Zero-Shot Recognition using Dual Visual-Semantic Mapping Paths

Yanan Li    Donghui Wang*    Huanhang Hu    Yuetan Lin    Yueting Zhuang
Institute of Artificial Intelligence, Zhejiang University
\{ynli, dhwang, huhh, linyuetan, yzhuang\}@zju.edu.cn

In this supplementary material, we provide below practical details of our implementation omitted in the main text.

1. Implementation Details

1. The choice of \( \Omega \) in Eq. 1. During the extraction of inter-class relationship by Eq. 1 in the main text, common choice for \( \Omega \) is \( \ell_1 \) norm or \( \ell_2 \) norm. When \( \Omega(\alpha_i) = ||\alpha_i||_2 \), Eq. 1 is a typical ridge regression problem and we exploit the global structure of \( X_s \) to reconstruct the inter-class relationship. When \( \Omega(\alpha_i) = ||\alpha_i||_1 \), where Eq. 1 becomes a sparse coding problem, the local structure of \( X_s \) is exploited. In our experiments, we choose \( \ell_2 \) norm for \( \Omega \).

2. The mapping function \( f_s \). Let us denote \( n \) labelled training data from \( k \) seen classes as \( X_s \in \mathbb{R}^{d \times n} \) and their ground truth labels are \( Y_s \in \{-1, 1\}^{n \times k} \), each row of which contains only one positive entry indicating the class it belongs to. Also, the label embeddings of seen classes are indicated by columns of \( K_s \in \mathbb{R}^{d \times k} \). We adopt the linear mapping function in [11] to learn the visual-semantic mapping \( f_s \). The objective function in Eq. 3 becomes:

\[
\arg \min_V ||X_s^T V K_s - Y_s||_F^2 + g(V),
\]

where \( V \in \mathbb{R}^{d \times p} \) is the parameter we learn and \( g(V) = \gamma||VK_s||_F^2 + \eta||X_s^T V||_F^2 + \gamma||V||_2^2 \). Thus its solution can be expressed in closed form:

\[
V = (X_s X_s^T + \gamma I)^{-1} X_s Y_s (K_s^T K_s + \eta I)^{-1}.
\]

3. Values of hyper-parameters. There are a few free hyper-parameters to be tuned in our approach, i.e. \( \lambda \) in Eq. 1 (in the main text), \( \gamma \) and \( \eta \) in Eq. 2. \( \lambda \) is set to \( 10^{-4} \). \( \gamma \) and \( \eta \) are chosen from range \( 10^{[1.2, 1.5]} \) and \( 10^{[4.2, 5.4]} \), respectively.

4. Dimensions of the image features and the semantic embeddings. We conduct experiments with deep features on all datasets, extracted by VGG [13], GoogLeNet [14] and ResNet [6]. For VGG and ResNet, we use the 1000-dimensional activations of last fully connected layer as features, and for GoogLeNet we extract features by the 1024-dimensional activations of the top-layer pooling unites. We choose two different types of word vectors in our experiments, i.e. skipgram [8] and glove [9]. They are trained on the Wikipedia corpus and their dimensions are set to 500 and 300, respectively.

2. Additional experimental results

We present in this section some additional experimental results on zero-shot recognition.

2.1. Visualization of the proposed DMaP-T

In addition to Fig. 5 of the main text, we further visualize our zero-shot recognition results of \( U \rightarrow T \) on CUB and \( U \rightarrow U \) on Dogs in Fig. 1 and Fig. 2, respectively.

2.2. Pre-inspection of Semantic Space \( \mathcal{K} \)

To demonstrate the necessity of the proposed pre-inspection step, we first split all classes into seen/unseen at different ratios. Then we extract the orthogonal projection of unseen classes on the subspace \( S \) spanned by seen class embeddings. Finally, we compute the Euclidean pairwise distances among all these projections. These pairwise distances on CUB and ImageNet datasets are visualized in Fig. 3 and Fig. 4.

We observed that when the number of seen classes is much smaller than that of unseen classes, a lot of pairwise distances tend to 0. This means \( f_s \), learned from seen classes is difficulty to discriminate among these unseen classes.

2.3. Comparison to the state-of-the-art methods

In addition to Tab. 3 of the main text, we display more details about the experimental setup of these methods in Tab. 1.

*Corresponding author
Table 1. cZSR ($\mathcal{U} \to \mathcal{U}$) comparison on AwA, CUB and Dogs. We compare ours (achieved using 2 iteration) with the state-of-the-art results using different $\mathcal{K}$, including word vector (W) and attribute (A). We only display the dimension of word vectors in the ‘Dim of $\mathcal{K}$’ column. In our DMaP, only skipgram is used for W. ‘L’ denotes low-level features. ‘T’ or ‘I’ denotes transductive or inductive methods. ‘+’ indicates the concatenation operation. ‘–’ means no result reported in the original paper.

