Hierarchical Visual Primitive Experts for Compositional Zero-Shot Learning
-Supplementary Materials-

Hanjae Kim\textsuperscript{1} Jiyoungh Lee\textsuperscript{2} Seongheon Park\textsuperscript{1} Kwanghoon Sohn\textsuperscript{1,3,*}
\textsuperscript{1}Yonsei University \textsuperscript{2}NAVER AI Lab \textsuperscript{3}Korea Institute of Science and Technology (KIST)
\{incohjk,saml21796,khsohn\}@yonsei.ac.kr lee.j@navercorp.com

This document presents supplementary materials for theAnonymous ICCV 2023 submission, “Hierarchical Visual Primitive Experts for Compositional Zero-Shot Learning”. In Sec. B, we describe the implementation details and applicability of minority attribute augmentation. We report more experimental results and discussion in Sec. C, and computational analysis in Sec. D.

A. Data statistics

Table 1 shows detailed data statistics of MIT-States [3], C-GQA [6] and VAW-CZSL [8]. Compared to MIT-States, the latest C-GQA and VAW-CZSL have a large number of attribute, object and composition labels, effective for discussing CZSL problems on realistic scenarios.

A.1. Long-tailed distribution

In Fig. 1, we visualize the distributions of composition class ids in the training set. All datasets, especially for C-GQA and VAW-CZSL having a large number of composition classes, show the long-tailed distribution of compositions. This is a natural effect because we can easily guess that ‘black dog’ is more frequent than ‘blue dog’ in the real world. Fig. 2 illustrates the imbalanced attribute composition (e.g., ‘white box’ are 8 times more frequent than ‘pink box’). However, this phenomenon makes it difficult to predict various and novel compositions. Therefore, we proposed minority attribute augmentation (MAA), which remedies a biased prediction caused by the imbalanced data distribution.

B. Implementation details of MAA

We summarize our training procedure of the proposed MAA in Algorithm 1. An auxiliary image $x_B$ is sampled with a sampling weight $\kappa$, which has a different attribute class to a given input $x_A$. Then, a virtual sample $(x_M, y_M)$ is generated by blending the input with the sampled auxiliary image. We first optimize the object and attribute experts in CoT with the generated virtual samples utilizing object and attribute losses (i.e., $L_{\text{obj}}$ and $L_{\text{att}}$) in the main paper. To align a virtual visual prototype $p_x$ from $x_M$ with a semantic prototype $p_y = g([w(o), w(a_M)])$, we simply modify the compositional loss $L_{\text{comp}}$ with the virtual label as follows:

$$L_{\text{comp}} = -\log \frac{\exp\{d(p_{x_M}, p_y)/\tau_c\}}{\sum_{y_k \in \mathcal{Y}_M} \exp\{d(p_{x_M}, p_{y_k})/\tau_c\}},$$

where $\mathcal{Y}_M = \mathcal{Y} \cup \{y_M\}$. We empirically find that directly applying the augmentation at the beginning of training leads to under-fitting. To remedy this, we schedule the training procedure for CoT and MAA. Specifically, we first train pure CoT with a backbone network at 2/3 of the total epochs. Then we freeze the backbone and apply MAA to CoT till the end of training.

C. More results

We provide more complementary results to validate CoT on MIT-States, C-GQA and VAL-CZSL benchmarks.

C.1. Hubness effect

We illustrate a distribution of k-occurrences ($N_k$) [7] to measure a hubness effect of visual features. Note that dif-

\begin{algorithm}[h]
\caption{Minority attribute augmentation}
\begin{algorithmic}
\Require Training dataset $D_{tr}$.
\Ensure Model parameter $\theta$.
\While{Training}
\For{$(x_A, a_A, o) \in D_{tr}$}
\While{$a_B \neq a_A$}
\State Sample $(x_B, a_B, o) \in D_{tr}$ with $\kappa(a_B, o)$
\State Get $p_{x_A}$ and $p_{x_B}$ from CoT
\State Sample $\lambda \sim \text{Beta}(1, 1)$
\State $p_{x_M} = \lambda p_{x_A} + (1 - \lambda) p_{x_B}$
\State $w(a_M) = \lambda w(a_A) + (1 - \lambda) w(a_B)$
\State $\theta \leftarrow \theta - \nabla L_{\text{total}}(p_{x_M}, w(a_M), w(o))$
\EndWhile
\EndFor
\EndWhile
\end{algorithmic}
\end{algorithm}
Table 1: The statistics of three benchmarks: MIT-States [3], C-GQA [6] and VAW-CZSL [8].

