Attention Based Simple Primitives for Open-World Compositional Zero-Shot Learning

Ans Munir
IT University
Lahore, Pakistan
ansmunir98@gmail.com

Faisal Z. Qureshi
Ontario Tech University
Oshawa, Canada
faisal.qureshi@ontariotechu.ca

Muhammad Haris Khan
MBZUAI
Abu Dhabi, UAE
muhammad.haris@mbzuai.ac.ae

Mohsen Ali
IT University
Lahore, Pakistan
mohsen.ali@itu.edu.pk

***Abstract*—Compositional Zero-Shot Learning (CZSL) aims to predict unknown compositions made up of attribute and object pairs. Predicting compositions unseen during training is a challenging task. We are exploring Open-World Compositional Zero Shot Learning (OW-CZSL) in this study, where our test space encompasses all potential combinations of attributes and objects. Our approach involves utilizing the self-attention mechanism between attributes and objects to achieve better generalization from seen to unseen compositions. Utilizing a self-attention mechanism facilitates the model's ability to identify relationships between attribute and objects. The similarity between the self-attended textual and visual features is subsequently calculated to generate predictions during the inference phase. The potential test space may encompass implausible object-attribute combinations arising from unrestricted attribute-object pairings. To mitigate this issue, we leverage external knowledge from ConceptNet to restrict the test space to realistic compositions. Our proposed model, Attention-based Simple Primitives (ASP), demonstrates competitive performance, achieving results comparable to the state-of-the-art. Code is available at https://github.com/ans92/ASP.**

***Index Terms*—Compositional Zero-Shot Learning, Self Attention, Attributes, Objects**

I. Introduction

Humans are capable of recognizing basic concepts in intricate, novel contexts. So, for example, a folded chair looks very different and works very differently from a folded paper, but humans do not have any problem in identifying both with the concept of *folded*. The traditional machine learning and deep learning approaches, however, fail to learn these semantics from the statistical associations that they are typically programmed for. The ability to reason about shared and discriminative properties of objects, and identify their attributes is a distinctively human ability, that is very ambitiously mimicked with the task of Compositional Zero-Shot Learning (CZSL).

CZSL aims to learn visual concepts as compositions of attribute and object primitives. For example, a composition *Red Car* consists of attribute *Red* and object *Car* as shown in Fig. 1. The formulation of the task of visual learning as compositions of primitives is a paradigm that mimics cognition and finds its applications in a variety of computer vision applications, including but not limited to image retrieval [2], [30], visual question-answering [16], [37], and image captioning [31], [32]. This compositionality and contextuality lays the foundation for learning of the long tail of visual concepts very efficiently with a small set of basic primitives.

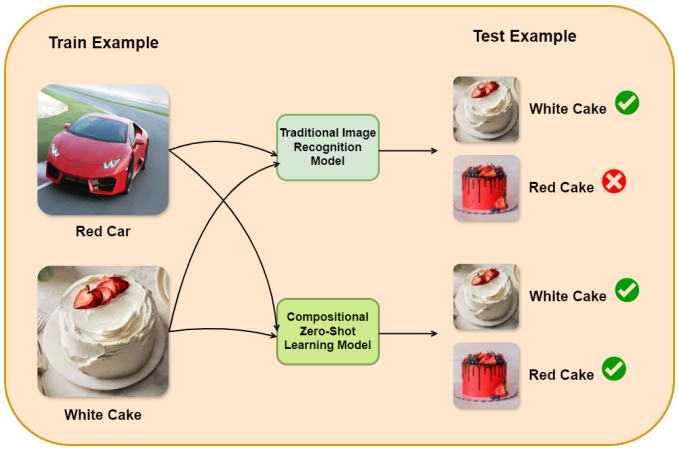


Fig. 1. In Compositional Zero-Shot Learning we have training set in the form of compositions that consist of attributes like "Red, White" and objects like "Car, Cake". Traditional image recognition models typically can only predict known compositions, but these struggle to compose new compositions during testing. In contrast, compositional zero-shot learning model effectively predicts new compositions during testing, as evidenced by the example of the "Red Cake".

The CZSL task has two settings: Closed-World and Open-World. Traditionally, closed-world setting has been used to evaluate the models. However, [22] proposed the open-world setting. In a closed-world setting, its assumed that test-time compositions, are known a-priori. In case of *generalized CZSL*, test-time compositions contain both seen or unseen during training. A closed-world situation is depicted in the Fig. 2. Training-set consists of three compositions (*Rusty Cycle, White Cake & Red Car*) and their respective images. However, during the testing-time two more compositions, *White Car* and *Red Cake*, and their images might be present. Therefore, CZSL model's task is to predict the "Red Cake" using the five compositions provided in both the training and testing sets.

