Learning Attention as Disentangler for Compositional Zero-shot Learning

Shaozhe Hao  Kai Han*  Kwan-Yee K. Wong
The University of Hong Kong
{szhao, kykwong}@cs.hku.hk  kaihanx@hku.hk

Abstract

Compositional zero-shot learning (CZSL) aims at learning visual concepts (i.e., attributes and objects) from seen compositions and combining concept knowledge into unseen compositions. The key to CZSL is learning the disentanglement of the attribute-object composition. To this end, we propose to exploit cross-attentions as compositional disentanglers to learn disentangled concept embeddings. For example, if we want to recognize an unseen composition “yellow flower”, we can learn the attribute concept “yellow” and object concept “flower” from different yellow objects and different flowers respectively. To further constrain the disentanglers to learn the concept of interest, we employ a regularization at the attention level. Specifically, we adapt the earth mover’s distance (EMD) as a feature similarity metric in the cross-attention module. Moreover, benefiting from concept disentanglement, we improve the inference process and tune the prediction score by combining multiple concept probabilities. Comprehensive experiments on three CZSL benchmark datasets demonstrate that our method significantly outperforms previous works in both closed- and open-world settings, establishing a new state-of-the-art. Project page: https://haoosz.github.io/ade-czsl/

1. Introduction

Suppose we have never seen white bears (i.e., polar bears) before. Can we picture what it would look like? This is not difficult because we have seen many white animals in daily life (e.g., white dogs and white rabbits) and different bears with various visual attributes in the zoo (e.g., brown bears and black bears). Humans have no difficulty in disentangling “white” and “bear” from seen instances and combining them into the unseen composition. Inspired by this property of human intelligence, researchers attempt to make machines learn compositions of concepts as well. Compositional zero-shot learning (CZSL) is a specific problem studying visual compositionality, aiming to learn visual concepts from seen compositions of attributes and objects and generalize concept knowledge to unseen compositions.

Learning attribute-object compositions demands prior knowledge about attributes and objects. However, visual concepts of attributes and objects never appear alone in a natural image. To learn exclusive concepts for compositionality learning, we need to disentangle the attribute concept and the object concept. As illustrated in Fig. 1, if we want to recognize the image of “yellow flower”, it is necessary to learn the “yellow” concept and the “flower” concept, i.e., disentangle visual concepts, from images of seen compositions. Previous works [22, 24, 25, 28–30, 36, 46] tackle CZSL by composing attribute and object word embeddings, and projecting word and visual embeddings to a joint space. They fail to disentangle visual concepts. Recently, some works [21,40,41,50] consider visual disentanglement but still have limitations despite their good performance. SCEN [21] learns concept-constant samples contrastively without constructing concept embedding prototypes to avoid learning irrelevant concepts shared by positive samples. IVR [50] disentangles visual features into ideal concept-invariant domains. This ideal domain generalization setting requires a small discrepancy of attribute and object sets and would degenerate on vaster and more...
complex concepts. ProtoProp [40] and OADis [41] learn local attribute and object prototypes from spatial features on convolutional feature maps. However, spatial disentanglement is sometimes infeasible because attribute and object concepts are highly entangled in spatial features. Taking an image of “yellow flower” as an example, the spatial positions related to the attribute “yellow” and the object “flower” completely overlap, which hinders effective attribute-object disentanglement.

