Discriminative Learning of Latent Features for Zero-Shot Recognition

Yan Li [1,2], Junge Zhang [1,2], Jianguo Zhang [3], Kaiqi Huang [1,2,4]
[1] CRIPAC & NLPR, CASIA [2] University of Chinese Academy of Sciences
[3] Computing, School of Science and Engineering, University of Dundee, UK
[4] CAS Center for Excellence in Brain Science and Intelligence Technology
yan.li@cripac.ia.ac.cn, jgzhang@nlpr.ia.ac.cn, j.n.zhang@dundee.ac.uk, kqhuang@nlpr.ia.ac.cn

1. How to Identify the Discriminative Region from an Image?

To search the discriminative region from an image in zero-shot learning (ZSL), two weakly supervised learning approaches can be considered: 1) directly regressing the locations of the identified region (*e.g.*, the proposed zoom scheme in our LDF model); 2) extracting multiple region proposals (*e.g.*, EdgeBox [7]) for the image and then selecting the most discriminative one. In this paper, we didn't utilize the latter region proposal method based on the following considerations. First, the goal of the region proposal algorithm [7] is to identify "objects". However, as shown in Figure 5 and claimed in Section 4.2, in ZSL, the identified region may contain context elements to match its user-defined attributes. Such region is not exactly equal to the "object" region and hard to be captured by EdgeBox. Second, processing multiple proposals (typically 2,000) for each image is quite inefficient, and selecting the proper region from 2,000 ones is also difficult in weakly supervised settings. We have conducted an experiment to test the region proposal approach for ZSL.

Specially, we first extract 2,000 EdgeBox proposals for each image. Then we replace the *pool5* layer in SS-BE-baseline (VGG19) with the RoI Pooling layer proposed in Fast RCNN [3]. The images with their region proposals are imported into the model, and the model could output the compatibility score for each region. Following the standard multiple instance learning (MIL) setting, the region with highest compatibility score is selected to compute the loss function as in (6). The network finally obtains 72.67% on AwA dataset. This result is even lower than SS-BE-Learned (Table 1, 78.35%), which directly extract image features from full-size images. Moreover, the runtime is 7$\sim$8 times longer than our zoom scheme.

2. The Bilinear Interpolation Operation

In Section 4.2, to obtain better representation for finer localized cropped region x^{crop}, the bilinear interpolation is utilized to adaptively zoom the cropped region to the same size with the original image. Concretely, for a point (i, j) of the zoomed region, its value $x^{\text{zoom}}_{(i,j)}$ can be computed by linearly combining the values of nearest four points in the cropped region. Formally,

$$\begin{aligned} x^{\text{zoom}}_{(i,j)} &= \sum_{\alpha,\beta} |1-\alpha-\{i/\lambda\}||1-\beta-\{j/\lambda\}| x^{\text{crop}}_{(m,n)}, \\ m &= \quad [i/\lambda] + \alpha + z_x - z_s, \quad \alpha = 0, 1 \\ n &= \quad [j/\lambda] + \beta + z_y - z_s, \quad \beta = 0, 1 \end{aligned} \tag{1}$$

where λ is the upsampling factor, *i.e.*, $\lambda = 1/z_s$. $[\cdot]$ and $\{\cdot\}$ is the integral and fractional part, respectively.

3. Experiments with Three Scales on AwA

As we have mentioned in Section 5.2, for AwA dataset, only one zoom operation is performed and the two-scale model is adopted. We claim the reason is that the objects in AwA images are usually large and centered. To verify this, in this section, we analyze the performance of three-scale MS-BE-Learned baseline on AwA. The experiment is conducted with GoogLeNet and all the experimental settings are the same as we described in Section 5.2. The performance of each single scale is shown in Table 1.

Additionally, the parameter z_s in Eq. (2) represents the length of the cropped regions. In scale 1 and scale 2, we respectively count the z_s values for all unseen images and show the mean value of the z_s in Table 1. It can be seen that when the

	ZSL performance on AwA	mean value of z_s
SS-BE-Learned	75.19	-
MS-BE-Learned (Scale 1)	75.47	0.87
MS-BE-Learned (Scale 2)	77.12	0.98
MS-BE-Learned (Scale 3)	77.05	-

Table 1: The detailed ZSL results (%) on each scale and the mean value of z_s parameter.

three-scale model is adopted on AwA, the performance of the second scale is higher than the first scale (77.12% *vs*. 75.47%). However, the performance of the third scale does not show the further improvement (77.05% *vs*. 77.12%). When we inspect the mean z_s values in the second scale, it can be found that the scale size of the cropped region is nearly 1 (0.98), that is, the zoom net in the second scale actually does not perform any cropping operation and directly send the original image to the third scale. As we have claimed, the objects in AwA images are large and centered. Through one time zoom operation, the network can capture the main object and the third scale is actually useless in the model.

