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 Ω in Eq.1. During the extraction of inter-class relationship by Eq. 1 in the main text, common choice for Ω is ℓ_1 norm or ℓ_2 norm. When $\Omega(\boldsymbol{\alpha}_i) = ||\boldsymbol{\alpha}_i||_2$, Eq. 1 is a typical ridge regression problem and we exploit the global structure of $\mathbf{X}_s$ to reconstruct the inter-class relationship. When $\Omega(\boldsymbol{\alpha}_i) = ||\boldsymbol{\alpha}_i||_1$, where Eq. 1 becomes a sparse coding problem, the local structure of $\mathbf{X}_s$ is exploited. In our experiments, we choose ℓ_2 norm for Ω.

2. The mapping function f_s. Let us denote n labelled training data from k seen classes as $\mathbf{X}_s \in \mathbb{R}^{d\times n}$ and their ground truth labels are $\mathbf{Y}_s \in \{-1, 1\}^{n\times k}$, 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 $\mathbf{K}_s \in \mathbb{R}^{p\times k}$. We adopt the linear mapping function in [11] to learn the visual-semantic mapping f_s. The objective function in Eq.3 becomes:

$$\arg\min_{\mathbf{V}} ||\mathbf{X}_s^T\mathbf{V}\mathbf{K}_s - \mathbf{Y}_s||_F^2 + g(\mathbf{V}), \quad (1)$$

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

$$\mathbf{V} = (\mathbf{X}_s\mathbf{X}_s^T + \gamma\mathbf{I})^{-1}\mathbf{X}_s\mathbf{Y}_s\mathbf{K}_s^T(\mathbf{K}_s\mathbf{K}_s^T + \eta\mathbf{I})^{-1}. \quad (2)$$

where $\mathbf{I}$ is the identity matrix.

3. Values of hyper-parameters. There are a few free hyper-parameters to be tuned in our approach, *i.e.* λ in Eq. 1 (in the main text), γ and η in Eq. 2. λ is set to 10^{-4}. γ and η 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.

*Corresponding author

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 $\mathcal{U} \rightarrow \mathcal{T}$ on CUB and $\mathcal{U} \rightarrow \mathcal{U}$ 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 $\mathcal{S}$ 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.

Table 1. cZSR ($\mathcal{U} \rightarrow \mathcal{U}$) 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.

Methods	$\mathcal{X}$	Dim of $\mathcal{X}$	$\mathcal{K}$	Dim of $\mathcal{K}$	T/I	AwA	CUB	Dogs
SSE [17]	vgg	4096	A	-	I	76.23	30.41	–
SJE [1]	goog	1024	A/W	1000	I	66.7	50.1	33.0
SynC [2]	goog	1024	A+W	100	I	72.9	54.7	–
LatEm [16]	goog	1024	A+W +H*	1000	I	76.1	51.7	36.3
RKT [15]	vgg+goog	2024	A+W	500	I	82.43	46.24	28.29
AMP [5]	OverFeat	4096	A+W	100	I	66	–	–
TMV-HLP [4]	OverFeat	4096	A+W	1000	T	73.5	47.9	–
	OverFeat + DeCaF	8192	A+W	1000	T	80.5	-	-
UDA [7]	OverFeat	4096	A	-	T	75.6	40.6	–
PST [10]	L	10940	A	-	T	42.7	–	–
DMaP	OverFeat	4096	A	-	T	80.35	51.01	-
			W	500	T	68.80	26.02	-
			A+W	500	T	83.50	50.8	-
	vgg	1000	A	-	T	85.66	50.45	–
			W	500	T	82.78	23.31	33.57
			A+W	500	T	**87.62**	**52.14**	–
	goog	1024	A	-	T	74.94	**61.79**	–
			W	500	T	67.90	31.55	38.92
			A+W	500	T	**78.61**	59.62	–
	res	1000	A	-	T	89.34	59.28	–
			W	500	T	85.70	29.97	40.18
			A+W	500	T	**90.15**	**60.90**	–
	vgg+goog	2024	A	-	T	87.52	**63.79**	–
			W	500	T	75.03	30.34	44.59
			A+W	500	T	**91.52**	62.62	–

[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

[1] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015.