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Attribute / #Object</th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT-States [3]</td>
<td>115 / 245</td>
<td>1262 / 30338</td>
<td>300 / 300</td>
<td>400 / 400</td>
</tr>
<tr>
<td>VAW-CZSL [8]</td>
<td>440 / 541</td>
<td>11175 / 72203</td>
<td>2121 / 2322</td>
<td>2449 / 2470</td>
</tr>
</tbody>
</table>

Figure 1: Composition class distribution on three datasets. x-axis (composition class id) is ordered by decreasing composition frequency. For better visualization, we plot the top 1000 frequent composition classes on (b) and (c).

Figure 2: Attribute class distributions on MIT-States and C-GQA. Every attributes in each plot are composed into a fixed object; ‘Silk’, ‘Bucket’, ‘Street’, and ‘Box’. (from left to right).

Table 2: Component analysis on MIT-States and C-GQA dataset.

<table>
<thead>
<tr>
<th>CoT</th>
<th>MAA</th>
<th>AUC</th>
<th>HM</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>8.89</td>
<td>22.9</td>
</tr>
<tr>
<td>✓</td>
<td>9.08</td>
<td>23.5</td>
<td></td>
</tr>
<tr>
<td>✓ ✓</td>
<td>10.26</td>
<td>25.2</td>
<td></td>
</tr>
<tr>
<td>✓ ✓</td>
<td>10.54</td>
<td>25.8</td>
<td></td>
</tr>
</tbody>
</table>

(a) MIT-States  (b) C-GQA

Table 3: Ablation study for different frequency types of MAA on VAW-CZSL.

<table>
<thead>
<tr>
<th>Frequency type</th>
<th>AUC</th>
<th>HM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/(ζ_o)</td>
<td>7.19</td>
<td>21.5</td>
</tr>
<tr>
<td>1/(ζ_o)²</td>
<td>7.07</td>
<td>21.2</td>
</tr>
<tr>
<td>1/(ζ_o) [default setting in paper]</td>
<td>7.20</td>
<td>21.7</td>
</tr>
</tbody>
</table>

C.2. Component analysis

In Table 2, we show the impact of each component (CoT and MAA) for CZSL performance on MIT-States and C-GQA datasets. The ablation results together with Table 3a in the main paper demonstrate that both CoT and MAA consistently give improvements in AUC and HM.

C.3. Other sampling weights

In Table 3, we conduct the ablation study for three different sampling weights leveraging an inverse attribute frequency 1/(ζ_o). Notably, the frequency of 1/(ζ_o)² yields worse performance. It is under-fitting because sampling few tail class samples with too high probabilities prevents learning with other majority classes. Sampling with square-root frequency, 1/(ζ_o)⁰.⁵, improves the performance on AUC and HM, but slightly below the result with 1/(ζ_o). We will include the above discussion in the paper.
C.4. Open World setting

To further analyze the generalization performance, we evaluate our CoT on Open World setting [5] in Table 4. Following [5], we compute the best seen (S), unseen (U) accuracies, area under curve (AUC) and the best harmonic mean (HM). Our method also performs well in Open world setting, outperforming previous state-of-the-art methods in all metrics except the best seen accuracy. This result demonstrates that enlarging visual discrimination with context modeling could also mitigate unfeasible compositions [5] from a large output space of Open World scenario.

C.5. Object-guided attention maps

We illustrate the object-guided attention maps in Fig. 4 with VAW-CZSL, and Fig. 5 with C-GQA. For visualization, we merge three attention maps from low, middle and high blocks through multiplication. The results demonstrate that the attention module could capture the contextualized regions for each attribute, enhancing the visual discrimination of attribute prototypes and its composition.

C.6. Top-3 prediction results

We visualize top-3 prediction results in Fig. 6 with VAW-CZSL, and Fig. 7 with C-GQA. CoT clearly outperforms the baseline [8] to retrieve the relevant composition labels. As discussed in Sec. 4.4 of the main paper, the qualitative results show the limitation of existing CZSL datasets [6, 8] including multi-label composition and multiple attribute-object interaction.

D. Computational Analysis

In Table 5, we compare computational complexity with the previous state-of-the-art OADis [8] by reporting the number of parameters and GFLOPs. Although CoT has more parameters induced from the ensemble of block features, it has almost the same model complexity in terms of GFLOPs, thanks to the parameter-efficient convolution based attention module.

References

Figure 4: Visualization of object-guided attention maps obtained on \textbf{VAW-CZSL}. The input image and attended region by its specific attribute are paired with (attribute, object) labels. (Attention weights: High to Low).

Figure 5: Visualization of object-guided attention maps on \textbf{C-GQA}. The input image and attended region by its specific attribute are paired with (attribute, object) labels. (Attention weights: High to Low).
Figure 6: Ground-truth and top-3 prediction results on VAW-CZSL. We compare CoT with baseline (OADis) [8]. Red box denotes the false positive, having irrelevant or opposite compared to ground truth.

Figure 7: Ground-truth and top-3 prediction results on C-GQA. We compare CoT with baseline (OADis) [8]. Red box denotes the false positive, having irrelevant or opposite compared to ground truth.