To overcome the above limitations, we propose a simple visual disentangling framework exploiting Attention as DisEntangler (ADE) on top of vision transformers [6]. We notice that vision transformers (ViT) have access to more sub-space global information across multi-head attentions than CNNs [38]. Therefore, with the expressivity of different subspace representations, token attentions of ViT may provide a more effective way for disentangling visual features, compared to using traditional spatial attentions across local positions on convolutional features [40, 41]. Specifically, it is difficult to disentangle the attribute-object composition “yellow flower” by spatial positions, but it is possible for ViT multi-head attentions to project attribute concept “yellow” and object concept “flower” onto different subspaces. Inspired by this property, we propose to learn cross-attention between two inputs that share the same concept, e.g., “yellow bird” and “yellow pear” share the same attribute concept “yellow”. In this way, we can derive attribute- and object-exclusive visual representations by cross-attention disentanglement. To ensure that the concept disentangler is exclusive to the specific concept, we also need to constrain the disentanglers to learn the concept of interest instead of the other concept. For example, given attribute-sharing images, the attribute attention should output similar attribute-exclusive features while the object attention should not. To achieve this goal, we apply a regularization term adapted from the earth mover’s distance (EMD) [13] at the attention level. This regularization term forces cross-attention to learn the concept of interest by leveraging the feature similarity captured from all tokens. Mancini et al. [24] propose an open-world evaluation setting, which is neglected by most previous works. We consider both closed-world and the open-world settings in our experiments, demonstrating that our method is coherently efficient in both settings. The contributions of this paper are summarized below:

- We propose a new CZSL approach, named ADE, using cross-attentions to disentangle attribute- and object-exclusive features from paired concept-sharing inputs.
- We force attention disentanglers to learn the concept of interest with a regularization term adapted from EMD, ensuring valid attribute-object disentanglement.
- We comprehensively evaluate our method in both closed-world and open-world settings on three CZSL datasets, achieving consistent state-of-the-art.

2. Related work

Visual attribute has been widely studied to understand how visual properties can be learned from objects. The pioneering work by Ferrari and Zisserman [10] learned visual attributes using a probabilistic generative model. The successive work by Lampert et al. [20] used visual attributes to detect unseen objects with an attribute-based multi-label classification. Similarly, Patterson et al. [35] proposed Economic Labeling Algorithm (ELA) to discover multi-label attributes for objects. Different from multi-label classification, other works [4, 8, 15, 23] learned attribute-object relationship to generalize attribute feature across all object categories based on probabilistic models. Visual attributes also benefit downstream tasks, e.g., object recognition [7, 16, 31], action recognition [1, 9, 26], image captioning [19, 33], and semi-supervised learning [42].

Compositional zero-shot learning (CZSL) is a special case of zero-shot learning (ZSL) [34, 39, 43, 47, 48], aims at recognizing unseen attribute-object compositions learning from seen compositions. Misra et al. [28] first termed and studied CZSL by projecting composed primitives and visual features to a joint embedding space. Nagarajan et al. [30] formulated attributes as matrix operators applied on object vectors. Purushwalkam et al. [36] introduced a task-driven modular architecture to learn unseen compositions by re-weighting a set of sub-tasks. Wei et al. [46] generated attribute-object compositions with GAN [11] to match visual features. Li et al. [22] proposed symmetry principle of attribute-object transformation under the supervision of group axioms. Naem et al. [29] and Mancini et al. [25] used graph convolutional networks to extract attribute-object representations. Recently, some works shift their interest from word composing to visual disentanglement. Atzmon et al. [2] solved CZSL from a causal perspective to learn disentangled representations. Ruis et al. [40] proposed to learn prototypical representations of objects and attributes. Li et al. [21] disentangled visual features into a Siamese contrastive space and entangled them with a generative model. Saini et al. [41] extracted visual similarity from spatial features to disentangle attributes and objects. Zhang et al. [50] treated CZSL as a domain generalization task, learning attribute- and object-invariant domains. A more realistic open-world CZSL setting was studied in [17, 24, 25], which considered all possible compositions in testing. Very recently, an inspiring work by Nayak et al. [32] introduced compositional soft prompts to CLIP [37] to tackle CZSL problem.