[2] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. *arXiv preprint arXiv:1603.00550*, 2016.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, pages 647–655, 2014.

[4] Y. Fu, T. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *PAMI*, pages 1–1, 2015.

[5] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, pages 2635–2644, 2015.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[7] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, pages 2452–2460, 2015.

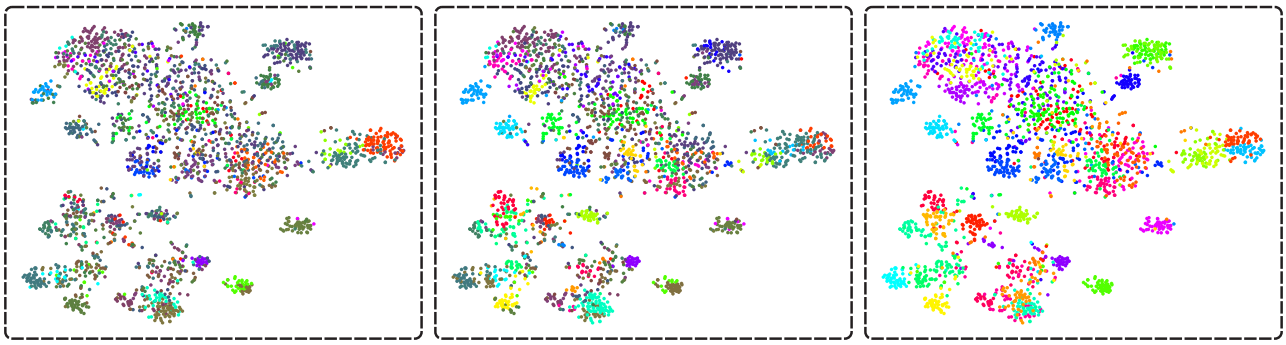
[8] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

[9] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[10] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *NIPS*, pages 46–54, 2013.

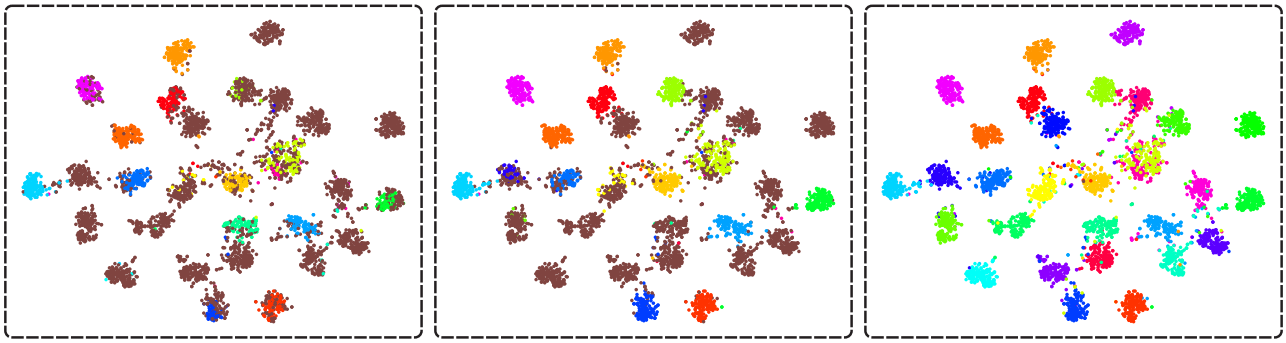
[11] B. Romera-Paredes, E. OX, and P. H. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015.

[12] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization



(a) (b) (c)

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.



(a) (b) (c)

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.

and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*, 2014.

[15] D. Wang, Y. Li, Y. Lin, and Y. Zhuang. Relational knowledge transfer for zero-shot learning. In *AAAI*, 2016.

[16] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. *arXiv preprint arXiv:1603.08895*, 2016.

[17] Z. Zhang and V. Saligrama. Classifying unseen instances by learning class-independent similarity functions. *arXiv preprint arXiv:1511.04512*, 2015.