	ZSL performance on AwA	The dimension of LA
SS-AE-Learned (LA)	75.75	k (85)
SS-AE-Learned (LA)	75.83	$2k$ (170)
SS-AE-Learned (LA)	76.01	$3k$ (255)

Table 2: The ZSL results (%) with different dimension of latent attributes.

4. The Effect of the Dimension of Latent Attribute

As we mentioned in Section 4.3.2, the dimension of the latent attributes (LA) is set to k, *i.e*., the same with the user-defined attributes (UA). In this section, we explore the effectiveness of the dimension of latent attributes and conduct experiments on AwA dataset with GoogLeNet. Specially, we train the SS-AE-Learned baseline with different dimensions of LA (*i.e*., k, $2k$ and $3k$), and perform ZSL prediction with the latent attributes only. The results are shown in Table 2. It can be seen that with the larger dimension of LA, the ZSL performance improves. But the improvement is slight and the performance in general is robust to the dimension of LA.

5. The Discriminativeness of the Learned Latent Attributes

In this section, we show more visualized examples to illustrate the discriminative property of latent attributes. For a latent attribute element, the images which have largest and smallest activations over this element are shown in Figure 1. Meanwhile, the examples selected with the learned UA features are shown in Figure 2 for comparison. From Figure 2, it can be seen that the user-defined attributes are shared in many objects. Another discovery is that the prediction results of user-defined attributes will be affected by mid-level cues, *e.g*., colors. For example, for UDA5 element, the *chimpanzee*, *whale* and *pig* objects are falsely predicted as *orange* due to the existing orange backgrounds. For UDA64 element, the *persian cat* and *pig* images are falsely predicted as *arctic*. It is possible that the two animals share white appearances.

6. Generalized Zero-Shot Learning Results

In conventional zero-shot learning (cZSL), ZSL methods are trained on seen classes and evaluated on unseen ones. The basic assumption in cZSL is that test instances always come from the unseen classes (denoted as $\mathcal{U} \rightarrow \mathcal{U}$), which is actually unrealistic in real-world applications. Motivated by this, recent ZSL works [2, 6] aim to measure the zero-shot performance in the generalized zero-shot learning (gZSL) setting. In gZSL, the test images are assumed to come from all target classes including both seen and unseen categories.

Method	AwA			CUB		
	$A_{\mathcal{U}\to\mathcal{T}}$	$A_{\mathcal{S}\to\mathcal{T}}$	H	$A_{\mathcal{U}\to\mathcal{T}}$	$A_{\mathcal{S}\to\mathcal{T}}$	H
DAP [4]*	2.4	77.9	4.7	4.0	55.1	7.5
IAP [4]*	1.7	76.8	3.3	1.0	69.4	2.0
ConSE [5]*	9.5	75.9	16.9	1.8	69.9	3.5
SynC$^{\text{o-vs-o}}$ [1]*	0.3	67.3	0.6	8.4	66.5	14.9
SynC$^{\text{struct}}$ [1]*	0.4	81.0	0.8	13.2	72.0	22.3
LDF (Ours)	**9.8**	**87.4**	**17.6**	**26.4**	**81.6**	**39.9**

Table 3: Generalized zero-shot learning results (%). All results are obtained with GoogLeNet features. $*$ means that the numbers of the method are cited from [2], since the original paper does not report the gZSL results. H denotes the harmonic mean.

Similar to [2], 20% of the images from seen classes are extracted and then merged with the images from unseen classes to form the new test set. We denoted the joint label space of seen and unseen classes as $\mathcal{T} = \mathcal{S} \cup \mathcal{U}$ and evaluate the proposed LDF model in terms of accuracy on $\mathcal{U} \to \mathcal{T}$ and $\mathcal{S} \to \mathcal{T}$, which are denoted as $A_{\mathcal{U}\to\mathcal{T}}$ and $A_{\mathcal{S}\to\mathcal{T}}$, respectively. $A_{\mathcal{U}\to\mathcal{T}}$ indicates the accuracies of classifying test images from unseen classes into the joint label space while $A_{\mathcal{S}\to\mathcal{T}}$ indicates the accuracies of recognizing seen objects into the joint label space. Moreover, similar to [6], the harmonic mean is computed to measure the ZSL methods with considering both the accuracy of seen classes and the accuracy of unseen classes. Formally,

$$H = \frac{2A_{\mathcal{U}\to\mathcal{T}}A_{\mathcal{S}\to\mathcal{T}}}{A_{\mathcal{U}\to\mathcal{T}} + A_{\mathcal{S}\to\mathcal{T}}} \tag{2}$$

The experiments are performed on both AwA and CUB datasets. The GoogLeNet model is utilized and the results are shown in Table 3. It can be seen that on both datasets, the proposed LDF model significantly outperforms previous methods on all the three metrics, which confirms the advantage of our method under the gZSL setting.