In an open-world context, the test-time, composition set C^t consists of all possible combinations of *attributes* and *objects*. The C^t is not only quite large, as it grows proportional unique attributes and objects, but might contain compositions that might be *unfeasible* in real-world. *Rusty Cake* is an example

979-8-3503-7903-7/24/$31.00 ©2024 IEEE

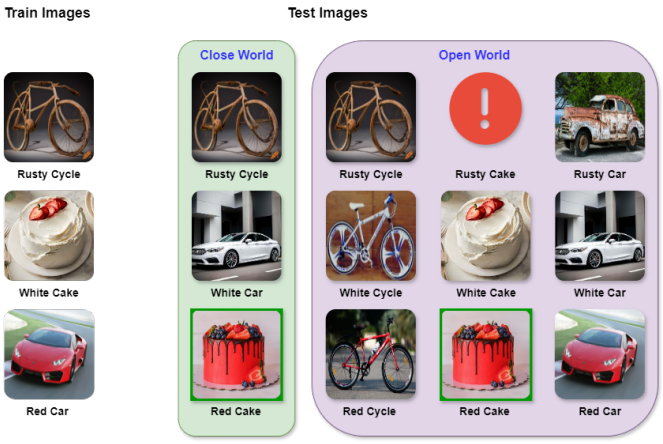


Fig. 2. Example demonstrating the difference between closed-world and open-world settings. In closed-world setting we assume that we know the test set *a priori*. However, in open-world setting we do not assume any knowledge of the test set. As such the test set contains all possible combinations of training attributes and objects. Composition "Red Cake" is highlighted to indicate the image that the model needs to predict.

of unfeasible composition, in the case depicted in Fig. 2.

CZSL demands two main capacities in a model: (1) the ability to compose and (2) the ability to contextualize. The concept of compositionality has been duly explored in the literature, with a composition of classifiers [26], modular networks [35], and composition of text embeddings with the attribute as operator [28] and graph neural networks [27]. But equally important is the ability to contextualize the attributes and objects to new compositions. Since these concepts are more semantic than visual, one object can appear drastically different under the influence of different attributes. Similarly, the same attribute can appear drastically different in the context of different objects. For example, the concept of old manifests with different visual features in the case of an elephant and totally unique visual features in the case of a car, see Fig. 3. Similarly, wet looks different in the context of the ground as compared to that of a cat (Fig. 3). Therefore, in addition to compositionality, we build a model that captures this contextuality between attributes and objects.

This paper makes the following contributions:

- develops Attention-based Simple Primitives (ASP) model for open-world compositional zero-shot learning that uses the self-attention mechanism between attributes and objects to capture (visual) interactions between these; and
- evaluates the proposed model on MIT-States, CGQA, and UT-Zappos benchmarks, showing that our method either outperforms or achieves results comparable to those reported by state-of-the-art methods.

II. Related Work

Compositional zero-shot learning: Recently, there has been significant attention given to Compositional Zero-Shot Learning (CZSL), which involves predicting joint labels for objects and attributes. Humans demonstrate a remarkable tendency to

Fig. 3. Demonstration of Visual Diversity in Primitives. The attribute Old looks drastically different in the context of Elephant (animate object) vs Car (inanimate object). Similarly the attribute Wet looks drastically different in the context of Cat (animate object) vs Ground (inanimate object).

recognize concepts in different contexts and generalize them to novel contexts. Machines have yet to master this ability.

The earliest approaches in CZSL tried to learn independent classifiers for attribute and object primitives but it failed to capture the joint context of the attribute-object compositions. [26] learns a transformation network to capture the relation between primitive concepts.

Many previous works build on the compatibility learning framework, with two major approaches. The first category builds on the idea that attributes lack independent, visual representation and thus treats them as operators on objects. It include works that treat attributes as a linear transformation of objects for learning a composite representation [28]. [20] treats attributes as modifiers of objects while adhering to group axioms, particularly symmetry of object representation under state transformations.

The second category where the aim is to learn the compatibility score of image embedding with composition embedding spawns a diverse array of approaches. [35] learns a modular representation of image embedding, with a gating mechanism conditioned on compositions, which allows modeling of the joint context of the image, object, and state. [41] develops a conditional generative approach for learning compositional embedding in an adversarial framework. [42] pools image features from various levels of CNN, and learns compatibility with quintuplet loss under an adversarial framework. CGE [27] learns globally consistent composition embedding, with graph regularization. Compcos [23] proposes an open-world scenario, where test time predictions are not limited to the set of predefined compositions. Co-CGE [22] combines both the methods, [27] and [23] and extends CGE [27] to open-world setting. Compcos [23] and Co-CGE [22] utilize training compositions to establish the feasibility of compositions in an open-world environment. [3] attempts to uncorrelated the object and attribute representations and deal it with a causal perspective for better generalization. [36] proposes to filter out spurious correlations between attribute and object primitives with independence loss and introduce the desired context in the composition embedding by following it with prototype graph propagation. [8] uses recurrent attractor networks in the linguistic and visual pathways to improve stability and generalization of composition prototypes, by exploiting the property of convergence and basin of attraction of attractor networks. [7] defines a linguistic graph encoder for information sharing