Attention mechanism has been well studied by non-local neural networks [45] in computer vision and transformers [44] in machine translation. Dosovitskiy et al. [6] adapted transformer architecture to computer vision field, proving its comparable efficiency over traditional CNNs. Inspired by the multi-head self-attention implemented by transformers, our work exploits efficient attention as disentangler for CZSL.
Figure 2. Method overview. Left (our framework ADE): Given one target image of “red bus”, we sample two auxiliary images of the same attribute “red wall” and of the same object “blue bus”. We feed the three images into a frozen ViT initialized with DINO [3]. We then input all encoded tokens (i.e., [CLS] and patch tokens) to three attention modules: (1) attribute cross-attention taking paired attribute-sharing tokens as inputs; (2) object cross-attention taking paired object-sharing tokens as inputs; (3) composition self-attention taking tokens of single target image as input. We then project the [CLS] tokens of attention outputs with three MLP embedders \( \pi_a, \pi_c, \) and \( \pi_o \). We finally compute cross entropy losses of embedded visual features with three learnable word embeddings: attribute embeddings, object embeddings, and their linear fused composition embeddings. Right: Illustration of five losses in attribute, object, and composition embedding spaces.

3. Preliminary

Compositional zero-shot learning aims at learning a model from limited compositions of attributes (e.g., yellow) and objects (e.g., flower) to recognize an image from novel compositions. Given a set of all possible attributes \( \mathcal{A} = \{a_0, a_1, \ldots, a_n\} \) and a set of all possible objects \( \mathcal{O} = \{o_0, o_1, \ldots, o_m\} \), a compositional class set \( \mathcal{C} = \mathcal{A} \times \mathcal{O} \) includes all attribute-object compositions. We divide \( \mathcal{C} \) into two disjoint sets, namely the set of seen classes \( \mathcal{C}_{\text{u}} \) and the set of unseen classes \( \mathcal{C}_{\text{s}} \), where \( \mathcal{C}_{\text{u}} \cap \mathcal{C}_{\text{s}} = \emptyset \) and \( \mathcal{C}_{\text{u}} \cup \mathcal{C}_{\text{s}} = \mathcal{C} \). Training images are only from classes in \( \mathcal{C}_{\text{u}} \) and testing images are from classes in \( \mathcal{C}_{\text{test}} = \mathcal{C}_{\text{u}} \cup \mathcal{C}_{\text{s}} \). For closed-world evaluation, \( \mathcal{C}'_{\text{u}} \) is a predefined subset of \( \mathcal{C}_{\text{u}} \), i.e., \( \mathcal{C}'_{\text{u}} \subset \mathcal{C}_{\text{u}} \). For open-world evaluation, all possible compositions are considered for testing, i.e., \( \mathcal{C}'_{\text{u}} = \mathcal{C}_{\text{u}} \). CZSL task aims to learn a model \( f : \mathcal{X} \rightarrow \mathcal{C}_{\text{test}} \) to predict labels in the testing composition set \( c \in \mathcal{C}_{\text{test}} \) for input images \( x \in \mathcal{X} \).

To learn a CZSL model, a common method is to minimize the cosine similarity score between the visually encoded feature and the word embedding of attribute-object compositions. The similarity score can be expressed as

\[
s(x, (a, o)) = \frac{\phi(x)}{||\phi(x)||} \cdot \frac{\psi(a, o)}{||\psi(a, o)||}
\]  

(1)

where \( \phi(\cdot) \) is the visual encoder, and \( \psi(\cdot, \cdot) \) is the composition function. The CZSL model \( f \) can be formulated as

\[
f(x; \theta_{\phi}, \theta_{\psi}) = \arg \max_{a_i \in \mathcal{A}, o_j \in \mathcal{O}} s(x, (a_i, o_j))
\]  

(2)

The model \( f \) is optimized by learning the model parameters \( \theta_{\phi} \) and \( \theta_{\psi} \) to match the input \( x \) with the most similar attribute-object composition. Visual encoder \( \phi \) and composition function \( \psi \), as two core components, take multiple forms in different methods. For visual encoder, works [40,41,50] manually design networks on top of a pretrained backbone (e.g., ResNet18 [12]) to disentangle attribute and object embeddings for visual features. For composition functions, the object conditioned network [41], the graph embedding [29,40], and the linear projector [24,50] are commonly used. Unlike employing complex composition functions, we adopt the simple linear projector in our method.