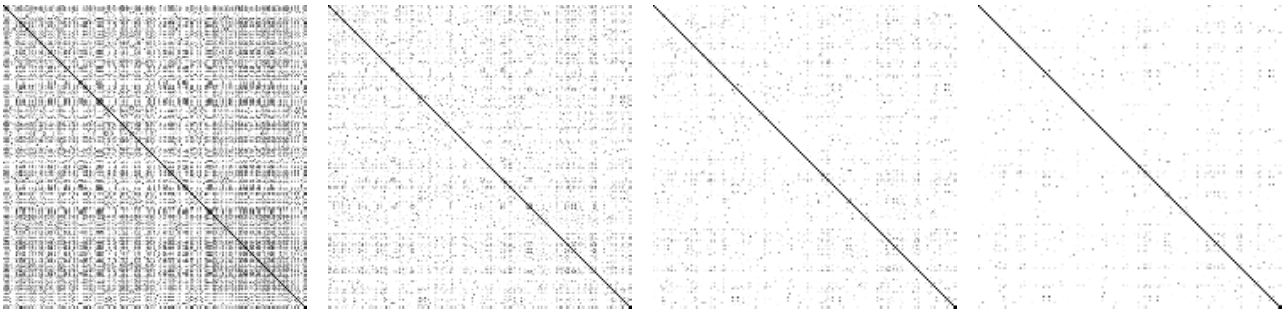


Figure 3. Visualization of pairwise Euclidean distances among orthogonal projections of unseen classes on CUB dataset. These pairwise distances are obtained by using different seen/unseen splits. (a) Results obtained on split 10/190. (b) Results obtained on split 20/180. (c) Results obtained on split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

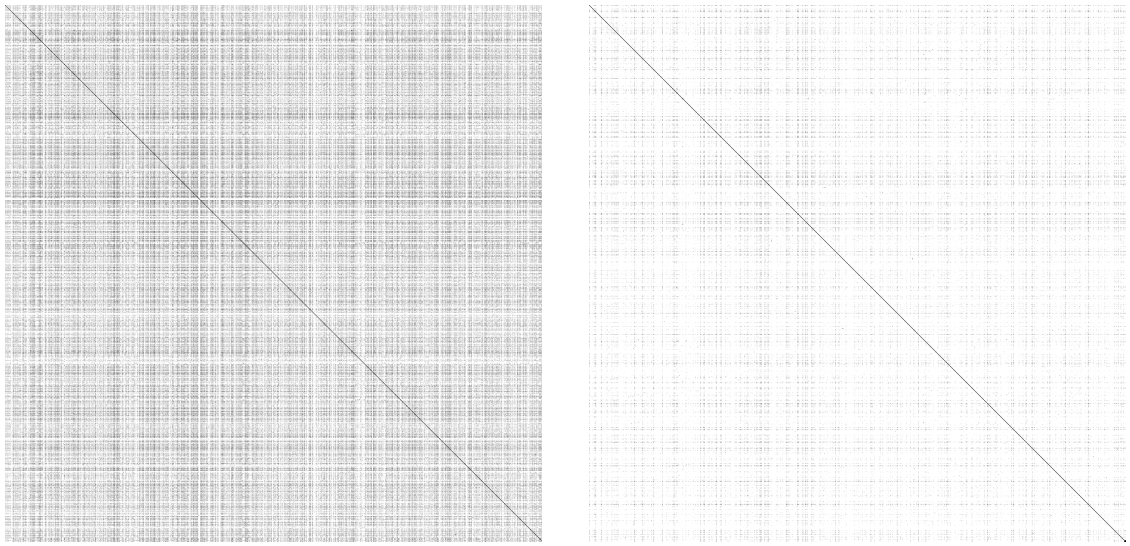


Figure 4. Visualization of pairwise Euclidean distances among orthogonal projections of unseen classes on ImageNet dataset. These pairwise distances are obtained by using different seen/unseen splits. (a) Results obtained on split 50/950. (b) Results obtained on split 100/900. Darker colors depicts closer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