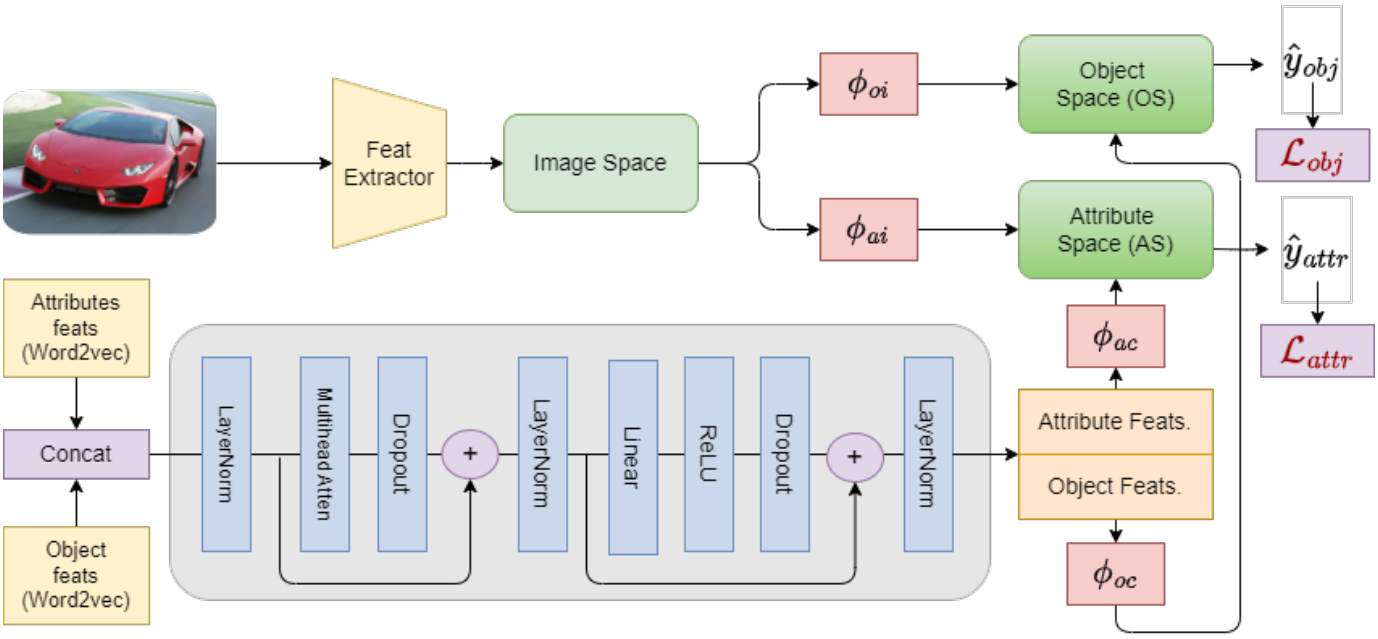


Fig. 4. ASP model architecture. After concatenating attribute and object features, we compute self-attention between attributes and objects to obtain the interactions between them. Then we get attribute as well as object features and project these to Attribute Space (AS) and Object Space (OS), respectively. MLPs are used for projection. Next, we compute the Cosine similarity between attribute image features and composition attribute features. Similarly, we obtain Cosine similarity between object image features and composition object features.

between attributes and objects, for smooth generalization to novel compositions. [44] introduces episodic training for zero-shot framework, and uses cross attention to better capture the joint context of compositions and images. From a contextual perspective, [12] addresses the CZSL issue by conditioning the attributes on objects and vice versa in order to capture the contextual nature of the composition.

Recent work [17] treats attributes and objects separately with independent classifiers to recognize attributes and objects in an open-world setting. ConceptNet [38] has been utilised by them to calculate the feasibility of compositions. This kind of independent classifier strategy, as previously mentioned in publications [17], performs better in an open-world context than in a closed one since there are less attributes and objects in an open-world context than there are combinations of those attributes and things.

Attribute classification and zero-shot learning: Very closely tied to the problem of CZSL is the problem of Attribute Classification. Attribute classification follows the route of learning to describe objects in terms of their attributes, rather than predicting their class labels. It finds use in image retrieval [34] and attribute-based zero-shot learning (ZSL) of objects. Prototype Learning [1], [45], [47] and Generative Models [11], [43], [48] are two prevalent approaches in ZSL. Score calibration [15], [19], [21] is also explored for better tuning to unseen classes in ZSL.