4. Our approach

We aim to disentangle attribute and object features for visual representation by learning from image pairs sharing the same attributes or objects. For example, by learning the object “flower” from the “purple flower” and the “red flower” and learning the attribute “yellow” from the “yellow bird” and the “yellow pear”, we can infer what a “yellow flower” looks like. To this end, we propose ADE to learn cross-attentions as concept disentanglers with the regularization of a novel token earth mover’s distance. The whole framework of our method is shown in Fig. 2.

4.1. Disentanglement with cross-attention

Multi-head self-attention in transformers [44] is powerful for extracting token connections. Attention maps each query token, and key-value token pairs to an output token. Every output token is a weighted sum of the values, where the weights is the similarity of the corresponding query token and the key tokens. The attention is formulated as:

\[
\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V
\]  

(3)
where three inputs are query tokens (Q), key tokens (K), and value tokens (V) and the scaling factor $d_k$ is the dimension of the query and key. Self-attention uses the same input for $Q, K, V$ and effectively captures relationships among [CLS] and patch tokens of a single input image. Normally, the [CLS] token is exclusively used as a global representation of the input image for downstream tasks, but exclusive [CLS] token highly entangles attribute and object concepts.

For CZSL task, we want to disentangle the exclusive concept by exploiting the semantic similarity between tokens from different concept-sharing features. For this purpose, we introduce a cross-attention mechanism to extract attribute-exclusive or object-exclusive features for paired images sharing the same attribute or object concept. The cross-attention has the same structure as the self-attention, but works on different inputs. Self-attention uses the same input $I$ for $Q, K, V$ (i.e., $Q=K=V=I$) while cross-attention uses one of the paired inputs $I$ for the query and the other $I’$ for the key and the value (i.e., $Q=I, K=V=I’$). In the cross-attention, the query input and the key input play different roles. The cross-attention output is the weighted sum of the value (also the key) input based on the similarity between the query and key inputs. For a pair of input images, we swap the query and key inputs in computing cross-attention to derive two cross-attended outputs for the respective inputs. Cross-attention with query-key swapping (QKS) is illustrated in Fig. 3.

The cross-attention outputs paired features of the exclusive concept with paired concept-sharing inputs. Taking the object-sharing pair as an example, object cross-attention leverages the shared object features of the key input to enhance the query input, thus making the outputs exclusive to object feature space, i.e., object-exclusive features $v_o$ and $v_o’$. The same applies to attribute-exclusive features $v_a$ and $v_a’$ extracted from paired attribute-sharing images. We also use a standard self-attention to extract composition feature $v_c$.

To train cross-attentions, we adopt common cross entropy loss after MLP embedders ($\pi_a, \pi_c, \pi_o$) for attribute-exclusive features ($v_a, v_a’$), object-exclusive features ($v_o, v_o’$), and composition feature ($v_c$). We denote any of these visual features as $v$ and the corresponding embedder as $\pi$. We denote the word embedding as $w_z$, where the concept $z \in Z$ can be any of the attribute $a \in A$, the object $o \in O$, or the composition $c \in C$. The class probability of L2-normalized $\pi(v)$ and $w_z$ can be computed as:

$$p_\pi(z \mid v) = \frac{\exp(\langle\pi(v) \cdot w_z\rangle / \tau)}{\sum_{z \in Z} \exp(\langle\pi(v) \cdot w_z\rangle / \tau)}, z \in Z$$

where $\tau$ is the temperature. We then compute the cross entropy to classify visual features:

$$H_\pi(v, z) = -\log p_\pi(z \mid v), z \in Z$$

Consdering different classification objectives of attribute, object, and composition, the general Eq. (5) can be applied to five visual features, i.e., $v_a, v_a’, v_o, v_o’, v_c$. Therefore, the training objective is composed of five cross entropy losses:

$$\mathcal{L}_{ce} = H_\pi(v_a, a) + H_\pi(v_a’, a) + H_\pi(v_o,c) +$$

$$\mathcal{L}_{attr} + \mathcal{L}_{attr} + \mathcal{L}_{com}$$

$$H_\pi(v_o, o) + H_\pi(v_o’, o)$$

$$\mathcal{L}_{obj}$$

where the attribute label $a \in A$, the object label $o \in O$, and the composition label $c = (a, o) \in C_{test}$.