References

[1] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016.

[2] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pages 52–68, 2016.

[3] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[4] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(3):453–465, 2014.

[5] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[6] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*, 2017.

[7] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, pages 391–405, 2014.

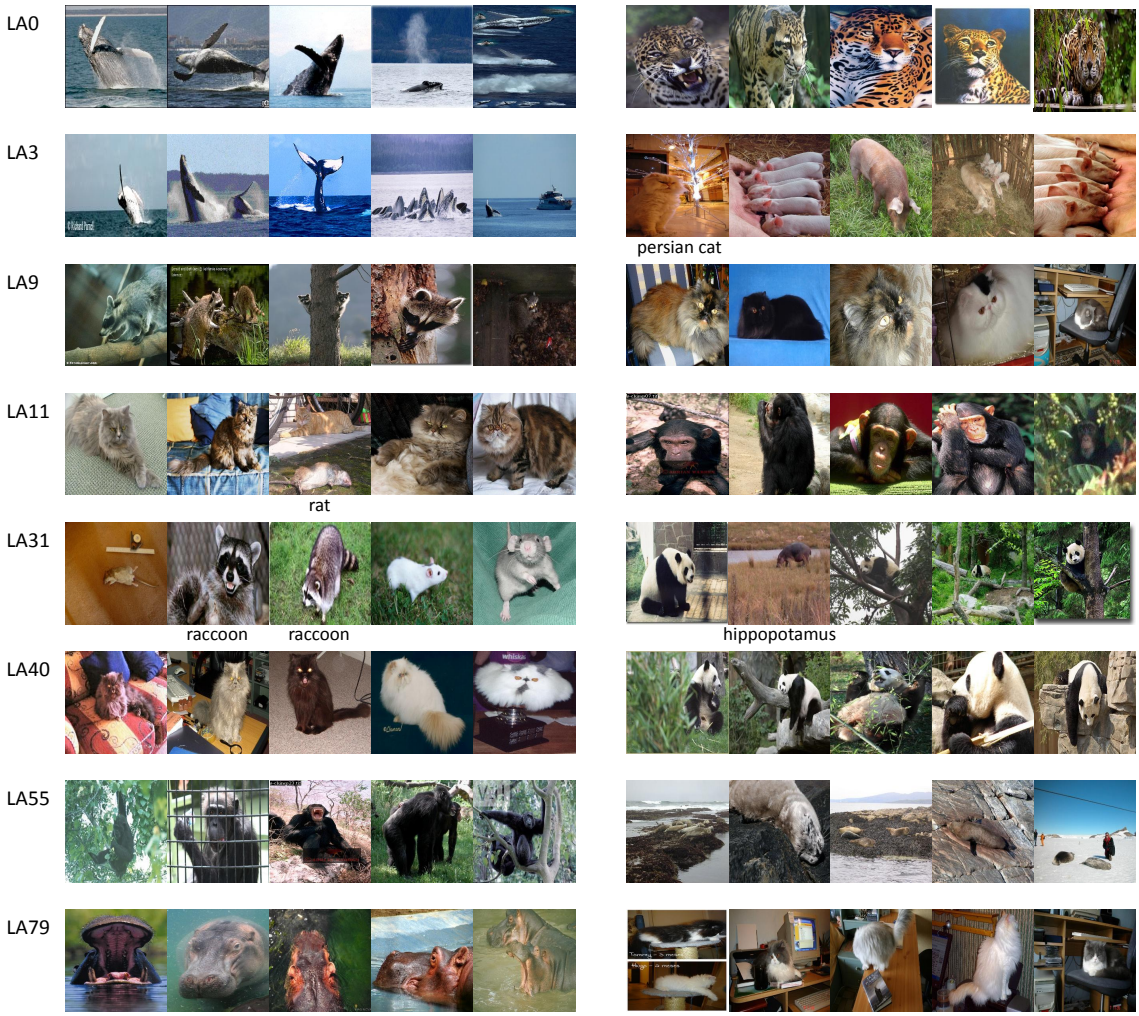

Figure 1: The visual examples of latent discriminative attributes (LA) on AwA. 'LAX' denotes the X-th element of the LA features. The LA features are obtained with the VGG19 SS-AE-Learned baseline. The first five images are top-5 images with largest activations over this element and the last five images are selected examples with smallest activations.

Figure 2: The visual examples of user-defined attributes (UA) on AwA. 'UDAX' denotes the X-th element of the UA features. The UA features are obtained with the VGG19 SS-AE-Learned baseline. The first five images are top-5 images with largest activations over this UA element and the last five images are selected examples with smallest activations.