Attention mechanism: Self-attention is a building block of Multi-head attention that has been used in transformer architecture and was introduced in [40]. Attention was originally proposed for Language related models. However, recent works [6], [9], [33], [10] uses attention mechanism for computer vision related tasks. [18] uses an attention mechanism for Compositional Zero-Shot Learning by computing attention between compositions. They train and test their model in a closed-world setting.

Our approach lies at the intersection of different approaches. Like [17] we have predicted attributes and objects independently but in addition to [17] we have also included semantic knowledge of composition in our model by allowing attention between attributes and objects. [18] also included attention mechanism but they only deal with closed-world settings. However, unlike [18] that has computed attention between compositions, we computed attention between attributes (states) and objects. As we are adding attention between attributes and objects in an open-world setting, this makes our approach different from the previous. Different from [12], our approach involves calculating similarities between attributes and objects using self-attention instead of computing conditional attributes and objects. This will enable us to capitalise on the interaction between attributes and objects.

III. Approach

Proposed CZSL model aims to understand the relationship between attributes, sometimes referred to as states, and objects using the training data. Subsequently, at test time, the model infers the attribute and object present in an image. What makes this problem challenging is the fact that the model may not have encountered this attribute/object pair during training time. Consequently, the model needs to be able to generalize to new

attribute/object pairs, which requires the model to capture how attributes and objects interact with each other in a particular visual context. It is important to note that training data includes all attributes and objects under consideration; however, it often does not contain all possible attributed/object pairs. See Fig. 4 for the overall layout of our model.

A. Problem Formulation

Given the set of attributes $A = \{a_1, a_2, ..., a_n\}$ and objects $O = \{o_1, o_2, ..., o_m\}$, our composition set consists of $C = A \times O$. We deal with the task of visual classification where each image $\boldsymbol{x}$ has a compositional label c. Specifically in our case the compositional label is a pair of an object primitive o and an attribute primitive a, i.e. $c = \{(a, o)|a \in A, o \in O\}$.

C_{train} is a subset of composition set C. The test set C_{test} is defined in three alternate ways in the CZSL literature. In the classic CZSL setting, C_{test} consists of all novel compositions, $C_{train} \cap C_{test} = \emptyset$. [35] proposed Generalized CZSL setting where test compositions consists of both C_{train} and C_{test} such that $C_{train} \cap C_{test} \neq \emptyset$. Third is an open-world setting [27] where test set consists of all the composition set, i.e. $C_{test} = C$.

B. Attention Based Simple Primitives

A feature extractor $w : X \mapsto Z$ computes image features $\boldsymbol{z}$ for a given image $\boldsymbol{x}$, i.e., $\boldsymbol{z} = w(\boldsymbol{x})$. Next, an image-object encoder $\phi_{oi} : Z \mapsto O$ projects image feature $\boldsymbol{z}$ to $\boldsymbol{o_i}$ within the object space O. Similarly, an image-attribute encoder $\phi_{ai} : Z \mapsto A$ projects image feature $\boldsymbol{z}$ to $\boldsymbol{a_i}$ within the attribute space A.

Word embeddings are used to map object labels and attribute labels into a common space. The key hypothesis of this work is that the relationship between objects and attributes is not independent of the visual appearance. We employ multi-headed attention block to capture the relationship between object labels and attributes. Multi-headed attention processes the object and attribute embeddings and construct transformed object and attribute embeddings. A object-context encoder ϕ_{oc} projects the transformed object label feature to $\boldsymbol{c_i}$ within the object space O. Similarly, an attribute-context encoder $\phi ac : Z$ projects the transformed attributed label feature to $\boldsymbol{a_i}$ within the attribute space A.

Cosine similarity between image object features and object context features is used to predict the object. Similarly, cosine similarity between image attribute and attribute context is computed to predict the attribute.

The proposed method predicts object labels and attribute labels independently; however, objects and attributes inform each other through the multi-headed attention mechanism mentioned above.

C. Knowledge Guidance for computing feasibility

To obtain the *composition score* probability of separate predictions of attribute and objects are multiplied together [17]. As an example, we multiply the scores of "Red" and "Apple" to obtain the composition's score of "Red Apple". We also have a large number of unfeasible compositions.

Not all of the compositions obtained by combining attributes and objects are feasible. Following [17] we employ ConceptNet [38] to remove unfeasible compositions. ConceptNet is a type of knowledge graph in which words and phrases are connected by labelled edges. Therefore, to determine a composition's score, our model compare the score of each attribute to the corresponding object as follows:

$$c_s^o = \rho_{KB}(s, o) \tag{1}$$

where $\rho_{KB}(s, o)$ denotes the score between s and o.