### 4.2. Disentangling constraint at the attention level

So far, we have enabled cross-attentions to disentangle concepts, but we want to further ensure that the attribute and the object disentanglers learn the corresponding attribute and object concepts instead of the opposite concepts. To this end, we introduce an attention-level earth mover’s distance (EMD) to constrain disentanglers to learn the concept of interest. The EMD is formulated as an optimal transportation problem in [13]. Suppose we have supplies of $n_a$ sources $S = \{s_i\}_{i=1}^{n_a}$ and demands of $n_d$ destinations $D = \{d_j\}_{j=1}^{n_d}$. Given the moving cost from the $i$-th source to the $j$-th destination $c_{ij}$, an optimal transportation problem aims to find the minimal-cost flow $f_{ij}$ from sources to destinations:

$$\text{minimize}_{f_{ij}} \sum_{i=1}^{n_a} \sum_{j=1}^{n_d} c_{ij} f_{ij}$$

subject to $f_{ij} \geq 0, \ i = 1, \ldots, n_a, \ j = 1, \ldots, n_d$$

$$\sum_{j=1}^{n_d} f_{ij} = s_i, \ i = 1, \ldots, n_a$$

$$\sum_{i=1}^{n_a} f_{ij} = d_j, \ j = 1, \ldots, n_d$$

where the optimal flow $f_{ij}$ is computed by the moving cost $c_{ij}$, the supplies $s_i$, and the demands $d_j$. The EMD can be further formulated as:

$$\text{EMD}(c_{ij}, s_i, d_j) = (1 - c_{ij}) f_{ij}$$

When the EMD is greater, the distributions $S$ and $D$ are closer. In Fig. 4, we show how to use the EMD as a feature similarity metric at the attention level in our cross-attention module. We can derive a cross-attention map from Eq. (3) and the counterpart cross-attention map after query-key swapping. We use the logits of [CLS]-to-patch attentions as supplies $s_i$ and demands $d_j$, which represent how similar the global feature of one image is to the local patches of the
We first fix $\beta = 1.0$ during training and then validate $\beta = 0.0$, 0.1, \ldots, 1.0 to choose the best $\beta$ on the validation set. Finally, we compute the composition prediction as Eq. (15) with the chosen $\beta$, making the best prediction trade-off between generalized composition and independent concept.

5. Experiments

5.1. Experimental details

Datasets We use three benchmark datasets in CZSL problem, namely Clothing16K [50], UT-Zappos50K [49], and C-GQA [29]. Clothing16K [50] contains different types of clothing (e.g., shirt, pants) with color attributes (e.g., white, black). UT-Zappos50K [49] is a fine-grained dataset consisting of different kinds of shoes (e.g., sneakers, sandals) with texture attributes (e.g., leather, canvas). C-GQA [29] is a split built on top of Stanford GQA dataset [14], composed of extensive common attribute concepts (e.g., old, wet) and object concepts (e.g., dog, bus) in real life. We follow the common data splits of these three datasets (see Table 1).

Table 1. Summary of data split statistics.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Composition</th>
<th>Train</th>
<th>Val</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clothing16K</td>
<td>9 8 72 18 7242 10 / 10 5515 9 / 8 3413</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UT-Zappos50K</td>
<td>16 12 192 83 25098 15 / 15 3214 18 / 18 2914</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C-GQA</td>
<td>413 674 278362 5592 26920 1252 / 1040 7280 888 / 923 5098</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Open-world setting In addition to the standard closed-world setting, we also evaluate our model on the open-world setting [24], which is neglected by most previous works. The open-world setting considers all possible compositions, which requires a much larger testing space than the closed-world setting during inference. Taking UT-Zappos50K as an example (see Table 1), the closed world only considers 36 compositions in the testing set while the open world considers total 192 compositions, in which \(\sim 81\%\) are ignored under the standard closed-world setting.