<table>
<thead>
<tr>
<th>Methods</th>
<th>$X$</th>
<th>Dim of $X$</th>
<th>$\mathcal{K}$</th>
<th>Dim of $\mathcal{K}$</th>
<th>T/I</th>
<th>AwA</th>
<th>CUB</th>
<th>Dogs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE [17]</td>
<td>vgg</td>
<td>4096</td>
<td>A</td>
<td>-</td>
<td>I</td>
<td>76.23</td>
<td>30.41</td>
<td>-</td>
</tr>
<tr>
<td>SJE [1]</td>
<td>goog</td>
<td>1024</td>
<td>A/W</td>
<td>1000</td>
<td>I</td>
<td>66.7</td>
<td>50.1</td>
<td>33.0</td>
</tr>
<tr>
<td>SynC [2]</td>
<td>goog</td>
<td>1024</td>
<td>A+W</td>
<td>100</td>
<td>I</td>
<td>72.9</td>
<td>54.7</td>
<td>-</td>
</tr>
<tr>
<td>LatEm [16]</td>
<td>goog</td>
<td>1024</td>
<td>A+W +H*</td>
<td>1000</td>
<td>I</td>
<td>76.1</td>
<td>51.7</td>
<td>36.3</td>
</tr>
<tr>
<td>RKT [15]</td>
<td>vgg+goog</td>
<td>2024</td>
<td>A+W</td>
<td>500</td>
<td>I</td>
<td>82.43</td>
<td>46.24</td>
<td>28.29</td>
</tr>
<tr>
<td>AMP [5]</td>
<td>OverFeat</td>
<td>4096</td>
<td>A+W</td>
<td>500</td>
<td>I</td>
<td>66</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TMV-HLP [4]</td>
<td>OverFeat + DeCaF</td>
<td>8192</td>
<td>A+W</td>
<td>1000</td>
<td>T</td>
<td>80.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>UDA [7]</td>
<td>OverFeat</td>
<td>4096</td>
<td>A</td>
<td>-</td>
<td>T</td>
<td>75.6</td>
<td>40.6</td>
<td>-</td>
</tr>
<tr>
<td>PST [10]</td>
<td>L</td>
<td>10940</td>
<td>A</td>
<td>-</td>
<td>T</td>
<td>42.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>OverFeat</td>
<td>4096</td>
<td>A+W</td>
<td>1000</td>
<td>T</td>
<td>73.5</td>
<td>47.9</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>vgg</td>
<td>1000</td>
<td>W</td>
<td>500</td>
<td>T</td>
<td>68.80</td>
<td>26.02</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>A+W</td>
<td>500</td>
<td>T</td>
<td>83.50</td>
<td>50.8</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>T</td>
<td>85.66</td>
<td>50.45</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>res</td>
<td>1000</td>
<td>W</td>
<td>500</td>
<td>T</td>
<td>67.90</td>
<td>23.31</td>
<td>33.57</td>
<td></td>
</tr>
<tr>
<td>A+W</td>
<td>500</td>
<td>T</td>
<td>87.62</td>
<td>52.14</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>-</td>
<td>T</td>
<td>74.94</td>
<td>61.79</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>goog</td>
<td>1024</td>
<td>W</td>
<td>500</td>
<td>T</td>
<td>85.70</td>
<td>29.97</td>
<td>40.18</td>
<td></td>
</tr>
<tr>
<td>A+W</td>
<td>500</td>
<td>T</td>
<td>90.15</td>
<td>60.90</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vgg+goog</td>
<td>2024</td>
<td>W</td>
<td>500</td>
<td>T</td>
<td>91.52</td>
<td>62.62</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

1 OverFeat and DeCaF denote deep features extracted from OverFeat [12] and DeCaF [3].
2 * Results obtained by using two types of word vectors, i.e. word2vec and glove. H denotes hierarchical semantic embeddings derived from WordNet.

References

Figure 1. Illustration of the results of $\mathcal{U} \rightarrow \mathcal{T}$ task on CUB dataset. (a) Results obtained by DMaP-I. (b) Results obtained by DMaP-T with one iteration. (c) Ground truth unseen class label. Dots with lower brightness denote unseen instances are mistakenly classified to the previously seen classes. The higher brightness of the whole image indicates the better recognition results. This figure is best viewed in color.

Figure 2. Illustration of the results of $\mathcal{U} \rightarrow \mathcal{U}$ task on Dogs dataset. (a) Results obtained by DMaP-I. (b) Results obtained by DMaP-T with one iteration. (c) Ground truth unseen class label. The brown color dots denote unseen instances are classified to wrong classes. This figure is best viewed in color.


Figure 3. Visualization of pairwise Euclidean distances among orthogonal projections of unseen classes on CUB dataset. These pairwise distances are obtained by using different seen/unseen splits. (a) Results obtained on split 10/190. (b) Results obtained on split 20/180. (c) Results obtained on split 30/170. (d) Results obtained on split 40/160. Darker colors depicts closer distances.

Figure 4. Visualization of pairwise Euclidean distances among orthogonal projections of unseen classes on ImageNet dataset. These pairwise distances are obtained by using different seen/unseen splits. (a) Results obtained on split 50/950. (b) Results obtained on split 100/900. Darker colors depicts closer distances.