D. Inference

During inference, once the image features are extracted, we use two multilayer perceptrons (MLPs), ϕ_{si} and ϕ_{oi}, to project image features into two distinct spaces: attribute space and object space. Similarly, the model project attribute and objects into textual embeddings and then pass them through the attention block to establish attention between attributes and objects. Subsequently, our model project these attributes and objects into the attribute space and object space using two MLPs, ϕ_{ac} and ϕ_{oc}, respectively. In the attribute space, cosine similarity is calculated between attribute visual features and attribute textual features. Likewise, in the object space, model also calculate the cosine similarity between object visual features and object textual features. The final prediction score is determined by multiplying the attribute and object scores, and the composition with the highest score is predicted as the image label. This process results in the overall prediction function.

$$f = \underset{(s,o)\in Y}{\arg\max} \left(\phi_{si}(\boldsymbol{x}).\phi_{ac}(s) \right) \times \left(\phi_{oi}(\boldsymbol{x}).\phi_{oc}(o) \right) \tag{2}$$

Equation (2) does not take into consideration the feasibility of the composition. We incorporate the feasibility score inferred from the ConceptNet for the final prediction. Updated equation is,

$$f = \underset{(s,o)\in Y, c_o^s > 0}{\arg\max} \left(\phi_{si}(\boldsymbol{x}).\phi_{ac}(s) \right) \times \left(\phi_{oi}(\boldsymbol{x}).\phi_{oc}(o) \right) \tag{3}$$

where $c_o^s > 0$ in above function means that we have only included the compositions that have positive feasibility.

E. Objective Function

We have used cross-entropy loss for objects as well as attributes to train our model.

$$\mathcal{L} = \sum_{i=1}^{N} \mathcal{L}_{state}(x_i, s_i) + \mathcal{L}_{obj}(x_i, o_i) \tag{4}$$

$$= - \sum_{(x,s)\in T} log \frac{\exp^{p(x,s)}}{\sum_{y\in S} \exp^{p(x,y)}} + \sum_{(x,o)\in T} log \frac{\exp^{p(x,o)}}{\sum_{y\in O} \exp^{p(x,y)}}$$

where T represents the test set, while s and o are part of the state S and object O sets, respectively.

TABLE I

THE TABLE COMPARES THE OUTCOMES OF PRIOR MODELS TO OURS, ASP. FF INDICATES THAT THE MODEL EMPLOYS A FIXED FEATURE EXTRACTOR FOR IMAGE FEATURES. THE TOP RESULTS ARE HIGHLIGHTED IN BOLD LETTERS, WHILE THE SECOND-BEST RESULTS ARE UNDERLINED. TOGETHER WITH IMPROVING RESULTS FROM KG-SP ON MIT-STATES, WE HAVE REACHED STATE-OF-THE-ART ON CGQA. (HIGHER RESULTS ARE BETTER). BEST RESULTS IN BOLD AND SECOND BEST UNDERLINED.

	MIT-States				UT-Zappos				CGQA			
Method	S	U	HM	AUC	S	U	HM	AUC	S	U	HM	AUC
SymNet [20]	21.4	7.0	5.8	0.8	53.3	44.6	34.5	18.5	26.7	2.2	3.3	0.43
CGE $_{ff}$ [27]	29.6	4.0	4.9	0.7	58.8	46.5	38.0	21.5	28.3	1.3	2.2	0.30
Compcos [23]	25.4	10.0	8.9	1.6	59.3	46.8	36.9	21.3	28.4	1.8	2.8	0.39
Co-CGE$_{ff}$ [22]	26.4	10.4	10.1	2.0	60.1	44.3	38.1	21.3	28.7	1.6	2.6	0.37
KG-SP$_{ff}$ [17]	23.4	7.0	6.7	1.0	58.0	47.2	39.1	22.9	26.6	2.1	3.4	0.44
ASP$_{ff}$	21.6	7.2	6.7	1.0	59.0	42.8	38.5	21.6	26.8	2.1	3.4	0.41
CGE [27]	**32.4**	5.1	6.0	1.0	61.7	47.7	39.0	23.1	**32.7**	1.8	2.9	0.47
Co-CGE [22]	30.3	**11.2**	**10.7**	**2.3**	61.2	45.8	40.8	23.3	32.1	3.0	4.8	0.78
KG-SP [17]	28.4	7.5	7.4	1.3	**61.8**	**52.1**	42.3	**26.5**	31.5	2.9	4.7	0.78
ASP	27.1	8.4	7.7	1.4	61.0	48.6	**43.1**	25.9	31.7	**3.2**	**5.0**	**0.80**

IV. EXPERIMENTS

In this section, we present the dataset details, evaluation protocols, and implementation details are included. Experimental results comparing previous state-of-the-art methods with proposed methods and ablation study are also included.