Evaluation metrics Since CZSL models have an inherent bias for seen compositions, we follow the generalized CZSL evaluation protocol [36]. To overcome the negative bias for seen compositions, we apply different calibration terms to unseen compositions and compute the corresponding top-1 accuracy of seen and unseen compositions, where a larger bias makes higher unseen accuracy and lower seen accuracy, and vice versa. We treat seen accuracy as $x$-axis and unseen accuracy as $y$-axis to derive an unseen-seen accuracy curve. We can then compute the area under curve (AUC), the best harmonic mean, the best seen accuracy, and the best unseen accuracy from the curve. In our experiments, we report these four metrics for evaluation, among which AUC is the most representative and stable metric for measuring CZSL model performance. Note that the attribute accuracy or the object accuracy alone does not reflect CZSL performance, because the individual accuracy on attribute or object does not necessarily decide the accuracy of their composition.
Table 2. Closed-world results on three datasets. We report the area under curve (AUC), the best harmonic mean (HM), the best seen accuracy (Seen), the best unseen accuracy (Unseen), the attribute accuracy (Attr), and the object accuracy (Obj) of the unseen-seen accuracy curve under the closed world-setting. AUC is the core CZSL metric. All models use the same DINO ViT-B-16 backbone.

Table 3. Open-world results on three datasets. Different from Table 2, open-world setting considers all possible compositions in testing.

Implementation details We use a frozen ViT-B-16 [6] backbone pretrained with DINO [3] on ImageNet [5] in a self-supervised manner as our visual feature extractor. The ViT-B-16 outputs contain 197 tokens (1 [CLS] and 196 patch tokens) of 768 dimensions. For three attention disentanglement modules, we implement one-layer multi-head attention following [44], changing the single input to paired inputs for cross-attentions. The embedders $\pi_a$, $\pi_r$, $\pi_o$ are the two-layer MLPs following the previous works [24, 50], projecting the 768-dimension visual features to 300-dimension word embedding space. The word embedding prototypes are initialized with word2vec [27] for all datasets and learnable during training. The composition function $\psi$ is one linear layer. We train our model using Adam optimizer [18] with a learning rate of $5 \times 10^{-6}$ for UT-Zappos50K and Clothing16K, and $5 \times 10^{-5}$ for C-GQA. All models are trained with 128 batch size for 300 epochs.

5.2. Comparison

To ensure a fair comparison and demonstrate that our improvement over baseline models is not merely by ViT, we adopt ViT backbone to state-of-the-art CZSL models and re-train all models. We compare our method with them: (1) OADis [41] disentangles attribute and object features from spatial convolutional maps; (2) SymNet [22] introduces the symmetry principle of attribute-object transformation and group theory as training objectives; (3) CompCos [24] extends CZSL to an open-world setting considering all possible compositions during inference, proposing a feasibility score based on data statistics to remove unfeasible compositions; (4) GraphEmb [29] and Co-CGE [25] propose to use graph convolutional networks (GCN) to represent attribute-object relationships and compositions; (5) SCEN [21] projects visual features to a Siamese contrastive space to capture concept prototypes, and introduces complex state transition module to produce virtual compositions; (6) IVR [50] proposes to disentangle visual features into concept-invariant domains from a perspective of domain generalization, by masking specific channels of visual features.

Closed-world evaluation In Table 2, we compare our ADE model with the state-of-the-art methods. ADE consistently outperforms others by a significant margin. ADE increases the core metric AUC by 1.8 on Clothing16K, 0.6 on UT-Zappos50K, and 1.4 on C-GQA (∼37% relatively). Similarly, ADE increases the best harmonic mean (HM) by 0.8% on Clothing16K, 1.9% on UT-Zappos50K, and 3.0% on C-GQA (Table 2). We notice that SymNet [22] and SCEN [21] perform badly on Clothing16K. The reason might be that not learning concept prototypes harms the word embedding expressivity on small-scale concepts. We also notice that IVR [50] performs very well on curated datasets Clothing16K and UT-Zappos50K but badly on larger-scale real-world dataset C-GQA. We hypothesize ideal concept-invariant domains might be difficult to learn from natural images and large-scale concepts of C-GQA. In contrast, our ADE model achieves state-of-the-art performance on all datasets.

