantly influenced by secondary objects: 'skateboarding' and clothing. Thus demonstrating the model's capability to identify diverse set of object instances.

6. Negative Societal Impact

Zero-shot learning (ZSL) is prominent research focus, offering promising robust solutions for real-world language and vision tasks. Enhancing the robustness of performance assertions is pivotal as it not only showcases attainable performance levels while identifing invalid solutions. Nonetheless, these assurances often may overlook different errors, such as generalization gaps resulting from domain shifts or training label inaccuracies. It is crucial to accurately interpret these bounds to avoid erroneous claims or unwarranted confidence in proposed ZSL models.

Table 8. Performance of the model in the presence of multiple object instances. Secondary object predictions are highlighted in red and final predictions are highlighted with black boxes.

GT: Huge Bear

huge	bear	huge bear
mossy	dog	young bear
young	forest	mossy bear

GT: Stone Stairs

skateboarding	jeans	skateboarding boy
skating	boy	skateboarding man
black	man	jumping boy

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[2] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 2, 3

[3] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9336–9345, 2022. 1

[4] Hanjae Kim, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Hierarchical visual primitive experts for compositional zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5675–5685, 2023. 3

[5] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5222–5230, 2021. 2, 3

[6] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017. 1

[7] Muhammad Ferjad Naeem, Yongqin Xian, Federico Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 953–962, 2021. 3

[8] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 1

[9] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3593–3602, 2019. 1

[10] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13658–13667, June 2022. 1, 2, 3

[11] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 1

[12] Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10248–10256, 2020. 1