A. Experimental Setup

Datasets: Following previous works [23], [22], [18] we tested our model on 3 datasets, MIT-States [14], UT-Zappos [42], [46] and CGQA [27]. MIT-States comprises a total of 63,440 images depicting 115 unique attributes and 245 unique object classes. The splits of the dataset are defined as follows, a training set of 30k images constituting 1262 compositions. A validation set of 10k images and a test set of 13k images with composition space equals 28,175 compositions. UT-Zappos is the smallest of the 3 datasets and has only 16 attributes and 12 object categories. The splits of UT-Zappos are as follows, a training set consisting of 23k images, and a validation and test set consisting of 3k images. Test time compositional space has 192 compositions. CGQA was recently proposed in [27] that has a larger number of attributes and objects as compared to MIT-States. It has 413 attributes and 674 object categories. The splits of CGQA are as follows, a training set consisting of 5592 pairs across 26k images; a validation set consisting of 4k images; and a test set consisting of 5k images. Test time compositional space has 278,362 compositions.

Evaluation protocol: In our evaluation, we conducted the experiment in an open-world setting [17], [22], [23], including all possible combinations of attributes and objects to form the test set. After following [17], [35], we introduced various scalar bias terms to the seen composition scores to address the bias for seen compositions. A large negative scalar bias results in high unseen accuracy and low seen accuracy, while a large positive scalar bias produces the opposite effect. By adjusting the scalar bias from large negative to large positive, we calculated the top-1 accuracy of both seen and unseen compositions. The area under the curve (AUC), best harmonic mean (HM), and the best seen and unseen accuracy were derived from the seen-vs-unseen accuracy curve.

Implementation details: To extract image features, we used ResNet18 [13] pre-trained on Imagenet as a feature extractor, similar to earlier methods. Two 2-layer MLPs have been employed as classifiers for attributes and objects. In the case of CGQA, we have obtained attributes and object embeddings using Word2Vec. In order to improve comparison with earlier SOTA models, we have employed both Word2Vec [24], [25] and FastText [5] for the two datasets, MIT-States and UT-Zappos. Then we used an Attention block with one multi-head attention layer followed by Dropout [39], Layer Normalization [4], Linear layer, and ReLU [29] activation. Finally, two 2-layer MLPs that map attention features to attribute and object space have also been used. After the first layer, we have utilised LayerNorm, ReLU, and Dropout in every MLP.

B. Quantitative Analysis

Results of Open-World Compositional Zero-Shot Learning (OW-CZSL) are shown in Table I. Subscript "$_{ff}$" in Table I indicates that a fixed feature extractor for image features was not fine-tuned throughout training. The top results shown in **bold**, while the outcomes that came in second are underlined.

Results on MIT-States Dataset are shown in Table I. [22] performs better than every other model in terms of Unseen, HM, and AUC scores. CGE [27] achieved the highest accuracy in terms of seen accuracy. Nonetheless, Unseen, HM, and AUC scores show that our model, ASP, performs better than KG-SP. Our model achieved particularly good results on the MIT-States dataset as compared to KG-SP. Compared to KG-SP's 7.5 unseen accuracy, ASP obtained 8.4. Even though the MIT-States dataset is thought to be noisy [3], our model performs better than KG-SP. Table II shows the results of our model on MIT-States dataset. First row of Table II shows the success examples while second row shows the failure examples of the model. In second row, black labels are actual labels while the red labels are wrong prediction made by our model.

The most recent dataset for the CZSL challenge is CGQA proposed in [27]. Compared to the other two datasets, this one is the largest and contains the most attributes and objects. In comparison with KG-SP and all other models, our model has attained the highest results. We surpassed all the models

Ancient Church

Ancient Jungle

Barren Island

Barren Lake
Steaming Lake

Ancient Town
Large House

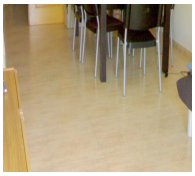
Ancient Clock
Small Clock

TABLE II
SUCCESS EXAMPLES IN FIRST ROW AND FAILURE EXAMPLES IN SECOND ROW OF MIT-STATES DATASET. THE GROUND TRUTH LABELS ARE REPRESENTED BY THE COLOR BLACK, WHEREAS THE MODEL'S PREDICTIONS ARE INDICATED BY THE COLOR RED.

White Shirt

Yellow Umbrella

Black Couch



Tan Floor
Hardwood Floor

Red Bush
Large Tree

Large Refrigerator
Red Window

TABLE III
SUCCESS EXAMPLES IN FIRST ROW AND FAILURE EXAMPLES IN SECOND ROW OF CGQA DATASET. THE GROUND TRUTH LABELS ARE REPRESENTED BY THE COLOR BLACK, WHEREAS THE MODEL'S PREDICTIONS ARE INDICATED BY THE COLOR RED.

including KG-SP in terms of HM, AUC and unseen accuracy. We also obtained a higher seen accuracy compared to KG-SP. Our improved unseen accuracy validates that the attention mechanism of ASP allows for better generalisation to unseen compositions.