Open-world evaluation In Table 3, we consider the open-world setting to compare our ADE with other methods. Likewise, ADE also performs the best among all methods under open-world setting. ADE increases AUC by 3.9 on Cloth-
Table 4. Ablate the components in ADE on open-world UT-Zappos50K. CA, AA, and OA denote composition, attribute, and object attention. Reg denotes the regularization term. We test self- or cross-attention for AA and OA.

<table>
<thead>
<tr>
<th>CA</th>
<th>AA</th>
<th>OA</th>
<th>Reg</th>
<th>AUC</th>
<th>HM</th>
<th>Seen</th>
<th>Unseen</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0)</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>23.8</td>
<td>41.1</td>
<td>59.0</td>
</tr>
<tr>
<td>(1)</td>
<td>self</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>25.3</td>
<td>42.3</td>
<td>61.1</td>
</tr>
<tr>
<td>(2)</td>
<td>self</td>
<td>self</td>
<td>self</td>
<td>✗</td>
<td>26.7</td>
<td>44.6</td>
<td>61.9</td>
</tr>
<tr>
<td>(3)</td>
<td>self</td>
<td>cross</td>
<td>cross</td>
<td>✗</td>
<td>26.9</td>
<td>44.5</td>
<td>63.4</td>
</tr>
<tr>
<td>(4)</td>
<td>self</td>
<td>cross</td>
<td>cross</td>
<td>✓</td>
<td>27.1</td>
<td>44.8</td>
<td>62.4</td>
</tr>
</tbody>
</table>

Table 5. Results on closed-world Clothing16K and C-GQA using different inference formulations. Rows (0)-(2) respectively represents the cases when $\beta = 0.0$, $\beta = +\infty$, and $\beta = 1.0$. Row (3) is our inference formulation, which applies an $\beta$ optimized on the validation set.

### Table 6. Compare ADE and OADis [41] with ResNet18 and ViT.

<table>
<thead>
<tr>
<th>Closed-world</th>
<th>Clothing16K</th>
<th>UT-Zappos50K</th>
<th>C-GQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backbone</td>
<td>AUC</td>
<td>HM</td>
<td>AUC</td>
</tr>
<tr>
<td>ResNet18</td>
<td>OADis [41]</td>
<td>85.5</td>
<td>84.7</td>
</tr>
<tr>
<td></td>
<td>ADE (ours)</td>
<td>87.2</td>
<td>85.1</td>
</tr>
<tr>
<td>ViT-B-16</td>
<td>OADis [41]</td>
<td>88.4</td>
<td>86.1</td>
</tr>
<tr>
<td></td>
<td>ADE (ours)</td>
<td>92.4</td>
<td>88.7</td>
</tr>
</tbody>
</table>

Table 7. Comparison of different regularization terms on open-world UT-Zappos50K.

### Table 7. Comparison of different regularization terms on open-world UT-Zappos50K.

#### Table 8. We propose an EMD-adapted regularization term at the attention level to force attentions to disentangle the concept of interest. We also investigate the effect of applying the regularization term at the feature level. Specifically, we compare our EMD-based distance to the cosine and euclidean feature distances. The results on open-world UT-Zappos50K are shown in Table 7. Our EMD-based regularization outperforms other distance forms, because our attention-level EMD distance considers token-wise similarity capturing the specific concept-related attention responses.