In open-world environments, the method of independently predicting primitives offers significant advantages. Consider CGQA, the largest dataset, where model only have to predict an image's attribute from 413 attributes and its object from 674 objects. If our model were to predict composition, it would need to select from 278,362 possible combinations, resulting in decreased accuracy as indicated in Table I CGQA results. While KG-SP [17] and ASP predict attributes and objects separately, Compcos [22] focuses on predicting compositions. Compared to the other two models, Compcos achieves a lower score.

Although it is be beneficial to independently predict basic elements in an open-world setting, the concept of composition remains influential. As demonstrated in Fig. 3, primitives are interdependent. Nevertheless, we argue that as the dataset expands and encompasses more features and objects, numerous improbable combinations emerge, leading to a decline in model accuracy, as evidenced in the CGQA scenario. As indicated in the results for Compcos in Table I, predictive accuracy could be higher when working with a relatively small dataset like MIT-States, but it decreases as the dataset size grows, as observed in the case of CGQA.

The UT-Zappos dataset is small and contains various styles of shoes. Because many types of shoes, such as leather, patent leather, and faux leather, are quite similar, our attention mechanism did not perform well on this specific dataset. Nevertheless, our model still surpasses all other models in the harmonic mean.

C. Qualitative Analysis

Within this section, we will be showcasing qualitative findings. The initial row in Table II depicts instances where our model succeeded, while the second row illustrates instances where it failed. As the images display, the model successfully predicted "Ancient Church", "Ancient Jungle", and "Barren Island". Nonetheless, it inaccurately predicted "Barren Lake" as "Steaming Lake", "Ancient Town" as "Large House", and "Ancient Clock" as "Small Clock". Despite these inaccurate predictions, the model's estimations are still in close proximity to the actual values. In three examples, two have accurately predicted objects, and the predictions are close to ground truth.

In Table III of the CGQA dataset, the first row showcases our model's successful predictions, including "White Shirt", "Yellow Umbrella", and "Black Couch". The second row, however, displays instances where the model's predictions were not successful. The model correctly identified the object in the first example. In the second example, our model misidentified "Bush" as a "Tree" which are closely related. In the third example, the model's prediction of "Refrigerator" as a "Window" seems reasonable upon observing the image, as there is a table and the "Refrigerator" resembles a window. We can conclude that attention plays a vital role in enabling the model to make more accurate predictions based on the given image.

D. Ablation Study

Number of heads in Multihead attention: Multi-head attention includes many heads. A self-attention system exists inside every head, as demonstrated in [40]. We have conducted multi-head attention trials with varying head counts. Our studies were conducted on MIT-States using a fixed feature extractor. Additionally, we seeded the values to ensure that the MLPs were initialised uniformly. Our model was trained once using 80 epochs for each experiment, and it was then evaluated using the test set.

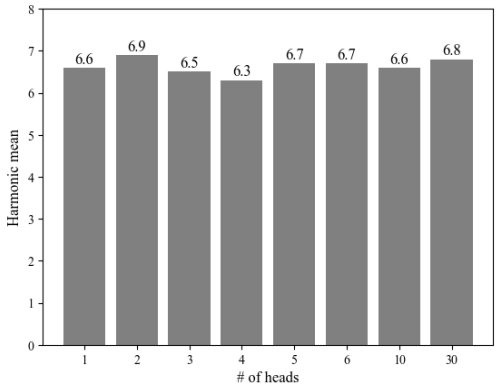


Fig. 5. Graph showing the effect of heads in Multihead Attention. On x-axis, there is the number of heads while on the y-axis there is the Harmonic mean.

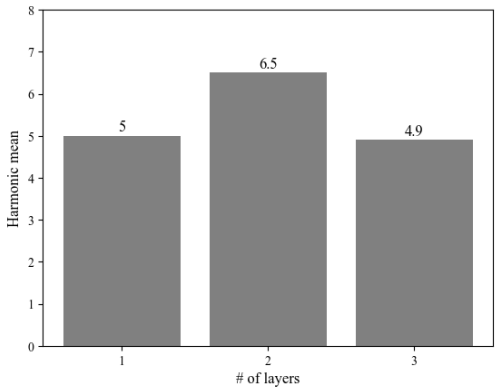


Fig. 6. Graph showing the effect of the number of linear layers in all the MLPs on the Harmonic Mean.

As shown in Fig 5 HM is almost the same in each experiment. The highest accuracy was achieved with two heads in the multi-head attention layer. HM is 6.9 when there are two heads, and 6.8 when there are thirty heads. As a result, accuracy does not increase as the number of heads in the multi-head attention layer grows.