#### Table 8. We propose an EMD-adapted regularization term at the attention level to force attentions to disentangle the concept of interest. We also investigate the effect of applying the regularization term at the feature level. Specifically, we compare our EMD-based distance to the cosine and euclidean feature distances. The results on open-world UT-Zappos50K are shown in Table 7. Our EMD-based regularization outperforms other distance forms, because our attention-level EMD distance considers token-wise similarity capturing the specific concept-related attention responses.
Image and text retrieval We first consider text-to-image retrieval. Given a text composition, e.g., “leather heels”, we embed it and retrieve the top-5 closest visual features based on the feature distance. We display four text compositions of the different objects sharing the same attributes and vice versa in Fig. 5a. We can observe that the retrieved images are correct in most of the cases. One exception is when retrieving “suede sandals”, the third closest image is “fake leather sandals”. Although “suede sandals” and “fake leather sandals” are not the same composition, they are quite visually similar. We then consider image-to-text retrieval, shown in Fig. 5b. Given an image, e.g., the image of a “brown zebra”, we extract its visual feature and retrieve the top-5 closest text composition embeddings. It is difficult to retrieve the ground-truth label in the top-1 closest text composition, but all top-5 results are all semantically related to the image. We take the image of “blond person” (row 3, col 2) as an example. Although the text composition “blond person” is not retrieved in the top-5 matches, the retrieved results “white shirts”, “white outfit”, “white shorts”, “white pants”, and “young girl” are all reasonable and actually present in the image. Image and text retrieval experiments validate that our ADE efficiently projects visual features and word embeddings into a uniform space.

Visual concept retrieval Because the attribute and the object are visually coupled in an image to a high degree of entanglement, it is challenging to visualize the disentanglement in feature space [41]. Saini et al. [41] retrieve single attribute or object text from test images. However, this process is the same as multi-label classification and insufficient to validate that disentangled visual concepts are learned from images. Based on the disentanglement ability of ADE, we construct a visual concept retrieval experiment to investigate the distances between visual concept features, i.e., the embedded attribute feature \( \pi_a(v_a) \) and the embedded object feature \( \pi_o(v_o) \), extracted from different images. Prior to our work, no existing models can do so, because none of them extracts concept-exclusive features like ADE. The results are shown in Fig. 5c. We first extract attribute features and object features from all seen images. Given an unseen image, we retrieve the top-5 closest images by measuring the feature distance between the attribute feature of the given image and that of all seen images, and the same goes for the object feature. For the image of “yellow skirt”, all retrieval results for the visual concept “yellow” are all “yellow [OBJ]”, and all retrieval results for “skirt” are “[ATTR] skirt”. For the “pink pants” image, our model also perfectly retrieves the visual concepts, i.e., the attribute “pink” and the object “pants”. Our experimental results demonstrate that our ADE model is effective to disentangle visual concepts from seen compositions and combine learned concept knowledge into unseen compositions.

6. Conclusion

In this paper, we tackle the problem of compositional zero-shot learning (CZSL). To disentangle visual concepts from the attribute-object composition, we propose ADE adopting cross-attentions to learn the individual concept from paired concept-sharing images. To constrain the disentanglers to learn the concept of interest, we employ a regularization term adapted from the earth mover’s distance (EMD), which is used as a feature similarity metric in the cross-attention module. Moreover, to exploit the attribute and object prediction ability of ADE, we improve the inference process by combining attribute, object, and composition probabilities into the final prediction score. We empirically demonstrate ADE outperforms the current state-of-the-art methods under both closed- and open-world settings. We also conduct a comprehensive qualitative analysis to validate the disentanglement ability of attention disentanglers in ADE.

Limitations Like existing CZSL methods, it is time- and computation-consuming to derive all composition embeddings when the numbers of attributes and objects are large. Moreover, it remains an open challenge to exploit concepts based on the actual semantics rather than solely on text; for instance, the “open” attribute in “open curtain” and “open computer” has completely different meanings.

Acknowledgements This work is partially supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27208022) and HKU Seed Fund for Basic Research.
References


[22] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In CVPR, 2020. 1, 2, 6, 7


[28] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In CVPR, 2017. 1, 2


[41] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In CVPR, 2022. 1, 2, 3, 6, 7, 8


[43] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In NeurIPS, 2013. 2