Number of layers for MLP: We have examined how each MLP's number of layers affects MIT-States using fixed feature extractors. The impact of the number of linear layers in an MLP on the harmonic mean is depicted in Fig. 6. Similar to Fig. 5 we have trained the model once for 80 epochs for each experiment and then evaluated on the test set. It is evident that a single layer prevents the model from understanding the intricate relationships between the attributes and objects. Nevertheless, the model is able to effectively learn the intricate relationships between them when the number of layers is increased to two. However, as we increased the number of layers, the model model likely overfits the data and hence its Harmonic Mean decreased.

V. Conclusion

We propose a novel solution for the Compositional Zero-Shot Learning (CZSL) in the Open-World setting. Unlike in closed-world CZSL, where composition set during test is known in advance, open-world setting assume no prior knowledge therefore include all possible combinations of attributes and objects. Primitives, such as attributes and objects, have been predicted independently. Our model has employed attention mechanism between primitives to capture their interactions, to generalize effectively from training compositions to test compositions. Our approach of independently predicting primitives performs better due to the significant reduction in our test space in open-world CZSL. In the open-world CZSL, there are also unrealistic compositions which we have addressed by utilizing external knowledge from ConceptNet. Our model achieved the top performance on the CGQA dataset, surpassing all previous models and achieving higher accuracy than KG-SP.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 819–826, 2013.

[2] Muhammad Umer Anwaar, Egor Labintcev, and Martin Kleinsteuber. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, pages 1140–1149, 2021.

[3] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017.

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020.

[7] Hui Chen, Zhixiong Nan, Jingjing Jiang, and Nanning Zheng. Learning to infer unseen attribute-object compositions. *arXiv preprint arXiv:2010.14343*, 2020.

[8] Hui Chen, Zhixiong Nan, and Nanning Zheng. Beyond supervised learning: Recognizing unseen attribute-object pairs with vision-language fusion and attractor networks. 2019.

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020.

[10] 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.

[11] Zongyan Han, Zhenyong Fu, and Jian Yang. Learning the redundancy-free features for generalized zero-shot object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12865–12874, 2020.

[12] Zoya Naseer Hashmi. Conditional embeddings for compositional zero shot learning. Master's thesis, Information Technology University, Lahore, 2022.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Phillip Isola, Joseph J. Lim, and Edward H. Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015.

[15] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1629–1636, 2014.

[16] Chenchen Jing, Yunde Jia, Yuwei Wu, Xinyu Liu, and Qi Wu. Maintaining reasoning consistency in compositional visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2022.

[17] 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.

[18] Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Alain Pagani, Didier Stricker, and Muhammad Zeshan Afzal. Learning attention propagation for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3828–3837, 2023.

[19] Yannick Le Cacheux, Hervé Le Borgne, and Michel Crucianu. From classical to generalized zero-shot learning: A simple adaptation process. In *International Conference on Multimedia Modeling*, pages 465–477. Springer, 2019.

[20] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325, 2020.

[21] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. *Advances in Neural Information Processing Systems*, 31, 2018.

[22] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *arXiv e-prints*, pages arXiv–2105, 2021.

[23] M Mancini, MF Naeem, Y Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *34th IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2021.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[25] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[26] 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.

[27] MF Naeem, Y Xian, F Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *34th IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2021.

[28] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018.

[29] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.

[30] Andrei Neculai, Yanbei Chen, and Zeynep Akata. Probabilistic compositional embeddings for multimodal image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4547–4557, 2022.

[31] Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. Compositional generalization in image captioning. *arXiv preprint arXiv:1909.04402*, 2019.

[32] George Pantazopoulos, Alessandro Suglia, and Arash Eshghi. Combine to describe: Evaluating compositional generalization in image captioning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 115–131, 2022.

[33] Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. Image transformer. In *International conference on machine learning*, pages 4055–4064. PMLR, 2018.

[34] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021.

[35] 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.

[36] Frank Ruis, Gertjan J Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

[37] Raeid Saqur and Karthik Narasimhan. Multimodal graph networks for compositional generalization in visual question answering. *Advances in Neural Information Processing Systems*, 33:3070–3081, 2020.

[38] R Speer, J Chin, and C ConceptNet Havasi. 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (December 2016)*, pages 4444–4451.

[39] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Xin Wang, Fisher Yu, Trevor Darrell, and Joseph E Gonzalez. Task-aware feature generation for zero-shot compositional learning. *arXiv preprint arXiv:1906.04854*, 2019.

[42] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3741–3749, 2019.

[43] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5542–5551, 2018.

[44] Guangyue Xu, Parisa Kordjamshidi, and Joyce Y Chai. Zero-shot compositional concept learning. *arXiv preprint arXiv:2107.05176*, 2021.

[45] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. *Advances in Neural Information Processing Systems*, 33:21969–21980, 2020.

[46] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5570–5579, 2017.

[47] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2021–2030, 2017.

[48] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1004–1013, 2018.